Research on Language Distribution Prediction Based on ARIMA Model

Yaling Ma[a], Hui Li, Zhexi Liu, Xia Wu

University of Science & Technology Beijing, Beijing, China

[a]mylove15082777155@163.com

Keywords: IMA; time series; language distribution

Abstract: order to predict language distribution in the next 50 years, we use the phenomena of immigrations among different continents to describe the changes of geographic distribution. We analysis changes of language distribution of a specific language only through the immigration status of countries whose native language ranks top ten. We assess the immigration data, and use two formulas to calculate net immigration rate and total immigration rate, then we use the ARIMA model to predict the corresponding rate of change in 50 years. Using the assessed distribution status of top ten languages in 2017, we will analysis changes in language geographic distribution in the future.

1. Introduction

In order to predict the language distribution in the next 50 years, we use the phenomenon of immigration between different continents to describe the changes in geographical distribution. We analyze the language distribution of a particular language only by immigration status in the top ten countries. First, this article considers geographic partitioning. Immigration rates are seen as a factor influencing language distribution because of its representation. In addition, we predict the distribution of language in the next 50 years. A related analysis was conducted.

The core of our model is two formulas, ARIMA model in time series, and bulk data processing. Based on the question, we separate language users into different segmentations by geographic distribution, and establish this model to predict net immigration rate of one language in different geographic partition.

2. Symbol Description

Table 1. Symbol Descriptio

Parameter	Meaning	Unit
T	temperature of water in the bathtub	K
T1	temperature of adding hot water	K
T3	temperature of wall surface of the bathtub	K
t1	time of adding water	s
t2	the moment of water began to overflow	s
H2	heat dissipation coefficient of bathtub wall	w/m^2k
f1	the flow of hot water into the bathtub	m^3 /s
ε	Boltzmann constant	-
σ_1	radiation coefficient of water	w/m^2k^4

3. Models

3.1 Model Overview

We first separates the problem into three phase.

Then, we set the prediction model of net immigration rate of native language speakers, and use related assumptions to predict the changes of distributions of top 10 languages in 50 years.

DOI: 10.25236/mmmce.2019.150

Via the three models above, we suggest some locations of new offices according to changes of language distribution in long-term and short-term.

First, we consider the system in natural state thermal energy conversion. Energy losses including water circulated by heat and lead to sporadic bath heat and energy loss.

Temperature decrease gradually. When the man feels cold, he turns on the tap. Allows the water to flow into the bathtub of hot water. But at the moment the bathtub water damage did not reach the overflow outlet. only need to add hot water inflow of heat . Of course, we cannot ignore quality flow into the bathtub.

Finally, when the water in the bathtub reaches the height of the overflow, a new energy change is created. That is the energy of the water flowing out. Our model three is based on the model two to account for the loss of energy.

Our model can description of the whole process of bathing. We can also solve mission temperature as close as possible to the initial temperature of the strategy.

According to the survey, common materials for acrylic bathtub, we selected the common parameters are as follows:

Table 2. Model Parameters

Parameter	Meaning	Value
T_0	initial temperature	312K
T_2	ambient temperature	289K
S_1	surface area of the bathtub’s upper surface	$0.66m^2$
S_2	surface area of the all inner wall of the bathtub	$2.29m^2$
D	thickness of bathtub wall	0.02m
h	overflow height of bathtub	0.48m
V_0	volume of the initial water in the bath tub	$0.21m^3$
V_r	the average volume of thehuman body inthe bathtub	$0.06m^3$
S_r	the average area of the human body into the bathtub	$1.45m^2$
V_m	maximum capacity of bathtub	$0.3168m^3$
m	the quality of initial water in the bathtub	210kg
m_1	the quality of water in bathtub when overflowing	256.8kg
H_1	heat dissipation coefficient of water	$300w/m^2k$
λ	heat thermal conductivity of bathtub’s wall	0.19w/mk
ρ	density of water	$1000kg/m^3$
T_{01}	the temperature of feeling cold	305k
c	specific heat capacity of water	4200J/(kg.℃)

3.2 Prediction model of immigration rate of native speakers

● Model Analysis

The core of this model is two formulas, ARIMA model in time series, and bulk data processing.

In order to predict language distribution in the next 50 years, we use the phenomena of immigrations among different continents to describe the changes of geographic distribution. We analysis changes of language distribution of a specific language only through the immigration status of countries whose native language ranks top ten.

Based on the former question, we separate language users into different segmentations by geographic distribution, and establish this model to predict net immigration rate of one language in different geographic partition.

We assess the immigration data in this part, and use two formulas to calculate net immigration rate and total immigration rate, then we use the ARIMA model to predict the corresponding rate of change in 50 years. Using the assessed distribution status of top ten languages in 2017, we will analysis changes in language geographic distribution in the future.

● Model Assumptions

① We divided the world into seven parts, east of Asia, west of Asia, Europe, North America, Latin America, Africa (the separation in Asia is rough, which aims to put India and Pakistani in the same segmentation and provides continent for selection of language partition later.

Table 3. geographic partition

Partition	Code
Africa	P_1
East of Asia	P_2
Europe	P_3
Latin America	P_4
North America	P_5
Oceania	P_6
West of Asia	P_7

② Make countries who use top ten languages as their native language the cradle. Original geographic language distribution is as follows(sort by letters)

Table 4. Geographic partition of languages(including countries)

Language	Code1	Code2	Distribution
Arabic	N_1	ara	West of Asia: Oma, Yemen, Iraq, Syria, Oman, United Arab, Emirates, Tajikistan, Uzbekistan European: Cyprus Africa: Algeria, Chad, Egypt, Libya, Morocco, Sudan, Tunisia
Bengali	N_2	ben	East of Asia: Bangladesh
Chinese	N_3	zho	East of Asia: China
English	N_4	eng	North of America: Canada, the United States Oceania: Australia, New Zealand Africa: South Africa Europe: England
Hindustani	N_5	hin	East of Asia: Pakistani, India
Japanese	N_6	jpn	East of Asia: Japan
Portuguese	N_7	por	Africa: Angola, Cape Verde, Guinea-Bissau., Mozambique, SAO Tome Latin America: Brazil Europe: Portugal East of Asia: Democratic Republic of Timor-Leste
Punjabi	N_8	pnb	East of Asia: Pakistani, India
Russian	N_9	rus	Europe: Russia, Belarus West of Asia: Kazakhstan, Kyrgyzstan
Spanish	N_{10}	spa	Africa: Equatorial Guinea Latin America: Cuba, Dominican Republic, El Salvador, Honduras, Costa Rica, Guatemala, Mexico, Nicaragua, Panama, Argentina, Chile, Venezuela, Peru, Ecuador, Paraguay, Uruguay, Bolivia, Colombia Europe: Spain

③ There exists one original phase when countries were set originally, and the population of the countries is assumed as the number of native language users originally.

④ As long as people settle down in one countries, they become one of the total users of native language of that country.

⑤ The only way to increase users of one language is to increase the number of people settled in that country. That is to say, immigration is the only way to change the geographic distribution of one language.

⑥ Displaced population is also considered users of native language of their former country.

● Model Building

1) Model formula building

According to the assumptions above, the languages rank top ten are named as$N_i(i = 1,2 \dots 10)$,the countries which use it as native language are named as$\mathrm{q}(\mathrm{q} = 1,2 \dots)$,, geographic partition of different countries are named as$P_k(k = 1,2 \dots 7)$, and calculate the net immigration rate of this language in year y by weight values of the population$NIR_y\ (N_i, R_k)$, $(\mathrm{y} = 1962,1967 \dots 2012)$:

$$NIR_y\ (N_i, R_k) = \frac{\sum_{q=1}^{n} IM_y(N_i, P_k, q) - \sum_{q=1}^{n} EM_y(N_i, P_k, q)}{\sum_{q=1}^{n} TP_y(N_i, P_k, q)}$$

$IM_y(N_i, P_k, q)$, $EM_y(N_i, P_k, q)$represent number of immigration and emigration respectively of country q.

Predict$NIR_y\ (N_i, R_k)$in 50 years with the former ARIMA model in time series, and calculate via the formula as follows$TPR(N_i, R_k)$:

$$TPR(N_i, R_k) = \prod_{y=y_0}^{y_n} [1 + NIR_y\ (N_i, P_k)]$$

When$TPR(N_i, R_k) > 1$, population distribution of $\mathrm{N_i}$ in $\mathrm{P_k}$ won't change in n years; u

When$0 < TPR(N_i, R_k) < 1$, population distribution of $\mathrm{N_i}$ in $\mathrm{P_k}$ will change in n years.

2) Model simulation and result

According to related data (https://data.worldbank.org.cn/, n.d.), we can obtain immigration status of different countries in 2017. Assumed that people's willing of immigration won't change over n years, we obtain population transferration of language $\mathrm{N_i}$ in $\mathrm{P_k}$（$0 < TPR(N_i, P_k) < 1$), which is also changes of geographic distribution of one language. As is shown in the following figure:

Therefore, we can have have changes of geographic distribution of the ten language. Analysis according to the size of immigration:

As can be seen in the wavy arc of immigration population, the distribution of Arabic will have the biggest change with the most people immigrant to Europe. Other changes can be seen in the figure below.

Table 5. Prediction of immigration rate and trends

Language	Partition	$TPR(N_i, R_k)$	Main destination	Numbers
Arabic	West of Asia	0.23	West of Asia	2119731
			East of Asia	64843
			Europe	2020225.428
			North America	227464
			Latin America	12321
			Africa	60398
Bengali	East of Asia	0.06	West of Asia	553939
			East of Asia	5552693
			Europe	243910
			North America	99771
Chinese	East of Asia	0.41	East of Asia	1051707
			Europe	104331
			North America	1353447

			Oceania	142780
Hindustani	East of Asia	0.15	West of Asia	4318089
			East of Asia	3381915
			Europe	890345
			North America	1350371
Portuguese	Africa	0.94	West of Asia	11239
			East of Asia	208991
Portuguese	Europe	0.98	Europe	1056927
			North America	368302
			Latin America	224624
			Africa	55520
Portuguese	East of Asia	0.98	Oceania	9389
			East of Asia	8450
			Europe	3937
Punjabi	East of Asia	0.15	West of Asia	4318089
			East of Asia	3381915
			Europe	890345
			North America	1350371
Russia	West of Asia	0.80	West of Asia	303976
			East of Asia	68859
			Europe	3028460
Spanish	Latin America	0.23	Europe	706779
			North America	12907074
			Latin America	2064062

3) Result Analysis

First and foremost, based on the prediction model above, this model takes geographic partition into consideration. The immigration rate is viewed as the factor influencing language distribution for it is representative.

However, related factors are out of consideration. It can’t conform to reality, simply thinking about immigration. With the trends of globalzation, increase of language total users isn’t necessary attribute to net immigration. And this is the main shortcoming of ths model.

3.3 Part two

● Problem restatement

Since there have been set national offices in Shanghai, China and New York, America, we can take the two country out of consideration when choosing the address.

Since the set of long-term office requires a long-term vision, we have different methods to choose the address, so do our suggestions. Then, we are going to analysis the problem both in short-term and long-term, and propose our suggestion accordingly.

● Short- term

1) Problem analysis

Since staff are required to master English and at least one extra language in a short-term office, we should first consider places where English has become more popular. That’s to say, it’s a place whose native language or second language is English or who attach more importance on English leaning. What’s more, in order to make interest in short-term, we must pay attention to the development degree of the local economic and technology as well. Obviously, a better developed place is preferred.

As shown in part1, We divided the world into seven parts, east of Asia, west of Asia, Europe, North America, Latin America, Africa(the separation in Asia is rough, which aims to put India and Pakistani in the same segmentation and provides convenient for selection of language partition later.

2) Problem assumptions

English usage: population of English users in different places, which is a number aggregated according to the number of regions in the world.

The development of other languages: developing trends of numbers of users in different languages.

GDP: development degree of different places.

Degree of globalization: there exist many factors influencing globalization which we can’t list thoroughly. So we can assume logically that number of international arrivals can represent it.

Development of technology: we can assume reasonably that the number of scientific papers published each year can represent the development of technology.

3) Problem solving

We first consider distribution of English nowadays in short-term prediction, and we can obtain the picture as follows: (We color it blue with different degree, dark blue, blue and grey according to the importance of English usage. The figure is from the internet.)

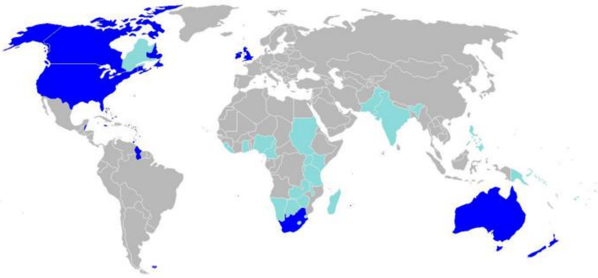

Figure 1 Distributions of countries who mainly use English.

We can obtain population distribution of different language from the following figure. Put the later development status aside, we first move into the second juding round.

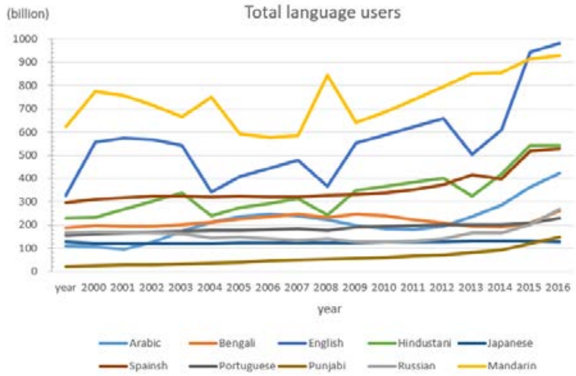

Figure 2 The line chart of top 10 language users from 2000 to 2016.

We can select the geographic partition to open new offices according to the current partition distribution of the top10 languages. The corresponding partition of each language has been shown . Therefore, integrating the former two steps, we have 5 of them, east of Asia, Europe, Africa, Oceania and North America.

Based on the selected partition, factors and reasonable assumptions showed in the flow chart and the site selecting model, we can do the calculation using the formula below. And then, we can get one or two weighted values of eligible countries in each partition.

$$m = \alpha \frac{\overline{G}}{\sum_{i=1}^{n_p} \overline{G}_{pi}} + \beta \frac{\overline{S}}{\sum_{i=1}^{n_p} \overline{S}_{pi}} + \theta \frac{\overline{T}}{\sum_{i=1}^{n_p} \overline{T}_{pi}} \quad (1)$$

Among them, $\overline{G}$、$\overline{S}$、$\overline{T}$ can be obtained as follows:

$$\overline{G} = \frac{\sum_{j=1}^{5} G_{2016-j}}{5}, \overline{S} = \frac{\sum_{j=1}^{5} S_{2016-j}}{5}, \overline{T} = \frac{\sum_{j=1}^{5} T_{2016-j}}{5} \quad (2)$$

α_1、α_2 、 β_1、β_2 、 θ_1、θ_2 :Regulated coefficient of experience, among them $\alpha_1+\alpha_2+\beta_1+\beta_2+\theta_1+\theta_2=1$

m' :Weighted values of site selecting in the long-term in different places

g :The growth rate of GDP

s :The growth rate of technology(growth rate of published scientific literat

t :The growth rate of international visitor arrivals

Others:The same as the short-term

Since the short-term considerations are limited and usage of English plays a decisive role, we can easily get the countries selected by several countries, for example, we choose Australia in Oceania, Canada in North America. As for other continents, we can get the outcome through rough selection and calculations by weighted values. Take Africa for example here, and we can get the weighted values of several countries avaiable.

Table 6. Short term weighted values of site selection in Africa

Country Name	**Weighted values/m**
South Africa	0.5507
Nigeria	0.3589
Namibia	0.0374
Ghana	0.0530

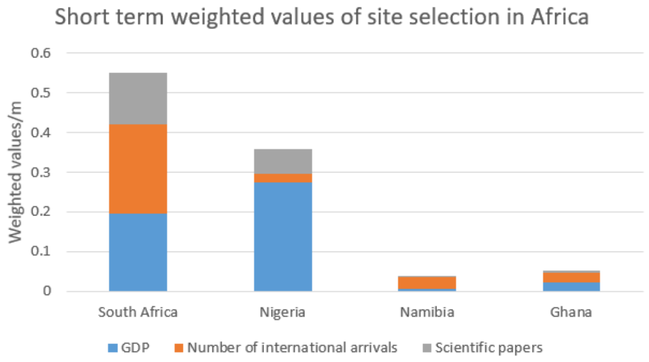

Figure 3 Short term weighted values of site selection in Africa

By comparison, South Africa was selected as the site for the African partition. The same in Asia, and we choose India and the Philippines In Europe, most of the countries have their basic language, so the ability of speaking English can not be put in the first place, but still a main factor. So we need take economic and tourism as auxiliary. Therefore, we choose France in Europe.

The six selected offices are as follows:

Table 7.

Place	Language	Place	Language
Austrila, Sydney	English, mandarin Italian	The Philippines	English, Filipino
Canada	English, Frank	South Africa	English, Filipino
English, Zulu	Hindi, English	French	French, English

● Long-term

1) Problem analysis

Choosing a long-term office requires a long-term consideration. When choosing the locations, we need to prioritize the ability of mastering English and at least one extra language in the office. As for the long-term, we can choose the ones in which English is promoted less widely, but the users of English is increasing more rapidly. And then, we can take the ones with better popularity of English and a better development degree. We needn’t choose the most developed sites as long as its development degree is acceptable, for there exists heavy competitiveness in highly-developed countries, which is bad for expansion of the company. Therefore, we can choose the places whose economy is not very developed, but the development potential is strong.

2) Problem assumptions

Population base of English usage: population base of each partition, but it accounts for a small percentage of the population because it mainly considers the rate of population growth.

Other languages: development of other languages except English.

The growth rate of GDP: we should mainly consider the growth rate of GDP which can reflect the development potential in each place.

The rest are reasonable assumptions and are the same as the short-term.

We use population base of English and population growth rate as the first selecting step. Since it’s in the long-term, we can first take population base out of account, and begin in the countries with smaller English users but larger immigration rate from English speaking countries. The distribution of immigrations from countries, except India, with lots of English users, to other countries is shown as follows (take the countries which has already appeared in the short-term prediction or has been a site for offices out of account):

Assume that the ratio of immigration from countries with larger English users to others is set, we can first think about those countries while selecting sites.

In the second step, we need to consider the trends of development and population base of other languages, as well as the growth rate of GDP of corresponding language. We have predicted the developing trends in the next 50 years of other languages in part 1, so we won’t talk about it here. Prediction of the growth rate of GDP is shown in the figure.

We can get the formula of weighted values, according to the analysis flow chart in the long-term.

$$m' = \alpha_1 \frac{\overline{G}}{\sum_{i=1}^{n_p} \overline{G}_{pi}} + \beta_1 \frac{\overline{S}}{\sum_{i=1}^{n_p} \overline{S}_{pi}} + \theta_1 \frac{\overline{T}}{\sum_{i=1}^{n_p} \overline{T}_{pi}} + \alpha_2 g + \beta_2 s + \theta_2 t$$

Based on the weighted values calculated from various factors, we can learn that Mexico and Spain are quite appealing.

There has been an office in New York, America. That’s to say, there has been two offices in one continent. Therefore, the developed Canada can hinder the long-term development of the client company. As a result, we choose Mexico between Canada and Mexico.

What’s more, while selecting in Europe, we can learn that compared with French, Spanish has a faster development and a larger population to use the language. In addition, Spain near the strait of Gibraltar, trade development may be prospectively. Therefore, we choose Spain to set the office in Europe.

As the others has no competitiveness, so we consider the locations chosen in the short term

To consider the development of global communication, we take the development of Internet for example. As we can see is a global Internet user growth trend chart and a smart phones users growth trends. We can find that the development of Internet is quite fast, and the development of electronic communications and social media has led to the gradual disappearance of global communication barriers. However, the use of some translation software reduces the communication barriers between different languages.

We suggest the company open less than 6 international offices in this global communication development trend..

For in the globalization trend, more and more people tend to travel abroad, and it would be a waste of time for service companies to open new offices that consume a certain amount of construction and labor costs. However, with the development of the Internet, it is advisable for service companies to expand their company's content and enhance their own competitive strength.

Here, we also need additional information: global communication development rate, related data of international popular tourism area distribution, the maturity rate of global service development.

● If we can get these information, we can select the office according to the local tourism development. And the maturity of service industry can help to distinguish the competitiveness and be used to measure the value of establishment of the office However, in the areas where global communication is highly-developed, we can think less on the language problem and focus on economic development to maximize profits.

References

[1] (https://data.worldbank.org.cn/, n.d.)https://data.worldbank.org.cn/

[2] (https://www.ethnologue.com/, n.d.)https://www.ethnologue.com/