For office use only	Team Control Number	For office use only
T1 _____	75104	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	B	F4 _____

2018
MCM/ICM
Summary Sheet

A Bottom-up Prediction Model for Multi-language Competition

Summary

With the rapid internationalization, people are getting more likely to communicate with each other, whether face to face or through a screen. Thus, the interacting environment of a certain language is becoming increasingly complicated.

The main target within this paper is to model such change of language speakers of first order (native), second order (or more) and in total. The data resource is so limited that besides the data of 2017, we only get the data of native speakers of year 2007, 2009, 2010, 2014 and 2015.

Considering the impact of social environment, migration and tourism or even the Internet. We first gather some supplementary data to quantify such factors. Then we derive two indexes: GI and NGI to reflect the geographic impact, such as social environment and the non-geographic impact, such as the Internet. After that, we further derive geographic language status and non-geographic language status for our model.

In order to do robust prediction, we apply LOWESS to deal with the missing language data of year 2008, 2011, 2012, 2013 and 2016. For we don't have the data of second (or third) language besides 2017, we only interpolate the data of native speakers, which also means that a historical data based prediction can only predict the change of native speakers. The basic prediction method is mainly based on the Markov Chain model, which is exactly a historical data based method. The prediction given by Markov Chain is relatively conservative.

To provide the prediction of second (or third) language and also include the impact of geographic factors and non-geographic ones, we propose our Bottom-up Prediction Model (BPM), of which the core part is an ensemble AS Model (eASM). The eASM includes a Geo-AS model and a NonGeo-AS model to deal with the geographic impact and non-geographic impact, respectively.

When dealing with short-term situation, we fix the migration and travel pattern, as well as the net growth rate of a certain country. When long-term situation is considered, we replace them with predictions based on ARIMA model.

During sensibility analysis, we mainly consider two potential events: the refugee trend like the one in Europe and the wall that might be built between Mexico and America. Both of the two events directly affect the migration pattern in short-term.

Finally, we compare the results from the basic prediction method and BPM, using each other as a validation. Our conclusions are drawn after full analysis about the results and the models.

Keywords: language; LOWESS; Markov Chain; ensemble AS model; travel; migration; refugee

1 A Memo for the Company

To: the Chief Operating Officer of the service company
From: a team hired by the company
Date: Feb. 12^{th}, 2018
Subject: Language Report Memo

Mr. Chief,

After receiving your e-mail, we've fully investigated in modeling the language dynamics, of which the whole structure can be found in Figure 1. Through several days' work, our team has finally got something exciting.

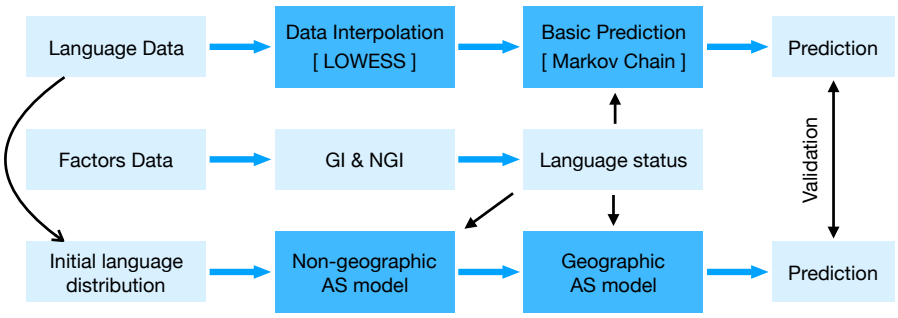

Figure 1: Structure of the integrated series of models

The numbers of native speakers and total speakers predicted 50 years later are shown in Figure 2.

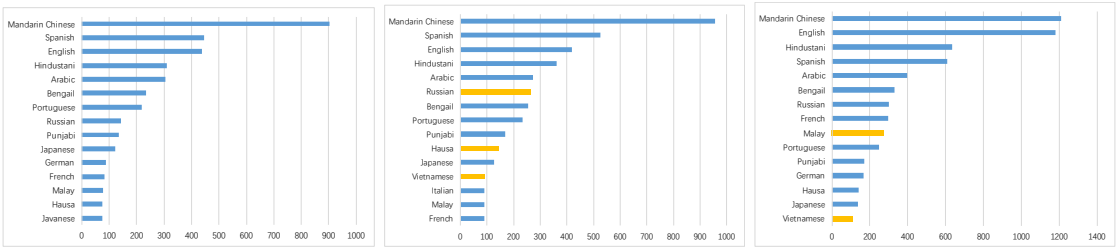

Figure 2: The top-15 rank list of year 2067. The first picture shows the result of native speakers from basic prediction model in Section 4. The second picture shows the result of native speakers from BPM in Section 5. The third picture shows the result of total speakers from the BPM in Section 5. The yellow ones are those languages whose rank differs from that of the original data. The possible reasons are explained in Section 6

The change of the geographic distributions of these languages over the next 50 years are shown in Figure 3

From Figure 3 we could see a trend from the multi-language environment to fewer languages. The typical area is East Asia, Europe and Latin America. And also we can find that the population size grows bigger in these areas.

Based on the change of geographic distributions as well as the change in rank list, we fully recommend the following 6 countries: (Note that there are already offices in America and China)

- Saudi Arabia, *Arabic*

  for its prosperity in economy and popularity in traveling

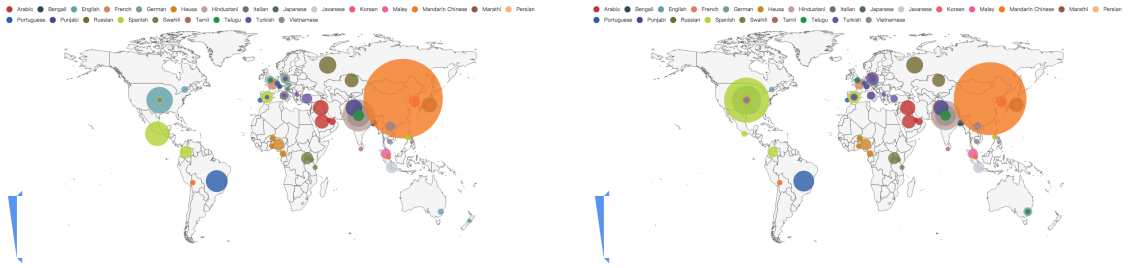

Figure 3: The change of the geographic distributions over the next 50 years. The left picture is the initial distribution in 2017. The right picture is the predicted distribution in 2067. Different color denotes different language. The size of the circle denotes the size of certain population.

- Brazil, *Portuguese*
  for its potentiality in development
- France, *French*
  for its rise in population size of native speakers
- India, *English*
  for its diversity in language and prosperity in economy
- Turkey, *Turkish*
  for its geographic advantage
- Vietnam, *Vietnamese*
  for its rapid growth in the rank list

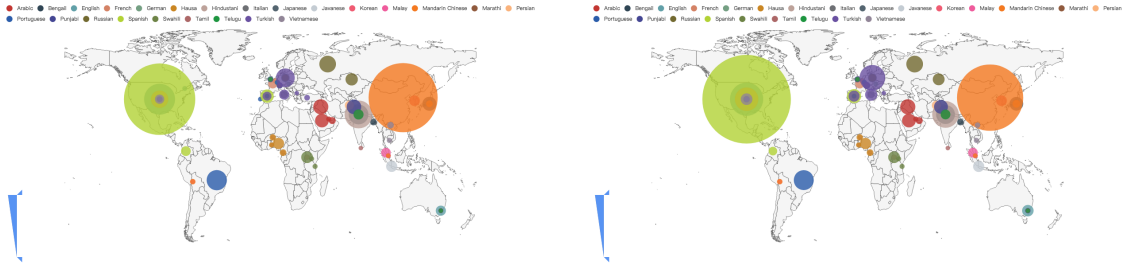

Figure 4: The left picture shows predicted geographic distributions after 25 years. The right picture shows predicted geographic distributions after 90 years.

For recommendations in short-term(25 years) or long-term(90 years), we can see from Figure 4 that the main trend barely changes over time. What's more, it has even been confirmed by long-term result.

However, this trend do help save the company's resources that fewer than 6 offices are also suitable for the world's situation. Thus, we recommend to remove Turkey and Vietnam from the 6 recommendations above. That is to say, we prefer 4 new offices in Saudi Arabia, Brazil, France and India.

All that mentioned above is included in our report. If you are interested in our model, please don't hesitate to read through our following paper.

Best,
MCM Team

Table 1: Symbols and Definitions

Symbols	Definitions
P_i	the population of the i^{th} country
P_A	the population of native speakers of language A
P_{Ai}	the population of people who speak language A in the i^{th} country
GI_i	the Geographic Index of the i^{th} country
NGI_i	the Geographic Index of the i^{th} country
T_i^{in}	the population that travel to the i^{th} country
T_i^{out}	the population that travel out of the i^{th} country
T_{ij}	the population that travel from i^{th} country to j^{th} country
M_{ij}	the population that migrate from i^{th} country to j^{th} country
M_i^{in}	the population that immigrate to the i^{th} country
M_i^{out}	the population that emigrate from the i^{th} country
$P(A \to AB)$	the probability of a native speaker of A learns language B
$P(AB \to A)$	the probability of a native speaker of A(with 2nd language B) forgets language B
$P(AB \to BA)$	the probability of a native speaker of A(with 2nd language B) exchanges the order of A& B
$P(stay)$	the probability of a native speaker of A stays unchanged
S_A^G	the geographic language status of A
S_A^N	the non-geographic language status of A
I_A	the language density of A online
$\mathcal{L}$	the set of the 24 languages
$\mathcal{L}_\mathcal{A}$	the set of countries where language A is the native language

2 Introduction

Nowadays, there are over 6,900 different languages spoken by living people. Among these languages, the top-10 languages in the number of native speakers mainly cover half of the world's population, while the top-10 list may be different when considering the total number of speakers, including 2nd(or 3rd) language. For instance, considering the number of native speakers, English(371 million) is fewer than Spanish(436) but English(983 million) in total is much more than Spanish(527 million).

The number of native (or total) speakers of a certain language will change over time, of which the potential factors are social environment, migration, tourism and the impact of widespread Internet.

Our main goal in this paper is to model the trends of these languages and help the company to locate their new offices.

2.1 Paper Structure

In our paper, we first introduce some of the related work in Section **2.2**. Then we claim some special definitions and additional assumptions in Section **2.3** and Section **2.4**. The analysis of the contributing factors and two indexes are proposed in Section **3**. In Section **4**, we construct a basic prediction method to deal with the missing data and then do the prediction of native speakers. We then propose our Bottom-up Prediction Model (BPM) in Section **5**, where we simulate the language behavior of people in a certain country as well as interactions between countries based on the AS model.[1] The result will be analyzed in Section **6**. The conclusions are given in Section **7**. The evaluation of our model is included

in Section **8**. There will also be some further discussions in Section **9**. The whole structure of our models and data is shown in Figrue 1. (The main symbols are listed in Table 1)

2.2 Related Work

The core part of this problem is to model the language competition among different group of people. The early research can be traced back to the work of *Daniel M. Abrams* and *Steven H. Strogatz* [1]. They modeled the dynamics of language death, which was restricted as a competition between two groups of people with only their native language, respectively.

Based on this, *Marco Patriarca* and *Teemu Leppn* [2] took the impact of space into consideration and they even extended the number of languages. However, they also excluded the case of bilingual speakers.

The work of *Zhang S, Yu Q* and *Bi G* [6] proposed a dynamic social network model for multi-language competition, which included bilingual speakers of a certain language and the agent based model provides an effective method to simulate such language dynamic.

2.3 Some Definitions or Explanations

- **Native speakers**

  According to the formal definition of *Native speakers* from the U.N., 'a Native speaker of a language is the one who learned it from birth, which is often used as a daily communication tool'. Here are two main points: learn from birth & use it as a communication tool. However, sometimes conflict occurs with the two points, especially when a person experienced migration. We find that the impact of a language focuses more on the number of people who actually speak it. Based on this, we prefer the latter point.

  So we define the Native speakers of language A as people who use it as a daily communication tool.

- **Geographic & Non-Geographic**

  There are mainly two models for the language change, of which one is based on face-to-face interaction, such as tourism and migration and the other one is opposite, such as online surfing and smart-phone applications. So we call them 'Geographic' & 'Non-Geographic', respectively.

- **AS model**

  The prototype of the critical models in our paper is the one proposed by *Abrams & Strogatz*[1]. And, we name it AS model for abbreviation.

2.4 Additional Assumptions

(1) **Wu Chinese & Yue Chinese are not included in our language list**, since there is already a department in Shanghai.

(2) **The data to establish our indexes can be predicted by Automatic Regression**, because we focus more on our models.

(3) **Surfing online is mainly a bilingual environment**, which will be discussed in Section 5.

(4) **The minimum time period is designed to be a year**, for the language level of a certain person changes slowly.

(5) **The language interaction between two nearby countries is ignored**, for this form of interaction will be included into the tourism and migration.

(6) **A person will have no more than two languages**, including a native language and a 2nd language. 3rd and other types will be discussed in Section 9.

3 Contributing Factors and Proposed Indexes

Recall our definition of Geographic and Non-Geographic interaction, to quantify this impact, we have introduced indicators called 'Geographic index' & 'Non-Geographic index'. By browsing the official data site such as The World Bank[1], United Nations Statistics Division[2], United Nations Economic Commission for Europe[3] and National Bureau of Statistics of China[4], we finally narrow down our concentration on data shown in Table 2:

Index	Aspects Quantified	Data Description
Geographic	Development	Merchandise trade, GDP, High-tech exports
	Landscape	Forest area, PM2.5 air pollution, International tourism Income
	Livability	Total health expenditure, Renewable electricity output, Life expectancy, Population density, International murder rate
Non-Geographic	the Internet	Weekly users of Internet, Secure Internet Server, Power rate
	Cultural worship	Science and technology journal articles, School enrollment, Research and Development Expenditure, Patent application, Population living in slums, Per Capita GDP

Table 2: Data collected from several official site which is used to quantify our index

Since poor-quality data will make it harder for our models to detect the underlaying patterns, the first thing is to clean up our collected data (*e.g*, discard the outliers and try to fix some errors manually, ignore the missing values and deal with NaN, do normalization). After having done that, for each aspect, we sum up corresponding data group and divide each data of this new summed column by its maximum to apply standardization. This is regarded as GI and NGI.

We can further define the language status based on the two indexes (1), which will be fully used in the following models. (see Figure 5 for the status in 2017)

$$S_A^G := \sum_{i \in \mathcal{L}_\mathcal{A}} GI_i, \; S_A^N := \sum_{i \in \mathcal{L}_\mathcal{A}} NGI_i \tag{1}$$

[1] `https://data.worldbank.org/`
[2] `https://unstats.un.org/home/`
[3] `http://w3.unece.org/PXWeb/en`
[4] `http://www.stats.gov.cn/english/`

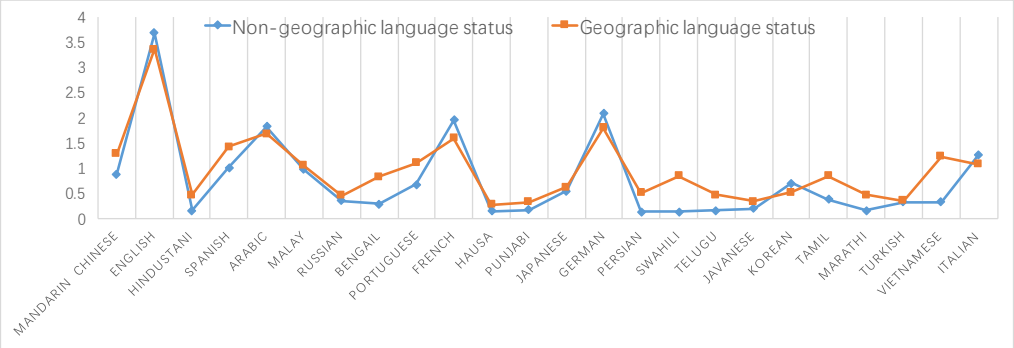

Figure 5: The geographic language status and non-geographic language status in 2017.

4 Basic Prediction Model

4.1 Data & preprocess

From the statistic available from *Ethnologue* [5] and *Wiki* [6], we can only collect the summary of native speakers all over the world of year 2007, 2009, 2010, 2014, 2015 and 2017. (see Figure 6)

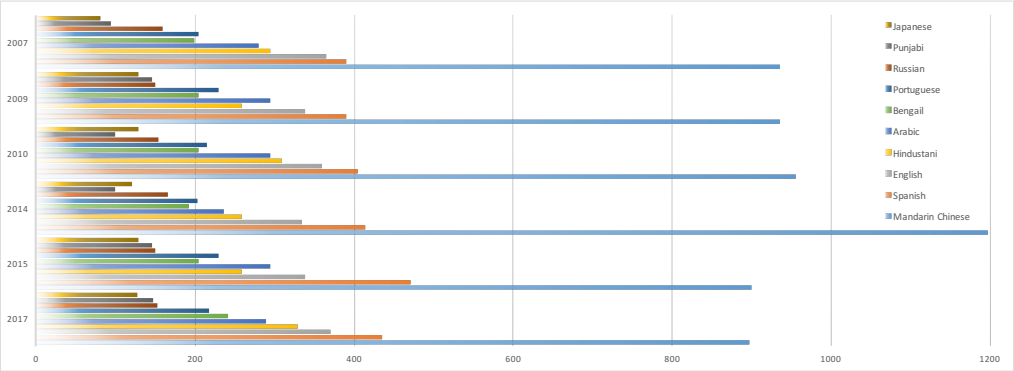

Figure 6: The top 10 languages of the number of native speakers (*million*)

Remarks:

- There is an obvious outlier in Figure 6 that the native speakers of Chinese in year 2014 is too much larger than that of 2010 and 2015. So, we simply reset it as *1000 million*.
- There are many languages such as 'Malay', 'Hausa' and 'Swahili', of which the statistics were not updated in some of the years or even all of the years in Figure 6.

So here, we will apply two different methods to deal with the missing data of year 2008, 2011, 2012, 2013 and 2016. One is based on the data itself and the other is based on some factors, including social pressures, migration and assimilation of culture groups.

[5] `https://www.ethnologue.com/`
[6] `https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers`

4.2 Data Interpolation — LOWESS

LOWESS (LOcal Weighted regrESSion) is frequently used as a method for data interpolation. The details of this method please refer to the work of *W.S. Cleveland.*[3] Here we apply it in $\mathcal{R}$, where we choose the *Epanechnikov*'s kernel of degree 2. The interpolation turns out to be good, see Figure 7 as an example.

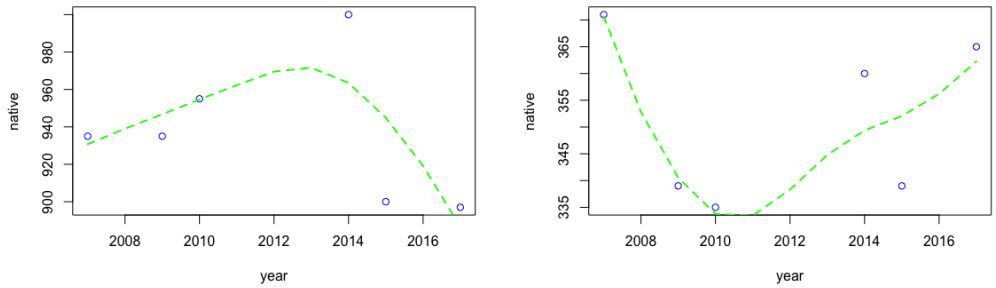

Figure 7: A demo of the data interpolated by LOWESS. The left one is Chinese and the right one is English. The blue points are known data and the green curve is regressed by LOWESS.

4.3 Data Prediction — Markov Chain

Here we use Markov Chain model to predict the distribution of native speakers in the next 50 years, based on the interpolated data from Section 4.2.

Here, our work is based on the estimation of markov chain transition probabilities and rates from partially observed data, which has been fully discussed in *Welton*'s paper.[4]

Note that the Markov Chain we use here is a little bit different, for we will first include the impact of language status of each year. The main reason is that the data of year 2008, 2010, 2011, 2012, 2013 and 2016 is interpolated by LOWESS, so we are more careful with the data of these years, of which the transition matrixes are constructed separately.

Also note that, for lack of the data for 2nd language speakers in any other year different from 2017, we can only predict the number of native speakers over time within this method. The prediction including 2nd language will be covered by BPM in Section 5.

Constrcution rule For a model of n languages (which is 24 in our paper), we construct a matrix $U_{n\times n}$. The element of U is determined by (2)

$$U_{BA}(t) := c \cdot \frac{(S_B^G(t) - S_A^G(t)) \cdot e^{S_B^N(t) - S_A^N(t)}}{\sum_{A\in\mathcal{L}} S_A^G(t)} \tag{2}$$

where S_A^G is the geographic language status of A language, S_A^N is the non-geographic language status of A language, c is a constant to minimize the MSE(Mean Square Error) of the error between the predicted distribution and the known distribution of next year, which is similar to the *bias* in linear regression.

From the available data in Section 3, we could do such construction year by year.

$$(P_A(t+1), P_B(t+1), \cdots)^T = U(t) \cdot (P_A(t), P_B(t), \cdots)^T \tag{3}$$

As a testament, we could compare the data predicted by Markov Chain with that interpolated by LOWESS. (see Figure 8)

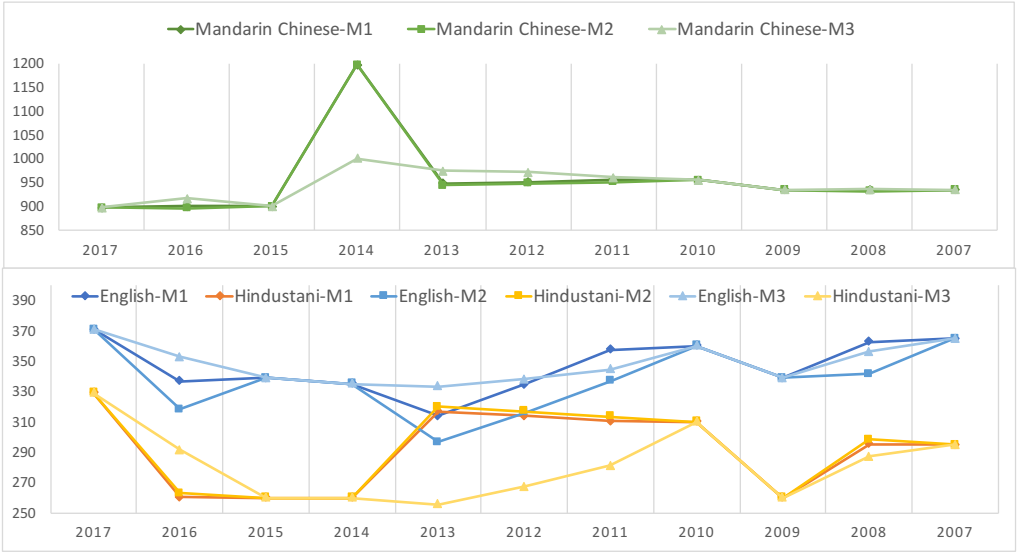

Figure 8: A demo of the comparison between interpolated data and predicted data. We actually compare three models here. M1 denotes the Markov Chain model with constant C = 1e-3; M2 denotes the Markov Chain model with constant C = 1e-2; M3 denotes the LOWESS model. We could see that the difference between the two different constant is little and considering that the status of languages shouldn't change greatly overtime, we prefer the constant C = 1e-3. And also, the data given by the two methods are similar, which provide a testament to each other.

Recall that we construct a transition matrix every year (2007, 2010, 2011, 2012, 2015) to get next year's (2008, 2010, 2011, 2012, 2013, 2016) prediction on native speakers distribution. Now, in order to do further prediction (next 50 years), we assume that there is an invariant transition pattern in such distribution, so we do the iterated prediction based on the five different transition matrix. Before that, we conduct the validation by predicting the 2017's distribution as follows.

$$
\begin{aligned}
(P_A(2017), P_B(2017), \cdots)^T &= U(2007)^{10} \cdot (P_A(2007), P_B(2007), \cdots)^T \\
(P_A(2017), P_B(2017), \cdots)^T &= U(2010)^{7} \cdot (P_A(2010), P_B(2010), \cdots)^T \\
(P_A(2017), P_B(2017), \cdots)^T &= U(2011)^{6} \cdot (P_A(2011), P_B(2011), \cdots)^T \\
(P_A(2017), P_B(2017), \cdots)^T &= U(2012)^{5} \cdot (P_A(2012), P_B(2012), \cdots)^T \\
(P_A(2017), P_B(2017), \cdots)^T &= U(2015)^{2} \cdot (P_A(2015), P_B(2015), \cdots)^T
\end{aligned}
$$

The prediction error is shown in Figure 9, from which we can see that the errors are all acceptable, except for some languages, such as Chinese, English and Hindustani, in certain years, such as 2013 as 2016.

In order to do further prediction, we derive a weighted sum of the five transition matrix as follows, the prediction error is the red line in Figure 9.

$$
U^* := \sum_t c_t \cdot U(t) \tag{4}
$$

where $c_t = \frac{t+1-2007}{2007+2010+2011+2012+2015-5*2007+5}$, $t \in \{2007, 2010, 2011, 2012, 2015\}$.
Based on the transition matrix above, we could have the change of number of native speakers over time. (see Figure 10)
The analysis of the language ranking will be included in Section 6, together with that of

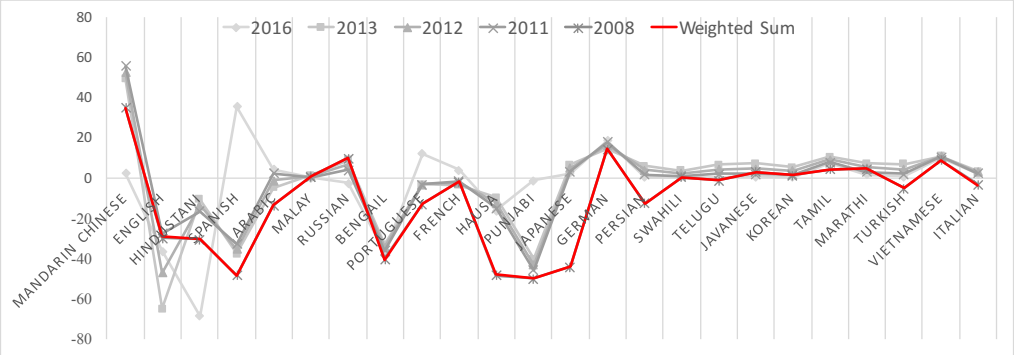

Figure 9: prediction error w.r.t the data of 2017

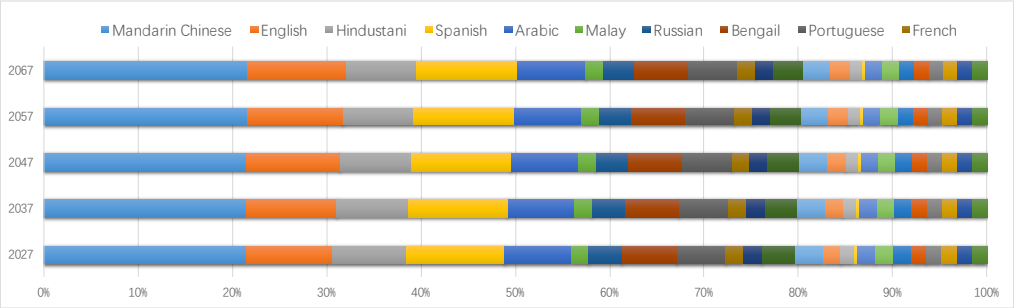

Figure 10: The percentage of native speakers of every language every 10 years.

BPM.

5 A Bottom-up Prediction Model

5.1 Language distribution

Given List of Languages by Total Numbers of Speakers[7], with references to each language's detailed wikipedia (*e.g*, English[8]) which made it possible for us to found out the distribution[9] of this language (*e.g*, L1 countries: United States, United Kingdom, Australia, New Zealand, Canada, Singapore; L2 countries: Malaysia, China, India, Germany, Luxembourg, Italy, Spain, Portugal), then we simply distribute L1 speakers and L2 speakers by the proportion of these countries' population. We use this data as initial input data of our model. (see Figure 11 as a demo)

5.2 Prediction of language status

Note that the data we gathered in the previous section are all time series data, such as GI, NGI and relating language status. To do the prediction of time series data, we apply the ARIMA model. The ability of predicting time series data of ARIMA (AutoRegressive Integrated Moving Average) model has been tested in the work of *Contreras J. et al.* [5] and others, which is also a classic model in the course of *Analysis of Time Series*.

There are three hyper-parameters in the model (p, d, q), where p denotes the hyper-parameter of the Moving Average procedure, q denotes the hyper-parameter of the Au-

[7] `https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers`
[8] `https://en.wikipedia.org/wiki/English_language`
[9] `https://en.wikipedia.org/wiki/English-speaking_world`

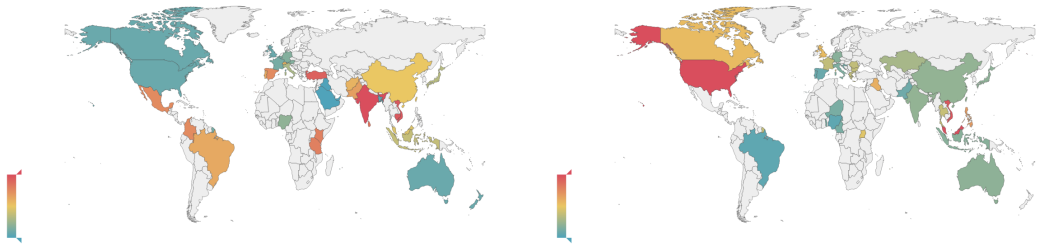

Figure 11: The initial language distribution (2017). The left picture shows the distribution of native language. The right picture shows the distribution of 2nd language. Note that the 'green-like' color denotes the Indo-European languages and the 'red-like' color denotes the Sino-Tibetan languages. For example, China in the left picture shows that its main native language belongs to Sino-Tibetan and China in the right picture shows that its main 2nd language belongs to the Indo-European.

toRegressive procedure and d denotes the degree that enables the difference sequence is stationary.

Take English for an example. (see Figure 12)

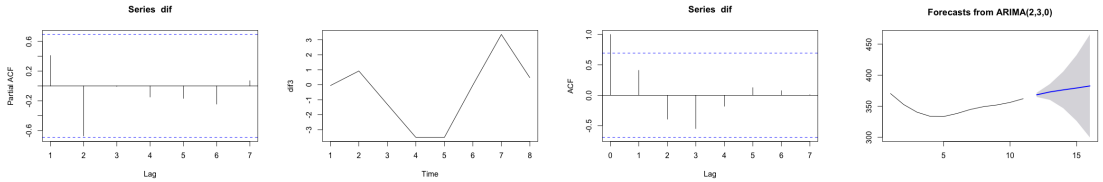

Figure 12: An example (English) of the procedure of ARIMA model. The first picture is the ACFE (Auto-Correlation Function Estimation) for $p = 2$. The second picture is the difference sequence of English native speakers, which is stationary and the degree $d = 3$. The third picture is the PACFE (Partial Auto-Correlation Function Estimation) for $q = 0$. The last picture is a demo for predicting the next 5 years using ARIMA(2,3,0), along with the 95% confidence interval.

5.3 Travel and migration pattern

Based on the migration population, we could build a human tourism and migration pattern.

For a certain country (let's say the r^{th} country), we have the population that migrate from it, M_r^{out}, and other countries' population that immigrate to them, M_i^{in} , $\forall i \neq r$. We then have the transition pattern.(5)

$$\forall i,\ M_{ri} = M_r^{out} \cdot \frac{M_i^{in}}{\sum_{j \neq i} M_j^{out}} \tag{5}$$

So, eq (5) gives out the pattern for population transition, but when dealing with the language interaction within a certain country [Assumption 5], we need take the impact of tourism into account. That is to say, based on eq (5), we can update the population over time. (6)

$$\forall i,\ P_i(t+1) = P_i(t) + M_i^{in} - M_i^{out} \tag{6}$$

However, the population in which language interaction happens should include the tourism.

(8)

$$\forall i,\ T_{ri} = T_r^{out} \cdot \frac{T_i^{in}}{\sum_{j \neq i} T_j^{out}} \tag{7}$$

$$\forall i,\ P_{Ai}(t) = P_i(t) + T_i^{in} - T_i^{out} \tag{8}$$

Remark: The item M_{ji}&T_{ji} seem to be unnecessary for the population update in eq (6)& eq (8), but they are essential because of the particular language they speak.

5.4 Ensemble AS model

After the population update, we have to deal with the language interaction. We conclude that the language influence is mostly based on two cases: 1) read from books or communicate online, which is not geographic. 2) build and maintain social relationship with people around, which depends on the location.

So respectively, we derive NonGeo-AS model and Geo-AS model to predict language changing trend with the impact from cases above, which are based on the work AS model of *Daniel M. Abrams, Steven H. Strogatz.*[1]

In that paper, *Daniel M. Abrams, Steven H. Strogatz* derived a bilingual competition model as eq (9) & eq (10).

$$\frac{dP_A}{dt} = \lambda_A \cdot P_B - \lambda_B \cdot P_A \tag{9}$$

$$\frac{dP_B}{dt} = \lambda_B \cdot P_A - \lambda_A \cdot P_B \tag{10}$$

where $\lambda_i = c_i s_i P_i^{\alpha}$, c_i is a constant, s_i is the language status index and $\alpha \in (1.05,\ 1.55)$ is also a constant.

5.4.1 NonGeo-AS model

However, according to this traditional model, all people are monolingual (A/B speaker). But the assumption for our AS models is that people all have the ability to acquire 2nd language($A/B/AB/BA$ speaker) but no person is trilingual or more.

The different part of NonGeo-AS model is that we assume a person who surf online will not forget his 1st language and will not exchange the 1st and 2nd languages' orders either. That is to say, assuming A is the 1st language, there are only three actions in this model.

- $A \to AB$. Learn a 2nd language:

$$P(A \to AB) = S_B^N \cdot I_B \tag{11}$$

- $AB \to A$. Forget a 2nd language:

$$P(AB \to A) = S_A^N \cdot I_A \tag{12}$$

- $A \to A$. Stay unchanged:

$$P(stay) = 1 - P(A \to AB) - P(AB \to A) \tag{13}$$

where S_A^N, S_B^N are the non-geographic language status, I_A, I_B are the language densities online. Which means, we can regard the world as one whole country and the probability of connecting with others depends on the population of every country.

We firstly show the validity of NonGeo Model in an ideal environment, where we set that three monolingual group $(A/B/C)$ in this tiny world, and then launch NonGeo Model. So they might learn and forget the 2nd language.

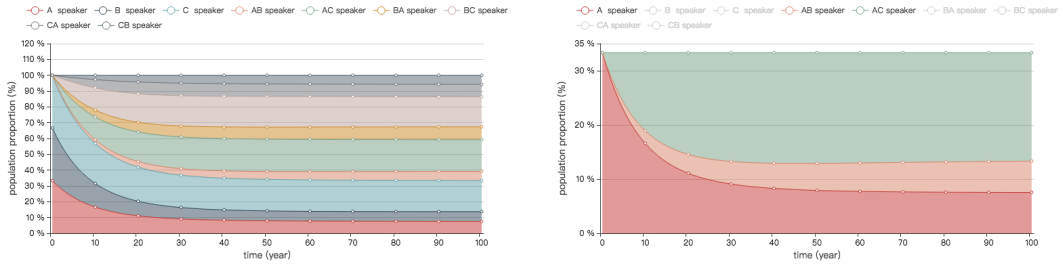

Figure 13: Stack statistics for result of NonGeo-AS model ideal experiment, where experiment settings are: $S = [0.2, 0.2, 0.6]$, the number of monolingual speakers are equal, no bilingual speakers. The left picture shows the proportion of all speaker. The right picture shows the proportion of A speaker

From the left picture in Figure 13 we know, there appear bilingual speakers with time going by. Though the number of monolingual A speaker(red) is declining, the proportion of people whose the 1st language is A is always 34% in the right picture. According to that we know A speaker will never forget his 1st language and might learn other language of B or C in NonGeo-AS model.

5.4.2 Geo-AS model

We consider the languages competition with people living around as geographic languages competition, so we propose Geo-AS model to address the problem.Comparing the assumptions with NonGeo-AS model, people in Geo-AS model are different in two ways: 1) they are only able to interact with people in the same country. 2) they can change their 1st language.

For example, if A speaker is going to change his 1st language as B, he must experience the process that is $A \to AB \to BA \to B$. And all the possibilities of states transfer are illustrated as follows.

- $A \to AB$. Learn a 2nd language:

$$P(A \to AB) = S_B^G \cdot I_B \tag{14}$$

- $AB \to BA$. Two languages compete:

$$P(AB \to BA) = S_B^G \cdot (I_B + I_{BA}) \tag{15}$$

- $AB \to A$. Forget a 2nd language:

$$P(AB \to A) = S_A^G \cdot I_A \tag{16}$$

- $A \to A$. Stay unchanged:

$$P(stay) = 1 - P(A \to AB) - P(AB \to BA) - P(AB \to A) \tag{17}$$

where S_A^G, S_B^G are the language status calculated by Geographic Index, which should be normalized by all kinds of languages. And I_A, I_B, I_{BA} are the language densities, which is the proportion of people with this certain languages.

Now, we will illustrate the validity of Geo-AS model using an experiment, which is of the same settings with NonGeo-AS model ideal experiment and three language groups are regarded living in the same country. The results show in Figure 14.

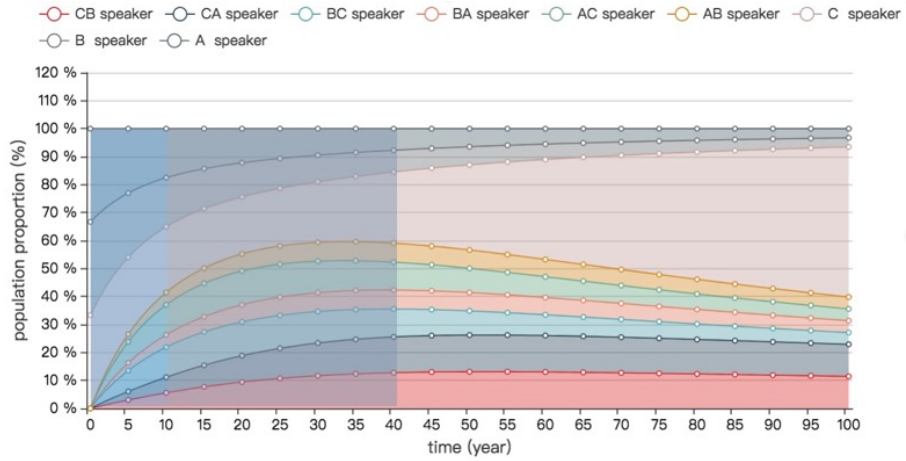

Figure 14: Geo-AS model ideal experiment result for $A/B/C$ languages competition, where experiment settings are: $S = [0.2, 0.2, 0.6]$, the number of monolingual speakers are equal, no bilingual speakers. **Stage** 1: year 0 $\rightarrow$ year 11. **Stage 2**: year 12 $\rightarrow$ year 42. **Stage 3**: year 43 $\rightarrow$ year 100

Geo-AS model simulates the multi-language competition well.

- With the country initialized, $A/B/C$ speakers begin to learn the 2nd language from each other, so bilingual speakers increase, enlarging these corresponding areas in Figure 14 Stage 1.
- The status of language C is higher than others, so more X speakers become XC speakers, and converted to C speakers, so other monolingual speakers are declining in Figure 14 Stage 2.
- The C language wins, other 1st languages are decreasing and will be dead in future, the trend shows in Figure 14 Stage 3.

5.4.3 Ensemble AS model

The Ensemble AS model is the combination of NonGeo AS model and Geo AS model. Moreover, the prediction model takes **migration**, **traveling** into consideration. For t year, the model runs as:

- Update the language population by the predicted migration and traveling ratio of t year.
- Update the language population by NonGeo AS model with the predicted S^N.
- Update the language population by Geo AS model with the predicted S^G.
- Update t to $t + 1$, continue or halt the model.

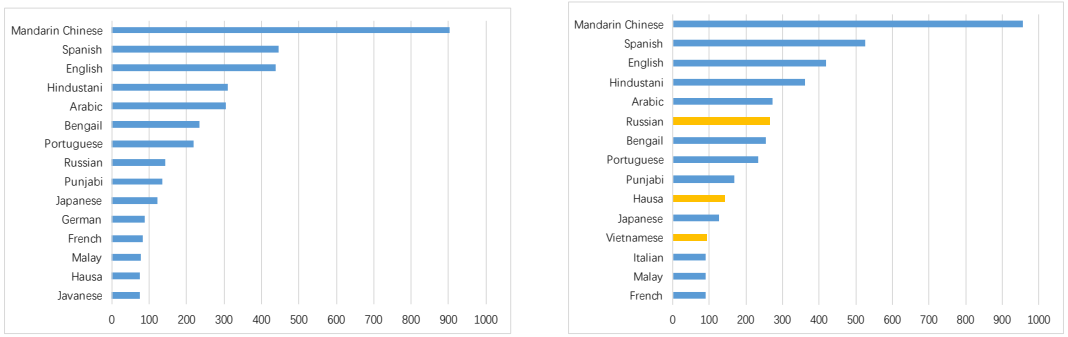

Figure 15: The rank of native speakers (top-15 list). The left comes from the basic prediction model in Section 4. The right comes from the BPM in Section 5

6 Results Analysis

6.1 Analyze the ranking of native speakers

From Figure 15 we can see that the predicted rank from the basic prediction model in Section 4 stays exactly the same as the rank of 2017. The reason may be that the transition matrix is based on the stationary pattern and what's more, the data in 2017 show that the native speakers of Japanese (rank 10^{th}, 128 million) are much more than those of Hausa (rank 11^{th}, 85 million).

By comparison, the rank predicted by BPM shows some difference from the original rank in 2017 that Russian and Hausa step into the top-10 list. Vietnamese steps into the top-15 list.

Our data show that the geographic language status of Vietnamese is high, which symbolize the popularity of travel and migration to the countries of Vietnamese. Speaking of Russian and Hausa, the language environment within those corresponding countries is relatively simple so that the people there are less likely to exchange the order of their native language and 2nd language, thus cementing the status of the two languages within those countries.

6.2 Analyze the ranking of 2nd and total speakers

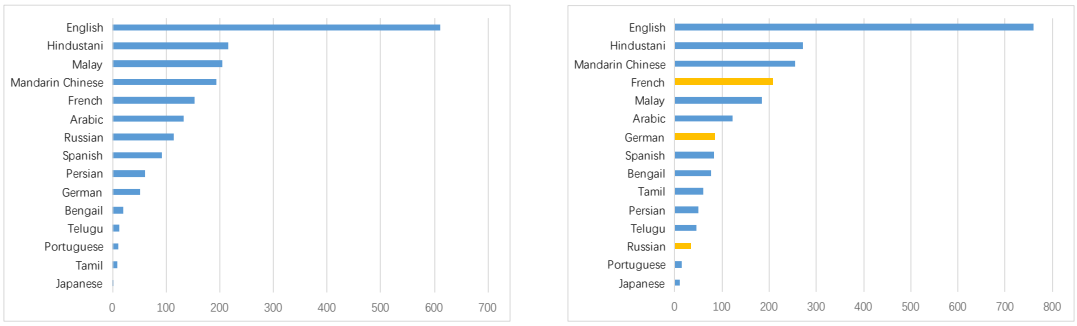

Figure 16: The rank of 2nd language speakers (top-15 list). The left comes from the original data in 2017. The right comes from the BPM in Section 5

From Figure 16 we can see that the rank predicted by BPM shows some difference from the original rank in 2017 that French and German step into the top-10 list. Russian steps into the top-15 list. The reason for French and German may be that the GI and NGI of the two languages are both high. (see Figure 5)

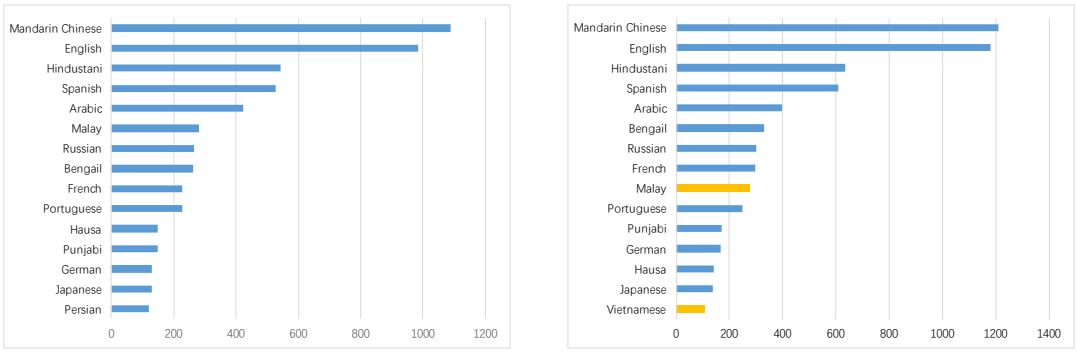

Figure 17: The rank of total language speakers (top-15 list). The left comes from the original data in 2017. The right comes from the BPM in Section 5

From Figure 17 we can see that Vietnamese again steps into the top-15 list, while Malay goes down. The reason for Vietnamese may be the growth in its native language speakers and actually, the number of Malay don't change greatly.

6.3 Analyze the geographic distribution change

The change of the geographic distributions of these languages over the next 50 years are shown in Figure 18

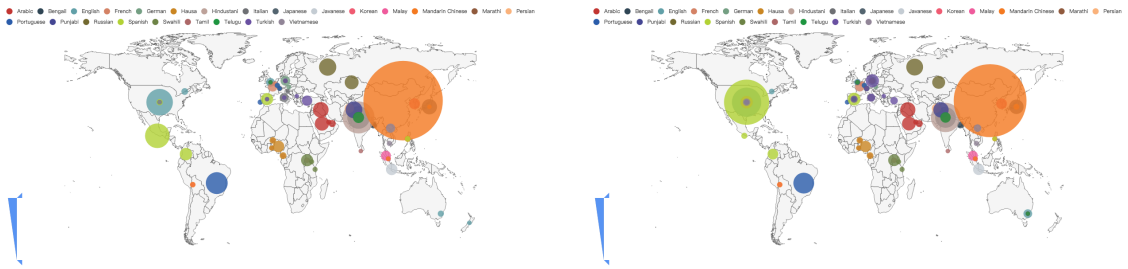

Figure 18: The change of the geographic distributions over the next 50 years. The left picture is the initial distribution in 2017. The right picture is the predicted distribution in 2067. Different color denotes different language. The size of the circle denotes the size of certain population.

From Figure 18 we could see a trend from the multi-language environment to fewer languages. The typical area is East Asia, Europe and Latin America. And also we can find that the population size grows bigger in these areas.

7 Conclusions

Based on the change of geographic distributions as well as the change in rank list, we fully recommend the following 6 countries: (Note that there are already offices in America and China)

- Saudi Arabia, *Arabic*

  for its prosperity in economy and popularity in traveling

- Brazil, *Portuguese*

  for its potentiality in development

- France, *French*
  for its rise in population size of native speakers
- India, *English*
  for its diversity in language and prosperity in economy
- Turkey, *Turkish*
  for its geographic advantage
- Vietnam, *Vietnamese*
  for its rapid growth in the rank list

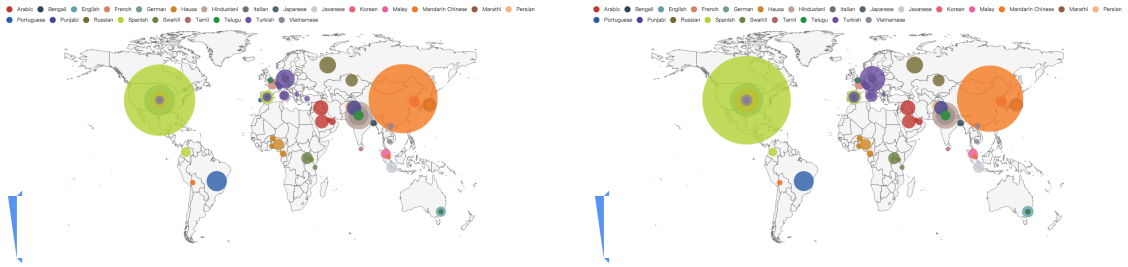

Figure 19: The left picture shows predicted geographic distributions after 25 years. The right picture shows predicted geographic distributions after 90 years.

For recommendations in short-term(25 years) or long-term(90 years), we can see from Figure 19 that the main trend barely changes over time. What's more, it has even been confirmed by long-term result.

However, this trend do help save the company's resources that fewer than 6 offices are also suitable for the world's situation. Thus, we recommend to remove Turkey and Vietnam from the 6 recommendations above. That is to say, we prefer 4 new offices:

- Saudi Arabia
- Brazil
- France
- India

8 Sensibility analysis of BPM

Here we investigate the two potential events in Figure 20.

9 Further Discussion

9.1 3rd or more language speakers

For lack of the data of 3rd (or more) language speakers, we cannot model such group of people. However, the integrated model in Figure 1 provides a complete workflow for such task.

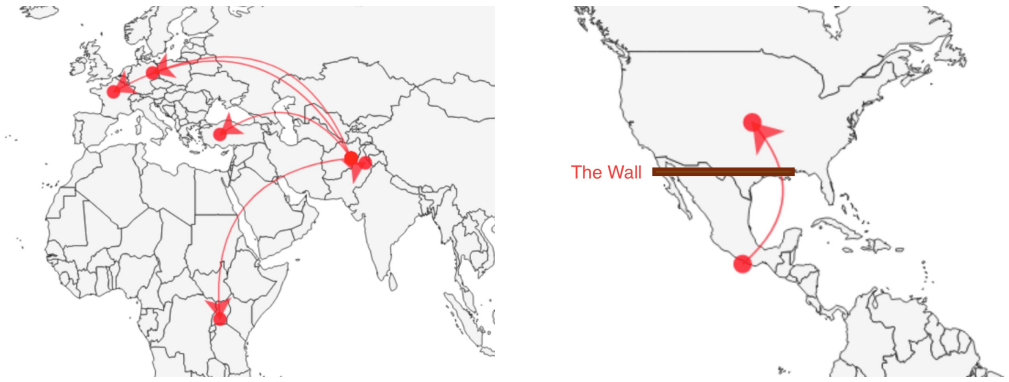

Figure 20: The two local events in the world. The left picture shows the refugee trend mainly from Afghanistan. The right picture shows the 'wall' which will probably be built by Trump.

9.2 Needed additional data

- detailed population data of each language in each country in each year
- detailed migration data of each country in each year
- detailed tourism data of each country in each year
- detailed population data of each country who surf online

References

[1] Abrams D M, Strogatz S H. Linguistics: Modelling the dynamics of language death[J]. Nature, 2003, 424(6951):900.

[2] Patriarca M, Leppn T. Modeling language competition[J]. Physica A Statistical Mechanics Its Applications, 2004, 338(1):296-299.

[3] Cleveland W S. LOWESS: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression[J]. American Statistician, 1981, 35(1):54-54.

[4] Welton N J, Ades A E. Estimation of markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis, and model calibration[J]. Medical Decision Making An International Journal of the Society for Medical Decision Making, 2005, 25(6):633-645.

[5] Contreras J, Espinola R, Nogales F J, et al. ARIMA models to predict next-day electricity prices[J]. IEEE Transactions on Power Systems, 2003, 18(3):1014-1020.

[6] Zhang S, Yu Q, Bi G. Complex Agent Network and Evolution Analysis for Multi-lingual Competition[J]. Journal of System Simulation, 2017.