For office use only	Team Control Number	For office use only
T1 _____	**73410**	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	**B**	F4 _____

2018
MCM/ICM
Summary Sheet

Language Population Projection and Location Optimizaion Model Based on Inhomogeneous Transition Matrix and Simulated Annealing Algorithm

Summary

With the advent of increasingly accelerated globalization, the intricate geographic distributions of languages start to hamper international business operations and cross-culture interactions. Comprehending the distribution dynamics has never been more crucial, yet projecting the distributions is difficult, particularly due to the complicated composition of speakers, the influence of various exogenous factors like the migration, government policies, economic development, and eagerness to learn. Therefore, we establish a new model to replace the projection model based purely on population, as it is not only inaccurate but also invalid faced with the scarcity of supporting data.

Our model focuses on native speakers and non-native speakers of languages. We introduce the transition matrix to describe the transition between native speakers and non-native speakers of different languages, because the population growth of a language doesn't solely come from natural births, but also migration and learning. Additionally, we introduce two new groups: learners and migrants to further analyze the transition. In the end, we introduce a set of parameters to represent the exogenous influences, a new variable to express time changes, and establish our inhomogeneous transition matrix.

It has been tested that our matrix only requires relatively little amount of data input to function well. We employ the model to successfully project the geographic distributions of languages in 2067, based on the data in 2017. In the end, we adopt the simulated annealing algorithm to help our client, a large multinational service corporation, select optimal location options for new offices.

Keywords: Population of languages, Transition, Office locations

Contents

1 Introduction

There are about 6,900 languages spoken on Earth nowadays. About half of the world's population take one of ten languages as their native language and much of the world population also speaks a second language. However, because of a variety of influences, the population of speakers of a language may increase or decrease over time. Our target is to investigate trends of global languages and location options for new offices.

In part I, we compared our problem with the idea of Markov chain, and add transition matrix to describe the population transition from native speakers of a language to second language speakers of another language to native speakers of another language. Considering that transitions are inhomogeneous, we finally built up inhomogeneous transition matrix and used this matrix to predict the populations of different languages and their geographical distribution. In part II, based on prediction of our model in part I, we used simulated annealing algorithm to find the best location options for new offices.

2 Preliminary Model

2.1 Notations and Symbol Description

2.1.1 Symbol Description

Symbol	Descrption
N	The number of languages in consideration
$Y_i^{(n)}(i = 1, 2, ..., N; n = 0, 1, 2...)$	The number of native speakers of language i in the year of n
$y_i^{(n)}(i = 1, 2, ..., N; n = 0, 1, 2...)$	The number of non-native speakers of language i in the year of n
$\mathbf{Y}^{(n)}$	the state vector of the model in the year of n
$\mathbf{A}$	The transition matrix
α_{ii}	The annual birth rate of native speakers of language i
$\alpha_{i(N+i)}$	The annual proportion of non-native speakers of language i giving birth to native speakers of this language
β_{ii}	The annual death rate of native speakers of language i
$\beta_{(N+i)(N+i)}$	The annual death rate of non-native speakers of language i
$\gamma_{(N+i)j}$	The annual proportion (learning rate) of native speakers of language j successfully starting to master language i
Γ_i	The total learning rate of language i

2.1.2 Notations

Native speakers of A are individuals whose first language is A.

Non-native speakers of A are individuals whose first language is not A, but who master A as a foreign language (Implying that the individual possesses advanced skills of language A and is fluent in both speaking and writing).

2.2 General Assumptions

1. Speakers of any particular language can be categorized into two groups: native speakers and non-native speakers.
2. The number of native speakers only increases out of natural birth, and all newborn babies remain native speakers at the year of their birth (Leaving out rare cases of prodigies who can instantly learn to speak foreign languages); The number of native speakers is only reduced by death (Leaving out rare cases of postnatal first language disability).
3. Native speakers of a certain language cannot be converted to non-native speakers of this language, but can become non-native speakers of other languages.
4. The number of non-native speakers only increases out of postnatal learning, and only decreases due to death (Leaving out rare cases of forgetting learnt foreign languages).
5. Non-native speakers of a certain language cannot be converted to native speakers of this language, but can become non-native speakers of other languages.
6. Once an individual has mastered a new language, he becomes a non-native speaker of this language, while remaining all previous identities.
7. No radical and unpredicted events will occur, causing utter shifting in the population structure.

2.3 Analysis of the Problem

It is apparent that native speakers and non-native speakers of the same language are more closely related, which mainly reflects in:

1. Parents usually raise their children to have the same first languages, or at least languages they master. Therefore, we can assume that new native speakers can only be given birth to by native speakers or non-native speakers of the same language (Leaving out refugees, asylum seekers etc.).
2. It's unlikely that native speakers of a certain language can abandon their first language. Therefore, we can assume that no native speakers of a certain language can be converted to non-native speakers of this language (Leaving out rare cases of postnatal first language disability).

Furthermore, A number of people who cannot master a certain language will be converted to non-native speakers of this language through learning every year. There will also be natural deaths causing the number of each group of people to drop.

2.3.1 N=2 Model

Take the simplest model with only two languages ($N = 2$) as an example, we are able to plot the transition between different groups of people (See Figure 1).

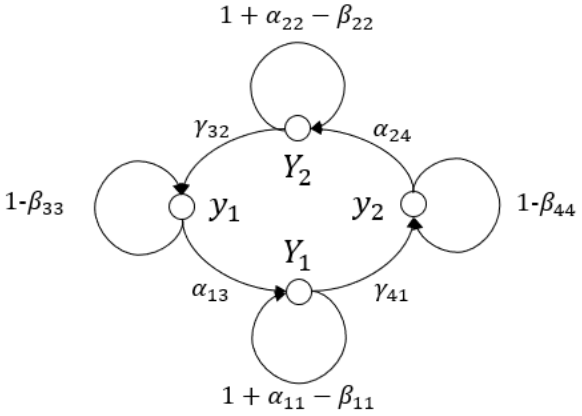

Figure 1: Transition Between Different Groups with Only Two Languages

It can be concluded that this model much resembles the model of homogeneous Markow chain (After converted to the new group, the element will remain in the previous groups). The transition follows the following rules:

$$\begin{aligned}\mathbf{Y}^{(n+1)} &= \begin{pmatrix} Y_1^{(n+1)} \\ Y_2^{(n+1)} \\ y_1^{(n+1)} \\ y_2^{(n+1)} \end{pmatrix} \\ &= \begin{pmatrix} 1+\alpha_{11}-\beta_{11} & 0 & \alpha_{13} & 0 \\ 0 & 1+\alpha_{22}-\beta_{22} & 0 & \alpha_{24} \\ 0 & \gamma_{32} & 1-\beta_{33} & 0 \\ \gamma_{41} & 0 & 0 & 1-\beta_{44} \end{pmatrix} \begin{pmatrix} Y_1^{(n)} \\ Y_2^{(n)} \\ y_1^{(n)} \\ y_2^{(n)} \end{pmatrix} \\ &= \mathbf{A}\mathbf{Y}^{(n)}\end{aligned} \tag{1}$$

In expression (1), parameters are set as follows:

1. α_{11}, α_{22} represents the annual birth rate of native speakers of language 1 and 2 respectively.
2. β_{11}, β_{22} represents the annual death rate of native speakers of language 1 and 2 respectively.
3. α_{13}, α_{24} represents the ratio of non-native speakers of language 1 and 2 giving birth to native speakers of corresponding language respectively (Assuming all births are single births).

4. γ_{32}, γ_{41} represents the ratio of native speakers of language 1 and 2 learn to speak the other language respectively.
5. β_{33}, β_{44} represents the annual death rate of non-native speakers of language 1 and 2 respectively.
6. It should be noted that we set zero the $(3, 4)$ and the $(4, 3)$ element in the matrix to prevent double counting.
7. Additionally, we also set zero the $(1, 4)$ and the $(2, 3)$ element to prevent double counting (e.g. offspring of people who are native speakers of language 1 and non-native speakers of language 2 will be counted twice).

If the transition reaches a steady state (Implying that each group takes up a steady split of total population, rather than in terms of absolute quantity), it can be derived:

$$\lim_{k\to\infty} \mathbf{A}^k \mathbf{Y}^{(0)} = \lambda^k \mathbf{Y} \tag{2}$$

λ is the growth rate of total population, $\mathbf{Y}^{(0)}$ is an arbitrary initial vector,$\mathbf{Y}$ is a constant vector.

When the transition reaches the steady state, we have:

$$\mathbf{AY} = \lambda \mathbf{Y} \tag{3}$$

Obviously, λ is the eigenvalue of matrix A, and $\mathbf{Y}$ is the eigenvector of this matrix. We can derive the proportion of each group through matrix normalization $\mathbf{Y}$.

To simplify the test of this model, let's consider a particular scenario with following parameters:

$$\begin{aligned} \alpha_{11} = \alpha_{22} &= 0.02 \\ \alpha_{13} = \alpha_{24} &= 0.05 \\ \beta_{11} = \beta_{22} = \beta_{33} = \beta_{44} &= 0.01 \\ \gamma_{41} = \gamma_{32} &= 0.01 \end{aligned} \tag{4}$$

The transition matrix is:

$$\mathbf{A} = \begin{pmatrix} 1.01 & 0 & 0.05 & 0 \\ 0 & 1.01 & 0 & 0.005 \\ 0 & 0.01 & 0.99 & 0 \\ 0.01 & 0 & 0 & 0.99 \end{pmatrix} \tag{5}$$

This matrix possesses four different eigenvalues, and only when $\lambda = 1.01049$, can the eigenvector be positive semidefinite vector (all elements are non-negative):

$$\mathbf{Y} = \begin{pmatrix} 0.500 \\ 0.500 \\ 1.124 \\ 1.124 \end{pmatrix} \tag{6}$$

This is the steady state of language population structure, which represents that no matter what the initial distribution is, non-native speakers of these two languages will both take up 1.124 of the total population.

Apparently, in the bilingual ($N = 2$) model, the number of non-native speakers of each language exceeds that of native speakers of each language, which poses an indirect constraint on the range of parameter γ (learning rate). By employing the Matlab, we derive the quantitative influence of learning rate γ on the proportion of non-native speakers.

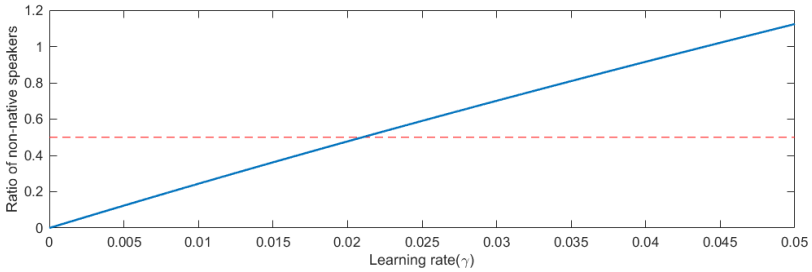

Figure 2: Correlation Between the Proportion of Non-native Speakers and Learning Rate (γ)

We can see in (Figure 2), when $\gamma > 0.021$, the proportion of non-native speakers exceeds that of native speakers. Therefore, this model cannot stay at a learning rate greater than 0.021 in the long run.

2.4 Calculating and Simplifying the Model

Since the top twenty-six languages cover most of the world population, we have reason to believe that the majority of verbal communications and information exchanges are conducted through these twenty-six languages. Due to limited data[1], we will only be analyzing twenty-two major languages out of these twenty-six languages in this subsection (We will come back to the twenty-six-language model in the following sections). Thus, we only consider a model with twenty-two languages $(N = 22)$, and the transition matrix for this model is:

$$\mathbf{A} = \begin{pmatrix} 1+\alpha_{11}-\beta_{11} & \cdots & 0 & \alpha_{1(N+1)} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1+\alpha_{NN}-\beta_{NN} & 0 & \cdots & \alpha_{N(2N)} \\ 0 & \cdots & \gamma_{(N+1)N} & 1-\beta_{(N+1)(N+1)} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{(N+N)1} & \cdots & 0 & 0 & \cdots & 1-\beta_{(2N)(2N)} \end{pmatrix} \tag{7}$$

To make a more accurate projection, we should set the values of parameters in this model to better fit the real scenario:

1. Given that native speakers of different languages have distinctive birth rates and death rates, we introduce a set of new data to reflect the difference[1].

   To derive the birth rate and death rate of different language groups, we first categorize nations by official language, and select several major nations to represent each group (Total number of native speakers in these countries taking up more than 90% of the total number of native speakers of this language worldwide). We have the following table in Appendix A.

We draw upon public sources for population, annual birth rate and death rate of aforementioned nations (2013). We then compute the weighted average annual birth rate and death rate of language groups, and list in the table in Appendix A)[4][5]

2. Since non-native speakers of a language can come from any nations (other than nations where this language is commonly regarded as first language), we may as well assume non-native speakers of different languages share the same birth rate and death rate (equal to world average death rate):

$$\begin{aligned} \beta_{(N+1)(N+1)} = \beta_{(N+2)(N+2)} = \cdots = \beta_{(N+N)(N+N)} &= \overline{\beta} \\ \alpha_{1(N+1)} = \alpha_{2(N+2)} = \cdots = \alpha_{N(N+N)} &= \alpha \end{aligned} \tag{8}$$

Drawing upon public sources, world average death rate: $\overline{\beta} = 0.0083$; world average birth rate: $\overline{\alpha} = 0.0193$, and the proportion of immigrants to non-native speakers of a certain language is: $r = 0.172$, Therefore, we may assume:

$$\alpha = r\overline{\alpha} = 0.0033 \tag{9}$$

3. Considering different languages bear different level of attractiveness, the learning rate for each language should also be different. To simplify the model, it is reasonable for us to assume a language is indifferently attractive to different groups of people, therefore leading to the same learning rate amongst different groups, though it may not be the real case.

$$\gamma_{i(N+1)} : \gamma_{i(N+2)} : \cdots : \gamma_{i(N+j)} : \cdots : \gamma_{i(N+N)};\ j = 1, 2, \cdots, N, j \neq i \tag{10}$$

The learning rate (shown above) is a constant rate. We can temporarily assume that the learning rate of a language equals the proportion of its non-native speakers to total non-native speakers, multiplied by the total learning rate.

$$\begin{aligned} &\gamma_{(N+1)i} : \gamma_{(N+2)i} : \cdots : \gamma_{(N+j)i} : \cdots : \gamma_{(N+N)i} \\ =&y_1^{(0)} : y_2^{(0)} : \cdots : y_N^{(0)} \\ =&178 : 510 : 90 : 214 : 26 : 30 : 20 : 87 : 41 : 8 : 0.2 : 5 : 8 : 4 : 0.2 : 3 : 0.3 \\ &: 1 : 5 : 1 : 6 : 2 \\ =&0.1491 : 0.4272 : 0.0754 : 0.1792 : 0.0218 : 0.0251 : 0.0168 : 0.0729 : 0.0345 \\ &: 0.0067 : 0.0002 : 0.0042 : 0.0067 : 0.0034 : 0.0002 : 0.0025 : 0.0003 : 0.0008 \\ &: 0.0042 : 0.0008 : 0.0050 : 0.0017 \\ =&\hat{y_1}^{(0)} : \hat{y_2}^{(0)} : \cdots : \hat{y_N}^{(0)} \end{aligned} \tag{11}$$

Note that the last equation is normalized.

Therefore, the annual learning rate $\gamma_{(N+j),i}$ from group i to group j (non-native speakers) is the normalization ratio $\hat{y_j}^{(0)}$ shown, multiplied by the group's annual total learning rate Γ_i, which is:

$$\gamma_{(N+j)i} = \hat{y_j}^{(0)}\Gamma_i \tag{12}$$

The modified transition matrix is too huge (44×44), so we will not present it here to save space:

2.5 The Model Results

We use data in 2013 to establish the initial vector $\mathbf{Y}^{(0)}$, and obtain the projected statistics in 2017 after four alterations. $\mathbf{Y}^{(4)} = \mathbf{A}^4\mathbf{Y}^{(0)}$ (Presenting only the top 10 figures) (Table 1):

Language	Native (predict)	Native (real)	Non-native (predict)	Non-native (real)
Mandarin	870	897	180	193
English	353	371	521	611
Spanish	422	436	92	91
Hindi/Urdu	340	329	219	215
Russian	170	153	27	113
Portuguese	206	218	31	11
Bengali	201	242	20	19
French	78	76	89	153
Japanese	129	128	?	?
German	78	76	8	52

Table 1: Projections for Population Statistics of Major Languages in 2017

There are deviations between projected and real statistics in a number of items, like the number of non-native German speakers. To quote the data in 2013, there are only eight million non-native German speakers, whereas the figure reaches fifty-two million in 2017. This apparently goes against common sense, so we believe that these deviations are mostly caused by the inaccuracy of the data (In fact the inaccuracy of the data is also pronounced by the data source itself).

The data in 2013 is in Appendix A[1].

Additionally we can compute the positive semidefinite eigenvector $\mathbf{Y}$ (Normalized according to statistics of the former 22 groups) and the eigenvalue λ of the transition matrix $\mathbf{A}$:

$$\begin{aligned} \mathbf{Y} =& (0.0050, \cdots, 0.9601, \cdots, 0.0003)^T \\ \lambda =& 1.0215 \end{aligned} \tag{13}$$

These two items have corresponding practical meanings:

1. Eigenvector: the equilibrium ratio (in the long run) of the population of each group to total population.

2. Eigenvalue: the equilibrium total growth rate (in the long run) (accumulated growth rate of all language groups).

We can find that native speakers of Panjabi will take up more than 96% of total population when reaching equilibrium, which obviously goes against common sense. The reason why this phenomenon takes place is that the birth rate of native speaker of Panjabi is the highest (0.0286) amongst all native speaker groups, and that the group with the highest growth rate will dominate in a homogeneous model.

Therefore, to conduct a more accurate projection in the long run, we must introduce time as a new variable into the model, or rather: an inhomogeneous model.

3 Modified Model

The aforementioned preliminary model yields quite accurate projections for population of different language groups in the short run. However, the model is lacking in accuracy when it comes to long-run projections. Defects of the model are:

1. No regard to the process of learning a language, which implies that it takes time for an individual to master a new language, and that the length of time is influenced by education level, language similarity (e.g. English is more similar to French than to Mandarin), and eagerness to learn
2. No regard to the change of birth rate and death rate over time
3. No regard to the economic development over time
4. Fail to reveal the exact impact of population migration
5. Fail to derive the distribution of English population group as requested by our client

Therefore, building upon the preliminary model, we introduce time as a new variable to better fit the real case. Furthermore, we introduce a new dimension of variables (language learners, migrants, a breakdown of English speakers) to observe the process of learning and migration.

Note: For convenience, we number English as 1, and Chinese as 2.

3.1 Notations and Symbool Description

3.1.1 Additional Notations

Learners of language B from language A are native speakers of language A who are starting to learn language B (not non-native speakers of language B).

Language group (region) A are regions where native speakers of language A take residence (Implying that the total population of language region A is the population of native speakers of language A, and that we use the economic data of aforementioned major countries to represent their corresponding language groups (regions)).

Migrants to language B region are non-native speakers of language B whose offspring is native speaker of language B, or rather the first-generation immigrants.

Decrease matrix is the matrix evaluating the natural death and migration of a language group.

3.1.2 Additional Symbol Description

Symbol	Description
$y_{j1}^{(n)}(j=2,...,N;n=0,1,2...)$	the number of non-native English speakers who are converted in the year of n, from language j
S_{ji}	the number of learners of language j from language i
W_j	the number of migrants to language j region
$p_{ji}(0<p_{ij}<1)$	the eagerness to learn language j of people from language i group (region)
d_{ji}	the similarity between language i and language j
σ_i	the average length of learning a foreign language of the same family in language i group (region)
a_i	the annual proportion of new learners to the total population of language region i
r_j	the proportion of migrants to non-native speakers of language j
$\mathbf{\Xi}$	the natural growth matrix of native speakers
$\mathbf{\Pi_1}$	the decrease matrix of non-native English speakers
$\mathbf{\Pi}$	the decrease matrix of non-native speakers
$\mathbf{\Omega}$	the death matrix of migrants
$\mathbf{S}$	the reproductive matrix of migrants
$\mathbf{G_1}$	the learning matrix of English
$\mathbf{G}$	the learning matrix of all languages combined
$\mathbf{R_1}$	the migration matrix of English region
$\mathbf{R}$	the migration matrix

3.2 Additional Assumptions

1. The birth rate and death rate of each group is not affected by the population of each group, and only changes over time.
2. An individual will not be born as a learner (Implying that the number of learners only increase due to learning, not reproduction), and will not die during learning (However, the natural death rate will not be affected, since learners are mostly young and unlikely to die).
3. Offspring of migrants will be considered as native speakers of the language region they migrate into.
4. Once an individual successfully migrates into a foreign language region, he will not be considered a non-native speaker of this language (Implying that the non-native speaker group in the aforementioned preliminary model can actually be broken down into two groups in this model: non-speaker group and migrant group of this language).

3.3 Analysis of the Problem

Example graph of transition between different groups in a bilingual model with regard to learners and immigrants is shown below (Figure 3):

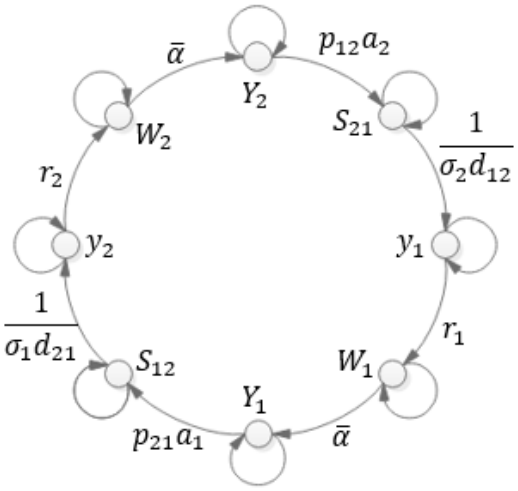

Figure 3: Example Graph of Transition Between Different Groups in a Bilingual Model

The transition matrix in the year of n is:

$$\mathbf{A}^{(n)} = \begin{pmatrix} \mathbf{\Xi}^{(n)} & \mathbf{0} & \mathbf{S}^{(n)} \\ \mathbf{G}^{(n)} & \mathbf{\Pi}^{(n)} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^{(n)} & \mathbf{\Omega}^{(n)} \end{pmatrix}$$

The population of a language in the year of $(n+1)$ can be expressed as:

$$\mathbf{Y}^{(n+1)} = \mathbf{A}^{(n)}\mathbf{Y}^{(n)} \tag{14}$$

When $\mathbf{Y}$ changes, it will, in turn, influence the transition matrix:

$$\mathbf{A}^{(n+1)} = F[\mathbf{Y}^{(n)}, n] \tag{15}$$

1. $\mathbf{\Xi}$ is the natural growth matrix of native speakers, and possesses the property of diagonality, with a diagonal entry equal to 1 plus natural growth rate (1 + natural growth rate).
2. $\mathbf{\Pi}$ is the decrease matrix of non-native speakers, and possesses the property of diagonality, with a diagonal entry equal to 1 minus death rate and migration rate (1 - death rate - migration rate).
3. $\mathbf{\Omega}$ is the death matrix of migrants, and possesses the property of diagonality.
4. $\mathbf{G}$ is the transition matrix from native speakers to non-native speakers of another language (learning matrix), and has a diagonal entry equal to 0.
5. $\mathbf{R}$ is the transition matrix from non-native speakers to migrants of the same language (learning matrix), and possesses the property of diagonality.
6. $\mathbf{S}$ is the reproduction matrix of migrants, and possesses the property of diagonality.

If we want to analyze the distribution of English speakers separately, we should modify the aforementioned matrix as follows:

$$\mathbf{A}^{(n)} = \begin{pmatrix} \mathbf{\Xi}^{(n)} & \mathbf{0} & \mathbf{0} & \mathbf{S}^{(n)} \\ \mathbf{G_1}^{(n)} & \mathbf{\Pi_1}^{(n)} & \mathbf{0} & \mathbf{0} \\ \mathbf{G}^{(n)} & \mathbf{0} & \mathbf{\Pi}^{(n)} & \mathbf{0} \\ \mathbf{0} & \mathbf{R_1}^{(n)} & \mathbf{R}^{(n)} & \mathbf{\Omega}^{(n)} \end{pmatrix}$$

1. $\mathbf{\Pi_1}$ is the decrease matrix of non-native English speakers and possesses the property of diagonality, with a diagonal entry equal to 1 minus death rate and migration rate (1-death rate-migration rate).

2. $\mathbf{G_1}$ is the transition matrix from native speakers of other languages to non-native English speakers (learning matrix of English), with all entries zero in the first column. The rest of the matrix is a diagonal matrix.

3. $\mathbf{R_1}$ is the transition matrix from non-native speakers of English to migrants of English region (migration matrix of English), with all entries zero other than entries in the first row.

It should be noted that we only consider the distribution of non-native English speakers in the aforementioned matrixes, but the distribution of English speakers can be influenced by emigrants of native English speakers. However, since the number of non-native speakers far exceeds that of emigrants', we'll leave out such an influence.

We have considered three exogenous factors in total:

1. Government policy factor will influence parameter p_{ij} and r_j.

2. Economic factor and initial population factor will influence parameter p_{ij} and r_j.

3. Education factor will influence parameter σ_i and a_i

3.4 Calculating and Simplifying the Model

We take the year of 2017 as the starting year. Since the top twenty-six languages cover most of the world population, we have reason to believe that the majority of verbal communications and information exchanges are conducted through these twenty-six languages. Therefore, we only consider a model with twenty-six languages $(N = 26)$.

Prior to determining the transition matrix, we should first set the initial vector, which stands for the initial distribution of non-native English speakers. We assume the initial distribution of non-native English speakers to be proportional to the distribution of native speakers in total, which is:

$$y_{21}^{(0)} : \cdots : y_{N1}^{(0)} = Y_2^{(0)} : \cdots : Y_N^{(0)} \tag{16}$$

To save time, we adopt the annual birth rate, death rate and migration rate projected for the next fifty years by United Nations. We employ the weighted average method to compute the birth rate, death rate and migration rate in the upcoming fifty years, and plot the trend in passing. See: Apendix B

Building upon these data, we can set: $\mathbf{\Xi}$,$\mathbf{\Pi_1}$,$\mathbf{\Pi}$,$\mathbf{\Omega}$,$\mathbf{S}$,$\mathbf{R_1}$,$\mathbf{R}$.

Then we should attempt to set $\mathbf{G}, \mathbf{G_1}$:

The learning rate $\gamma_{ji}^{(n)}$ from language group i to language j satisfies that:

$$\gamma_{ji}^{(n)} = \frac{p_{ji}^{(n)} a_i^{(n)}}{\sigma_i^{(n)} d_{ji}^{(n)}} \tag{17}$$

We assume the eagerness $p_{ji}^{(n)}$ to learn is proportional to the number of non-native speakers of language j (Assuming that people's eagerness to learn a language in the past reflects people's eagerness to learn this language in the future):

$$p_{1i}^{(n)} : \cdots : p_{Ni}^{(n)} = y_1^{(n)} : \cdots : y_N^{(n)} = \hat{y_1}^{(n)} : \cdots : \hat{y_N}^{(n)}$$

We assume all language groups share the same level of education:

$$a_1 = \cdots = a_N = 0.01$$

$$\sigma_1 = \cdots = \sigma_N = 3$$

We assume that any two languages from the same language family have the first-level similarity (Implying a similarity parameter equal to 1), whereas any two languages from different language families have the second-level similarity (Implying a similarity parameter equal to 2), which can be expressed as:

$$d_{ji} = \begin{cases} 1, & \text{If language i and j belong to the same language family} \\ 2, & \text{If language i and j belong to different language families} \end{cases}$$

This implies that one in a hundred people start to learn a foreign language annually in each country. On average, it takes them three years to master a foreign language from the same language family as their first language, and six years if from different language families.

3.5 The Model Results

3.5.1 Graph of Population of Different Language Groups

We will skip the computation process to save space. Through repeated alterations, we can plot the graph of population of different language groups in the next fifty years (Figure 4).

Based on the projections, we can align the top ten languages in terms of population in 2017 and 2067: (Table 2):

According to this ranking, Portuguese and French both drop out of the top ten, whereas Hausa and Punjabi make it to the top ten. We observe the following patterns for languages whose ranking elevates:

1. High birth rate, e.g. Hausa and Punjabi
2. For language groups whose migration is mostly immigration, belonging to a bigger family (more languages from the same family) would be better, whereas it is the opposite for language groups whose migration is mostly emigration, e.g. Arabic
3. High attractiveness (Implying people's stronger eagerness to learn this language), e.g. English

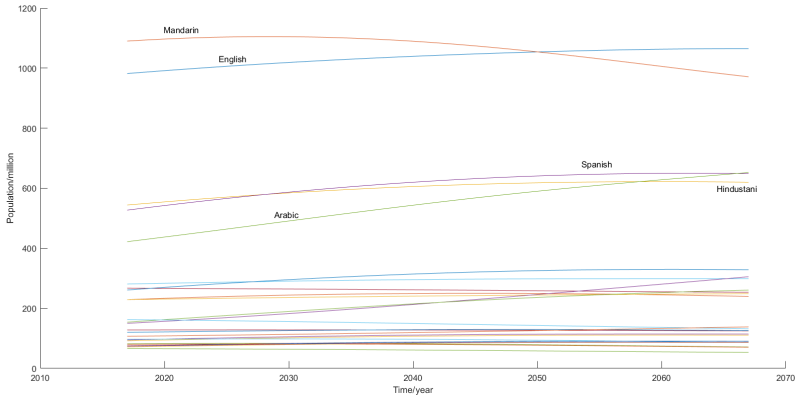

Figure 4: Graph of Population of Top Five Language Groups in the Next Fifty Years (See in Apendix B for the rest twenty-one language groups)

Ranking	2017	2067
1	Mandarin	English
2	English	Mandarin
3	Hindi/Urdu	Arabic
4	Spanish	Spanish
5	Arabic	Hindi/Urdu
6	Malay	Bengali
7	Russian	Hausa
8	Bengali	Malay
9	Portuguese	Punjabi
10	French	Russian

Table 2: Ranking of Languages by Population in 2017 and 2067

3.5.2 Graph of the Scale of Migration

Additionally, we can easily derive the graph of the scale of migration starting from the year of 2017 (Figure 5)

We only label in the graph the top four attractive language regions for immigrants, and it's apparent that they are all developed countries except for Russia (Mostly due to the historical factor that the population is not clearly divided amongst neighboring countries after the disintegration of USSR). Furthermore, the quantity of immigrants to English region (Over one hundred million) far exceeds that of other language regions', mostly driven by the extensive acceptance of this language.

3.5.3 Distribution of Non-native English Speakers

We can also easily derive the distribution of non-native English speakers (Figure 6):

The scale of non-native English speakers is unparalleled. According to our projections, there are more than ninety million non-native English speakers amongst all native Chinese speakers.

What intrigues us is that despite most English nations (e.g. U.S.A., UK) being isolated

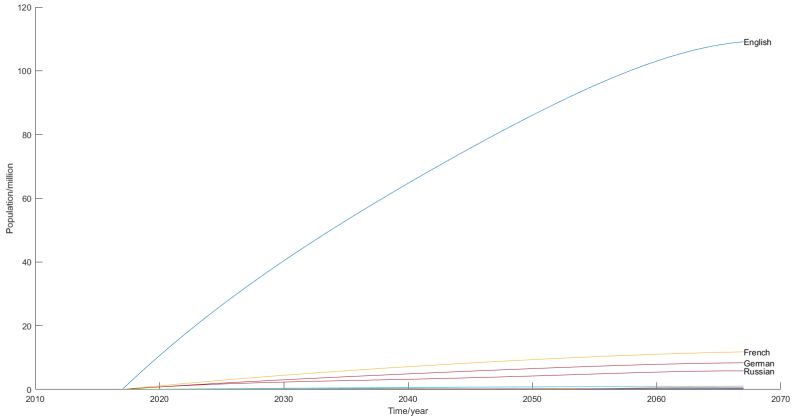

Figure 5: Graph of the Scale of Migration in the Next Fifty Years (Starting from the year of 2017)

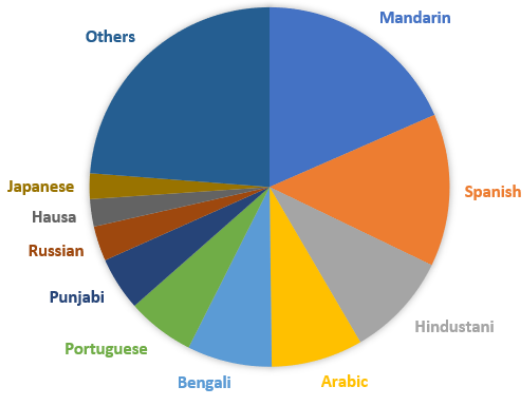

Figure 6: Distribution of Non-native English Speakers in 2067

in terms of geography, English regions have the largest scale of immigrants and non-native speakers. This implies that geographic factors will not pose significant impact on migration in the next fifty years.

3.6 Sensitivity Analysis

In this subsection, we will be mainly analyzing the sensitivity of our model to the uniformity of eagerness to learn of different language groups (Implying that and that if a language group has a barely uniform eagerness to learn, people in this language group tends to learn different foreign languages).

3.6.1 High Uniformity Scenario

If a language group has high uniformity in terms of eagerness to learn, people's probability of choosing to learn each language is more uniformly distributed. The major corresponding real cases are:

1. Governments impose no restrictions over second language, and people in this lan-

guage group can choose which foreign language to learn at their discretion.

2. People are not highly purposeful when it comes to learning a foreign language, which means that they are not learning foreign languages merely for the purpose of attaining more opportunities.

To reflect the characteristics of a high uniformity scenario, we modify expression (17) to:

$$\gamma_{ji}^{(n)} = \frac{(y_i^{(n)})^k a_i^{(n)}}{\sigma_i^{(n)} d_{ji}^{(n)}}$$

The higher k is, the lower the uniformity is. The initial model adopts $k = 1$, and when $k = 0.1$, the graph of population of different language groups are as follows: (Figure 7).

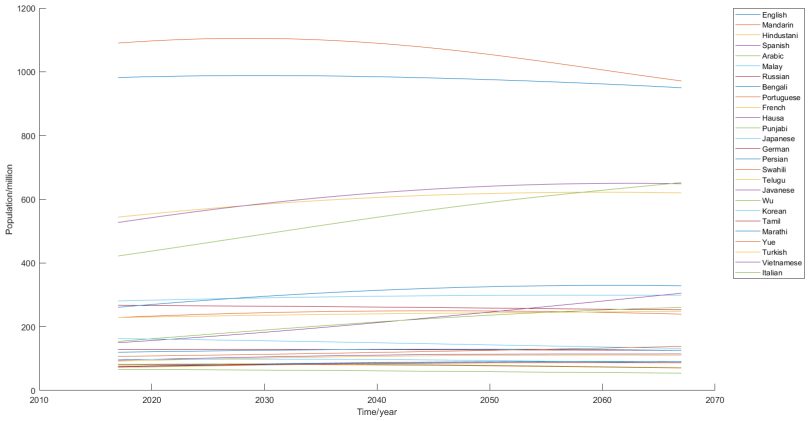

Figure 7: Graph of Population of Different Language Groups When k=0.1

We can find that the population of English doesn't exceed that of Mandarin's. The attractiveness of English is impaired.

3.6.2 Low Uniformity Scenario

If a language group has high uniformity in terms of eagerness to learn, people's probability of choosing to learn each language is less uniformly distributed (concentrated at widely accepted foreign languages). The major corresponding real cases are:

1. Governments introduce restrictions over second language, e.g. all people should learn a specified foreign language.

2. People are highly purposeful when it comes to learning foreign languages, eager to attain more opportunities through language learning.

When $k = 10$, the graph of population of different language groups is as follows (Figure 8).

We can find that the population of English is rocketing, surpassing that of Mandarin's tens of years earlier. The attractiveness of English is further reinforced.

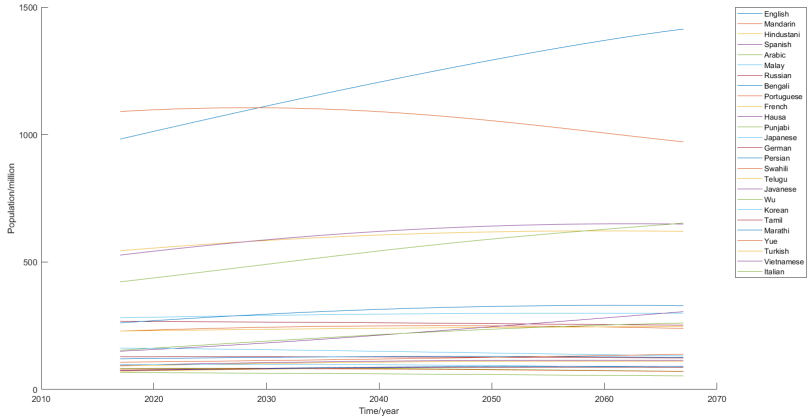

Figure 8: Graph of Population of Different Language Groups When k=10

4 Application of Our Model

4.1 Assumptions

With regard to our client company being a large multinational service company, we derive their extrinsic and intrinsic needs for additional international offices:

1. These six new offices should cover the population of major languages (twenty-three major languages listed by our client) in a mutually exclusive and collectively exhaustive way (Implying that each language can only be covered by one of the six offices except for English).

2. The basic working languages of an office are English and local language.

3. These offices should cover languages with the largest population and the most robust economy (Evaluated based on the GDP of different language regions). The transportation costs between each office and the language regions they cover should also be the smallest.

4. One office shouldn't cover too many different languages, or else there will be an impairment on profit due to lower operation efficiency and higher costs.

5. There must be English speaking residents near each office (Each language region is projected to have English speaking residents by our previous models)

Drawing upon the analysis above, we establish our rating model for office locations:

$$\begin{aligned} S_i &= \epsilon^{k_i} \sum_{j=0}^{k_i} \frac{P_{ij}^{\eta_1} E_{ij}^{\eta_2}}{D_{ij}^{\eta_3}} - h_i \\ h_i &= \frac{10^{\eta_1} E_{i0}^{\eta_2}}{D_{i0}^{\eta_3}} \end{aligned} \tag{18}$$

4.2 Symbol Descrption

Symbol	Description
S_i	The score of office location i
k_i	The number of additional languages used by office i (Other than English and the local language)
ϵ	The marginal impairment rate on profit when an office covers one more language
P_{ij}	The total population of language j group (All covered by office i, and evaluated based on accumulated population of major nations in the group) (Unit: million)
E_{ij}	The accumulated GDP of language j group (All covered by office i, and evaluated based on accumulated GDP of major nations in the group) (Unit: ten billion USD)
D_{ij}	The transportation costs from office i to language j region (Office i covering language j, and evaluated based on current weighted average transportation costs by air, water and land route (if possible), from the site to major nations in the region)
η_1, η_2, η_3	Weight coefficient
h_i	The establishment cost of office i (e.g. rent)

Note: $j = 0$ represents the local language of the office.

4.3 Calculating and Simplifying

To start with, we propose a list of candidate cities for further examination and selection. These cities cover major nations of all twenty-three languages, except for English, Mandarin and Wu (Because our client has already set up offices in cities covering those language regions):

Bombay, Dubai, Madrid, Jakarta, Moscow, Rio de Janeiro, Dhaka, Paris, Abuja, Islamabad, Tokyo, Berlin, Tehran, Dodoma, Seoul, Hong Kong, Istanbul, Ho Chi Minh City, Rome.

4.3.1 The Year of 2017

We conduct simulation on the data of 2017, with weight coefficients set as (Set to fit the real case and their orders of magnitude):

$$\epsilon = 0.98\,,\ \eta_1 = \eta_2 = \eta_3 = 1,$$

Given the enumeration method taking too much time, we decide to adopt the simulated annealing algorithm to derive the model. We set the initial temperature: $T_0 = 97$, terminal temperature: $T_t = 3$, attenuation coefficient: $\alpha = 0.99$, Markoff chain length: $l = 10000$, and language-city transition ratio: $\rho = 8 : 2$. We assume the model satisfies Metro Polis-Hastings algorithm.

After repeating for ten times, we have the stable solution with the highest accumulated scores:

Office location	Covering languages (Exclusive of English)
Tokyo	Japanese, Korean, Javanese, Malay
Bombay	Hindustani, Telugu, Tamil, marathi
Madrid	Spanish, Italian
Rio de Janeiro	Portuguese
Moscow	Arabic ,Russian ,Bengali, Hausa, Punjabi, Persian, Swahili, Yue, Vietnamese, Turkish
Paris	French, Germany
Accumulated scores	43.0

If we set up only five offices, then the optimal stable solution is as follows:

Office location	Covering languages (Exclusive of English)
Tokyo	Japanese, Korean, Javanese, Malay
Bombay	Hindustani, Telugu, Tamil, marathi
Madrid	Spanish, Germany
Moscow	Arabic ,Russian ,Bengali, Hausa, Punjabi, Persian, Swahili, Yue, Vietnamese, Turkish
Paris	French, Portuguese, Italian
Accumulate scores	39.7

We can see that despite the costs of opening more offices, opening six offices is still superior to opening five offices.

We also simulate other scenarios with different weight coefficients if our client attaches great importance to regional economy or population. See: Appendix A

4.3.2 The Year of 2067

Given an intrinsic change in communications and transportation in 2067 (Compared to 2017), the geographic factor is less significant. We set the coefficients as:

$$\epsilon = 0.98\,,\ \eta_1 = \eta_2 = 1\,, \eta_3 = 0.1,$$

We make projections about the economy (GDP) of each language region (Evaluated based on major nations in each region), by establishing a cycle projection model. We integrate the Kitchin cycle, Juglar cycle, Kuznets swing and Kondratiev wave, and simulate the integrated cycle with a trigonometric function. We then use the GDP data of each region in the past forty-six years to solve for the coefficients, and then employ the function to forecast GDP of each language region in 2067. See Appendix A[6] for our projected growth rate in each cycle and terminal GDP (Starting at 2015, with each cycle representing approximately thirteen years).

Combined with the projected distribution of population of different languages in 2067 by our previous models, we can derive the optimal stable solution:

Office location	Covering languages (Exclusive of English)
Tokyo	Japanese, Swahili, Yue, Vietnamese
Bombay	Hindustani, Telugu, Tamil, marathi
Madrid	Spanish, French, Germany
Jakarta	Javanese, Malay, Hausa, Korean
Dubai	Arabic ,Russian ,Bengali, Punjabi, Persian, Turkish
Paris	French,Italian
Accumulated scores	236.2

If we open only five offices, then the optimal stable solution is:

Office location	Covering languages (Exclusive of English)
Bombay	Hindustani, Telugu, Tamil, marathi
Madrid	Spanish, French, Germany
Jakarta	Javanese, Malay, Japanese, Korean, Yue
Dubai	Arabic ,Russian ,Bengali, Hausa, Punjabi, Persian, Turkish
Rio de Janeiro	Portuguese, Swahili, Vietnamese,Italian
Accumulated scores	225.2

We can see that opening six offices is still superior to opening five offices, yet the gap is narrowing. If the management cost is taken into consideration, opening five offices can be a viable option for our client.

5 Strengths and Weaknesses

- **Strengths**

  1. We preserve sufficient interfaces in our model, for further analysis and consideration of more variables.
  2. We leave out irrelevant (Or weakly correlated) variables, making the model simple and concise.
  3. We make our model more legible and applicable, by introducing the transition matrix and transition graph.
  4. We make our model fit the real case better, by separating learners, migrants, and English speakers from other groups.

- **Weaknesses**

  1. Our consideration of the influence of economic development is not sufficient.
  2. We regard natural growth, migration, and language learning as major factors influencing the language structure. However, the change in language structure can influence these three factors in turn. We only consider its influence on language learning in our model.
  3. The first language of migrants can also influence the language structure of the regions they migrate to. This model doesn't take this into consideration.
  4. The distinctions of education level in different language regions are not sufficiently considered in our model.

6 Memo

Dear Sir or Madame,

In accordance to your requests, we establish two customized models to develop accurate projections and help your company deliver successful global solutions at the lowest possible costs.

First, we investigate trends of global languages through our inhomogeneous transition model, and cast a projection for population of different languages in the next 50 years (With regard to the uneven distribution of population amongst languages, we only examine languages with a population larger than 50 million in 2017).

The projected distribution is rather different from that in 2017, and the key findings are:

1. The population of English rises to surpass Mandarin, and dominates as the international language with most speakers, mostly due to its extensive acceptance by people across the world. The population of non-native English speakers will reach approximately 490 million, with most of being native Chinese speakers (91 million).

2. The population of Arabic / Punjabi / Hausa increases greatly (The number of Arabic speakers only falls behind that of English's and Mandarin's), mostly due to the high birth rate.

3. The population of Mandarin / Japanese / Russian / Korean / Yue / Wu decreases sharply, mostly due to the declining birth rate and aging population.

4. The geographic distribution of migrants is extremely uneven in 2067, with the English region hosting most immigrants (More than 100 million), mostly due to the imbalanced migration amongst developed and under-developed regions.

Building upon our language population model, we establish the site rating model. If your company value cost savings and revenues expansion equally, we hereby propose the optimal location options for new offices:

Office location	Covering languages (Exclusive of English)
Tokyo	Japanese, Korean, Javanese, Malay
Bombay	Hindustani, Telugu, Tamil, marathi
Madrid	Spanish, Italian
Rio de Janeiro	Portuguese
Moscow	Arabic ,Russian ,Bengali, Hausa, Punjabi, Persian, Swahili, Yue, Vietnamese, Turkish
Paris	French, Germany

If your company offers necessities (Implying that prices of your offerings vary little to consumption power / GDP), then total population of covered languages has a greater impact on the revenues of each office. We hereby propose:

Office location	Covering languages (Exclusive of English)
Tokyo	Japanese, Korean, Hausa
Bombay	Hindustani, Telugu, Tamil,marathi
Madrid	Spanish, Portuguese
Jakarta	Javanese, Malay, Bengali
Moscow	Arabic, Russian, Punjabi, Persian, Swahili, Yue, Vietnamese, Turkish
Paris	French, Italian, German

If your company offers non-necessities (Implying that prices of your offerings vary greatly to consumption power / GDP), then accumulated GDP of covered language regions has a greater impact on the revenues of each office. We hereby propose:

Office location	Covering languages (Exclusive of English)
Tokyo	Japanese, Korean, Malay, Javanese
Bombay	Hindustani, Telugu, Tamil, marathi
Madrid	Spanish, Portuguese
Beilin	German, Hausa
Moscow	Arabic, Russian, Punjabi, Persian, Swahili, Yue, Vietnamese, Turkish, Bengali
Paris	French, Italian

Note that:

1. We select these locations from a list of candidate cities that are major cities or pivot cities in different language regions. For example, we choose Berlin to represent the German region, and Rio de Janeiro to represent the Portuguese region (Because Brazil has a higher GDP than Portugal).

2. With regard to your established offices in New York, U.S.A. and Shanghai, China, we exclude English, Mandarin, and Wu in our model (Shanghai office can cover Mandarin and Wu, and English is mandatory for all offices).

3. In the light of your globalization strategy, we assume that these 6 new offices can cover all 23 languages left of major languages in a mutually exclusive and collectively exhaustive way to save resources.

Drawing upon both models, we project our proposal to be still valid even after 50 years. However, due to economic development of Middle East and Southeast Asia, as well as the population growth of Arabic and multiple languages in Southeast Asia, we suggest that your company should move the offices in Moscow and Paris to Dubai and Jakarta. Meanwhile, the languages each office cover should also be adjusted.

We do not recommend cutting down the number of new offices below 6 in the short term, but our model projects that over time the returns of opening 5 new offices are catching up with those of opening 6 new offices, due to the development of communications and transportation. In 2067, the difference is around 5%, so we recommend your company may consider to shut down one office, if the saved costs can compensate the lost 5% revenues.

Sincerely,

Team# 73410

References

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

[2] UN Data. Net migration rate.
http://data.un.org/Data.aspx?q=migration&d=PopDiv&f=variableID%3a85

[3] UN Data. GDP by Type of Expenditure at current prices - US dollars.
http://data.un.org/Data.aspx?q=GDP&d=SNAAMA&f=grID%3a101%3bcurrID%3aUSD%3bpcFlag%3a0

[4] UN Data. Crude death rate.
http://data.un.org/Data.aspx?q=death+rate&d=PopDiv&f=variableID%3a65

[5] UN Data. Crude birth rate.
http://data.un.org/Data.aspx?q=birth+rate&d=PopDiv&f=variableID%3a53

[6] 2018 | 22nd ANNUAL EDITION LONG-TERM CAPITAL MARKET ASSUMPTIONS, by J.P. Morgan Asset Management

Appendices

Appendix A Tables

Language	Countries
Mandarin	China, Singapore
English	United Kingdom, United States of America, Canada, Australia, New Zealand, South Africa
Spanish	Spain, Mexico, Colombia, Argentina
Hindi/Urdu	India, Pakistan
Russian	Russia, Belarus, Kyrgyz Republic, Kazakhstan
Portuguese	Portugal, Brazil
Bengali	Bangladesh
French	France, Canada, Belgium, Switzerland
Japanese	Japan
German	Germany, Austria, Switzerland

Table 3: Major Countries of Different Language Groups

Language	Weighted average birth rate α	Weighted average death rate β
Mandarin	0.0121	0.0071
English	0.0136	0.0086
Spanish	0.0206	0.0078
Hindi/Urdu	0.0161	0.0058
Russian	0.0140	0.0141
Portuguese	0.0141	0.0067
Bengali	0.0193	0.0057
French	0.0113	0.0082
Japanese	0.0082	0.0094
German	0.0097	0.0093

Table 4: Birth Rate and Death Rate of Major Language Groups

Language	Native	Non-native
Mandarin	850	178
English	353	510
Spanish	400	90
Hindi/Urdu	324	214
Russian	170	26
Portuguese	200	30
Bengali	190	20
French	76	87
Japanese	130	?
German	78	8

Table 5: Population Statistics of Major Languages in 2013

Office location	Covering languages (Exclusive of English)
Tokyo	Japanese, Korean, Hausa
Bombay	Hindustani, Telugu, Tamil,marathi
Madrid	Spanish, Portuguese
Jakarta	Javanese, Malay, Bengali
Moscow	Arabic, Russian, Punjabi, Persian, Swahili, Yue, Vietnamese, Turkish
Paris	French, Italian, German

Table 6: Office Location Selections When $\eta_1 = 2, \eta_2 = \eta_3 = 1$ (More regard to population)

Office location	Covering languages (Exclusive of English)
Tokyo	Japanese, Korean, Malay, Javanese
Bombay	Hindustani, Telugu, Tamil, marathi
Madrid	Spanish, Portuguese
Beilin	German, Hausa
Moscow	Arabic, Russian, Punjabi, Persian, Swahili, Yue, Vietnamese, Turkish, Bengali
Paris	French, Italian

Table 7: Office Location Selections When $\eta_1 = 2, \eta_2 = \eta_3 = 1$ (More regard to population)

	Cycle 1	Cycle 2	Cycle 3	Cycle 4	2017 GDP	2067 GDP
India	1.77093244	1.79285437	1.20503007	1.18984108	2.11624E+12	9.63382E+12
UAE	2.34739506	2.39043403	1.32380487	1.29887136	3.70296E+11	3.57275E+12
Spain	1.77925415	1.80145680	1.20690268	1.19156508	1.19296E+12	5.49389E+12
Malaysia	2.04379203	2.07530650	1.26381944	1.24388686	8.61934E+11	5.74723E+12
Rassia	1.17979323	1.18406734	1.05484027	1.05099867	1.32602E+12	2.05362E+12
Bengaladesh	1.74035172	1.76124883	1.19810151	1.18346095	1.94466E+11	8.45179E+11
Spain	1.93249998	1.96000801	1.24046279	1.22243396	1.77259E+12	1.01811E+13
France	1.63293805	1.65032250	1.17315091	1.16046588	2.41895E+12	8.87464E+12
Nigeria	1.80587369	1.82897954	1.21285675	1.19704555	4.94583E+11	2.37168E+12
Pakistan	1.68267193	1.70166563	1.18482564	1.17122932	2.66458E+11	1.05876E+12
Japan	1.70175575	1.72137488	1.18924877	1.17530547	4.38308E+12	1.79463E+13
Germany	1.62696655	1.64415981	1.17173442	1.15915950	3.3636E+12	1.222078E+13
Iran	1.90019888	1.92656774	1.23352935	1.21606083	3.98563E+11	2.18869E+12
Tanzania	1.67179450	1.69043368	1.18229065	1.16889277	4.56282E+10	1.78202E+11
Korea	2.09597423	2.12940922	1.27450158	1.25369008	1.37787E+12	9.82618E+12
Hong Kong	1.94604458	1.97403352	1.24334890	1.22508618	3.09236E+11	1.80953E+12
Turkey	1.81628507	1.83974633	1.21517075	1.19917504	7.17888E+11	3.49557E+12
Vietnam	2.02929384	2.06027943	1.26082183	1.24113496	1.93241E+11	1.26428E+12
Italy	1.63241448	1.64978215	1.17302684	1.16035146	1.82158E+12	6.67733E+12

Table 8: Economic Outlook for the Next Fifty Years

Appendix B Figures

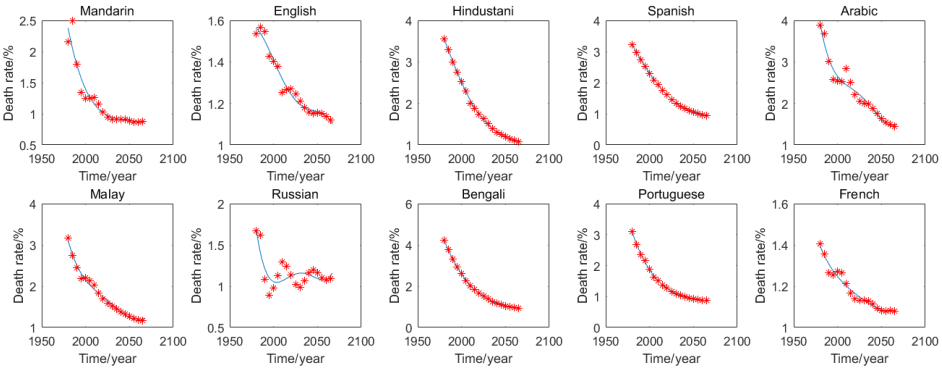

Figure 9: Graph of Birth Rate in the Next Fifty Years

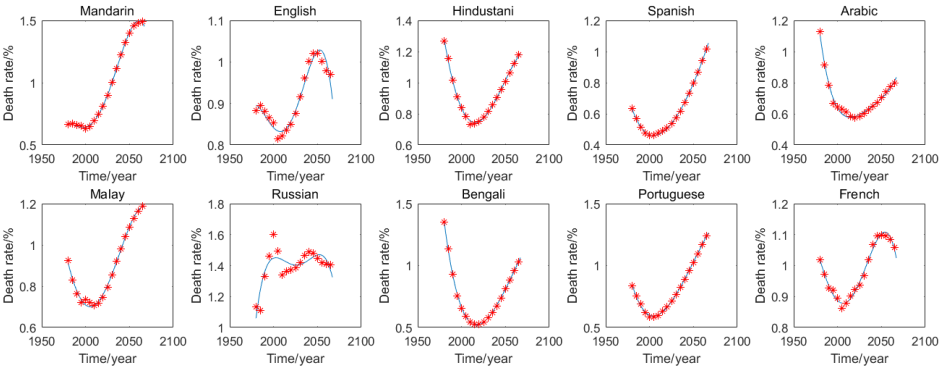

Figure 10: Graph of Death Rate in the Next Fifty Years

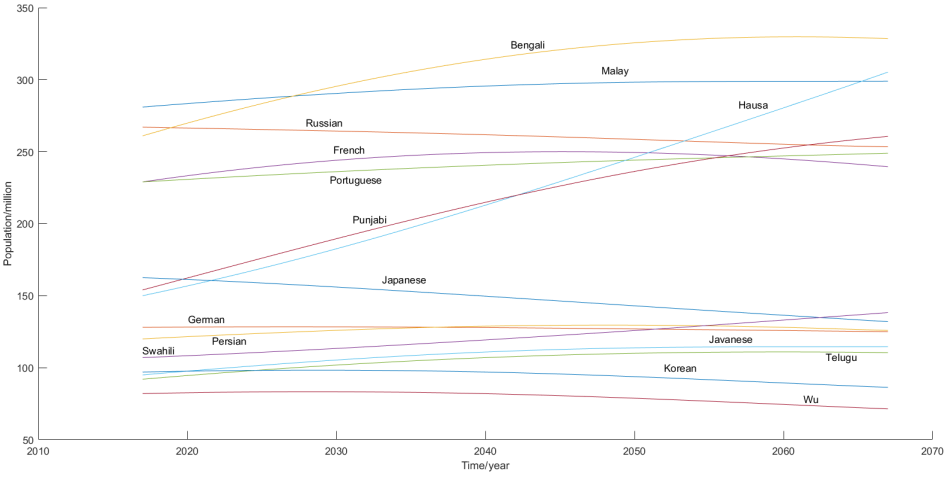

Figure 11: Graph of Population of Top 6 to 16 Language Groups in the Next Fifty Years

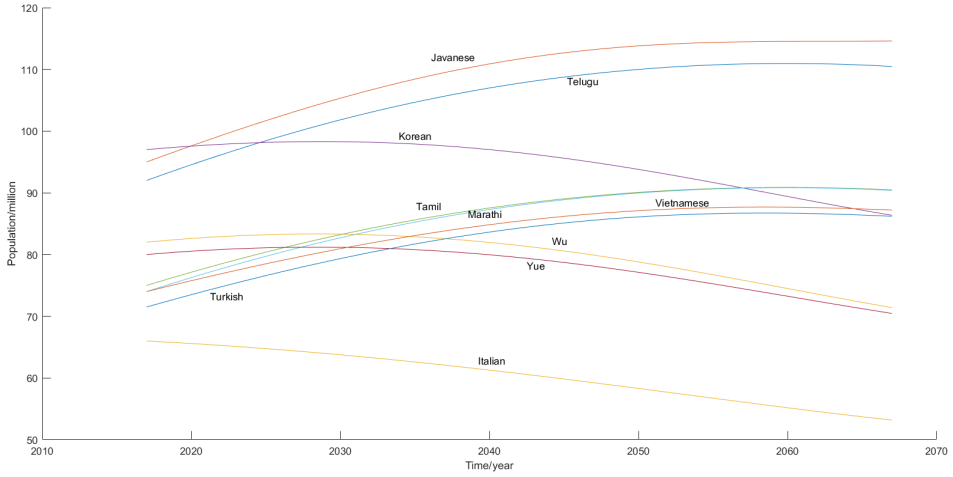

Figure 12: Graph of Population of Top 17 to 26 Language Groups in the Next Fifty Years