Predicting Trends of Languages and Determining Office Locations by Statistics Models

Haonan Duan, Saiyue Lyu, and Zhenyuan Zhang

Abstract

Our work consists of the four parts:

Firstly, we investigate and predict the future trend of the number of native language speakers in 50 years, for various languages. We employ the ARIMA model in time series analysis, and the projected trend shows that Spanish will surpass Mandarin, ranking first in 2067. In the rankings, we also expect the rise of Arabic, Punjabi, and French, as well as the drop of English, Bengali, and Russian.

Secondly, various factors leading to the trend are analyzed. Here multiple linear regression method is used to study the effects of migration, import & export, population growth rate, and GDP per capita. Our model gives a formula showing that the change in the number of native speakers is closely related to emigration, import, and population growth rate.

Thirdly, we calculate the projected geographic distributions of the main languages by continents. It is shown that a precise solution is given by analytic derivation, given an assumption we have shown to be reasonable. One sees that minor changes will occur except for Arabic and French, where the former shifts from Asia to Africa and the latter from North America and Europe to Africa.

Finally, we study the office location problem: how to wisely locate six or less international offices around the world, over both the short term and the long term. Our approach is quantitative. We consider multiple factors: future language trends, geographic distributions, GDP per capita, and minimum wage. From those factors, we derive the ASI (Aggregate Suitability Index) for each possible location to make the comparison. The result shows that Singapore, India, Brazil, Australia, Switzerland, and Egypt are optimal locations over the short term, while Egypt is replaced by Germany over the long term.

A Memo to the Chief Operating Officer

To: Chief Operating Officer
From: Team 89176
Date: Feb 12, 2018
Subject: Modeling Results for International Office Locations

Dear officer:

We have built a powerful mathematical model to help you analyze the optimal locations of international offices over both the short term and long term. Denoting $|l^j|$ the number of speakers of a fixed language j worldwide, we present the outline of our model as below:

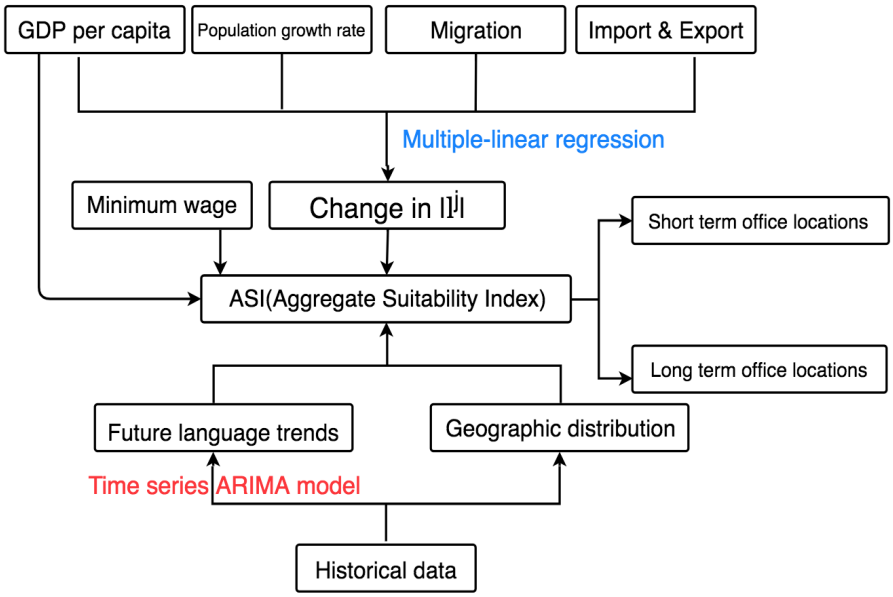

Figure 1: Long-term office locations based on ASI

Our modeling results show that there will be a significant change in the distribution of various languages in 50 years. In particular, Spanish ranking is projected to increase and surpass Mandarin to become the language with most native speakers in the world. In the rankings, we also expect the rise of Arabic, Punjabi, Wu, French and the drop of English, Bengali, Russian, Japanese. For details, see Table 2 in our paper (P10).

We also predict that minor changes will occur regarding the geographic distribution of languages, except for Arabic and French, where the former shifts from Asia to Africa and the latter from North America and Europe

to Africa. You may see Tables 3&4 in page 16 of our paper.

We define Aggregate Suitability Index (ASI) by assigning appropriate weights to multiple factors which have the major impact on the location choices: minimum wage (cost), GDP per capita (prosperity), and most importantly, current and projected number of people in the adjacent countries who speak one of the office languages. The following tables give the locations and office languages we suggest:

Table 1: International Office Locations

Country	Suggested City	Language(s)	Short Term	Long Term
Singapore	Singapore	Malay	✓	✓
India	New Delhi	Hindi, Bengali, Punjabi	✓	✓
Brazil	Rio	Spanish, Portuguese	✓	✓
Australia	Melbourne	Malay	✓	✓
Switzerland	Geneva	French,German,Italian	✓	✓
Egypt	Cairo	Arabic	✓	
Germany	Berlin	German		✓

We propose that over the short term, it is not wise to remove any of the six offices because their ASIs are close and they are uniformly distributed around the world. However, over the long term, we recommend removing either the office in Germany or in Switzerland since they are geographically close to each other. Practically, to decrease cost one may want to remove the Swiss office for its high minimum wage. Alternatively, to save client resources one may wish to remove the German office since it covers fewer people speaking the office languages than its Swiss counterpart.

We hope that our modeling and results can inspire you at some point when deciding where to open your offices. We look forward to your response.

Yours sincerely,
Team 89176

Contents

1 Introduction

1.1 Background and restating the problem

Language diversity has changed substantially over the past few decades. Currently, there are 6900 languages worldwide. Most people speak Mandarin, Spanish, English, Hindi, Arabic, Bengali, Portuguese, Russian, Punjabi, and Japanese as native languages while some people also speak a second language. The change of the language is driven by multiple factors, ranging from economics to social sciences.

We are interested in building a mathematical model to predict the trend for the number of various languages speakers based on all kinds of factors such as human development, socialization, migrant flow and other global factors. Using a tripartite division as a starting point for analysis, the problem is divided into three parts:

• Using mathematical criterion to build models to simulate the language speakers changing process and to predict the possible trend of each language

• Based on the results from the above models to re-rank the languages after 50 years and to present how the geographic distributions of these languages change with the predicted human migration patterns.

• According to the results above, we determine the locations and the number of the new international offices with possible additional information, for both short term and long term. We also determine whether it is feasible to remove one of the offices.

The languages we are interested in are those with a large number of native speakers, and we concentrate our on those with more data supporting, such as English, French, Arabic, Mandarin, Spanish, so that our conclusion can be more precise. We also give rough estimates about the other languages with fewer data.

1.2 Notations and definitions

Symbol	Definition
$\mathbb{N}$	set of positive integers
Y	a fixed year, in most cases Y=2017
I,J,K	subsets of $\mathbb{N}$ where $K = \{1, 2, 3, 4, 5, 6\}$
i,j,n,k	elements in $\mathbb{N}$
λ	a coefficient in [0,1] that is near 1
$l := l^j$	a fixed language indexed by j
$\|l\| := \|l^j\|$	number of people speaking l^j as first language
$\Delta\|l^j\|$	change in the number of speakers for l^j in two years
$\{d_k\}_{k\in K}$	set of continents
$\{c_i\}_{i\in I}$	set of countries that speak a fixed language j
$\|l_{c_i}\| := \|l^j_{c_i}\|$	number of people in country c_i that speak language l^j
W_i	the minimum wage of country i
E_i	population in i that speak English
P_i	population in both i and its adjacent countries that speak one of (or more) official languages of i
$a^{l^j}_{c_i}$	proportion of people in country c_i that speak language l^j
$b^{l^j}_{c_i}$	proportion of people speaking language l^j in country c_i
L_1, L_2	first and second language speakers, respectively
$IM^j_i(\times 10^5)$	number of immigration in year Y of c_i speaking language l^j
$EM^j_i(\times 10^5)$	number of emigration in year Y of c_i that speak language l^j
$IP^j_i(\times 10^7)$	number of import in year Y of c_i that speak language l^j
$EP^j_i(\times 10^7)$	number of export in year Y of c_i that speak language l^j
$GR^j_i(\%)$	population growth rate of country c_i that speak language l^j
$G^j_i(\$)$	GDP per capita of country c_i that speak language l^j

To be specific, $d_1, ..., d_6$ represents North America (NA), Asia, Europe, Oceania, Africa, Latin America (LAC), respectively. The only place we consider the order of these six lies in our migration flowchart. Also, we ignore the Antarctica where residents are rare.

In the last six symbols when we remove the index i, we mean the higher hierarchy: the same quantity considered language-wisely instead of country-wisely. For example, EM^j represents the total migration of a certain language j. In mathematical language,

$$EM^j = \sum_{i\in I} \frac{|l_{c_i}|EM^j_i}{|l|}$$

Also when we replace i by k we mean the same quantity considered continent-

wisely instead of country-wisely. For example in mathematical form,

$$EM_k^j = \sum_{i,c_i \in d_k} \frac{|l_{c_i}| EM_i^j}{|l_k|}$$

where $|l_k|$ is the number of people speaking language l in continent k.

1.3 Assumptions

- We assume every language is spoken in only one or two countries, with a few exceptions: English, Hindi, Spain, Arab, Bengali, and Punjabi, where we consider only the countries with the major contribution. We call the languages spoken in more than 2 countries Tier 1, and the rest languages belong to Tier 2. This assumption will not lead to large error due to the languages we have chosen to study.
- Third languages are regarded the same as second language, so are fourth languages and so on.
- We assume no influence from the family of a language on the future trend of the number of people speaking it, nor the geographic distribution of that language.
- We ignore unpredictable, or high effect but low probability events, such as financial crisis and asteroid collisions.
- The proportion of people in a continent that speaks a certain language remains the same over time. See explanations in section 4.
- For any country, every citizen must be able to speak one of the official languages, either native or not. This simplifies our model in section 5.

1.4 Overview of our models

With the above assumptions on hand, we are able to set out to the problem. We use two different models: ARIMA model of time series analysis and the multiple linear regression model. ARIMA model is used to predict the future trends of different languages in the world, while regression model is to analyze the main factors driving this trend, which gives us a deeper understanding of this change. Note that, the data of some factors, such as import, are approximated by main countries with more contributions to the language speakers. After that, we can predict the change of geographic distributions of the languages in a rough sense, by showing our "strong assumption" actually only leads to small errors.

Based on this time model and geographical model, we are able to recommend the locations for new international offices. Our ideal office locations are supposed to be multicultural (multiple official languages and large English speaking population), prosperous (high GDP per capita) and economical (low minimum wage). Aggregate Suitability Index is introduced to help us pick up the best location. We consider the above quantities now for short-term analysis, and the predicted quantities for the long term. We also analyze qualitatively whether to remove one of the offices based on geographic and economic reasonings.

The prediction of total language speakers (i.e. first and second language speakers) is only briefly studied in section 4 where we consider continent-wisely. This is mostly due to the lack of data, which will lead to big errors and even contraries.

2 ARIMA model: forecasting number of L1 speakers in 50 years

To predict the number of native speakers in the next 50 years, we choose a time series model to project the trend.

2.1 Data visualization

For each language, we collect the data of the number of native speakers from 1967 to 2017. The data is shown in Figure 2. From Figure 2, it can be concluded that the number of native speakers is growing for most languages. For Mandarin and Spanish, the increasing rates are especially large.

2.2 ARIMA model

Time series analysis is a powerful statistics technique that deals with data in a set of time periods or intervals. This method can be used to predict future values based on previously observed values. Different from the regular regression model, time series model doesn't make any independent assumption about the observations.

The specific time series model we choose is called Autoregressive integrated moving average (ARIMA) model. This model can be decomposed into 3 parts, AR, I and MA. AR stands for Auto Regression model, which represents the dependent relationship between an observation and some number of lagged observations. I represents integrated model, used when differencing the raw observations in order to make the time series stationary. MA standards for Moving Average model, which uses the dependency

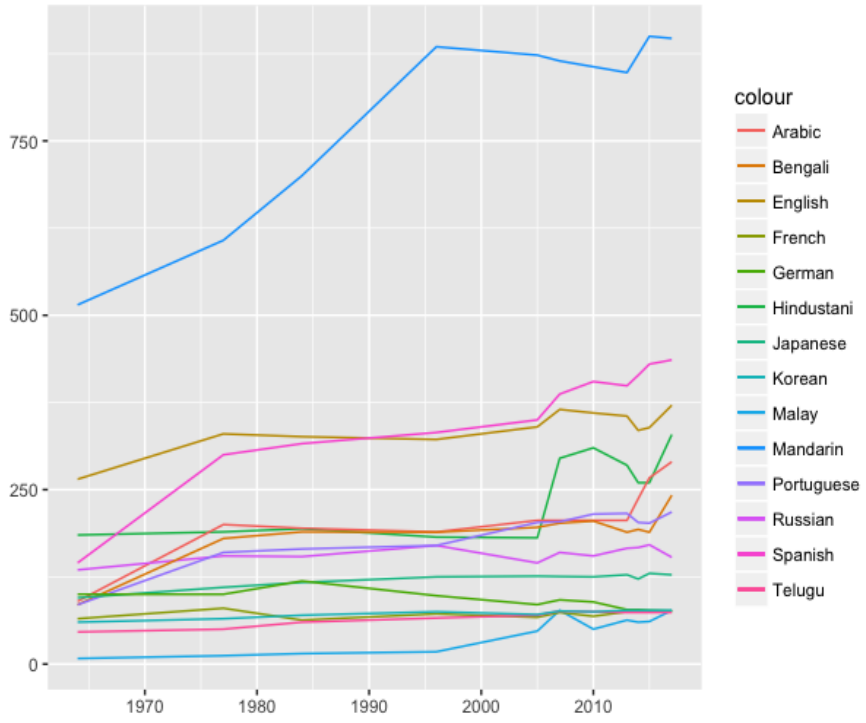

Figure 2: $|l^j|$ vs. time

between an observation and a residual error from a moving average model applied to lagged observations. The result is shown below:

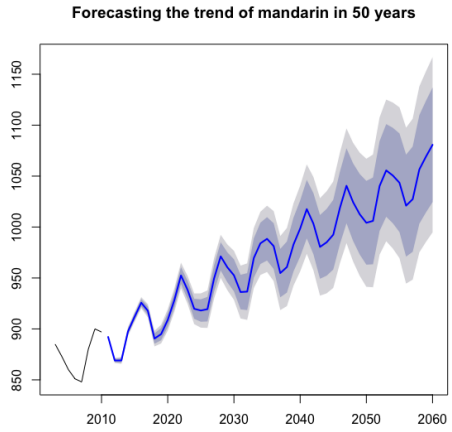

Figure 3: Forecasting for Mandarin

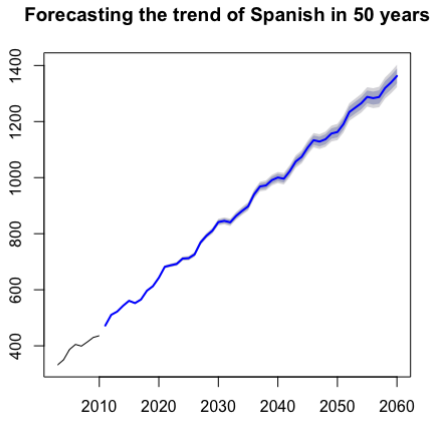

Figure 4: Forecasting for Spanish

For Spanish, we can see in Figure 4 that the number of L1 speakers will increase sharply for the next few decades. According to our prediction, the number of Spanish L1 speakers reaches 1402.3 million in 2067, which surpasses Mandarin and becomes the language spoken by most people.

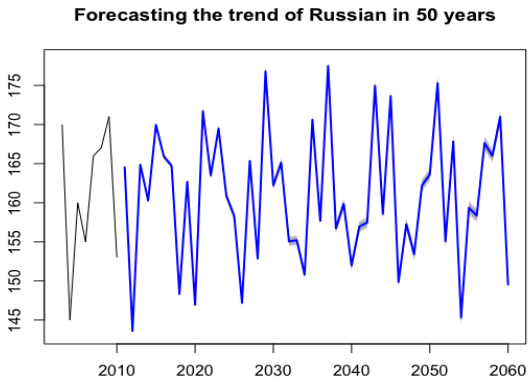

Figure 5: Forecasting for Russian

The whole forecasting result can be seen in the Table 2, Figures 6 and 7. Spanish, Arabic, Punjabi, Malay, Wu and French all go up on the ranking list in the next 50 years, while Mandarin, English, Bengali, Russian and Japanese drop on the list.

Table 2: Forecasting L1 VS Present L1

Language	Forecasting L1	Future Rank	Present L1	Present Rank
Spanish	1402	1	436	2
Mandarin	1166	2	897	1
Arabic	956	3	290	5
Hindu	888	4	329	4
Punjabi	739	5	148	9
English	584	6	436	3
Portuguese	549	7	218	7
Malay	416	8	77	12
Bengali	274	9	242	6
Chinese, Wu	273	10	90	11
French	175	11	76	13
Russian	150	12	153	8
Japanese	126	13	128	10
German	68	14	76	14

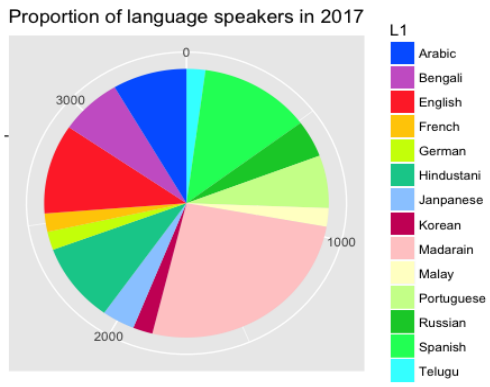

Figure 6: Proportion of each language speakers in 2017

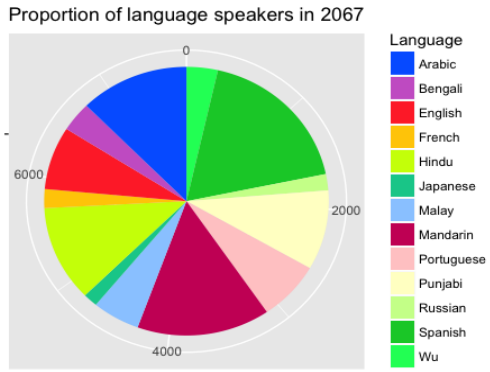

Figure 7: Proportion of each language speakers in 2067

3 Multiple linear regression Model: analyzing the explanatory variables

Our interest is to analyze how different factors may affect change of the number of speakers of a language. The response variable is the change in the number of language speakers in two years. Multiple variables we consider are immigration & migration, net immigration rate, import & export, population growth rate, and GDP per capita. The difficulty here is to col-

lect data of a certain quantity of people speaking a certain language. We overcome this by approximation.

3.1 Approximation setup

Given a fixed language l, we approximate using the countries that contribute the major part (at least λ, where recall λ is a constant between 0 and 1) of people who speak that language, and then normalize it to consider the contribution of other countries who contribute the minor part. The reason behind is that the major countries dominate the "behavior" of people who speak a certain language, while the "behavior" of those other countries with few contribution converges to the behavior observed by the dominating countries almost surely, according to the strong law of large numbers.

However, the strong law of large numbers may malfunction if the number of minor countries is too small (that is, Tier 2 languages). Hence, in this case, we may ignore the minor countries, for example, see the second assumption. In fact, the languages we choose are "well-behaved" in the sense that the contributions are largely based on only one or two countries.

Now we explain our approximation method in the first case. Choose among the set of countries $\{c_i\}_{i\in I}$ a subset of countries $\{c_i\}_{1\le i\le n_j}$ (where n_j depends on the language j) such that

$$|l_c| := \frac{\sum_{i=1}^{n} |l_{c_i}|}{|l|} \ge \lambda$$

where letting $\lambda := 0.7$, $|l_c| \ge 0.7$ acts as a normalizing factor which will be used later. The idea here is to choose n large enough such that the countries $\{c_i\}_{1\le i\le n_j}$ dominate the contribution to the language. We have

$$IM^j \approx \frac{\sum_{i=1}^{n_j} a_{c_i}^{l^j} IM_i^j}{|l_c|}$$

And note that by our previous definition,

$$a_{c_i}^{l^j} = \frac{|l_{c_i}|}{|c_i|}; b_{c_i}^{l^j} = \frac{|l_{c_i}|}{|l|}$$

The same approximations work for the quantities EM^j, IP^j, EP^j. Note that here we must use $a_{c_i}^{l^j}$ as weights for each country, because these quantities depend largely on the population. On the other hand, the quantities G^j, GR^j don't depend on the population at least in an obvious sense. Thus we use $b_{c_i}^{l^j}$ as weights:

$$G^j \approx \frac{\sum_{i=1}^{n_j} b_{c_i}^{l^j} G_i^j}{|l_c|}; GR^j \approx \frac{\sum_{i=1}^{n_j} b_{c_i}^{l^j} GR_i^j}{|l_c|}$$

3.2 Multiple linear regression specifications

Multiple linear regression is a generalized form of the simple linear regression. It contains one continuous response variable and more than one predictor. The model can be specified as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon, \epsilon \sim N(0, \sigma^2)$$

We should note that linear regression is linear in terms of the coefficient, not the explanatory variable. That means, $y = \log(x) + x^2$ is a legit linear model.

To estimate the coefficient, we should specify the model in the matrix form.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Using least square estimator, the coefficient can be calculated as

$$\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

3.3 Model fitting

First, we try to regress the response variable over all explanatory variables. According to the summary of this full model, we drop all covariates that have p-value larger than 0.05. The remaining covariates are GDP, import, and immigration. They are all statistics significant.

Next, we look at the residual plot. It shows a heteroscedasticity pattern, which implies that the constant variance assumption of the linear regression model is violated. Therefore, logarithmic transformation is applied to the response variable and some explanatory variables. The final model has adjusted R^2 to 0.82.

3.4 Final model presentation

With the approximating method illustrated in the above section, we may derive the quantities for a certain language to study the relationship between the number of speakers and those quantities.

We derive the following fitting result:

$$\log(\Delta|l^j| + 20) = 2.7867 - 0.4448\log(EM^j) + 0.3019\log(IP^j) + 1.6390GR^j$$

in alternative form,

$$\Delta|l^j| = \frac{16.2274(IP^j)^{0.3019}5.15^{GR^j}}{(EM^j)^{0.4448}} - 20 \tag{1}$$

This agrees with our intuition:

- For a language with a larger emigration, people speaking the original language will possibly abandon their original language, leading to a smaller number of speakers of that language.
- For a language with a larger import, its export amount will also be larger (according to our statistics), thus with more trading taking place, the requirement for that language will be higher.
- For a language with a higher growth rate, it is evident that the number of speakers will be larger, according to an exponential rate. In the following figure, we illustrate this point:

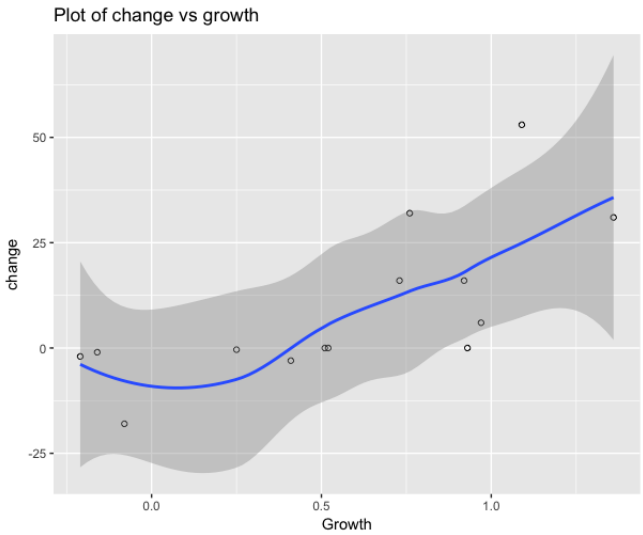

Figure 8: $\Delta|l^j|$ vs. GR^j

4 Predicting the geographic distributions

4.1 First attempt using regression equation from 3.4

The first natural attempt is to make use of equation (1) to approximate $\Delta|l^j|$ using the integral form

$$|l^j(2067)| - |l^j(2017)| = \int_{2017}^{2067} (\frac{16.2274(IP_k^j(t))^{0.3019}5.15^{GR_k^j(t)}}{(EM_k^j(t))^{0.4448}} - 20)dt$$

within a fixed continent k, then it suffices to predict the expressions of $IP_k^j(t), GR_k^j(t), EM_k^j(t)$. However, two problems arise:

- It is not absolutely perfect to apply the above equation (1) continent-wisely. Remember in the derivation of (1) we were considering the

speakers all around the world. It is, in fact possible, to do a similar multiple linear regression on each continent, but our attempt on this was prevented by lack of data.

- There are extremely big errors regarding the approximation values of $IP_k^j(t), GR_k^j(t), EM_k^j(t)$, even though we have tried regression out to fit the model. The reason is that the values $IP_k^j(t), GR_k^j(t), EM_k^j(t)$ are too complicated and involve lots of uncertainties, such as financial crisis.

It turns out that the projected values are wild, thus this attempt is considered not feasible in our settings.

4.2 Second approach using simplified model analytically

Now we take another approach to simplify our model: we are interested in the evolution of proportion of people speaking the language that is in a certain continent, i.e.

$$b_{d_k}^{l^j} := \frac{|l_k|}{|l|} = \frac{|l_k|}{\sum_{k=1}^{6} |l_k|} \tag{2}$$

Remember we are analyzing continent-wisely, so the migration within a continent is not considered. As stated in our assumption, we assume that the proportion in a continent that speaks a certain language,

$$a_{d_k}^{l^j}(Y) := \frac{|l_k(Y)|}{|d_k(Y)|}$$

doesn't change with Y. That is, when considering two different years Y,Y' (we choose Y=2017,Y'=2067 that is 50 years later) we have

$$\frac{|l_k(Y)|}{|d_k(Y)|} = \frac{|l_k(Y')|}{|d_k(Y')|} \tag{3}$$

The reason behind this assumption is the following network flowchart:

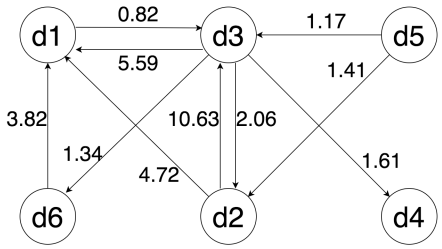

Figure 9: Migration flow between continents

From the chart we see that the number of migration is relatively small compared to the total number of people in that continent, leading to our assumption. Hence we may conclude that for languages that are mostly focused on one continent (that is Tier 2 and part of Tier 1), the evolution of geographic distribution is too small to be considered. For this reason, we only consider four languages: English, Spanish, Arabic, and French.

From (3) we have

$$|l_k(Y')| = \frac{|l_k(Y)||d_k(Y')|}{|d_k(Y)|} = \frac{b_{d_k}^{lj}(Y)|l(Y)||d_k(Y')|}{|d_k(Y)|}$$

The quantities appearing on the right-hand side can be calculated directly from our data. By plugging the result into (2) we have the following tables for 50 years later and now respectively:

Table 3: Distribution of language speakers by continents (2017)

Language	NA	Asia	Europe	Oceania	Africa	LAC
Spanish	27.2%	0	10.78%	0	0	62.02%
Arabic	0	32.59%	0	0	67.41%	0
English	71.3%	5.6%	16.5%	5.26%	1.71%	0
French	10.71%	0	20.05%	0	69.24%	0

Table 4: Distribution of language speakers by continents (2067)

Language	NA	Asia	Europe	Oceania	Africa	LAC
Spanish	28.46%	0	8.13%	0	0	63.41%
Arabic	0	18.09%	0	0	81.91%	0
English	75.63%	6.72%	12.61%	8.35%	3.73%	0
French	6.4%	0	8.63%	0	84.98%	0

From the tables, it is evident that changes of geographic distribution by continents are wild for Arabic and French but not Spanish or English. For instance, the population speaking Arabic is expected to shift significantly from Asia to Africa. Besides, we expect more weights of most languages on Africa, mostly because its population is projected to rise dramatically compared to the other continents. On the other hand, it is shown that Europe will take on less weight for the same reason.

5 Office location problem

5.1 Introduction and setup

In this section, we study how to locate six or fewer offices for an international company, whose headquarters are located in Shanghai and New York.

The analysis is largely based on the number of official languages speakers, GDP, minimum wages, the potential of the official languages (we represent it by the sum of $\Delta|l_i|$) and English-speaking population. We consider first nine possible locations: UK, France, German, Switzerland, India, Singapore, Nigeria, Egypt, Brazil. For every fixed country, we calculate the corresponding Aggregate Suitability Index (ASI) according to the above four quantities.

The reason why we choose those nine countries is that they have international metropolises. Rio de Janeiro in Brazil, London in the UK, Paris in France, and Geneva in Switzerland are very convincing examples. Also, we choose countries with a large number of English speakers since English is the most widely spoken language nowadays.

To calculate same language speakers near the country, we consider all the adjacent countries instead of countries within a certain distance, because the latter approach is hard to execute (possibly having a large territorial area). We also assume that in every country, a citizen must be able to speak one of its official languages.

5.2 Over the short term

Let $c_1, ..., c_9$ be the nine countries we are interested in. To make sure that our data are of the same scale, we need to normalize our data such that the sum of every quantity of the nine countries is 1. Then we get new normalized values $W_i', E_i', P_i', G_i', \Delta|l_i|'$. For example,

$$W_i' := \frac{W_i}{\sum_{i=1}^{9} W_i}$$

The formula of ASI is given by

$$ASI_i := \frac{3}{20}\Delta|l_i|' + \frac{8}{20}G_i' + \frac{11}{20}P_i' + \frac{2}{20}E_i' - \frac{1}{20}W_i'$$

because we are putting most our consideration on GDP and the total number of language speakers of the language we set for the office. Besides, we assign a negative weight to W_i' since it impacts our choice of that country negatively.

We have the following table to illustrate how we derived the ASI:

Table 5: Derivation of ASI for the nine countries

Country	PPP	Wage Py	ES(10^6)	SLS(10^6)	$\Delta\lvert l^j\rvert$	ASI
India	7200	743	125.345	738	40.9	0.203547
Singapore	90500	8700	4.22	191	33.7	0.150973
Brazil	15500	3491	10.542	492.66	23.4	0.138748
Australia	49900	26862	24.13	208.99	29.5	0.120601
Switzerland	61400	45398	4.68	223.15	-6	0.093708
Egypt	13000	820.8	28.1	142.32	59.8	0.091992
UK	43600	20063	59.6	86.95	20.5	0.086453
Germany	50200	21036	46.273	95.14	-10	0.070302
France	43600	20071	23	90.453	4.78	0.067277
Nigeria	5900	1122	79	85	20.5	0.057617

where PPP is GDP per capita, Wage Py is Wage per year, ES is the number of English speakers, SLS is the sum of the numbers of the official language speakers in the country and its adjacent countries. Note that

$$\frac{\sum_{i=1}^{9} P_i}{\sum_{i=1}^{9} E_i} = 5.67 \approx 5.5$$

and that is why we chose the weights to be $\frac{11}{20}$ and $\frac{2}{20}$. The following table shows which languages we choose for the offices:

Table 6: Office languages besides English

Country	Official Languages Besides English
India	Hindi, Bengali, Punjabi
Singapore	Malay
Brazil	Spanish, Portuguese
Australia	Malay
Switzerland	French, German, Italian
Egypt	Arabic

A figure to visualize our locations:

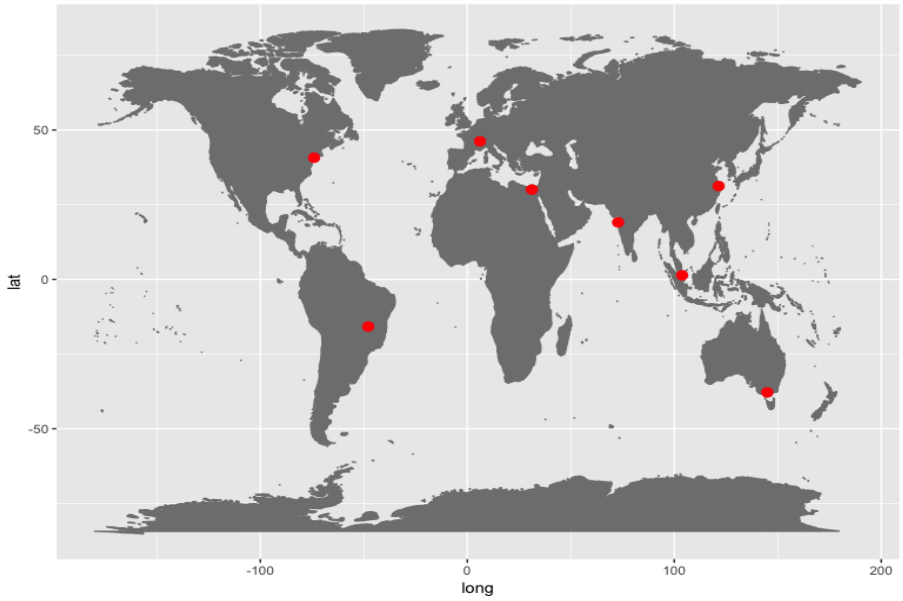

Figure 10: Office locations over short term based on ASI

One observes geographically that the offices are somehow distributed in a uniform sense. Thus it may not be a wise choice to cut the number of offices to five. However, if one insists to do so we recommend replacing Switzerland since its ASI ranks the sixth and is not near the fifth.

5.3 Over the long term

We use our result in section 3 and section 5 to analyze the change in the geographic distribution of the languages.

The change in population factor has been eliminated since forecasting gives large errors. E_i and W_i are kept the same since we are going to normalize it eventually. P_i is approximated using our result in section 5: we have

$$P_i(2065) = P_i(2015) \times \frac{b_{d_k}^{lj}(2065)}{b_{d_k}^{lj}(2015)}$$

where k is the continent where country c_i belongs to and j is a fixed language. Given also the GDP per capita projection in 50 years (G_i) we can calculate the ASI by

$$ASI_i := \frac{5}{10}G_i' + \frac{5}{10}P_i' + \frac{1}{10}E_i' - \frac{1}{10}W_i'$$

Note that we don't consider the change $\Delta|l_i|'$ which is not being well-predicted, and is actually a repetition of our predicted value P_i'.

Our result indicates that Singapore, India, Brazil, Australia, Germany, and Switzerland will become the final six locations. Figure 11 is a world map to visualize the geographic distributions of these offices:

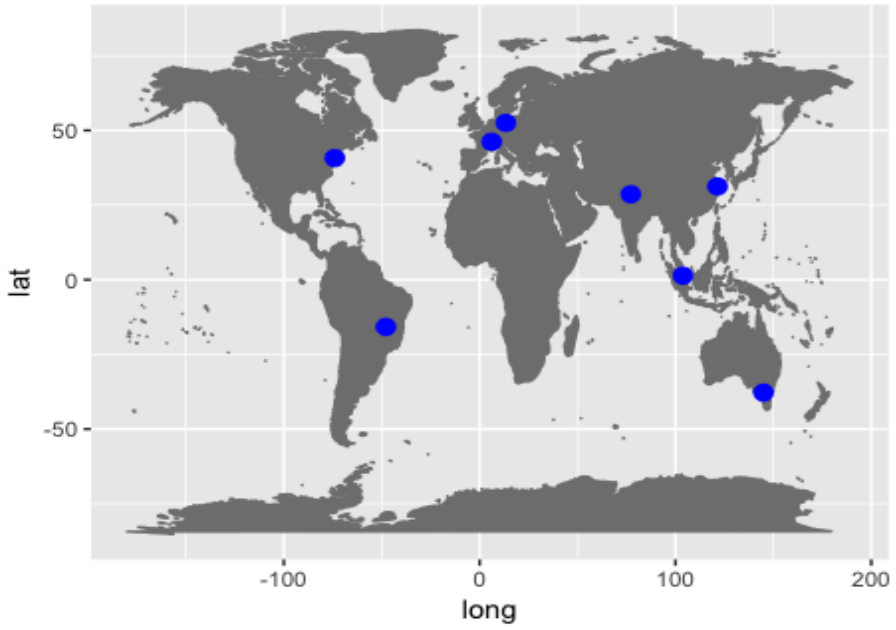

Figure 11: Office locations over long term based on ASI

Compared to the short term result, Egypt has been replaced by Germany and Arabic is replaced by German. This is because in the long run, the advantage in GDP per capita of Germany over Egypt has got comparatively large.

However, this means that the geographic distribution of those offices gets less uniform. In particular, Germany and Switzerland are near each other. Their ASI's are also very close. However, Switzerland has an advantage on diverse languages, while German's low minimum wage can save cost for the company. Hence based on our objective, to save clients or to save cost, we conclude that replacing one of the six offices is feasible. In the former case, we remove German and Switzerland.

6 Conclusions

We use time series analysis to model the projected trend of major languages in the world after 50 years. Our result shows that Spanish will surpass Mandarin to become the language with most native speakers in the world. Also, we expect the rise of ranking of Arabic, Punjabi, Malay, French

and fall of English, Bengali, Russian, Japanese. In particular, Malay and Wu will replace Russian and Japanese to become the top 10 language in 2065.

Also, we analyze how factors in social sciences and economics will affect the change of language speakers $\Delta|l^j|$, using multiple linear regression. Our result shows the following relation:

$$\Delta|l^j| = \frac{16.2274(IP^j)^{0.3019}5.15^{GR^j}}{(EM^j)^{0.4448}} - 20$$

where $IP^j(\times 10^7)$ is the import amount, $GR^j(\%)$ is population growth rate, and $EM^j(\times 10^5)$ is the number of emigration.

By making an appropriate assumption we can derive analytically the geographic distributions of each language by continents. It turns out that Arabic and French possess significant change, precisely, with the weight of Arabic shifting from Asia to Africa, of French from North America and Europe to Africa.

We tackle the office location problem by considering short term and long term separately. In the short term case, we suggest that the six offices should be located in Singapore, India, Brazil, Egypt, Australia, and Switzerland. It is not recommended in this case to remove any of the above since their geographic distribution is approximately uniform.

In the long term, we suggest that Germany replace Egypt due to predicted prosperity. In addition, we recommend that one of Germany and Switzerland should be removed from the list since the two countries are near. The final decision should base on our objective: compared to the recommendation list in the short term, we shall include Germany to attract more clients and remove Switzerland to save cost.

7 Testing the model

7.1 Sensitivity analysis

In this section, we analyze how sensitive our approximation model in 4.1 is on evaluating the parameters of languages (e.g. IM^j) using the same parameter of countries (e.g. IM_i^j). We illustrate this in terms of English and Spanish:

Table 7: Approximated quantities language-wisely, $\lambda = 0.6$

Language	$\Delta\lvert l^j\rvert$	GDP	IM^j	EM^j	IP^j	EP^j	GR^j
English	32	49.97	669.85	89.35	293.64	189.18	0.85
Spanish	6	21.97	158.67	217.28	131.43	142.71	0.83

Table 8: Approximated quantities language-wisely, $\lambda = 0.7$

Language	$\Delta\lvert l^j\rvert$	GDP	IM^j	EM^j	IP^j	EP^j	GR^j
English	32	56.49	651.43	102.2	313.38	202.29	0.76
Spanish	6	20.47	149.6	266.46	121	120.62	0.97

Table 9: Approximated quantities language-wisely, $\lambda = 0.75$

Language	$\Delta\lvert l^j\rvert$	GDP	IM^j	EM^j	IP^j	EP^j	GR^j
English	32	54.61	666.90	103.54	306.72	206.93	0.79
Spanish	6	22.78	132.79	258.97	114.45	132.76	0.96

From the above tables we see that taking $\lambda := 0.7$ is feasible since the values we approximate are within an acceptable range in our analysis, comparing to the data for $\lambda = 0.6$ and $\lambda = 0.75$. Hence we chose $\lambda := 0.7$ in section 4.1 when performing approximations.

7.2 Strengths

- **Our models are statistically significant (section 2)**
  In our multiple linear regression model, we achieved an adjusted R^2 value of 0.82, and all the p values regarding the factors we study are all less than 0.05, implying that our model can explain the data well.

- **The method of approximating data for languages by countries (section 3)**
  It is extremely hard to find the objective quantities for a single language, thus we approximate the quantity for the language by giving suitable weights to the main countries contributing speakers of that country. This not only controls error but also simplifies our model a lot.

- **Choosing the right model to predict the geographic distributions (section 4)**
  We abandon the model that uses our previous equation which relates the change in the number of speakers to emigration, import, and growth rate. This is primarily because the error is too large in predicting the above three quantities. Instead, we change our model

to an analytic one, based on a verified assumption. The new model turns out to decrease the error significantly.

- **More factors are considered in determining the office locations (section 5)**
  The office location problem is studied by taking consideration of not only the language population and its change, but also some economic factors: GDP and minimum wage. This avoids the scenario where only places with a larger population are selected, which is indeed not rational.

7.3 Weaknesses

- **Lack of data for some of the languages**
  Given more data, we could have made our time series model more precise. Also, it restricts the number of choices of languages in our multiple-linear regression analysis.

- **Some non-numerical factors were not considered for analysis**
  We didn't manage to consider the effect of the following factors to the number of speakers: political environment, cultural effect, and tourism. In fact, not considering the above factors greatly simplifies our model, and the factors we choose to study are supported by much more data and more feasible to predict numerically than those factors not considered.

- **Lack of information about the company**
  Since we are only given that the company is a service company, we can only roughly consider the effects of GDP and housing price of the chosen locations. However, our model would perform better if we had more information on how the service company operates, and what kinds of services it performs.

8 Appendices

8.1 List of figures

Table 10: Figure list

8.2 Tables

Table 11: Possible factors for change in language speakers

Language	$\Delta\|l^j\|$	GDP	IM^j	EM^j	IP^j	EP^j	GR^j
Mandarin	-3	16.6	9.78	63.48	105.6	139.49	0.41
English	32	56.49	651.43	102.2	313.38	202.29	0.76
Hindustani	69	7.2	12.89	62.91	8.77	6.4	1.17
Spanish	6	20.47	149.6	266.46	121	120.62	0.97
Arabic	23	17.29	151.17	116.37	40.75	36.93	1.58
Malay	16	14.23	8.43	17.14	4.56	5.01	0.92
Russian	-18	27.9	116.43	106	18.23	28.55	-0.08
Bengali	53	5.33	17.28	85.89	7.02	4.81	1.09
Portuguese	16	15.5	18.47	15.44	13.76	18.52	0.73
Punjabi	31	5.89	21.53	40.32	3.66	1.91	1.36
Japanese	-2	42.7	20.44	7.97	60.69	64.49	-0.21
German	-1	50.2	120.06	40.45	106.07	134.08	-0.16
Persian	-1	14.57	28.32	59.67	0.66	0.07	1.58
Korean	0	26.83	13.74	24.59	40.62	49.54	0.51
Turkish	0	26.5	29.65	31	19.86	14.25	0.52
Vietnamese	0	6.9	0.68	26	16.58	16.2	0.93
Italian	-0.4	40.93	82.28	17.88	67.4	76.7	0.25

Table 12: Historical data for native language speakers worldwide

L1 (million)	1964	1977	1984	1996	2005	2007	2010	2013	2014	2015	2017
Mandarin	515	607.5	700	885	873	865	856	848	874	900	897
English	265	330	326	322	340	365	360	355.5	335	339	371
Hindustani	185	189.5	194	182	181	295	310	285	260	260	329
Spanish	145	300	316	332	350	387	405	399	414	430	436
Arabic	90	200	194.7	189.4	206	206	206	206	237	267	290
Malay	8	12	15	17.6	47.3	77	50	63	60	61	77
Russian	135	155	154	170	145	160	155	166	167	171	153
Bengali	85	180	189.5	189	196	202	205	189	193	189	242
Portuguese	85	160	165	170	203	204	215	216	203	202	218
French	65	80	63	72	67	74	68.5	75	75	76	76
Japanese	95	110	117	125	126	125.5	125	128	122	130	128
German	100	100	119	98	85.2	92	89	78	78	77	76
Telugu	46	50	60	66	70	76	75	74	74	74	74
Korean	60	65	70	75	71	76	75	77	77	77	77

Table 13: Calculation of ASI for long term

Country	Gi'	Wi'	Ei'	Pi'	ASI
India	0.03631	0.00501	0.30981	0.35396	0.22561
Singapore	0.27958	0.05866	0.01043	0.18753	0.22873
Egypt	0.03923	0.00553	0.06945	0.06979	0.06090
Brazil	0.04750	0.02354	0.02606	0.26903	0.15852
UK	0.10746	0.13528	0.14731	0.03519	0.07253
Australia	0.10892	0.18112	0.05964	0.18197	0.13329
Switzerland	0.13018	0.30611	0.01157	0.08075	0.07601
Nigeria	0.01489	0.00757	0.19526	0.08893	0.07068
France	0.10979	0.13533	0.05685	0.03254	0.06332
Germany	0.12614	0.14184	0.11437	0.04563	0.08314
Total	1.00000	1.00000	1.00000	1.00000	1.00000

References

[1] List of sovereign states and dependent territories by immigrant population. Retrieved from https://en.wikipedia.org/wiki/List_of_sovereign_states_and_dependent_territories_by_immigrant_population

[2] Languages used on the Internet. Retrieved from https://en.wikipedia.org/wiki/Languages_used_on_the_Internet

[3] The world factbook - Central Intelligence Agency. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/rankorder/2004rank.html

[4] Trade statistics by country - WITS. Retrieved from https://wits.worldbank.org/countrystats.aspx

[5] List of countries by net migration rate, wikipedia. Retrieved from https://en.wikipedia.org/wiki/List_of_countries_by_net_migration_rate

[6] List of Languages by Total Numbers of Speakers. Retrieved from https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

[7] Ethnologue:Languages of the world. Retrieved from https://www.ethnologue.com/

[8] Maurais J.,Morris M.A. Languages in a globalising world, Cambridge: University Press

[9] Ortman J.M., Shin H.B., Language Projections: 2010 to 2020 (2011) U.S. Census Bureau

[10] Graddol D., The future of English? (1997) The British Council

[11] The world factbook - Central Intelligence Agency. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/fields/2002.html

[12] Department of Economic and Social Affairs Population Division, United Nations. *International Migration Report (2017).*

[13] World Trade Organization. *World Trade Statistical Review (2016).*

[14] List of countries by English-speaking population. Retrieved from https://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population

[15] Political Map of the World. Retrieved from https://geology.com/world/world-map.shtml

[16] Geographical distribution of French speakers. Retrieved from https://en.wikipedia.org/wiki/Geographical_distribution_of_French_speakers

[17] List of minimum wages by country. Retrieved from https://en.wikipedia.org/wiki/List_of_minimum_wages_by_country

[18] List of official languages by country and territory. Retrieved from https://en.wikipedia.org/wiki/List_of_official_languages_by_country_and_territory#I

[19] List of countries where Arabic is an official language. Retrieved from https://en.wikipedia.org/wiki/List_of_countries_where_Arabic_is_an_official_language

[20] List of countries by future population. Retrieved from https://en.wikipedia.org/wiki/List_of_countries_by_future_population_(United_Nations,_medium_fertility_variant)

[21] Eberly College of Science, Forecasting with ARIMA Models. Retrieved from https://onlinecourses.science.psu.edu/stat510/node/66