Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
B		000111

An MCM Paper Made by Team 0000111

Summary

Here is the abstract of your paper.

Firstly, that is ...

Secondly, that is ...

Finally, that is ...

Contents

1 Introduction

1.1 Problem Background

It is generally taken for granted that language, as a concomitant of culture, can spread. With the trend of globalization and the worlds cultural exchange, language transfer and integration are also more common. Nowadays more and more people can speak two or even more languages. The shift and spread of language can be seen through the amount of speakers, including native speakers plus second or third, etc. language speakers. However, the total number of speakers of a language fluctuates under the influence of various complicated factors. These factors involve political, economic, diplomatic, social relations and other aspects, such as:

- government-mandated official languages.
- tourism among nations.
- migration and population movements
- the promotion of new social media (facebook, Twit-ter, etc.)

and so on.

As known to all, nearly 7000 languages are spoken over the world, and they make up the communication network through hundreds of countries and regions. Languages are essential to construct foreign trade, develop tourism and promote scientific and technological progress, which makes it an indicator and an effective tool to measure a countrys comprehensive power. Also, a measurement of the utility of a particular language is the number of speakers who use it as native or the second or third language. Therefore, it should be taken attention that the number of speakers of a particular language would change over times with the languages rise and fall as it may be coincident with the economic and political development of its main country. For now, ten languages are claimed to use by half the worlds population, which includes Mandarin (incl. Standard Chinese), Spanish, English, Hindi, Arabic, Bengali, Portuguese, Russian, Punjabi, and Japanese. And the number of speakers of one language would be influenced by migration, social pressures, business relations, social media and so on. It is necessary for us to find out its variation and trends in the future to expect their rankings and make better use of them.

1.2 Restatement of the problem

We are required to predict the spread and development of languages all over the world under the influence of several factors and help a large multinational service company to determine the locations of new offices. The problem can be analyzed into three parts:

1. Develop a model of the distribution of various language speakers over time based on impact factors and predict what will happen to the number of speakers of each language in the next 50 years.

2. Use the model to predict the geographic distributions of languages in the next 50 years.

3. Determine the locations of new international offices and the languages used in the new offices based on the modeling results.

2 Preparation of the Models

2.1 Assumptions

2.2 Notations

The primary notations used in this paper are listed in Table 1.

Table 1: Notations

Symbol	Definition
A	the first one
b	the second one
α	the last one

3 The Models

3.1 Model 1

3.1.1 Detail 1 about Model 1

The detail can be described by equation (1):

$$\frac{\partial u}{\partial t} - a^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) = f(x,y,z,t) \tag{1}$$

3.2 Location Model

To determine the location of the new offices, we build a location model. Since international branch offices are usually set in well developed countries and countries with potential. On the basis of countries' GDP and their potential, we selected 50 countries as candidates.

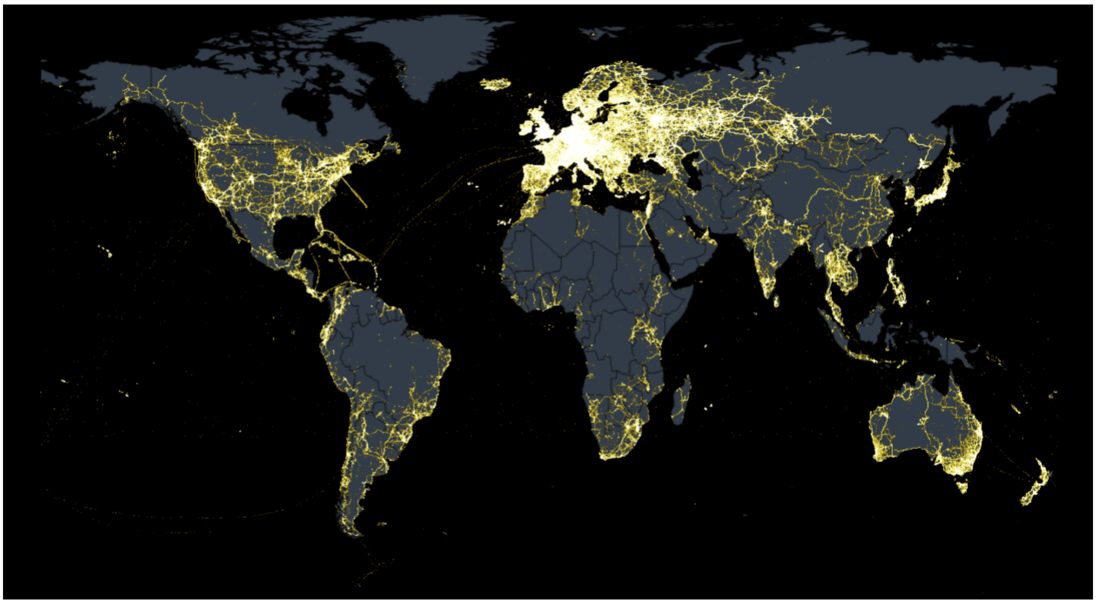

Figure 1: Development in different parts of the world

This figure reflects the development of each area in the world.

1	China	11	Mexico	21	France	31	Ukraine	41	Malaysia
2	India	12	Philippines	22	United Kingdom	32	Argentina	42	Venezuela
3	United States	13	Ethiopia	23	Italy	33	Algeria	43	Peru
4	Indonesia	14	Vietnam	24	Burma	34	Poland	44	Uzbekistan
5	Brazil	15	Egypt	25	South Africa	35	Uganda	45	Nepal
6	Pakistan	16	Iran	26	Tanzania	36	Iraq	46	Saudi Arabia
7	Nigeria	17	Congo, Dem. Rep.	27	Korea, South	37	Sudan	47	Yemen
8	Bangladesh	18	Germany	28	Spain	38	Canada	48	Ghana
9	Russia	19	Turkey	29	Colombia	39	Morocco	49	Mozambique
10	Japan	20	Thailand	30	Kenya	40	Afghanistan	50	Korea, North

Figure 2: 50 candidate countries

Since China and USA had been selected as the office locations, our model is for the remaining 48 countries.

We use the topsis method to determine the office location, and use the entropy method, CRITIC method, and Standard deviation method to set the corresponding weights for each indicator. After the office location is initially determined, we use k-means clustering analysis to determine the office service area. After that, a series of indicators were used to re-score the service effect of different office locations and office areas, and comprehensively determine the best office locations and number. Here are the flow chart of our model:

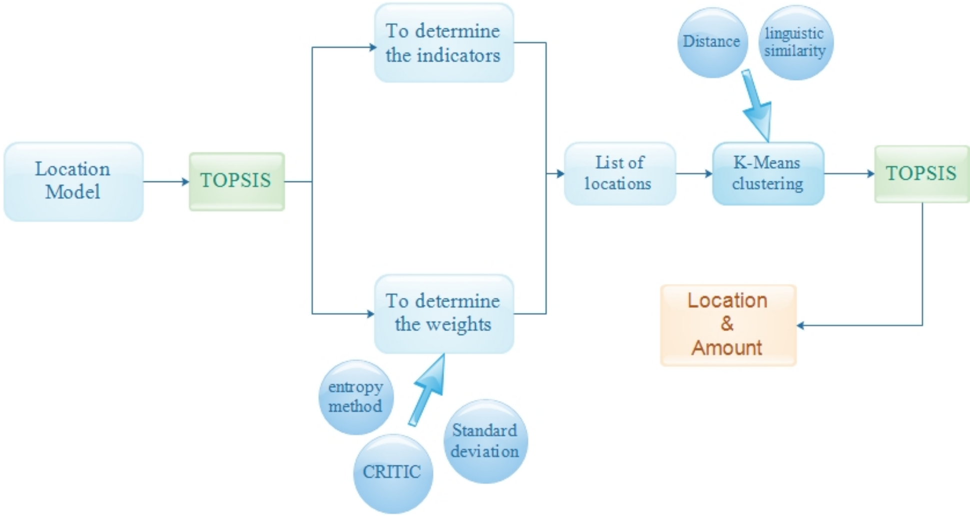

Figure 3: flow chart of our model

3.2.1 Weighted-Topsis

The Topsis method was first proposed by C.L.Hwang and K.Yoon in 1981. The TOPSIS method is based on the closeness of a limited number of evaluation objects to the idealized target. It is a relatively good evaluation of the existing objects. The TOPSIS method is a sorting method that approximates the ideal solution. This method only requires that the utility functions be monotonically increasing (or decreasing). TOPSIS method is a commonly used effective method in multi-objective decision analysis, also known as the pros and cons solution distance method. The basic principle is to sort by detecting the distance between the evaluation object and the optimal solution and the worst solution. If the evaluation object is closest to the optimal solution and farthest from the worst solution, it is the best; otherwise it is not optimal. The value of each index of the optimal solution reached the optimal value of each evaluation index. The values of each index of the worst solution reached the worst value of each evaluation index.

In TOPSIS method, "ideal solution" and "negative ideal solution" are the two basic concepts of TOPSIS method. The so-called ideal solution is a conceived optimal solution (scheme), each attribute value of which reaches the best value of each alternative; the negative ideal solution is a conceived worst solution (scheme), it Each attribute value of is the worst value in each alternative. The rule for ordering the schemes is to compare the alternatives with the ideal solution and the negative ideal solution. If one of the schemes is closest to the ideal solution and at the same time is far away from the negative ideal solution, the scheme is the best solution among the alternatives.

In the Weighted-Topsis method in our model, there are 2 types of the indicators:

Table 2: 2-types of the indicators in our model

Benefit indicator	The goal is to maximize the indicator
Cost indicator	The goal is to miniimize the indicator

According to the indicators given in the title and our considerations, we finally chose 6 indicators. They reflect a series of comprehensive factors such as market, cost, language, service range, etc., and can well consider whether an address is suitable as an office. Here are our indicators:

Indicators	Implication	type
Proportion of national service workers	Measuring the vitality of the service industry	Benefit indicator
Average monthly salary of workers	Measuring office operating costs	Cost indicator
National GDP	Measuring the size of a country's market	Benefit indicator
Proportion of English speakers	Measure the suitability of office location	Benefit indicator
Proportion of office-selected language speakers	Measure the suitability of office location	Benefit indicator
Size of the office-service area	Measure the suitability of office location	Cost indicator

Figure 4: indicators of our model

The Proportion of national service workers, Average monthly salary of workers, National GDP and Proportion of English speakers are used in our first layer of the topsis, and the Proportion of office-selected language speakers and Size of the office-service area are added in the second layer of the topsis after the area has been clustered by K-means method.

The following paragraphs describe the steps to use topsis on our model.

1. **Unified indicator type.**

   Turn all indicators into benefit indicators. The second indicator and the last indicator of our model are the cost indicator, we need to turn them into benefit indicator by using:

$$\max - x \tag{2}$$

   After this, we can get forward matrix X.

2. **Standardized forward matrix**

   In order to eliminate the influence of different index dimensions, the matrix needs to be processed.

   Assume that there are n evaluation objects and m evaluation indexes that have been forwarded. The forward matrix formed is as follows:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{3}$$

   The matrix to which it is normalized is called Y, and for each element y_{ij} in Y:

$$y_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} {x_{ij}}^2}} \tag{4}$$

3. **Calculate weighted matrix**

Because the impact of each indicator is different, we need to assign a value to each indicator. The weighted matrix is called Z, and each element in Z:

$$z_{ij} = w_i \times y_{ij} \tag{5}$$

We will discuss the method of empowerment later.

4. Calculate score and normalize

After the above steps, we get the weighted normalized matrix Z:

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{bmatrix} \tag{6}$$

Define maximum value Z^+ and minimum value Z^-:

$$\begin{aligned} &Z^+ = ({Z_1}^+, {Z_2}^+, \cdots {Z_m}^+) \\ &= (\max\{z_{11}, z_{21}, \cdots, z_{n1}\}, \max\{z_{12}, z_{22}, \cdots, z_{n2}\}, \cdots, \max\{z_{1m}, z_{2m}, \cdots, z_{nm}\}) \\ &Z^- = ({Z_1}^-, {Z_2}^-, \cdots {Z_m}^-) \\ &= (\min\{z_{11}, z_{21}, \cdots, z_{n1}\}, \min\{z_{12}, z_{22}, \cdots, z_{n2}\}, \cdots, \min\{z_{1m}, z_{2m}, \cdots, z_{nm}\}) \end{aligned} \tag{7}$$

Define the distance between the i-th $(i = 1, 2, \cdots, n)$ evaluation object and the maximum value as ${D_i}^+$:

$${D_i}^+ = \sqrt{\sum_{j=1}^{m} \left({Z_j}^+ - z_{ij}\right)^2} \tag{8}$$

Define the distance between the i-th $(i = 1, 2, \cdots, n)$ evaluation object and the minimum value as ${D_i}^-$:

$${D_i}^- = \sqrt{\sum_{j=1}^{m} \left({Z_j}^- - z_{ij}\right)^2} \tag{9}$$

Then we can calculate the score of the i-th $(i = 1, 2, \cdots, n)$ evaluation object S_i:

$$S_i = \frac{{D_i}^-}{{D_i}^- + {D_i}^+} \tag{10}$$

Next we explain how to weight indicators.

Entropy Method, Standard Deviation, CRITIC—The principle of these three methods is based on the degree of variation of the indicator. When the degree of variation of the indicator is smaller, the amount of information reflected is less and the corresponding weight is lower.

Since the weights given in one way may have large deviations, we combine the three methods to give weights, and analyze the deviations to give the index weights in two ways, which give less deviation.

- **Entropy Method**

Let m indicators of n evaluation objects have been normalized as $y_{ij}\,(i=1,2,\cdots,n; j=1,2,\cdots m)$. The information entropy of the j-th index is:

$$E_j = -\frac{\sum\limits_{i=1}^{n} p_{ij}\ln p_{ij}}{\ln n}\,(i=1,2,\cdots,n; j=1,2,\cdots,m) \tag{11}$$

and $p_{ij} = \frac{y_{ij}}{\sum\limits_{i=1}^{n} y_{ij}}$. When E is smaller, the difference between the data is larger, so the larger the information provided, the greater the weight of the indicator, and vice versa.

Then we can get the calculation formula of objective weight:

$$w_j = \frac{1-E_j}{m-\sum\limits_{j=1}^{m} E_j}(j=1,2,\cdots m) \tag{12}$$

- **Standard Deviation**

  Unlike the calculation of information entropy, in the standard deviation method, we use the method of calculating standard deviation to measure the amount of information provided by an indicator. When the standard deviation is large, we think that it provides more information. When the standard deviation is small, we think it provides less information.

  The calculation as following:

$$w_j = \frac{\sigma_j}{\sum\limits_{j=1}^{m} \sigma_j}(j=1,2,\cdots,m) \tag{13}$$

  where σ_j represents the standard deviation.

- **CRITIC**

  The CRITIC method weights indicators based on two basic concepts: The first is contrast. When the standard deviation is larger, the weight is relatively larger. The second is to evaluate the conflict between indicators. Here we introduce the correlation coefficient r between indicators. When there is a strong positive correlation between indicators, it means that the conflict between the two indicators is low, and the information reflected by the two indicators is relatively similar. When there is a strong negative correlation between the two indicators, it means that the conflict between the two indicators is large, and the information reflected by the two indicators is quite different.

  The calculation as following:

  The amount of information contained in the j-th indicator is:

$$c_j = \sigma_j \sum_{i=1}^{m} (1-r_{ij})(j=1,2,\cdots,m) \tag{14}$$

  where r_{ij} represents the correlation coefficient between the i-th indicator and j-th indicator.

  Then we get the j-th indicator's weight:

$$w_j = \frac{c_j}{\sum\limits_{i=1}^{m} c_i}(j=1,2,\cdots,m) \tag{15}$$

3.2.2 Weighted K-Means Clustering

After the first usage of the Weighted-Topsis, we get the score of each country among the 50-countries' list. Considering that each office needs to serve an area, we introduce Weighted K-Means Clustering to determine the area that one office needs to serve.

K-means is a certain distance from the data point to the prototype as the objective function of the optimization, and the method of calculating the extreme value of the function is used to obtain the adjustment rule of the iterative operation. The K-means algorithm uses Euclidean distance as a similarity measure. It seeks the optimal classification corresponding to a certain initial clustering center vector MC, so that the evaluation index D is minimized. The algorithm uses the sum of squared error function as the clustering criterion function. Eventually, the obtained clusters satisfy the similarity of objects in the same cluster, and the cluster center and the objects assigned to them represent a cluster.

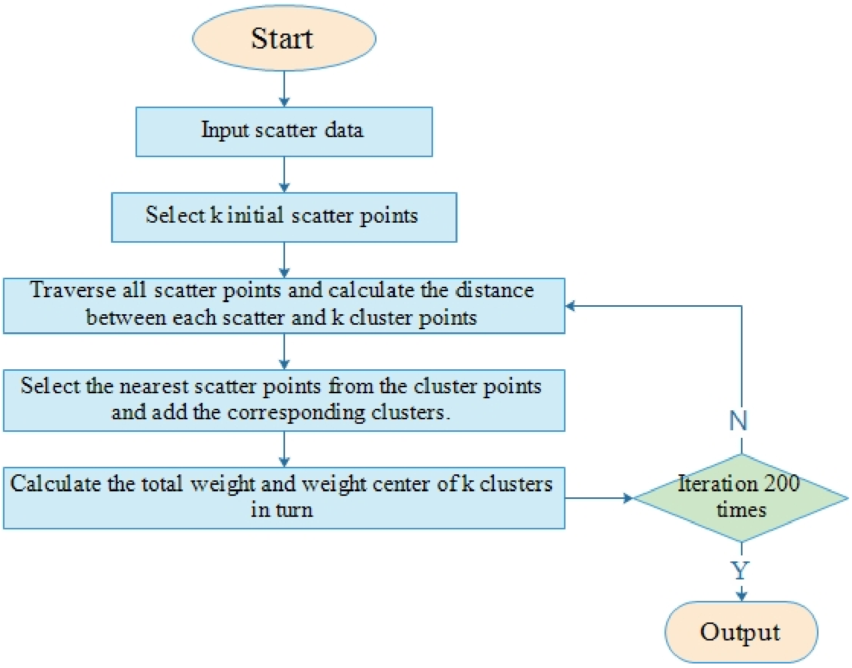

Figure 5: flow chart of K-Means Clustering

The core formula of the K-means algorithm:

$$D\left(\{\pi_c\}_{c=1}^{k}\right) = \sum_{c=1}^{k} \sum_{a,ex_c} \|a_i - m_c\|^2, \text{where} m_c = \frac{\sum_{a,\in\pi_c} a_i}{|\pi_c|} \tag{16}$$

$a_{ij}, (i = 1,2,...,n; j = 1,2,...,m)$ refers to all of the elements. And here m=2, means the longitude and latitude, which is the location of the center point(At first is the office selected.). π_c means the cluster group and the total number is k. m_c is the cluster center of the element a_i in group π_c. After continuous iteration we can get the final clustering result.

Through the k-means algorithm, we can obtain a good regionalization scheme from the initial center point and continuous iteration. Here is an example:

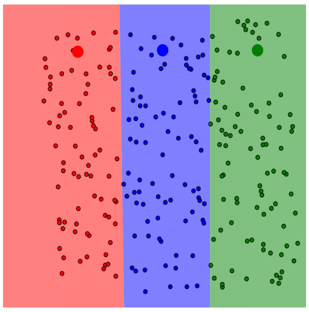

Figure 6: Init

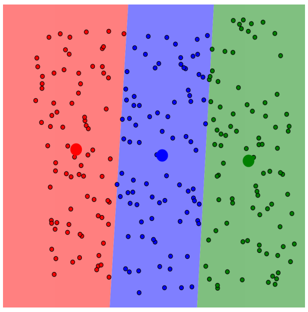

Figure 7: After 10 iterations

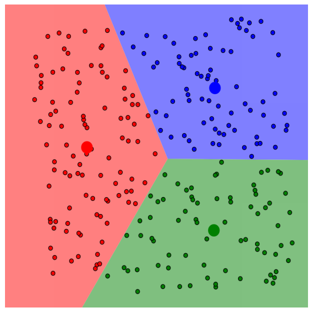

Figure 8: After 50 iterations

From the example we can see the cluster center in the final cluster compared to the original has a large offset. This means the K-Means clustering is not enough to solve our location model. Since the K-means clustering model uses a 1: 1 ratio for the distance comparison of each scattered point set, but for us, we need to consider not only the straight line distance between the capital and the cluster center, but also to the language similarity and other factors. If we use distance as the sole criterion, then it is better to define the service area according to the radiation range centered on the country. Because then the office location we originally selected will not change. However, Based on our considerations, the reason we apply the k-means algorithm is as follows:

- Introduce language similarity as weights to help us flexibly divide regions.
- Judging the correctness and appropriateness of the initial choice by observing the change in the center point. If the center point has a large deviation, it means that the initially selected point is not suitable as an office. Then we can consider changing the location or country for better results.

Based on the above ideas, we apply the Weighted K-Means Clustering to solve our model.

The core formula of the Weighted K-means Clustering:

$$D\left(\{\pi_c\}_{c=1}^{k}\right)=\sum_{c=1}^{k}\sum_{a_i\in\pi_c} w_i^y\|a_i-m_c\|^2, \quad \text{where} m_c=\frac{\sum\limits_{a_i e\pi_c} w_i^y a_i}{\sum\limits_{a_i\in\pi_c} w_i^y} \tag{17}$$

The calculation method of this formula is very similar to the k-means algorithm, except that for different sample points, a corresponding weight wi and a weight attenuation coefficient y are multiplied. Here we set the y=1.

3.2.3 Solution and Analysis of Location Model

Via the data we collected, we quantified six indicators and normalized these indicators. Running the Entropy Method, Standard Deviation, CRITIC respectively. We get three different weight vectors as following:

$$\text{Entropy Method:} \begin{bmatrix} 0.22 \\ 0.25 \\ 0.31 \\ 0.22 \end{bmatrix}^{-1}$$

$$\text{Standard Deviation: } \begin{bmatrix} 0.14 \\ 0.39 \\ 0.32 \\ 0.15 \end{bmatrix}^{-1}$$

$$\text{CRITIC: } \begin{bmatrix} 0.30 \\ 0.15 \\ 0.35 \\ 0.20 \end{bmatrix}^{-1}$$

We calculated the rankings obtained by topsis under the three kinds of weights, and finally selected 12 countries. Considering that a country's capital often best reflects its economic level. At the beginning, we assumed that the national capital was the office location. As for the choice of language: we assume the most spoken language in the region other than English as the language the office selected to use.

Country	China	US	Australia	Brazil	Germany	Russian
City	Shanghai	New York	Canberra	Brasilia	Berlin	Moscow
Language	Mandarin	Spanish	Spanish	Spanish	German	Russian
Country	Italy	India	Saudi Arabia	Nigeria	Canada	France
City	Rome	New Delhi	Riyadh	Abuja	Ottawa	Paris
Language	German	Hindi	Arabic	Arabic	French	French

Figure 9: Selected location by topsis

Because China and the United States have been selected as office locations, we just need to consider the other 10 regions. The selection of China and the United States also proved the correctness of the first site selection in our model.

After the Weighted-Topsis, we tried the choices of 6 different countries among lists and brought them into the k-means cluster with distance and language consistency and country relations as weights to divide the service area for the selected office.

Later, the regional division was completed, we brought the next two indicators to recalculate the score of topsis, and selected the highest 6 locations in different divisions by clustering.

After multiple assignments we found that when we use Germany, Brazil, Australia, Nigeria, Russian and India as the locations of the office, we can receive the highest benifit and lowest cost. The deviation of the cluster center is also little, which means the capital is suitable for the location. And we find that when we use different combination, the 6 highest average scores are also these countries. Here is the score in average:

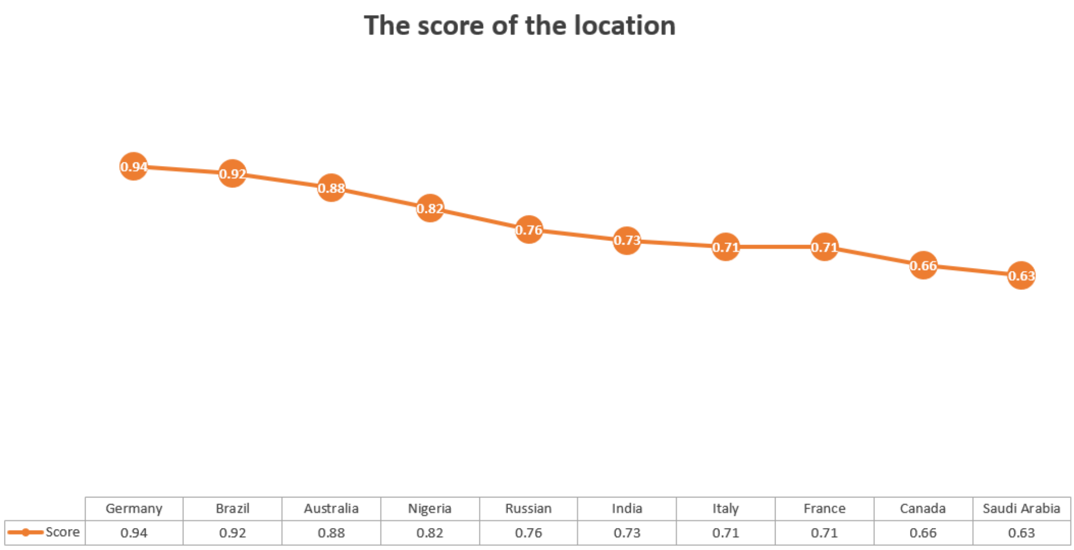

Figure 10: Score of the location

For determing the best amount of the offices, we just need to make minor changes to our model, since the amount does not affect the score. We can know from the supervisor's analysis that when we cut down the amount of the offices, the office construction costs and operating costs will decrease, but revenue and service intensity will increase. Our indicators take these points into account, so we can continue to use our model rather than change the indicators. Finally, by running our models, we find that building 6 offices in these countries is still the most suitable. At this point, our model is solved, and we give the distribution of our offices in World Map:

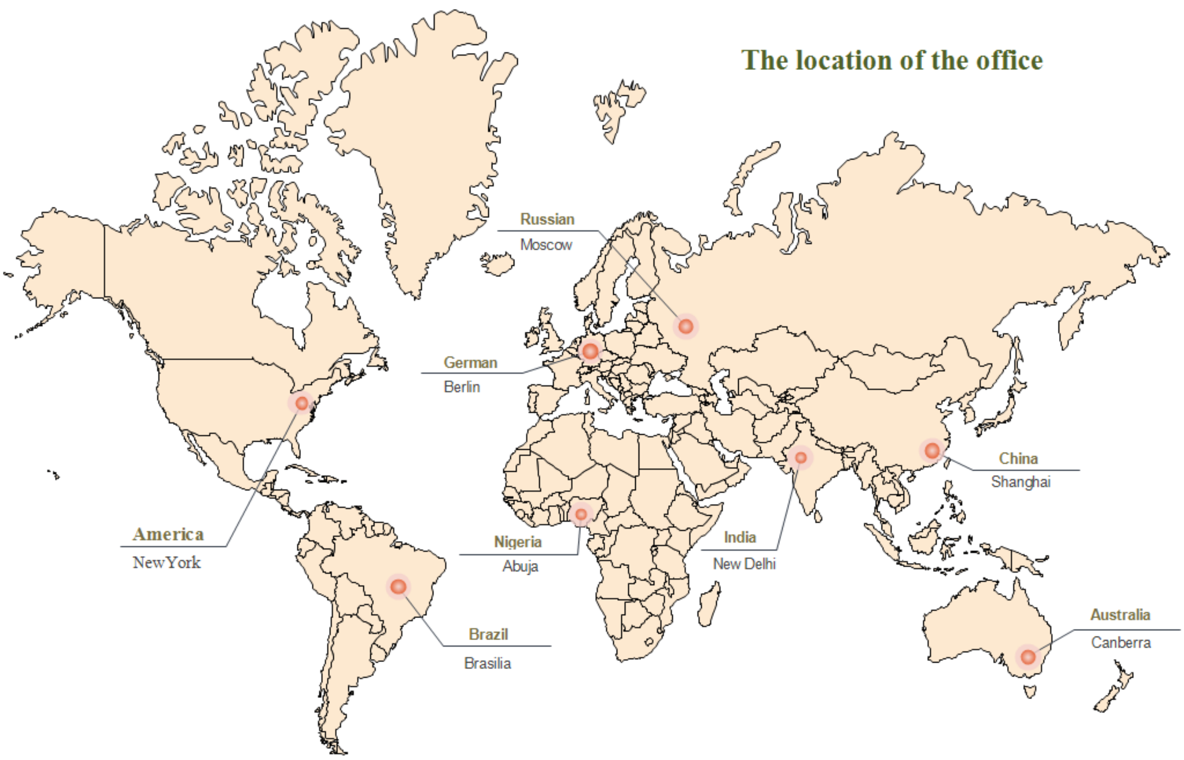

Figure 11: location of the offices

3.2.4 Sensitivity Analysis

For problem 2, because our model considers more comprehensive factors, it can be consistent in the short and long term. Here we give a sensitivity analysis on time.

Based on our initial scores, the values and scores of the six selected countries are shown below.

Indicator / Location	Proportion of national service workers	Average monthly salary of workers($)	National GDP(trillion $)	Proportion of English speakers	Score
Brasilia	0.44	713	1.87	0.72	0.92
Canberra	0.23	1023	1.43	0.98	0.88
New Delhi	0.42	587	2.27	0.89	0.73
Abuja	0.38	332	0.39	0.12	0.82
Berlin	0.32	1872	4.03	0.78	0.94
Moscow	0.25	982	1.66	0.67	0.76

Figure 12: Short-term Score

According to our analysis, over time, when a country begins to develop, its number of English speakers will increase, and GDP will increase. These are all contributing factors, but the decline in service practitioners and the increase in per capita wages will be restrained. Effect, according to our prediction, this promotion and inhibition effect will maintain the stability of the score. This effect is predictable, because in the initial country selection, we not only considered the current development level of this country, but also considered the potential development potential of developing countries.

Indicator / Location	Proportion of national service workers	Average monthly salary of workers($)	National GDP(trillion $)	Proportion of English speakers	Score
Brasilia	0.44--0.32	713--1092	1.87--3.87	0.72--0.87	0.92--0.95
Canberra	0.23--0.25	1023--1230	1.43--2.37	0.98--0.98	0.88--0.84
New Delhi	0.42--0.31	587--988	2.27--3.32	0.89--0.91	0.73--0.74
Abuja	0.38--0.34	332--562	0.39--0.52	0.12--0.23	0.82--0.76
Berlin	0.32--0.33	1872--1928	4.03--4.22	0.78--0.83	0.94--0.92
Moscow	0.25--0.35	982--1129	1.66--2.94	0.67--0.74	0.76--0.82

Figure 13: Long-term Score

As can be seen from the above figure, in the long-term development process, our best choice is still these six countries, and the scores have not changed much.

4 Strengths and Weaknesses

4.1 Strengths

- First one...
- Second one ...

4.2 Weaknesses

- Only one ...

Memo

To: Heishan Yan
From: Team XXXXXXX
Date: October 1st, 2019
Subject: A better choice than MS Word: LaTeX

In the memo, we want to introduce you an alternate typesetting program to the prevailing MS Word: **LaTeX**. In fact, the history of LaTeX is even longer than that of MS Word. In 1970s, the famous computer scientist Donald Knuth first came out with a typesetting program, which named TeX . . .

Firstly, . . .

Secondly, . . .

Lastly, . . .

According to all those mentioned above, it is really worth to have a try on LaTeX!

References

[1] Einstein, A., Podolsky, B., & Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete?. *Physical review*, 47(10), 777.

[2] *A simple, easy LaTeX template for MCM/ICM: EasyMCM*. (2018). Retrieved December 1, 2019, from `https://www.cnblogs.com/xjtu-blacksmith/p/easymcm.html`

Appendix A: Further on LaTeX

To clarify the importance of using LaTeX in MCM or ICM, several points need to be covered, which are ...

To be more specific, ...

All in all, ...

Anyway, nobody **really** needs such appendix ...