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

Team Control Number

82898

Problem Chosen

B

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

2018
MCM/ICM
Summary Sheet

Summary

Currently, there exist quite a lot of languages in the world. The distribution of speakers of different languages may vary over time and it is meaningful to forecast it. In this paper, we construct two models to do prediction about it, Total Speakers Growth Model (TSG-M) and Language Distribution Growth Model (LDGM). TSGM focuses on the changes in number of speakers of different languages over time, while LDGM focuses on the changes in geographic distribution of languages.

Considering that total speakers of some language consist of native speakers and second (or third, etc.) language speakers, we construct two sub-models, Native Speakers Growth Model (NSGM) and Second (or third, etc.) Language Speakers Growth Model (SLSGM), to forecast the number of native speakers and the number of second (or third, etc.) language speakers respectively. We apply logistic growth model and function fitting in NSGM. In SLSGM, we apply curve fitting and Multiple Regression Analysis. Based on the forecasting results, we predict that tiny change will occur to the top 10 list.

Then we construct LDGM by divide the world into six continents (except Antarctica) and forecast the number of speakers of each language in each continent. In this model, we apply Analytic Hierarchy Process to forecast migration patterns by determining the attractiveness and suitableness of each continent for living. Taking the influences of global population growth, human migration and the development of education into account, we construct a sequence to forecast number to do prediction. In this sequence, we introduce three parameters, , and . Later we do sensitivity analysis on them and justify that our choices of values of , and in forecasting are reasonable. Based on the current and forecasting distribution, we offer advice on locations of new offices for the company both in the short run and the long run.

Furthermore, we give some conclusions and state the strengths and weaknesses of our models.

Finally, we write a letter to the Chief Operating Office to summarize results and suggestions.

Contents

1 Restatement of the problem

Firstly, we are asked to construct a model to forecast the number of speakers of languages over time, considering the influences of government, migration, international business relations, the development of technology and so on, then to forecast the number of speakers in the next 50 years with the model that we have constructed.

Additionally, we are asked to model the geographic distributions of languages over time to find whether some changes will occur, given the influences of global population, human migration and so on. Then we required to give advice on the locations of new offices, both in the short term and in the long term.

To solve the problems, we will proceed as follows:

1.Construct TSGM to forecast the number of total speakers of each language, which consists of two sub-models, NSGM and SLSGM, since total speakers have two parts: native speakers and second (or third, etc.) language speakers. NSGM is constructed to forecast the number of native speakers and we apply **logistic growth model** and **function fitting** in it, while SLSGM is constructed to forecast the number of second (or third) language speakers and we apply **function fitting** and **multiple regression analysis** in it.

2.Forecast number of total speakers of each language 50 years later with NSGM, then find out whether the top 10 popular languages have changed. Besides, we also do sensitivity analysis on TSGM.

3.Construct LDGM to forecast geographic distribution of several languages. We construct a sequence and apply **Analytic Hierarchy Process** in this section.

4.Based on the forecasting results of LDGM, we offer advice on the locations of new offices for the company both in the short term and in the long term.

5.At last, we write a letter to the Chief Operating Officer of the company to summarize our results and suggestions.

2 Assumptions and Notations

2.1 Assumptions

- **Everyone only has one native language which is decided by his (or her) parents, considering the definition of native language.** Further explanation is in 3.1.
- **A couple shares the same native language.** Since one‘s native language is decided by his (or her) parents‘ native language and one person has only one native language, for simplicity we assume a couple shares the same

native language. In fact, most of couples share the same native language in real life.

- **Native language cannot be altered even after people emigrate to a new country, but they have opportunities to master another language.** Though Immigrants use official languages of a different country when they work (probably different from their native language), they still speak their original language, namely their native language, at home, which means that they remain native speakers of their original language.
- **People aged above 40 will not receive tertiary education, which means they will not learn any other languages.**

2.2 Notations

Table 1: Notations

Symbol	Definition
TS	number of total speakers
NS	number of native speakers
SLS	number of second(or third, etc) language speakers
x_1	number of employees in multinational companies
x_2	number of people who can learn other languages from territory education
x_3	number of migrants
x_4	number of people who use social networks
n	number of people under 40
p	population of people under 40 that receive territory education
SN	stable number of habitants for a continent
SW	stable weight(population) of habitants for a continent
TNP	total number of people on earth
N	number of people in a continent
$CNSM$	important extent of a language
IE	important extent of a language
λ	natural population growth rate in a continent
AIR	adjusted increase rate

3 Total Speakers Growth Model

In this section, we construct **Total Speakers Growth Model (TSGM)** to predict the growth of total speakers of any specific language. For a specific language

A, total speakers of A consist of native speakers and second (or third, etc.) language speakers. So, we do prediction of native speakers of A and second (or third, etc.) language speakers of A separately.

In this part, we construct **Native Speakers Growth Model (NSGM)** to predict the number of native speakers of A and **Second (or third, etc.) Language Speakers Growth Model (SLSGM)** to predict the number of people who can speak more than one languages over time.

3.1 Construction of NSGM

In this part, we construct **Native Speakers Growth Model (NSGM)** to predict the number of native speakers of language A.

Firstly, we look through the definition of native language. Native language is defined by one‘s first language, learned in early childhood. Thus, one‘s native language largely depends on his (or her) family, since he (or she) spends the whole childhood with his (or her) families. And usually, the language one is exposed to first after birth is his (or her) parents‘ native language, and this language tends to become his (or her) native language. Therefore, we can assume that one‘s native language is the same as his (or her) parents‘ native language (A1), and for simplicity we assume a couple shares the same native language (A2). Besides, one person can only be native speaker of one language, so population growth largely accounts for the growth of native speakers. Therefore, we consider logistic growth model, which is often used in prediction of population growth.

Introduction of logistic growth model

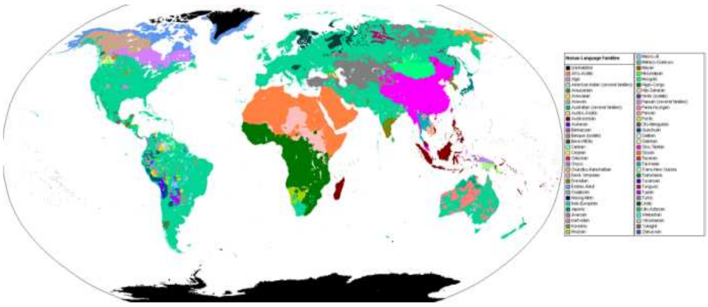


Figure 1: Current distribution of human language families[1]

‘Verhulst considered that, for the population model, a stable population would consequently have a saturation level characteristic; this is typically called the carrying capacity, K, and forms a numerical upper bound on the growth size.‘[2] Logistic growth model can be applied to predicting the population growth since unrestricted growth is unrealistic due to limited environmental resources. To construct NSGM, we

apply logistic growth model. We can divide the whole population into several groups by native language. Each group is characterized by one language and people in the same group share the same native language. For each group, we apply logistic growth model to predict the number of native speakers of its characterized language. Since people share the same native language tend to live together and thus form a community (Figure 1), so limited environmental resources of communities accounts for the restricted growth of native speakers of a specific language. Besides, Language Projections: 2010 to 2020 by Jennifer M. Ortman[3] also demonstrates it is reasonable to model growth of native speakers with logistic growth model by doing projections and comparing results.

For a specific language A, by logistic growth model, we have

$$NS_t^A = \frac{a_A}{1 + b_A e^{-c_A^t}} \tag{1}$$

Where NS_t^A represents the number of native speakers of A in year t. Then we use data to do function fitting to get these parameters a_A, b_A, c_A. Thus, we get the logistic growth function NS_t^A for language A, and we can predict the native speakers of language A with it.

3.2 Construction of SLSGM

We have discussed the case of native speakers above. In this section, we will construct the Second (or third, etc.) Language Speakers Growth Model (SLSGM) to predict the growth of the number of second (or third, etc.) language speakers over time, denoted by SLS.

3.2.1 Factors considered in SLSGM

To begin with, we consider four functions to reveal some factors which affect SLS. Those factors include international business relations, tertiary education, migration and the use of electronic communication and social media. Now we explain all of them respectively.

Business relations

With economic globalization, international business relations develop rapidly. At the same time, multinational companies are key elements of international business relations, so most of employees who work for them will be paid well.

To give a description of the influence of worldwide business on SLS, we use the number of employees in multinational companies as its parameter, denoted by x_1(t). We assume that the relative growth rate of x_1(t) is constant. Then, we have equation

$$\frac{dx_1(t)}{dt} = kx_1(t) \tag{2}$$

where k is a constant. After that, we get the solution of (2) as follows

$$x_1(t) = ke^{kt} + c \tag{3}$$

where c is a constant.

Tertiary education

In our school or for studying further, we are usually required to learn other languages, while some of people may not be able to learn another language due to the poor educational conditions. However, because of the requirement of globalization and better educational resources, worldwide education will be improved constantly as time goes by, and tertiary education will be popularized.

To explain the effect of tertiary education to SLS, we find that it is mainly related to the number of people under 40 based on assumption and the proportion of them who receive tertiary education. Both of them are in relation to time t, so we derive two functions n(t) and p(t) for each. The total number of people who can learn other languages from tertiary education is denoted by x_2(t). Then, we get the equation as follows

$$x_2(t) = n(t)p(t) \tag{4}$$

From Figure 2[4],we get the data for n(t).

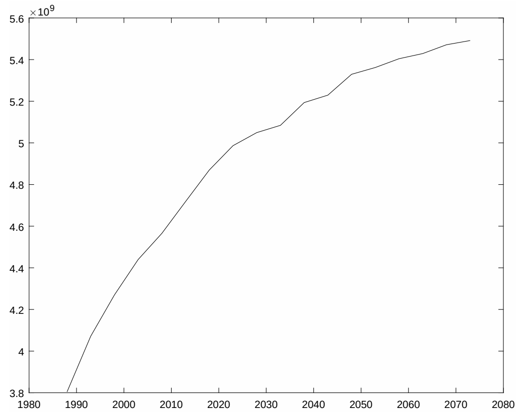


Figure 2: the number of people under 40 over time

Because of globalization and educational condition improved incessantly, more and more people will receive tertiary education and learn more languages aside from native language. Worldwide communication has begun since human civilization appeared, which means we can assume that p(0)=p_0, p_0 ∈[0,0.4). With the development of education, we also can assume that

$$\lim_{t\to+\infty} p(t) = 1 \tag{5}$$

Meanwhile, the growth rate of p(t) decreases as t increases because it is really difficult to realize that everyone can receive higher education as p(t) increases. Generally, p(t) is similar to the curve in Figure 3. So we define

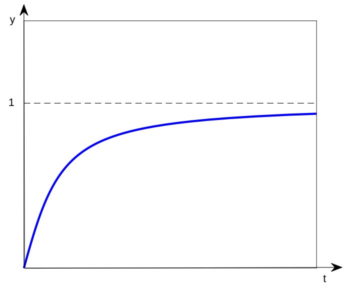


Figure 3: Rough picture function p(t)

$$p(t) = \frac{2}{\pi} arctan(\lambda t + t_0) \tag{6}$$

where t_0, λ are constants.

To simplify calculation, we make p_0=0. Then t_0=0.

From (4)(5)(6), we get the equation

$$x_2(t) = \frac{2}{\pi} n(t) arctan(\lambda t) \tag{7}$$

Migration

Due to various factors such as studying more academic knowledge, better or poor economic conditions, national policies, many people may learn second or more languages to go abroad. Therefore, migration plays such an important role in N that we cannot ignore it.

To describe how migration affects SLS, we use the number of migrants as variable of multiple regression analysis, denoted by x_3(t). Figure 4 shows the number of migrants in the world 1960-2015[5].

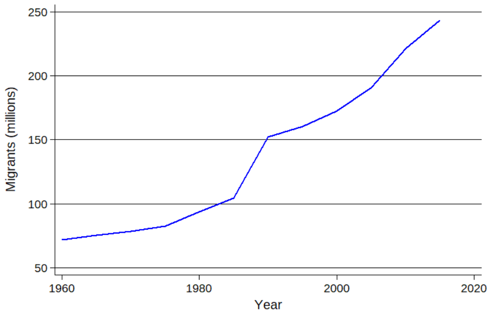


Figure 4: Migrants in the world 1960-2015[5]

To simplify calculation, we use simple linear regression to get the fitting curve and predict the trend of the number of worldwide migrants in the next 50 years. Let x_3(t)=ax+b. Based on Figure 4, we can calculate the value of a,b.

The use of electronic communication and social media

In the recent years, the Internet and the rapid development of electronic products have significantly promoted communication among different countries. To explain SLS under the influence of global communication in the Internet, we consider the number of people who use social networks, denoted as x_4(t).

According to the statistics from The Statistics Portal[6], we can get some data for x_4(t).

At first, we know Facebook is the most popular international social network, and we assume that the percentage of Facebook in the world, denoted by h. From Figure 5 which shows most popular social networks worldwide as of January 2018, we can define h=0.1616.

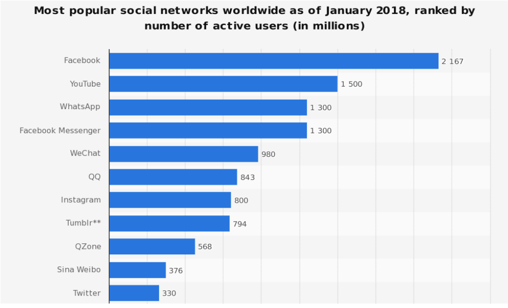


Figure 5: Most popular social networks worldwide as of January 2018, ranked by number of active users(in millions)[7]

Secondly, the period of 2008-2017 experiences a considerable increase, as Figure 6 shows

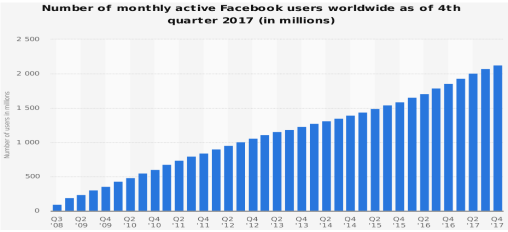


Figure 6: Number of monthly active Facebook users worldwide as of 4th quarter 2017(in millions)[8p]

Denote the number of active users of Facebook in relation to time t as the function g(t). While observing the data from figure 6, we find that g(t) increases constantly over time. Although it is difficult to forecast the accurate values after 2018 with such little information, we also find out the growth rate will decrease over time due to decreasing natural population growth rate. In order to simplify the calculation and estimate the number of active users of Facebook in the future, we consider using function y=$a\sqrt[3]{b(t-2006)}$ to approach it and predict the trend of g(t) in the next 50 years. Finally, we have the equation x_4(t)=g(t)/h

3.2.2 The function SLS(t)

From the influence of factors we discussed above, the number of second(or third,etc.) language speakers SLS(t) will increase while x_i(t) increase. However, the data among x_i(i=1,2,3,4) may be counted repeatedly. Moreover, the extent of their effect on SLS(t) are different. To simplify the calculation, we seek to find the method to consider all of those factors. Therefore, we assume that

$$SLS(t) = \beta_0 + \sum_{i=1}^{4} \beta_i x_i(t) \tag{8}$$

where β_i are constants.

We have gotten the number of second(or third,etc.) language speakers SLS(t). In this case, we assume that the proportion of language A in SLS(t) will not be changed in the next 50 years, denoted as p_A. Based on the statistics from question, we can acquire p_A for top-twenty-six languages. As a result, the number of people who consider language A as second (or third, etc.) language SLS_t^A is equal to p_A*SLS(t).

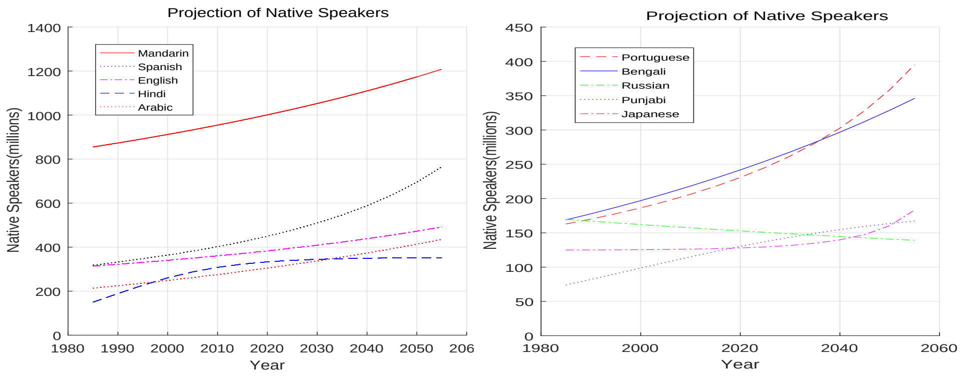
3.3 Construction of TSGM

Through the discussion of NSGM and SLSGM for a specific language A, we have known that we can get the data about language A including the number of native speakers NS_t^A and the number of second (or third, etc.) language speakers SLS_t^A. All in all, we will know the number of total speakers TS_t^A=NS_t^A+SLS_t^A.

3.4 Forecasting Results

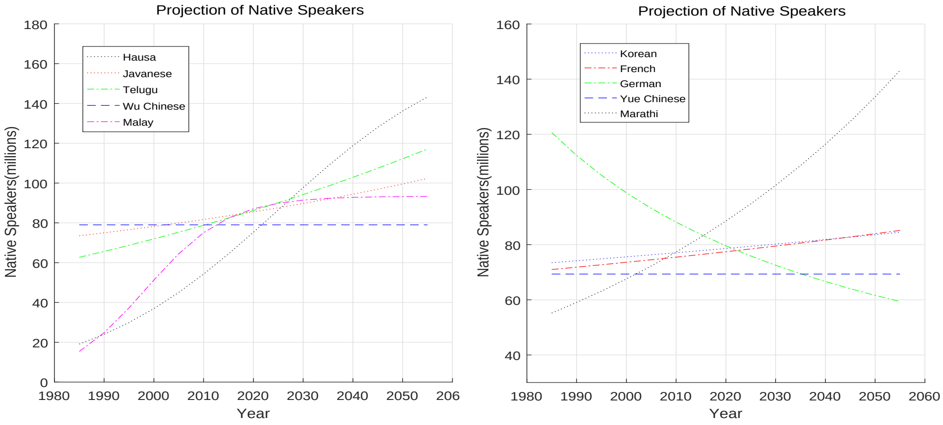
NSGM

The following figures show the result of function fitting of NSGM.



(a) Language Ranked 1-5

(b) Language Ranked 6-10



(c) Language Ranked 11-15

(d) Language Ranked 16-20

Figure 7: Projection of Native Speakers

SLSGM

In order to predict the number of second (or third, etc.) language speakers for languge A over time, we have to determine SLS(t) and p_A.

At first, we consider the first function x_1(t). Using predicted or projected values in [9], we get the data x_1(2016)=63115489. And we rationally assume that x_1(2006)=56798389. Then, the solution can be acquired as follow

k=0.01112, c=2569728

So, we get the function $x_1(t) = 0.01112 * e^{0.01112t} + 2569728$. Secondly, we think about function x_2(t). We assume that 40% of people in the world have the opportunities to receive tertiary education in 2018, which means p(2018)=0.4. Let p_0=0. Then we have $\lambda = 3.6 * 10^-4$ and the function $x_2(t) = \frac{2}{\pi}n(t)arctan(3.6 * 10^{-4} * t)$

Thirdly, based on figure 4, we can find out the concrete numerical value of a,b by Matlab. Then, we get a=3.3*10^6, b=6.4077*10^9, and the number of migrants x_3(t)=(3.3*t-6407.7)*10^6. And the fitting curve follows

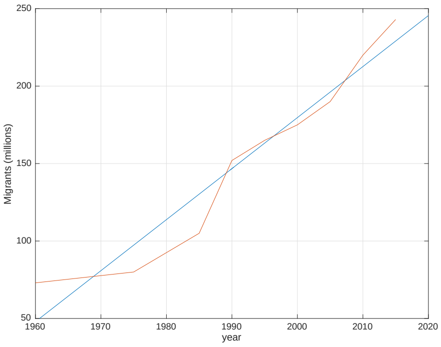


Figure 8: The fitting curve $x_3(t)$

Fourthly, from figure 5, we have gotten h=0.1616. As for the number of active users of Facebook g(t), g(t)=$\frac{2}{\pi}n(t)arctan(3.6*10^{-4}*t)$ is constructed to approach the data in figure 6 and more accurate estimation by Matlab. Then, we get the calculating results are a=124.7, b=157.7. And the fitting curve(figure 9) follows

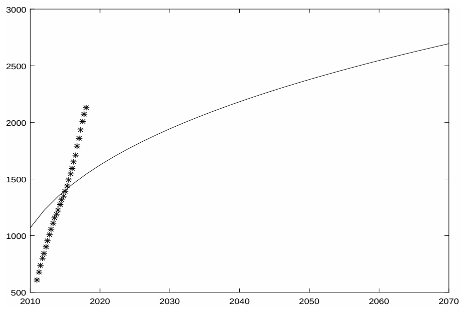


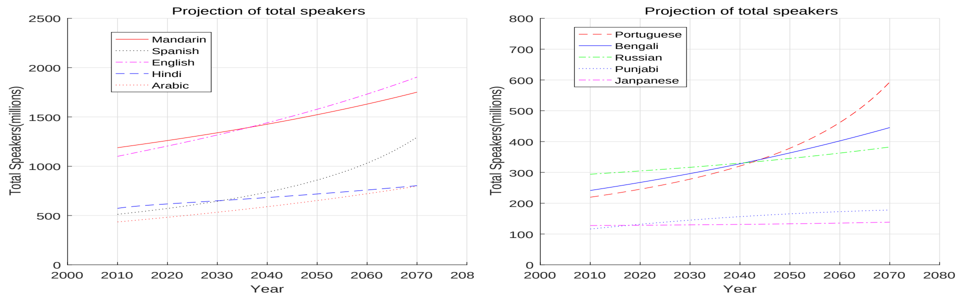
Figure 9: The fitting curve $x_4(t)$

Furthermore, to summarize the four factors we have discussed above, we calculate β_i(i=1,2,3,4) by the method of multiple linear regression(In statistics, multiple linear regression is a linear approach for modelling the relationship between a scalar dependent variable y and more explanatory variables (or independent variables) denoted X.). Then, the solutions are β_0=169100000, β_1=2.6647, β_2=0.1785, β_3=0.0329, β_4=0.0001. And we will get the function SLS(t).

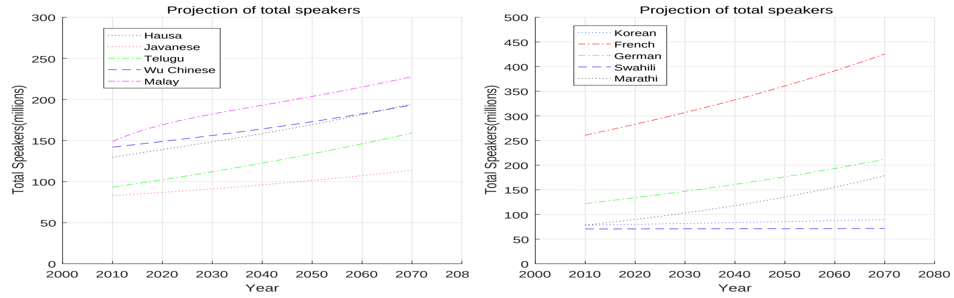
Finally, for a specific language A, its percentage p_A can be calculated by the data in subject(A*). In this case, we discuss top-twenty languages as second(or third, etc.) language. The figure(figure 20) about top-twenty languages corresponding to their percentage will be in the appendix. Therefore, based on function SLS_t^A=SLS(t)*p_A, we can get the projection of second(or third, etc.) language speakers for those languages.

TSGM

We have gotten both of the projection of native speakers and sencond(or third, etc.) language speakers. And based on TSA(t)=SLS_t^A+NS_t^A, we can estimate the number of total speakers of A. The figures containing the projection of total speakers of top-twenty languages follow



(a) Language Ranked 1-5 (b) Language Ranked 6-10



(c) Language Ranked 11-15

(d) Language Ranked 16-20

Figure 10: Projection of total speakers

3.5 Sensitivity Analysis

In TSGM(which include NXGM and SLSGM), the number of total speakers is corresponding to time so that we can estimate the situation about each language in the next 50 years. While finishing constructing those models, we look back to think about different value of some parameters.

According to our assumption and calculating process, we find that TSGM depends on the selection of parameter β_0. We choose 4 different values, which are 0, 0.1, 0.2,0.3 respectively, and get the t-TS(t) curve about English by Matlab.

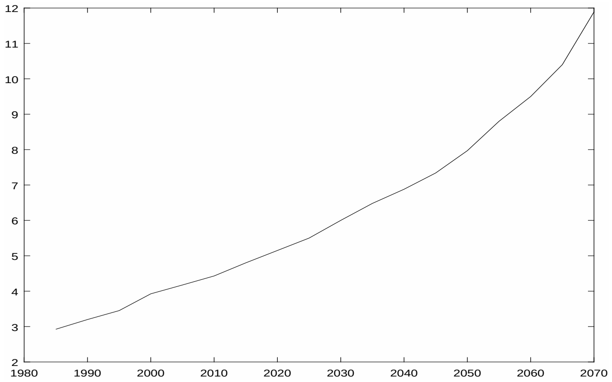


Figure 11: Fitting curve x_3(t)

3.6 Discussion

According to the derivation above, it is easy to find that the degree of impact in TSGM from those four factors are different a lot. Through observing the forecasting results and sensitivity analysis, we can discuss further and draw two conclusions below:

1.From figure 10, we can predict the top-ten lists in the next 50 years. As those figures show, Japanese will replace Malay in the future, and the other remain in the top-ten list.

2.Business relations is the most important factors of them. The reason for this is likely to say that business cooperation and economic communication promote people to learn more languages. Meanwhile, people prefer to learn other languages when it can earn money rather than other situation.

3.Based on the sensitivity of parameter p_0 in TSGM, we can find that the number of total speakers is related to p_0. As p_0 increases, the number of total speakers we estimate may decrease for a specific language. Therefore, we can say TS is sensitive with the value of p_0.

4 Language Distribution Growth Model(LDGM)

4.1 Construction of LDGM

In this section, we predict changes in geographic distributions of languages and focus on the geographic differences over time. Global population growth and human migration are the main reasons for the changes. So, we focus on these two reasons in this section. Besides, for simplicity we divide the world into six continents (except Antarctica) and consider these six continents separately. For each specific continent, we construct **language distribution growth model (LDGM)** to predict the number of total speakers of each language.

We know that human migration, global population growth and the enhancement of level of knowledge largely account for changes in geographic distributions of languages. So, firstly we consider the influences of migration. We use **analytic hierarchy process (AHP)** to determine the stable weight (SW) of each continent. 0<SW<1 and the larger SW of a continent is, the more attractive for living the continent is. Economic factors and environmental (climate) factors are the main determinants of whether the continent is attractive and suitable for living. So we consider economy and environment in our AHP.

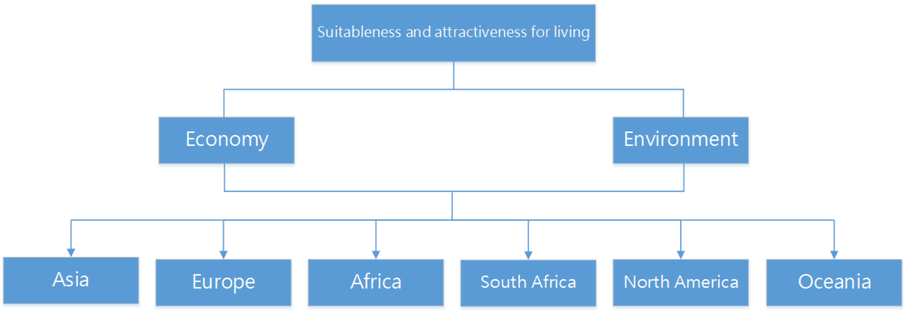


Figure 12: Analytical hierarchy process

We get SW for each continent and the sum is 1. So SW of a continent can be viewed as the stable proportion of the whole population to live in this continent. Then we define Stable Number (SN) for each continent:

$$SN = SW * TNP \tag{9}$$

where TNP is total number of people on earth in year t and we can estimate it with world natural population growth rate. If the number of people who live in a continent is larger than SN of that continent, then some people will migrate out to other continents, due to high living expenses or limited resources. Conversely, if the number of people who live in a continent is smaller than SN of that continent, then some people from other continents will migrate in. Then we consider its influence on number of total speakers. That People move out means fewer speakers are left in that continent. Let N_t represent total number of people in that continent in year t and we can use average natural population growth rate of that continent to estimate it. Let t_0 be the base year.

$$N_t = N_{t_0} * (1+\lambda)^{(t-t_0)} \tag{10}$$

where λ is the average natural population growth rate of the continent. Then we define Change in Number of Speakers caused by Migration(CNSM) in year t:

$$CNSM_t = (N_t - SN) * \alpha \tag{11}$$

where α is a parameter and determines how much migration will affect each language. For simplicity, we give each language equal weight so is the same for each language, and we do sensitive analysis on α later. If CNSM>0, people migrate out of this continent and thus the total speakers in this continent decrease.

Then we consider the influences of population growth and the enhancement of the level of knowledge. The population growth mainly account for the increase in native speakers, while the enhancement of level of knowledge mainly account for the increase in second (or third, etc.) language speakers. We use average natural population growth rate of a continent to account for population growth, and use growth rate caused by **importance extent (IE)** of that language to account for the enhancement of level of knowledge. Consider the status of each language will not change within 50 years, we assume that IE of each language don't change over time. The larger IE is, the more important the language is and the more people use it, which means more people are willing to learn it and thus the faster the total speakers of this language increase. So we can get IE of each language by the current number of SLS of that language over total number of SLS of all languages.

$$IE^A = \frac{current\ number\ of\ SLS\ of\ A}{current\ number\ of\ SLS\ of\ all\ languages} \tag{12}$$

We define adjusted increase rate (AIR) of language A:

$$AIR^A = (1+\lambda)(1+\frac{IE^A}{\beta} - 1) \tag{13}$$

where λ is the natural population growth rate of that continent and β is the adjusting parameter (β>1). We do sensitivity analysis on β later.

Let ${TN_t}^A$ donate total number of speakers of language A in year t. Taking all the factors above into account, we have:

$${TN_{t+1}}^A = {TN_t}^A * (1 + AIR^A) - CNSM_t \tag{14}$$

namely:

$${TN_{t+1}}^A = {TN_t}^A * (1 + \lambda)(1 + \frac{IE^A}{\beta}) - (N_t - SN) * \alpha \tag{15}$$

4.2 Forecasting Results

For simplicity we only consider current top 10 languages.

Firstly, we calculate SW for each continent by AHP.

Consider that economy and environment are both important factors for living, they should have roughly equal weight. But consider that nowadays people focus a lot on health, so we think that environment is slightly more important than environment.

Table 2: AHP_1

	economy	environment	weight
economy	1	1/3	0.25
environment	3	1	0.75

Then we construct comparison matrix between continents, according to the general rank of economic development.

Economy:

Table 3: AHP_2

	Asia	Europe	Africa	South America	Nouth America	Oceania	Weight
Asia	1	1/3	4	2	1/5	1	0.1045
Europe	3	1	7	5	1/2	2	0.2504
Africa	1/4	1/7	1	1/2	1/9	1/5	0.0318
South America	1/2	1/5	2	1	1/	1/3	0.0531
North America	5	2	9	7	1	4	0.4313
Oceania	1	1/2	5	3	1/4	1	0.1288

CI=0.0186. When n=6, RI=1.24, so CR=CI/RI=0.0150<0.1. Therefore, the examination of consistency is passed.

Environment:

Table 4: AHP_3

	Asia	Europe	Africa	South America	Nouth America	Oceania	Weight
Asia	1	1/7	2	1/2	1/5	1/6	0.0465
Europe	7	1	9	5	3	2	0.4011
Africa	1/2	1/9	1	1/3	1/6	1/7	0.0314
South America	2	1/5	3	1	1/3	1/4	0.0766
North America	5	1/3	6	3	1	1/2	0.1787
Oceania	6	1/2	7	4	2	1	0.2656

CI=0.0263. When n=6, RI=1.24, so CR=CI/RI=0.0212<0.1. Therefore, the examination of consistency is passed.

Then we can calculate SW for each continent.

Table 5: Stable Weight

	Asia	Europe	Africa	South America	Nouth America	Oceania
SW	0.0900	0.2881	0.0317	0.0590	0.3682	0.1630

The results show North America, Europe and Oceania are popular continents for living, which is consistent with the case in real life.

Then we find natural population growth rate for each continent in the Statista Database.

Table 6: Natural population growth rate

	Asia	Europe	Africa	South America	Nouth America	Oceania	World
λ	0.011	0.00	0.025	0.012	0.004	0.011	0.012

Using the data of SLS from Wikipedia[9] we can calculate IE value for each language.

Table 7: Importance Extent

	Mandarin Chinese	Spanish	English	Hindustani	Arabic	Bengali	Portuguese	Russian	Punjabi	Japanese
IE	0.1392	0.0657	0.4408	0.1551	0.0952	0.0137	0.0079	0.0815	0	0.0007

With the data of SLS from Wikipedia [10] and the language distribution picture [11] we can estimate the number of speakers of each language for each continent.

Then we apply the model (LDGM), and get the predicting results. In the forecasting process, we let α=0.1, β=15 and θ=0.5. Later we will do sensitivity analysis on the three parameters and justify their values.

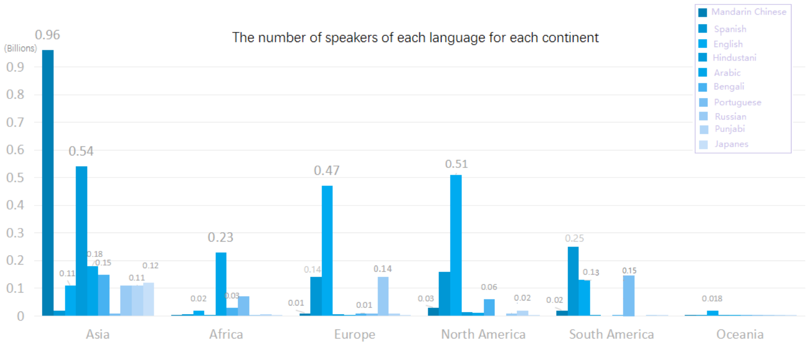


Figure 13: Current geographic distribution of languages

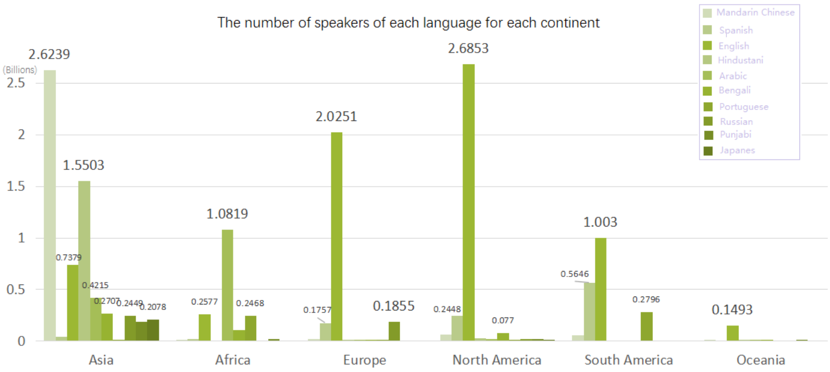


Figure 14: Forecasted geographic distribution of languages

From the forecasting results, we find the number of speakers of each language grows more or less, which results from the population growth and enhancement of level of knowledge. We can see that Chinese, English and Hindustani will be the most popular languages. Large numbers of people speak Chinese and Hindustani due to the large population of China and Hindu. While for English, English is the official language of a lot of countries and play an irreplaceable role in international communication, so many people will choose to learn English first if English is not their first language. That also accounts for the great speed of growth of English speakers.

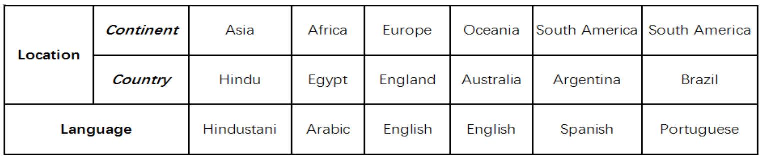
Comparing the geographic distribution currently and that of 50 years later, we can find the percentage of speakers of many languages in each continent remain almost unchanged, except English. From the forecasting results, we can find that English will be dominant around the world and large numbers of people can speak English in each continent. The important status of English and the development of education account for the rapid growth of speakers of English. Since English is of vital importance in international communication and cooperation, many people will rank English first if their native language is not English. So, the enhancement of level of knowledge will affect number of English speakers most, which results in considerable increase in number of English speakers. Additionally, we can find that numbers of speakers of languages all rise a lot. Aside from population growth, the development of education and people's awareness of importance of learning second or even third language may account for it.

4.3 Discussion of the locations of the offices

In this part, we talk about appropriate locations for new offices to offer advice for the company. We consider it both in the short term and in the long term.

In the short term

Firstly, for the further development and globalization of the company, we should set at least one office in each continent (except Antarctica). Considering that we already have one office in China and one in the US and we are going to set six more offices, after giving each continent (except Antarctica), we have two offices left. So, we have to choose two continents to give each of them one extra office. Based on the results in Figure 13, we know the dominant languages for each continent. If one continent has two dominant languages, then we set one extra office there. Also taking economic factors into account, we set offices and specify the languages that should be spoken there. The results are as follows. (Two offices in Shanghai, China and New York, the US will not appear in the form.)

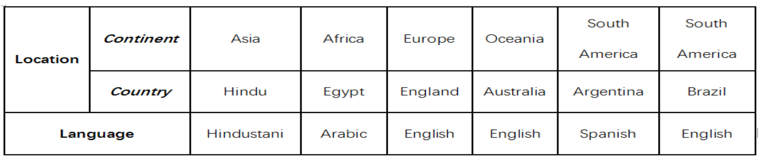


Location	*Continent*	Asia	Africa	Europe	Oceania	South America	South America
	Country	Hindu	Egypt	England	Australia	Argentina	Brazil
Language		Hindustani	Arabic	English	English	Spanish	Portuguese

Figure 15: Suggestions in the short term

In the long term

Based on our forecasting results, we know that English will be popular and even dominant in each continent. Besides, large number of people speak Chinese and Hindustani in Asia. Similar to our analysis in the short term and based on the results in Figure 14, the results are as follows. (Two offices in Shanghai, China and New York, the US will not appear in the form.)



Location	*Continent*	Asia	Africa	Europe	Oceania	South America	South America
	Country	Hindu	Egypt	England	Australia	Argentina	Brazil
Language		Hindustani	Arabic	English	English	Spanish	English

Figure 16: Suggestions in the long term

4.4 Sensitivity Analysis

In this part we do sensitivity analysis on α, β, θ respectively. Those are the parameters we introduce in LDGM.

Sensitivity analysis on α

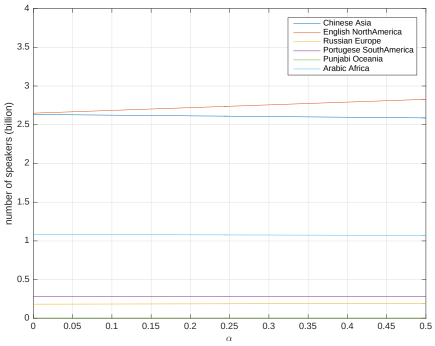


Figure 17: Sensitivity analysis on α

From Figure 17 we can see that more number of speakers lead to larger sensitivity. Number of speakers of English is most sensitive to the value of since English is the most popular language currently and will be popular for a long time. As we mention before, denotes the migration effect on number of speakers. Since English is most widely spoken language around the world, the change in value of will affect the number of speakers of English most. But generally the sensitivity of is acceptable. Therefore it is reasonable for us to choose α=0.1 to forecast.

Sensitivity analysis on θ

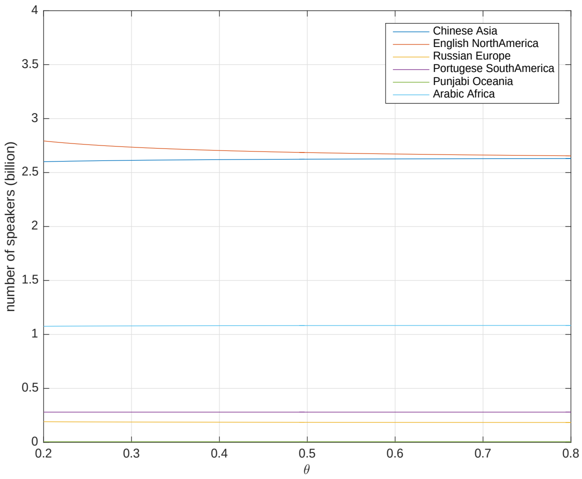


Figure 18: Sensitivity analysis on θ

From Figure 18 we can find that more speakers result in larger sensitivity. But generally the change in number of speakers caused by the change of value of is very tiny, so the sensitivity of θ is also acceptable. Therefore it is reasonable for us to choose θ=0.5 to forecast.

Sensitivity analysis on β

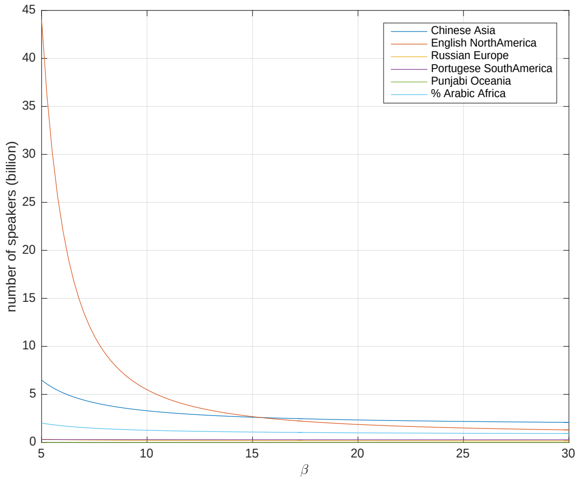


Figure 19: Sensitivity analysis on β

From Figure 19 we can find that the number of speakers that we predicted is sensitive to the value of β, especially when β is small. Also, the more number of speakers, the more sensitive. As we mention before, β is set to adjust IE, the important extent of each language. So, the change of value of β will affect the most important languages, namely most widely spoken languages. We can see from figure 19 that English and Chinese are sensitive to β, especially English. Since English is the official language for many countries and play an irreplaceable role in international business, majorities of people will choose to learn English if English is not their native language. So, it the value of β is small, which means the influence of enhancement of level of knowledge is significant on the increase of number of speakers, the number of speakers of English will definitely rise rapidly. Therefore, we need to be careful when choosing the value of β and value of β that make results relatively robust is preferred. This accounts for the reason why we choose β=15 to do forecasting.

5 Conclusions

In this paper, we construct two main models to predict number of speakers of each languages and changes of geographic distribution of languages over time respectively. Firstly, we construct TSGM to predict total speaker growth, which consists of two sub-models, NSGM and SLSGM. NSGM is designed to predict native speaker growth and SLSGM is designed to predict second (or third, etc.)

language speaker growth. We apply logistic growth model in NSGM and multiple regression analysis in SLSGM, which do well in projection. Then we get TSGM after combing NSGM and SLSGM. With these models, we forecast the number of native speakers and total language speakers in the next 50 years and we predict that Japanese will replace Malay to be in the top 10 list.

Then we construct LDGM to forecast changes in geographic distribution of languages, taking human migration, global population growth and the enhancement of level of knowledge into account. With this model, we forecast the number of speakers of top 10 languages in each continent and do comparisons between continents to find changes in geographic distribution. We predict that English will be the dominant language around the world in the future. Based on forecasting results, we offer advice on the locations of the offices both in the short term and in the long term.

Strengths

1.TSGM consists of two sub-models, which forecast number of native speakers and the number of second (or third, etc.) language speakers respectively. Considering that total speakers are made up of these two parts, and these two parts are influenced by different factors, our TSGM is more precise in forecasting.

2.TSGM and LDGM both use available data on the Internet to forecast, and perform quite well.

3.TSGM and LDGM both take quite a lot of factors that will influence the number of speakers into account, which means our models almost reach every aspect of this matter.

Weaknesses

1.In our LDGM model, we divide the whole world into six continents and do forecasting respectively to find the changes of geographic distributions of languages. But a continent is still very large, which means we dismiss the internal changes in each continent.

2.We make many assumptions in forecasting, which may deviate from reality. Without some assumptions, our models may perform not so good.

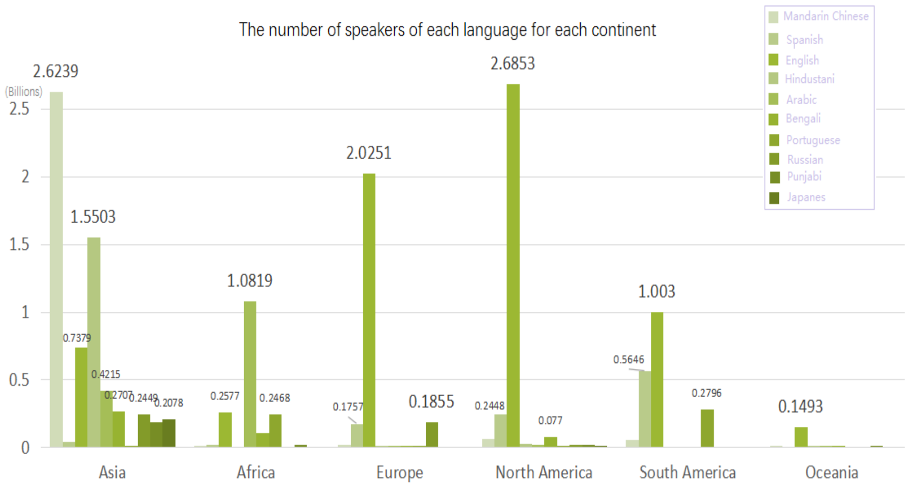
6 Letter

Letter

Dear Chief Operating Officer,

It is our pleasure to offer your company suggestions. We have constructed two models to forecast the distribution of speakers over time. These two models focus on time and geographic distribution respectively. We are going to offer some advice based on the forecasting results of our models in the following passage.

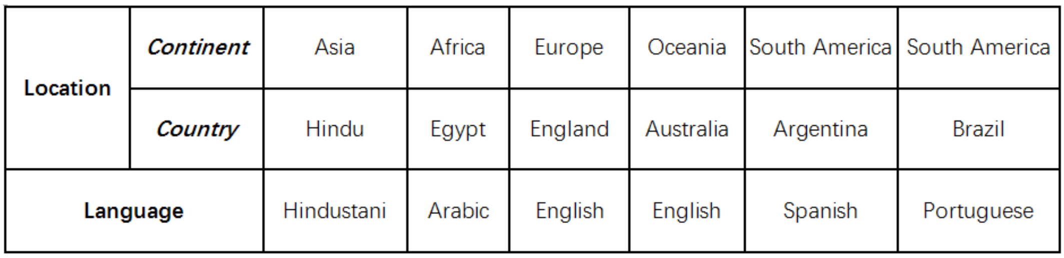
Based on the forecasting results, English will definitely become the dominant language and the most widely spoken language around the world. The following figure is our prediction of number of speakers of each languages in each continent.



From the figure, we can see that English is popular in each continent. Since English is the official language of many countries and plays an irreplaceable role in international communication, with the acceleration of globalization, increasing people will learn English if their native language is not English, especially considering the development of education and people'ps awareness of importance of learning a second or third language. Besides, quite a large number of people speak Mandarin Chinese and Hindustani in Asia, which results from the large population and rapid population growth rate of China and Hindu.

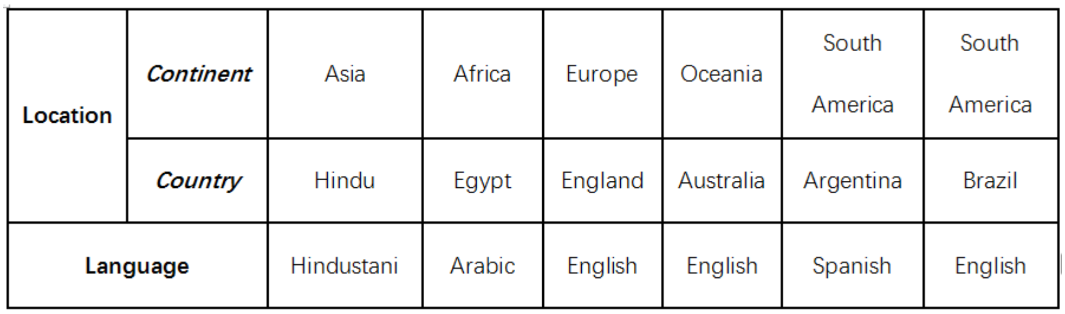
For the locations of new offices, we offer different suggestions in the short term and in the long term. Suggestions in the short term are mainly based on current geographic distribution of languages, while suggestions in the long term are mainly based on forecasted geographic distribution of languages 50 years later. Considering that we already have two offices in Shanghai, China and New York, the US respectively, the results are as follows.

In the short term:



Location	*Continent*	Asia	Africa	Europe	Oceania	South America	South America
	Country	Hindu	Egypt	England	Australia	Argentina	Brazil
Language		Hindustani	Arabic	English	English	Spanish	Portuguese

In the long term:



Location	*Continent*	Asia	Africa	Europe	Oceania	South America	South America
	Country	Hindu	Egypt	England	Australia	Argentina	Brazil
Language		Hindustani	Arabic	English	English	Spanish	English

There is only slight difference between suggestions in the short run and in the long run. However, when thinking in the long term, we suggest that fewer than six offices will be set to save resources and cost, considering the dominance of English. Some additional information about the company is needed. We need to know the production of the company, the structure of the company, the objects of the company to determine the appropriate number of offices to be set. With the information, We are able to analyze whether one location is suitable for the development of the company and the prospects of setting an office there, so as to determine the number of offices to be set and where to set.

Thank you for reading our suggestions. We sincerely hope that our suggestions will be helpful.

Sincerely,
Team #82898

References

[1] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers/

[2] https://www.sciencedirect.com/science/article/pii/S0025556402000962#BIB1/

[3] Jennifer M. Ortman, and Hyon B. Shin, 2011, Language Projections: 2010 to 2020, the Annual Meetings of the American Sociological Association, 2011

[4] https://www.populationpyramid.net/world/2068/

[5] https://en.wikipedia.org/wiki/Human_migration/

[6] https://www.statista.com/

[7] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[8] https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

[9] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

[10] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

[11] https://en.wikipedia.org/wiki/Template:Distribution_of_languages_in_the_world

Appendices

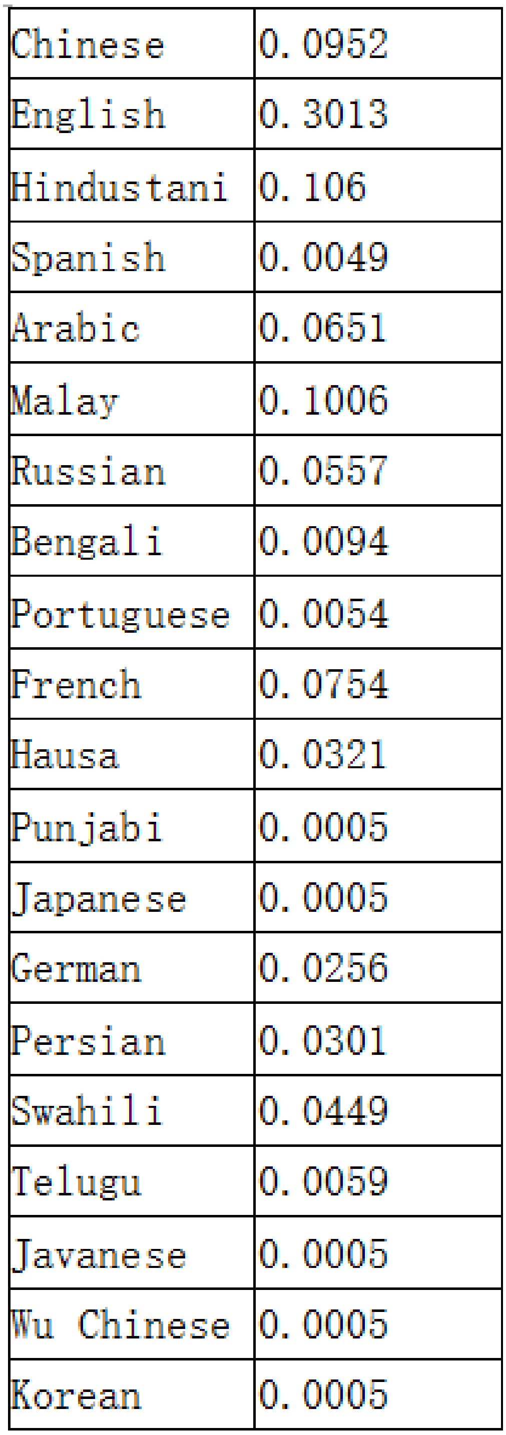
Unit: billion

Table 8: Current geographic distribution of languages

	Asia	Africa	Europe	South America	Nouth America	Oceania
Mandarin Chinese	0.96	0.002	0.01	0.03	0.02	0.001
Spanish	0.02	0.005	0.14	0.16	0.25	0.001
English	0.11	0.02	0.47	0.51	0.13	0.018
Hindustani	0.54	0.001	0.007	0.013	0.001	0.001
Arabic	0.18	0.23	0.001	0.011	0.001	0.001
Bengali	0.15	0.03	0.01	0.06	0.001	0.002
Portuguese	0.01	0.07	0.01	0.01	0.15	0.001
Russian	0.11	0.001	0.14	0.01	0.001	0.001
Punjabi	0.11	0.007	0.009	0.02	0.001	0.003

Table 9: Forecasted geographic distribution of languages

	Asia	Africa	Europe	South America	Nouth America	Oceania
Mandarin Chinese	2.6239	0.0063	0.0196	0.0628	0.0573	0.0044
Spanish	0.0396	0.0196	0.1757	0.2448	0.5646	0.0027
English	0.7379	0.2577	2.0251	2.6853	1.0030	0.1493
Hindustani	1.5503	0.0003	0.0160	0.0320	0.0027	0.0049
Arabic	0.4215	1.0819	0.0037	0.0212	0.0023	0.0033
Bengali	0.2707	0.1076	0.0107	0.0770	0.0019	0.0037
Portuguese	0.0174	0.2468	0.0104	0.0127	0.2796	0.0018
Russian	0.2449	0.0022	0.1855	0.0183	0.0022	0.0030
Punjabi	0.1901	0.0241	0.0090	0.0244	0.0018	0.0052
Japanese	0.2078	0.0034	0.0020	0.0049	0.0018	0.0017



Chinese	0. 0952
English	0. 3013
Hindustani	0. 106
Spanish	0. 0049
Arabic	0. 0651
Malay	0. 1006
Russian	0. 0557
Bengali	0. 0094
Portuguese	0. 0054
French	0. 0754
Hausa	0. 0321
Punjabi	0. 0005
Japanese	0. 0005
German	0. 0256
Persian	0. 0301
Swahili	0. 0449
Telugu	0. 0059
Javanese	0. 0005
Wu Chinese	0. 0005
Korean	0. 0005

Figure 20: Proportion of top-twenty languages