

For office use only

T1 _____
T2 _____
T3 _____
T4 _____

Team Control Number

74316

Problem Chosen

B

For office use only

F1 _____
F2 _____
F3 _____
F4 _____

2018
MCM/ICM
Summary Sheet

The purpose of this paper is to analyze the time and geographical distributions of various languages and forecast the development and distribution of the next 50 years, providing the theoretical basis and advice for the company's offices locating decision.

Firstly, we analyze the quantitative temporal distribution of language users. Various kinds of influences and factors that affect the number of languages users are considered, and concluded as ten indicators, such as GDP per capita, average years of schooling. By Principal Component Analysis, they are combined into four primary components: level of economic development, level of social equality, level of national welfare, and cultural exchanges. On this basis, the short-term difference models are established for native speakers and non-native speakers. First-order autoregressive model (AR(1)) is used to fit the time distribution of native speakers in order to reflect the autocorrelation characteristics. Most native speakers are consistent with non-stationary unit root process. Then, we construct the co-integration relationship between the principal components and the second language users. The error correction model is established and it is found that the random error and the error correction term all achieved stability. In the co-integration space, the influence of the principal component on L2 has a first-order differential stationary nature.

Based on the short-term model, the long-term differential model is further established. Considering the change process of native speakers as a logistic model similar to the natural population growth, the system is stable for the coefficients in the normal range, and the stable equilibrium is given maximum capacity under the current conditions. The time distribution largely synchronizes with the natural change process of the population. Besides, due to the differential smoothness of the linear combinations of the various factors, the influence on second-language users is regarded as a constant for a long time. Therefore, the model of L2 is a constant coefficient differential equation whose time path is determined by the strength of language influence. Therefore, we sum up L1 and L2 to calculate the total number of language speakers, which is a non-stationary dynamic system. It means that the driving forces of a particular language are the endogenous growth of native speakers and the external influence as a second language. The time distribution of the total number of languages is on the rise.

Secondly, we use the long-term model to predict the situation of each country in the next 50 years. The number of influential language speakers increase significantly, while the growth pattern is driven by the second language transmission. The number of less influential language speakers grow less obviously or even decreased. The growth pattern is endogenous to the native speakers. In Top 10, there is a possibility that the number of native speakers will drop significantly or the number of non-native speakers may not grow enough, thus the future rankings may be superseded. Sensitivity analysis and Monte Carlo robustness simulations show that our model is robust and predictable.

Thirdly, we build a Markov Model to analyze the geographical distribution of languages and their changes. This paper constructs a transition matrix of immigrants. Based on the information of population growth, natural growth rate and language distribution, the distribution of the total number of each language in each country is inferred. Then we center on and visualize the national capitals. The prediction shows the geographical distribution of languages tends to be intertwined and spread as second-language in the future.

Next, we locate the new offices by Cluster Analysis and think that the ability of speaking English and Chinese is needed and that the development of economy is considered as well. So, we construct the 4 indicators, the ratio of English speakers and Chinese speakers, GDP per capita and net immigrants. The short-term and long-term models and the Markov model of the forecast results calculate the value. According to the 4 indicators, we analyze the 224 countries by cluster analysis separately. Due to the quantified result of grades of each country, we use the Multiple Objective Decision Making (MODM) of Fuzzy Evaluation to calculate the grade and choose 6 national capitals with maximum grades, as the location of new offices.

Finally, Using MINE model, the latitude and longitude coordinate grids are meshed. And the grid density index is calculated to identify the areas with over-dense distribution of offices. Eliminating appropriate number of offices allowed for reduction costs by serving the largest global scale with minimal office locations. The result shows that it is more suitable to set up 4 new offices. Thus, as for the short term, we recommend to build 4 new offices in London, Singapore, Ottawa and Canberra. And Singapore, London, Canberra and Paris are recommended in the long run.

Key Words: Autoregressive Model, Co-integration, Differential System, Markov Chain, Cluster Analysis, MINE Model

更多数学建模资料请关注微店店铺“数学建模学习交流”

<https://k.weidian.com/RHO6SPa>

Content

1 Introduction.....	2
1.1 Background.....	2
1.2 General Assumptions.....	2
2 Model 1 – Language Speakers’ Quantitatively Distribution.....	2
2.1 Introduction and Assumptions.....	2
2.1.1 Introduction.....	2
2.1.2 Assumptions.....	2
2.2 Variables and Parameters.....	2
2.2.1 Notations.....	2
2.2.2 Dimension Reduction - PCA.....	3
2.3 Short-term Models.....	4
2.3.1 Native Speaker Model - Autoregressive model.....	4
2.3.2 Non-native Speaker Model – Error Correction Model.....	6
2.4 Long-term Models – Logistic Model.....	8
2.4.1 Assumptions.....	8
2.4.2 Model Design.....	8
2.5 Predictions.....	10
3 Validation of Model 1.....	11
3.1 Sensitivity Analysis.....	11
3.1.1 fixed growth rate δ	11
3.1.2 External Promoting Coefficient μ	12
3.1.3 Initial Value.....	12
3.2 Robust Analysis.....	12
3.2.1 Monte-Carlo Simulation.....	12
4 Model 2 - Geographic distribution.....	13
4.1 Migration Patterns – Markov Process.....	13
4.1.1 Assumptions and Variables.....	13
4.1.2 Model Settings.....	13
4.1.3 Solution and Visualization.....	14
4.2 Language Distributions.....	15
4.2.1 Prediction.....	15
4.2.2 Conclusion.....	16
4.3 Comparison of Model 1&2.....	16
5 Model 3 – Offices Location Decisions.....	16
5.1 Model in Short-term and Long-term - Cluster Analysis & MODM.....	16
5.1.1 Assumption.....	16
5.1.2 The Settings and Solutions of Model.....	16
5.2 Comparison.....	18
5.3 Resource-saving Suggestions – MINE Model.....	18
6 Strengths and Weaknesses.....	20
6.1 Strengths.....	20
6.2 Weaknesses and Improvements.....	20
Memo.....	21
Work cited.....	23
Appendix.....	24

Notes: Due to space limitations, the appendix does not show all the data.

1 Introduction

1.1 Background

As known to all, nearly 7000 languages are spoken over the world, and they make up the communication network through hundreds of countries and regions. Languages are essential to construct foreign trade, develop tourism and promote scientific and technological progress, which makes it an indicator and an effective tool to measure a country's comprehensive power. Also, a measurement of the utility of a particular language is the number of speakers who use it as native or the second or third language. Therefore, it should be taken attention that the number of speakers of a particular language would change over times with the languages' rise and fall as it may be coincident with the economic and political development of its main country.

For now, ten languages are claimed to use by half the world's population, which includes Mandarin (incl. Standard Chinese), Spanish, English, Hindi, Arabic, Bengali, Portuguese, Russian, Punjabi, and Japanese. And the number of speakers of one language would be influenced by migration, social pressures, business relations, social media and so on. It is necessary for us to find out its variation and trends in the future to expect their rankings and make better use of them.

1.2 General Assumptions

- (1) Following models only consider the top 26 languages used in the world as the number of speakers of them nearly amounts to 98% population of the world.
- (2) There is no unexpected collision of other planets and no other disasters disrupting people's normal life.
- (3) In addition to the differences of details, numbers of speakers used particular languages are following the same model settings.
- (4) All stochastic error terms can be expressed as white noises or its time-weighted form.
- (5) All individuals are homogeneous. Their choices of the second or third languages are completely influenced by factors of countries, societies and cultures. And the contribution to individual language using from personal interests, life plans are negligible.
- (6) The macro-level factors such as culture and migration have the same impact on the number of total speakers using one particular language in the long run.

2 Model 1 – Language Speakers' Quantitatively Distribution

2.1 Introduction and Assumptions

2.1.1 Introduction

In Part 1, a modeling analysis is acquired to analyze the different languages users' time distribution, including native speakers and non-native speakers, and to predict the situation in the next 50 years. Firstly, we discuss the native speakers and non-native speakers separately. The use of native language is mainly determined by the environment and is hardly related to those factors, for example, the society, media, technology and tourism. And the use of language over time is highly likely to be positively related to the changing number of native people, which, moreover, mostly depends on the autocorrelation. But the factors like migration and culture shock primarily influence the use non-native language. As a consequence of that, we take account of those factors to set the model of non-native language.

2.1.2 Assumptions

- (1) The time distribution of some native speakers is determined by the pattern of the population distribution. The factors besides population changing are neglected.
- (2) The time distribution of non-native speakers is determined by the serial correlation, population migration and those factors, including international business relations, increased global tourism, the use of electronic communication and social media, and the use of technology to assist in quick and easy language translation.
- (3) There are some factors that have the first co-integration relationship with the non-native speakers.

2.2 Variables and Parameters

2.2.1 Notations

<i>VARIABLES</i>	<i>DEFINITION</i>
	The Number of Native Speakers

	The Number of Non-native Speakers
N_T	The Number of Total Speakers
ΔS_t	$S_t - S_{t-1}$
$X_{1t} \dots X_{10t}$	Variables as following
$Y_{1t} \dots Y_{4t}$	Primary Components
ΔY_t	$Y_t - Y_{t-1}$
	Stochastic Error Term
ECM_{t-1}	Error Correction Model term, which indicates the extent to which the explained variables deviate from the long-term equilibrium in the previous period
N_M	Maximum population
	One-step Transition Matrix
P	Possibility from One Statement to Another
	Steady-state probabilities
PARAMETERS	DEFINITION
	Parameters for AR(p)
	Parameter for long-term model
	Parameter of variable ECM
	Natural Growth Rate of the Number of Native Speakers

Table 1: Variables and Parameters for Model 1

2.2.2 Dimension Reduction - PCA

Using PCA to derive metrics for the measurement of language evolving and reduce variables dimension by create new variables that are linear combinations of the original variables. New linear combinations are uncorrelated and only a few of them contain most of the original information, which are called principal components.

Following variables are derived from the country where people use the particular language. And the final values of each observation variables are weighted by the number of speakers in the country.

a. GDP (per capita)/\$

GDP is a monetary measure of the market value of all final goods and services produced in a period.

b. Crop yield/\$

Crop yield refers to both the measure of the yield of a crop per unit are of land cultivation.

c. Average years of schooling/Years

Average years of schooling reflect the educational attainment among age groups and genders.

d. Gini coefficient

The Gini coefficient measures the inequality among values of a frequency distribution.

e. Gross National Happiness Index

GNH is a measurement of the collective happiness in a nation, which can imply social pressure.

It can be formulated as follows:

$$GNH = \frac{\Delta \text{Income}}{\text{Gini coefficient} \times \text{unemployment rate} \times \text{inflation rate}}$$

f. Number of Migrants/10000 people

The number of migrants has an effect on the use of language.

g. Labor Productivity/\$10000

Labor productivity presents how many products have been produced.

h. Consumer Price Index/%

Consumer Price Index presents the price level of goods, which can reveal the relationship between demand and supply.

i. Income of Tourism/\$

With the prosperity of tourism in one country, its language is used more frequently and broadly, as foreign tourists flood in and hung around the interest.

j. The amount of translation or directory softwares that record the language/unit

More softwares record the particular language, the more popular it is among the world.

Metrics for the language using measurement

As for the **non-native speakers**, we take the serial correlation and other factors into consideration. From the macro-view, those factors, including international business relations, global tourism, social media and technology of language translation, can have an effect on the use of non-native language. We collect 10 micro-factors that might affect the number of non-native speakers, which form a time series from 2008 to 2016. Due to the limit of data, we only collect the 6 official languages of UN, Mandarin Chinese, English, Spanish, Arabic, Russian, and French, as the representative territorial data to form a 6x9 Panel. (Detailed in Appendix 2).

Principal component analysis (PCA) is statistical procedure that transforms the statics to orthogonal linear equations to set up a new evaluating system. The first principal component has the largest possible variance and the resulting vectors ($Y_1, Y_2, Y_3, \dots, Y_{10}$) are an uncorrelated orthogonal basis set. So, we got the principal below:

$$\begin{cases} Y_1 = \lambda_{11}X_1 + \lambda_{12}X_2 + \lambda_{13}X_3 + \dots + \lambda_{110}X_{10} \\ Y_2 = \lambda_{21}X_1 + \lambda_{22}X_2 + \lambda_{23}X_3 + \dots + \lambda_{210}X_{10} \\ Y_3 = \lambda_{31}X_1 + \lambda_{32}X_2 + \lambda_{33}X_3 + \dots + \lambda_{310}X_{10} \\ \dots \\ Y_{10} = \lambda_{41}X_1 + \lambda_{42}X_2 + \lambda_{43}X_3 + \dots + \lambda_{1010}X_{10} \end{cases}$$

$$\begin{aligned} \text{var}(Y_i) &= \text{var}(\lambda_i' X) = \lambda_i' \Sigma \lambda_i \\ \lambda_i' \lambda_i &= 1 \end{aligned}$$

Y_i, Y_j should be uncorrelated ($i \neq j$).

has the largest possible variance.

In this case, we use 10 columns of data matrix $X = (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10})$, to represent ten factors including GDP (per capita), crop yield, average years of schooling, GNH, Gini coefficient, number of migrants, labor productivity, CPI, income of tourism and the number of software.

After the factors rotating, the feature vectors involved with the 4 main components are,

$$\begin{cases} \eta_1 = (3.16 & 2.95 & 0.63 & 0.87 & 0.42 & 0.75 & 2.76 & 0.44 & 2.58 & 0.37)^T \\ \eta_2 = (0.92 & 0.77 & 0.63 & 3.78 & 0.40 & 0.53 & 0.46 & 0.29 & 0.66 & 0.29)^T \\ \eta_3 = (0.82 & 0.41 & 1.89 & 2.33 & 2.26 & 0.38 & -0.07 & -0.34 & 0.65 & 1.42)^T \\ \eta_4 = (0.67 & 0.95 & 0.58 & 0.33 & 0.26 & 3.89 & 1.01 & -0.88 & 0.61 & 3.37)^T \end{cases} \quad (1)$$

Namely, $Y_1 = X\eta_1, Y_2 = X\eta_2, Y_3 = X\eta_3, Y_4 = X\eta_4$

Y_1 is the economic factor, including GDP per capital, Crop Yield, Tourism Income. Y_2 is the social equality factor. Y_3 is the national welfare factor. Y_4 is the culture-shock factor.

So, we got a new metrics system by PCA, Y_1, Y_2, Y_3, Y_4 inherits 87.6% possible variance from X , which is more than 85%. Thus, it's effective and properly reflects the original factors. (Y_1, Y_2, Y_3, Y_4) is defined as the metrics system for language using measurement model.

2.3 Short-term Models

2.3.1 Native Speaker Model - Autoregressive model

According to the short-term model below, we solve the native speaker and non-native speaker models separately, then sum up the number of total target language speakers.

For native speaker, due to hypotheses, the number of some native speakers is influenced by autocorrelation in time series. As a consequence, we choose p^{th} -order Auto-Regression Model to fit curve, namely AR(p), as $N_t = \varphi_1 N_{t-1} + \varphi_2 N_{t-2} + \dots + \varphi_p N_{t-p} + \varepsilon_t$ (2)

N represents the number of native speakers in year t .

$\varphi_1, \dots, \varphi_p$ represent the influential coefficients of different lag orders.

ε_t represents the error term whose mean value is 0 and variance is σ^2 . The distribution matches White Noise Process $WN(0, \sigma^2)$

Firstly, we collect the time-series data of those language speakers. As for the top 10 languages, we use the statistics of native speakers, non-native speakers and total speakers from 2003 to 2017

Mandarin Chinese			English			Hindustani			Spanish		
L1 Speakers	L2 Speakers	Total	L1 Speakers	L2 Speakers	Total	L1 Speakers	L2 Speakers	Total	L1 Speakers	L2 Speakers	Total
2003	768	81	839	385	591	946	185	214	399	418	426
2004	764	111	875	364	596	950	194	215	409	418	426
2005	772	131	903	385	597	952	201	217	418	419	428
2006	779	147	926	385	600	955	209	218	427	421	448
2007	788	154	942	386	602	958	216	223	439	422	461
2008	786	175	971	386	607	965	221	224	447	423	470
2009	807	178	985	360	609	969	237	217	454	426	481
2010	818	183	1011	362	610	972	253	218	463	425	488
2011	832	195	1027	361	611	972	265	219	475	430	493
2012	841	197	1038	365	606	971	282	201	483	431	501
2013	846	193	1039	366	608	974	294	206	500	432	512
2014	858	188	1046	368	609	977	303	206	508	436	524
2015	865	198	1063	370	608	978	310	211	521	434	535
2016	887	190	1077	372	609	981	318	215	533	434	547
2017	897	193	1090	371	612	983	329	215	544	436	557

Arabic			Malay			Russian			Bengali		
L1 Speakers	L2 Speakers	Total	L1 Speakers	L2 Speakers	Total	L1 Speakers	L2 Speakers	Total	L1 Speakers	L2 Speakers	Total
2003	247	99	346	69	211	167	106	271	168	64	212
2004	260	101	361	64	210	167	106	271	168	64	220
2005	264	104	368	68	209	164	105	269	161	45	226
2006	266	109	375	72	207	161	109	270	165	40	236
2007	266	111	367	77	204	161	107	268	164	37	241
2008	260	112	372	80	205	159	105	264	162	34	246
2009	265	109	374	85	205	167	107	264	161	29	260
2010	267	118	385	91	206	153	114	267	166	25	253
2011	270	120	390	93	210	160	118	268	166	25	261
2012	274	122	396	96	211	146	118	266	164	23	271
2013	276	126	402	102	216	147	118	265	164	26	280
2014	281	122	403	109	219	145	120	265	164	28	292
2015	285	126	411	113	223	145	118	263	163	29	300
2016	288	128	416	121	226	143	119	262	162	24	306
2017	290	132	422	126	227	141	118	259	163	17	310

Portuguese			French		
L1 Speakers	L2 Speakers	Total	L1 Speakers	L2 Speakers	Total
2003	191	7	198	66	172
2004	195	6	201	69	180
2005	200	5	205	67	185
2006	204	6	210	68	190
2007	210	4	214	69	194
2008	215	4	219	69	194
2009	213	6	219	70	203
2010	221	7	221	73	207
2011	215	9	224	72	211
2012	217	10	227	71	214
2013	218	10	228	76	214
2014	220	10	230	77	221
2015	223	11	232	77	225
2016	220	10	230	78	228
2017	218	11	229	75	229

Table 2: List of Languages by Numbers of Speakers (Unit: Millions)

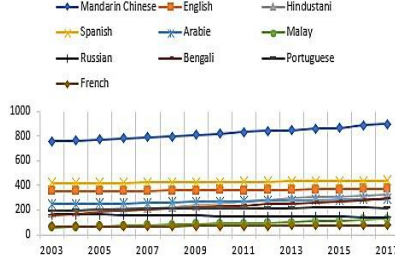


Figure 1: Number of L1 Speakers (Unit: Millions)

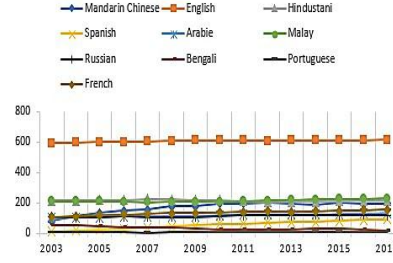


Figure 2: Number of L2 Speakers

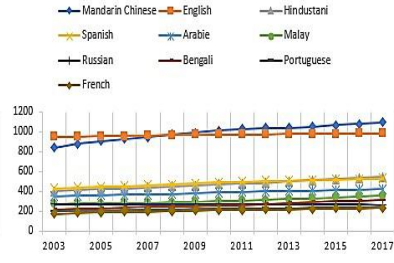


Figure 3: Number of Total Speakers

Then we solve the autoregressive model of the number of native speakers and calculate the time-series autocorrelation function and the partial autocorrelation function of L1 speakers to identify the order, specifically valuing the parameter p .

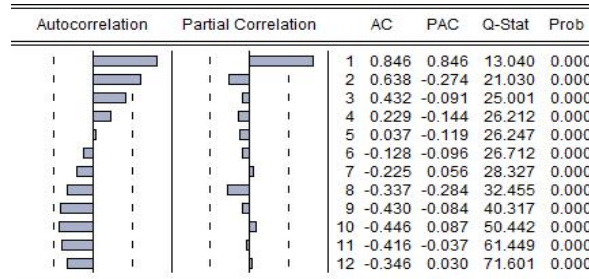


Figure 4: Autocorrelation and Partial Correlation of English

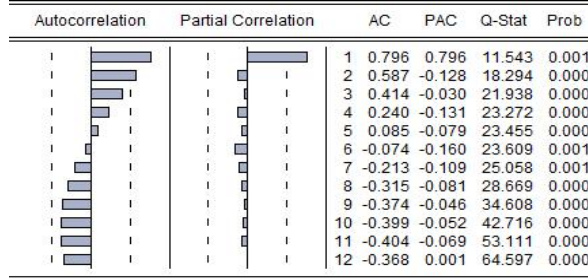


Figure 5: Autocorrelation and Partial Correlation of Mandarin Chinese

The functions of the other 8 native speakers are included in Appendix 1.

According to those graphs, the serial correlation of 10 representative languages represents an obvious autoregressive structure, namely AR1, which means the involved functions don't have the nature of censoring and the partial function has the nature of censoring. So, we conclude $p=1$ and all the native speakers have the same model setting,

$$N_t = \phi_1 N_{t-1} + \varepsilon_t \quad (3)$$

The mean variance of the coefficient ϕ_1 and the error term ε is different with the language changing. By calculation, we use the MLE to estimate the coefficients of the ten native languages above. The results are as following.

Languages	ϕ_1	σ	P-Value	R^2
Mandarin Chinese	1.012	4.073	0.000	0.991
English	1.003	1.459	0.000	0.947
Hindustani	1.040	3.332	0.000	0.995
Spanish	1.003	1.640	0.000	0.930
Arabic	1.012	1.325	0.000	0.941
Malay	1.053	1.609	0.000	0.993
Russian	0.987	1.449	0.000	0.971

Bengali	1.041	1.254	0.000	0.943
Portuguese	1.009	1.988	0.000	0.951
French	1.010	1.951	0.000	0.962

Table 3: AR(1) Model for Different Languages

Due to the results, the setting of AR(1) model is universal and applicable for the 10 representative languages. The results of coefficients testing are significant and the overall fitting is excellent with a high R-squared. The mean variance of error term is basically stable between 1 to 2. Excerpt Russian, most of the native speakers show a non-stationary growth over time, which mostly accords with the regular development of population and language. However, the Russian native speakers are declining due to the serious population decrease.

Then we hold an over-fitting and under fitting test to ensure the accuracy of $p=1$. The test uses the Algorithm of information guidelines to take both the error term and the complexity of parameters into account. The algorithm includes 3 sub-rules, FPE and AIC raised by Akaike and BIC raised by Haman.

$$FPE = \frac{\text{size of sample} + p}{\text{size of sample} - p} \quad (4)$$

$$AIC = \ln \sigma^2 + \frac{2p}{\text{size of sample}} \quad (5)$$

$$BIC = \ln \sigma^2 + \frac{p}{\text{size of sample}} \ln(\text{size of sample}) \quad (6)$$

And the appropriate p should minimize the Algorithm of information guidelines.

	FPE			AIC			BIC		
	P=0	P=1	P=2	P=0	P=1	P=2	P=0	P=1	P=2
Mandarin Chinese	23.54	22.16	22.88	31.63	28.79	29.1	30.24	27.58	27.61
English	26.81	23.77	29.67	33.78	30.19	32.55	31.91	30.02	33.05
Hindustani	41.69	26.74	35.37	43.58	36.55	38.49	36.05	31.98	32.33
Spanish	20.54	15.55	16.78	21.61	16.98	19.84	23.49	21.04	22.66
Arabie	43.54	23.54	33.88	48.67	44.35	45.62	43.71	38.5	40.13
Malay	31.56	24.86	25.74	23.34	22.09	18.89	23.79	21.05	19.93
Russian	34.54	31.45	37.43	24.54	20.54	24.99	28.65	25.73	26.49
Bengali	34.56	31.98	32.54	46.91	43.32	44.55	39.54	36.73	37.9
Portuguese	22.1	19.86	20.88	43.54	37.94	40.56	45.58	40.59	42.38
French	25.84	21.09	23.45	22.4	19.01	18.99	29.08	24.77	25.92

Table 4: Akaike Information Criterion Test

The testing shows that the model of $p=1$ mostly optimize the fitting except the indicators of AIC. But considering the fault of AIC itself, and the optimization of FPE and BIC, we conclude that the fitting of AR(1) is reasonable.

2.3.2 Non-native Speaker Model – Error Correction Model

For non-native speaker, both autocorrelation and macro-level factors are considered in sequence linear regression. To avoid producing false regression under autocorrelation, we perform the co-integration analysis based on the traditional least-squares linear regression. Let a certain number of non-native language speakers be sequence elements $\{S_t\}$, looking for 10 macro factor sequences $\{X_{1t}\}, \{X_{2t}\}, \dots, \{X_{10t}\}$ to form a vector sequence $X_t = (S_t, X_{1t}, X_{2t}, \dots, X_{10t})$. And After PCA, we have a vector sequence of elements S and Y , X s' primary components, $Y = (S_t, Y_{1t}, Y_{2t}, Y_{3t}, Y_{4t})$. (7)

It is a co-integration matrix. So, we use error correction model to describe its evolving process. As known, the number of non-native speaker(S) is related to its lag value and its independent variables. So, the theoretical co-integration relationship is as follows.

$$S_t = \alpha + \varphi S_{t-1} + \beta_{11} Y_{1t} + \beta_{12} Y_{1t-1} + \beta_{21} Y_{2t} + \beta_{22} Y_{2t-1} + \beta_{31} Y_{3t} + \beta_{32} Y_{3t-1} + \beta_{41} Y_{4t} + \beta_{42} Y_{4t-1} + \epsilon_t \quad (8)$$

$\beta_{11}, \beta_{12}, \beta_{21}, \dots, \beta_{42}$ are the coefficients of co-integration relationship, extracted from the basis

$(1, -\beta_{11}, -\beta_{12}, \dots, -\beta_{42})$ of the co-integration space. In the short run, the equilibrium state in formula (8) is difficult to be fully satisfied. Considering the autocorrelation of short-term variables, the lagged form of first order distribution.

Take a first-order difference to both ends of equation (8) simultaneously and we produce Error Correction Model (ECM) based on co-integration.

$$\begin{aligned}\Delta S_t &= \alpha + (\varphi - 1)S_{t-1} + \beta_{11}\Delta Y_{1t} + (\beta_{11} + \beta_{12})Y_{1t-1} + \beta_{21}\Delta Y_{2t} + (\beta_{21} + \beta_{22})Y_{2t-1} + \beta_{31}\Delta Y_{3t} \\ &\quad + (\beta_{31} + \beta_{32})Y_{3t-1} + \beta_{41}\Delta Y_{4t} + (\beta_{41} + \beta_{42})Y_{4t-1} + \varepsilon_t \\ \Delta S_t &= \beta_{11}\Delta Y_{1t} + \beta_{21}\Delta Y_{2t} + \beta_{31}\Delta Y_{3t} + \beta_{41}\Delta Y_{4t} - (1 - \varphi)[S_{t-1} - \frac{\alpha}{1 - \varphi} - \frac{(\beta_{11} + \beta_{12})}{1 - \varphi}Y_{1t-1} - \frac{(\beta_{21} + \beta_{22})}{1 - \varphi}Y_{2t-1} \\ &\quad - \frac{(\beta_{31} + \beta_{32})}{1 - \varphi}Y_{3t-1} - \frac{(\beta_{41} + \beta_{42})}{1 - \varphi}Y_{4t-1}] + \varepsilon_t \\ \Delta S_t &= \beta_1^T \Delta Y_t - \lambda ECM_{t-1} + \varepsilon_t\end{aligned}\tag{9}$$

$$\beta_1^T = (\beta_{11}, \beta_{21}, \beta_{31}, \beta_{41}), \Delta Y_t = \begin{pmatrix} \Delta Y_{1t} \\ \Delta Y_{2t} \\ \Delta Y_{3t} \\ \Delta Y_{4t} \end{pmatrix}, \lambda = 1 - \varphi, (\varphi < 1), \lambda \text{ is an adjustment factor.}$$

β_1^T represents the influence coefficient of cointegration relationship in the current period,

$$ECM_{t-1} = S_{t-1} - \frac{\alpha}{1 - \varphi} - \frac{(\beta_{11} + \beta_{12})}{1 - \varphi}Y_{1t-1} - \frac{(\beta_{21} + \beta_{22})}{1 - \varphi}Y_{2t-1} - \frac{(\beta_{31} + \beta_{32})}{1 - \varphi}Y_{3t-1} - \frac{(\beta_{41} + \beta_{42})}{1 - \varphi}Y_{4t-1}$$

ECM_{t-1} is the error correction term, which means the correction of the t-1 period by the t-th model driven by the equilibrium relation, where $\lambda > 0$ indicates that the correction is a negative feedback process, so that our model approaches towards a stable equilibrium.

And here comes the calculation process (which will be explained by taken the number of speakers using English for example). We use Engle-Granger two-step methods to construct the co-integration.

First, run a unit root test for dependent variable and independent variables to see if the five sequences satisfy the co-integration relationship. Based on the principle of least squares algorithm, ADF-t statistic was constructed by computer regression and compared with the critical value to determine the co-integration of the four principal components in the languages.

Languages	ADF-t Statistics	Is $\{\varepsilon_t\}$ a Unit Root ?	Co-integration
Mandarin Chinese	-1.76	No	Exist
English	-2.23	No	Exist
Spanish	-0.88	No	Exist
Arabic	-1.68	No	Exist
Russian	-1.35	No	Exist
French	-1.42	No	Exist

Table 5: EG Co-integration Test

From the results shown, the t value of all variables of difference are smaller than at least one critical level of significance, which infers that the sequence of difference does not have unit root and it can be called as stationary time series. So, the co-integration of the model is established.

Then we have $S, Y_1, Y_2, Y_3, Y_4 \sim I(1)$.

Next, according to the co-integration model (4), the first step least squares estimation is performed.

$$S_{t-1} = \gamma_0 + \gamma_1 Y_{1t-1} + \gamma_2 Y_{2t-1} + \gamma_3 Y_{3t-1} + \gamma_4 Y_{4t-1} + ECM_{t-1}\tag{10}$$

Languages					
Mandarin Chinese	-0.48	1.26	1.61	0.92	0.73
English	0.55	1.12	1.24	0.67	0.91
Spanish	0.95	1.43	1.88	0.41	1.02
Arabic	-0.54	1.56	1.94	1.04	0.12
Russian	0.46	1.32	1.54	0.66	0.54
French	-0.46	1.65	1.47	0.78	1.06

Table 6: Consequence of The First Step of EG Method

Except the constant parameter, others are all positive, which means that for all language models, its number of non-native speakers has positive relationships with the economic development level, social equality, national welfare and the effort for cultural exchanges.

$$ECM_{t-1} = S_{t-1} - (\gamma_0 + \gamma_1 Y_{1t-1} + \gamma_2 Y_{2t-1} + \gamma_3 Y_{3t-1} + \gamma_4 Y_{4t-1})\tag{11}$$

The specific results will be seen in appendix 4.

Then perform the second step of estimation. According to formula (9), we obtained ECM sequence and estimated error correction model

$$\Delta S_t = \beta_{11}\Delta Y_{1t} + \beta_{21}\Delta Y_{2t} + \beta_{31}\Delta Y_{3t} + \beta_{41}\Delta Y_{4t} - \lambda ECM_{t-1} + \varepsilon_t\tag{12}$$

Languages	β_{11}	β_{21}	β_{31}	β_{41}	λ
-----------	--------------	--------------	--------------	--------------	-----------

Mandarin Chinese	1.87	0.76	0.65	2.91	-0.66
English	4.87	0.57	0.31	4.41	-0.96
Spanish	0.47	0.02	3.56	1.04	-0.63
Arabic	3.29	4.17	3.05	1.45	-0.95
Russian	3.97	0.85	3.46	0.5	-0.31
French	3.57	0.96	2.56	0.2	-0.02

Table 7: Consequence of The Second Step of EG Method

All the first-order difference coefficients in the model are positive, reflecting that the speed of the propagation of a language as a non-native language in the world is positively related to the speed of economic development, the improvement of social equality, the growth of welfare and the improve of cultural exchanges in a cultural circle. While the ECM coefficient is stable at (-1,0), confirming that the process has a trend of negative feedback.

To sum up, we build short-term dynamic models for the number of native speakers (formula 3) and non-native speakers (formula 12), however, which cannot be exerted in the long-term. Because the prediction error of model (formula 3) will increase exponentially with the step size (Detailed in Appendix 5, but data disclosure is highly demanded in model (12), which is not applicable for most languages. As a result, we need to build a more stable and universal long-term model.

2.4 Long-term Models – Logistic Model

2.4.1 Assumptions

Except for the assumptions mentioned in page 2, another 2 assumptions are come up as follows

1. The number of all language speakers is extremely large and is differentiable in the infinite time.
2. Languages won't be extinct unless there are no speakers.

2.4.2 Model Design

Firstly, we build a long-term differential model for native speakers. According to the difference equation from formula (3) and Table 2, the number of native speakers in most languages follows a non-stationary unit root process. So, there is the same increasing pattern. As the number of native speaker is highly related to the population in the reign, we quantify the pattern of endogenous growth by Logistic Model,

$$\frac{dN}{dt} = \delta N \left(1 - \frac{N}{N_m} \right), N(0) = N_0 \quad (13)$$

δ is the coefficient of inherent growth rate, which represents the inner growth rate of the number of native speakers without limitations. N_m is the maximum capacity of native speakers in the country, which shows the increase of native speakers is confined by the gross population. N_0 is the initially calculating value.

Then the differential model is established for the number of **non-native speakers**. According to the model (12) and its EG test, the residual term is a stationary process. According to its white noise properties, the mean of long-term value can be approximated as zero. The ECM coefficient is between (0,1). The long-term error correction ECM also tends to zero. Model can be rewritten as

$$\Delta S_t = \beta_{11}\Delta Y_{1t} + \beta_{21}\Delta Y_{2t} + \beta_{31}\Delta Y_{3t} + \beta_{41}\Delta Y_{4t} \quad (14)$$

Further, we test the first-order differential linear combination model of the right side of the equation (14). In connection with the six estimations in Table 5, we construct a linear combination and perform six first-order difference tests, respectively, and determine whether the first-order difference is stable by the Dickey-Fuller test. (See Appendix 6 for details)

Languages	PP-statistics	Is first order stationary?
Mandarin China	12.3488	Yes
English	3.68	Yes
Spanish	7.97	Yes
Arabic	6.60	Yes
Russian	2.89	Yes
French	1.78	Yes

Table 8: Consequence of Dickey-Fuller Test

At the 5% significance level, the combinations of the data of six representative languages all have first-order differential stability. Because of the similarities in growth patterns of different languages, we infer that the long-term average increment of non-native speakers in one language is constant, thus, the differential equation is written as following form.

$$\frac{dS}{dt} = \mu \quad (15)$$

is a real number of any value. Add formula (13) and formula (15), a motivation system for speaker using particular language(s) will be established.

$$\begin{cases} \frac{dN_T}{dt} = \delta N \left(1 - \frac{N}{N_m}\right) + \mu \\ N_T = N + S \end{cases} \quad (16)$$

N_T is the number of total speakers using one particular language.

Formula (16) shows the dynamic system of language development includes two parts, namely, the endogenous growth system and the external impact, of which the endogenous growth system is characterized by the natural increase of the number of native speakers and the external shock is manifested by the fact that the language spread in the world as a non-native language. In the process, a series of external factors such as economy, society, national welfare and cultural communication play a role in promoting its development. Solve equation (16), we get

$$\frac{dN_T}{dt} = \delta(N_T - \mu t) \left(1 - \frac{N_T - \mu t}{N_m}\right) + \mu \quad (17)$$

The differential equation is inhomogeneous and there is no global stable equilibrium. We explore its local stable equilibrium by linear approximation. It equilibrium condition is

$$\frac{dN}{dt} = 0 \quad (18)$$

Two equilibrium solutions are $N = 0$ and $N = N_m$.

Use Lyapunov method to analyze its stability, we have

$$\begin{cases} \frac{d^2 N}{dt^2}(0) = \delta \\ \frac{d^2 N}{dt^2}(N_m) = \delta - 2 \end{cases} \quad (19)$$

Therefore, the stability of endogenous growth ultimately depends on the intrinsic growth rate coefficient δ . When $\delta < 0$, the global stability formula is $N = 0$, but the path of approach is determined by the initial value.

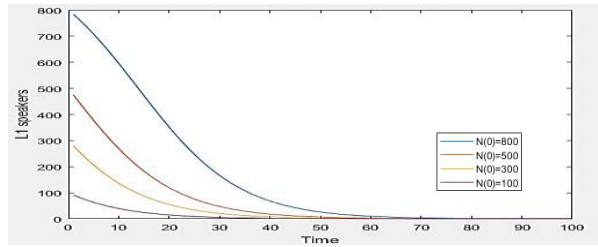


Figure 6: L1 Speakers Versus Time when $N(0) > N_m$ ($\delta = -0.1$, $N_m = 1000$)

Notes: In this section, all units of graphics abscissa are years, and all units of ordinate are millions.

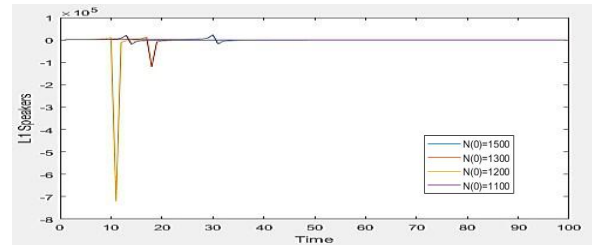


Figure 7: L1 Speakers Versus Time when $N(0) > N_m$ ($\delta = -0.1$, $N_m = 1000$)

When $0 < \delta < 2$, there is the only stable equilibrium solution of global system, and this equilibrium is not sensitive to the initial value.

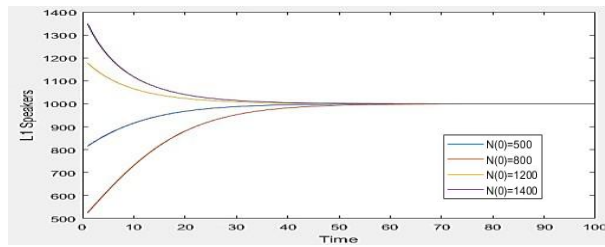


Figure 8: L1 Speakers Versus Time ($\delta = -0.1$, $N_m = 1000$)

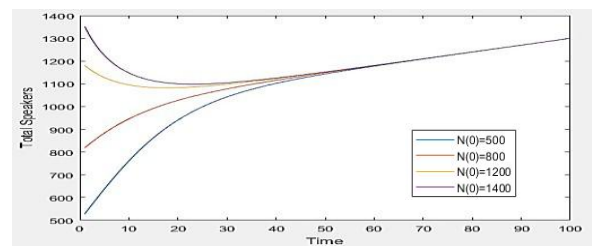


Figure 9: Total Speakers Versus Time for Different Initial Value ($\delta = 0.1$, $N_m = 1000$, $\mu = 3$)

At this point, the final result of the long-term model is

$$N_T = N_m + \mu t + \mu(0) \quad (20)$$

Now estimate the specific parameters corresponding to different languages. The inherent growth rate δ is mainly determined by the natural growth of the population in the official native speaker area. We use the natural growth rate of weighted average population of all the official mother tongue using areas in the most recent disclosure of a certain language as the estimate of δ . N_m is approximately equivalent to the weighted sum of the maximum population capacity in official language using areas of native speakers; μ approximates the increase in number of non-native speakers. We use the annual average increment (in

millions) of non-native speakers in the past 15 years as the estimate. The model shows that μ is, essentially, determined by economic, educational, political, cultural and other welfare factors.

Languages	δ	Nm	μ
Mandarin Chinese	0.005	1600	3
English	0.007	2100	9
Hindustani	0.016	1100	-2
Spanish	0.006	900	5
Arabic	0.009	700	-1
Malay	0.002	300	-1
Russian	-0.006	1700	3
Bengali	0.012	400	-2
Portuguese	0.004	600	2
French	0.001	1000	7

Table 9: Approximation of Parameters in Long-term Models

Take Mandarin Chinese as an example for stable equilibrium analysis. Its native language usage model is

$$\frac{dN}{dt} = 0.005N \left(1 - \frac{N}{1600}\right), N(0) = 897 \quad (21)$$

The model has a stable equilibrium solution $N=1600$. It takes about 300 years to reach an equilibrium state from now on. Based on the model, we add a non-native model to obtain the total model.

$$\frac{dN_T}{dt} = 0.005N \left(1 - \frac{N}{1600}\right) + 3 \quad (22)$$

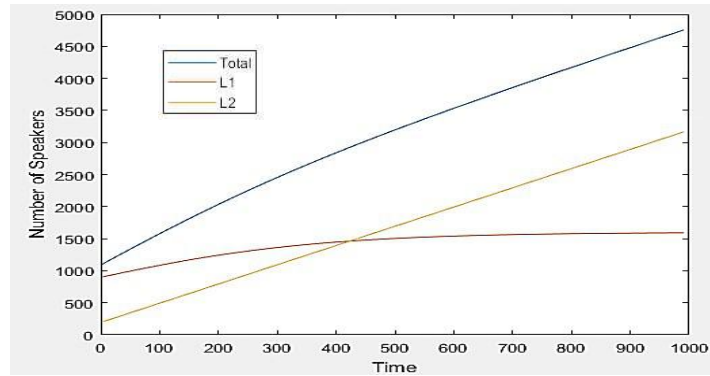


Figure 10: Number of Speakers Versus Time for Mandarin Chinese

The model shows that due to the government's control of population growth, the population growth in China gradually reaches the bottleneck. The impetus of the native Chinese users to Mandarin Chinese is weakened and the use of non-native speakers (external force) is the main driving force for language development. The remaining languages changes in future are shown in Appendix 7.

2.5 Predictions

Based on the estimated language model results, the number of top 10 languages currently in use is predicted over the next 50 years (2067).

Languages	L1 Speakers	Total Speakers	Contribution of Endogenous growth
Mandarin Chinese	994	1337	28.3%
English	490	1551	20.1%
Hindustani	536	651	193.4%
Spanish	503	753	29.6%
Arabic	368	450	278.5%
Malay	83	237	-
Russian	116	379	-33.0%
Bengali	248	160	-
Portuguese	231	262	39.3%
French	79	274	6.7%

Table10: Predictions for current Top 10 Languages in Next 50 years

According to the forecast results, the future development of languages with greater cultural influence, such as English, Mandarin Chinese and French, is mainly driven by the spread as second languages and not dependent on the growth of native speakers. In this model, language users are growing rapidly. For less culturally-influential languages, their future development depends primarily on the spontaneous growth of native speakers, which compensates for the loss of users due to declining language power under the slow growth rate of the language users and easy to fall into the bottleneck. For countries such as Russia and other

countries with negative population growth, the growth of mother-tongue users stagnated, so the second language communication has become the only way of language development.

We think Malay and Russian will not be able to maintain Top 10 in Total Speakers in the future. Malay has fewer native speakers and relies solely on second language users to promote its development. However, the users of second language reflect negative growth trends, and the general prospects of language development are not optimistic. As native speakers of Russian language continue to decline with the population reducing, although political and cultural influence ensure its dissemination as a second language, the uncertainty of the economic outlook makes it difficult for the second language as a long-term support for the development of Russian.

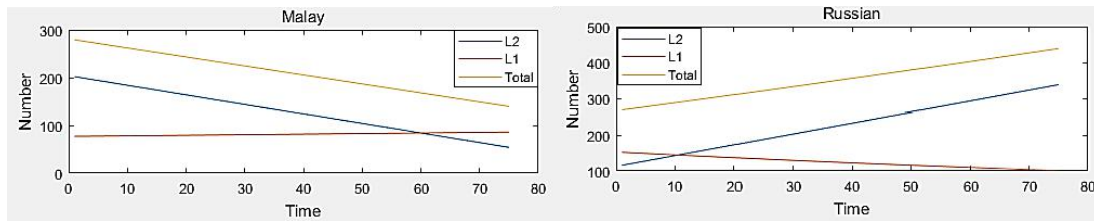


Figure 11: Predictions for Malay and Russian

Hausa and German most likely to become the alternative of Malay and Russian. The weighted average population growth rate of more than 10% in Hausa region (Africa) provided powerful support for the growth of native speakers amount, with a population of about 600million in all regions. Hausa's communication process is less likely to be bottlenecked and its coverage as a second language is on the basis of a wide population. As a second language, German has a very wide range and rapid growth. It is typical of German to rely on external cultural transmission to promote the development, which can bring about rapid growth.

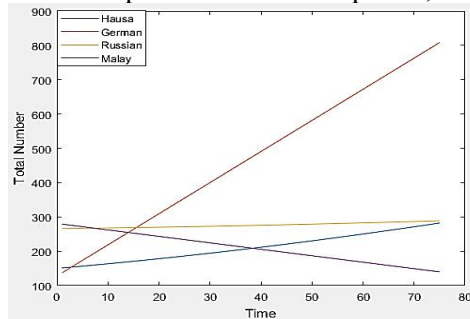


Figure 12: Number of Total Speakers for 4 Special Languages

3 Validation of Model 1

3.1 Sensitivity Analysis

We validate the long-term differential model to refrain from changes of coefficients and the turbulence of predicted results for estimated error. The capacity N_m is stable in the long run and has no effect on the number of speakers, so we won't discuss it.

3.1.1 fixed growth rate

The fixed growth rate δ influence the inner variety speed of the number of native speakers. We consider it as the natural growth rate of population approximately, which may induce error. So, we test the sensitivity by valuation. We test it, Mandarin Chinese, for example, in steps of 0.001 from 0.002 to 0.008, and estimated δ is 0.005.

δ	Equilibrium of L1 Speakers	Prediction of L1 Speakers (50 Years)	Bias of Equilibrium	Bias of Prediction
0.002	1600	936	0	6.2%
0.003	1600	952	0	4.2%
0.004	1600	980	0	1.4%
0.005	1600	994	-	-
0.006	1600	1002	0	0.8%
0.007	1600	1011	0	2.1%
0.008	1600	1026	0	3.8%

Table 11: Sensitivity Analysis on

The graph shows that the coefficient δ is greatly stable. Given the range, the harmonious situation of native speakers is the same as before for its insulation of the variety of δ . And the 50-year prediction bias is below 6.2%.

3.1.2 External Promoting Coefficient μ

The promoting coefficient μ from culture exchange is the main reason for the long-run development of languages. In the Mandarin Chinese model, let μ be 3, then we test it in steps of 0.1 from 2.7 to 3.3.

μ	Prediction of L2 Speakers (50 Years)	Prediction of Total Speakers (50 Years)	Bias of Prediction for Total Speakers
2.7	328	1322	1.1%
2.8	333	1327	0.8%
2.9	338	1332	0.4%
3.0	343	1337	-
3.1	348	1342	0.4%
3.2	353	1347	0.8%
3.3	358	1352	1.1%

Table 12: Sensitivity Analysis on μ

The graph shows the model is stable as μ changes. Given the range, the prediction bias of the total speakers is below 1.1%.

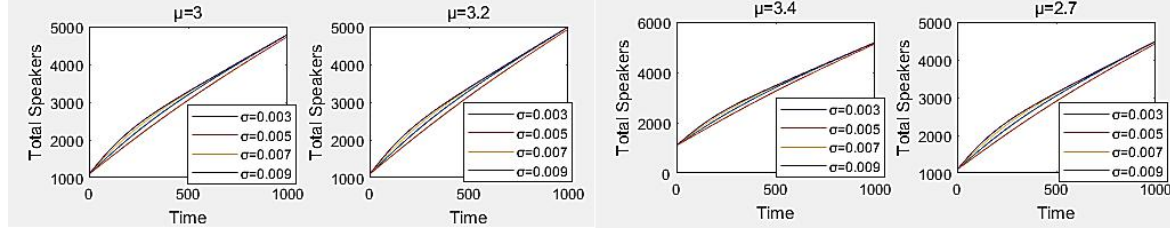


Figure 12: Comprehensive Sensitivity Analysis for Mandarin Chinese

For all those reasons, the long-term model is stable for the inner coefficients, which is applicable for the prediction of this problem. The analysis of the other languages is detailed in Appendix 8.

3.1.3 Initial Value

The number of native Russian speakers is special. The path towards the equilibrium point is more sensitive than other models towards the initial value, for which we have to analyze the sensitivity of its initial value.

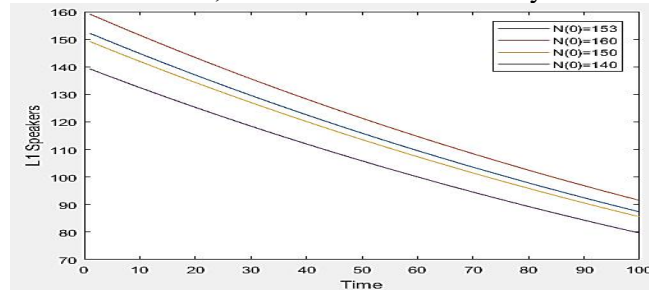


Figure 13: Sensitivity Analysis on Initial Value of N for Russian Model

The graph shows the time distribution path moves in a relatively long interval (100 years), when the initial value fluctuates within the range of 20 million, which have little effect on the prediction. So, the Russian model is available, the number of native speakers in the next 50 years will be stable between 108 million and 120 million

3.2 Robust Analysis

3.2.1 Monte-Carlo Simulation

In the previous situation, we transform the short-term time series differential model into the long-term differential model. Now we introduce random error, establish long-term stochastic model and visualize it with Monte-Carlo simulation.

Firstly, in the recursion time equation of the number of native speakers, the change from the t^{th} period to the $t+1^{\text{th}}$ period will be affected by other random factors except the self-correlation. Assuming that it forms a normal distribution, namely $N(0, \sigma_N^2)$, then the equation is

$$N_{t+1} = N_t + \delta N_t \left(1 - \frac{N_t}{N_m}\right) + N(0, \sigma_N^2) \quad (23)$$

Similarly, in the process of changing the number of non-native speakers, there are also random factors other than economic, culture, immigrant and other factors. Assuming that the mean is 0 and the variance is positive, the recurrence equation is, $S_{t+1} = S_t + \mu + N(0, \sigma_S^2)$ (24)

The recurrence equation of the total number is,

$$N_{Tt+1} = N_{Tt} + \delta N_t \left(1 - \frac{N_t}{N_m}\right) + \mu + N(0, \sigma^2) + N(0, \sigma^2) \quad (25)$$

We simulate the models (23/25) of different variance, Mandarin Chinese, for example.

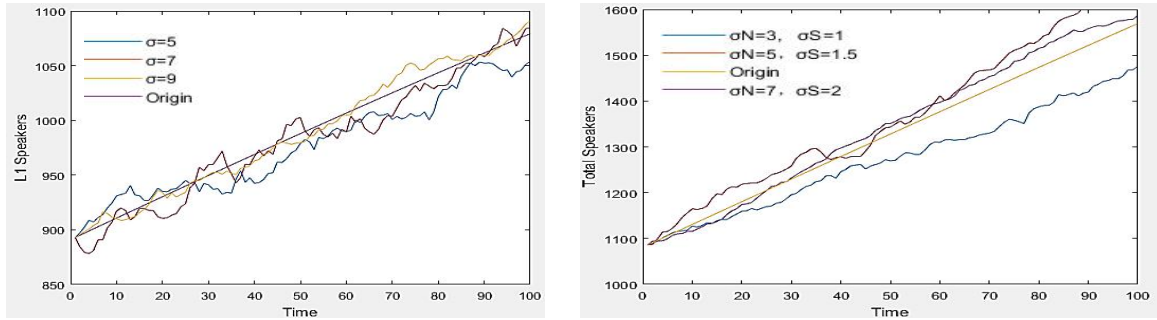


Figure 14: Monte Carlo Simulation on L1 Speakers and Total Speakers for Mandarin Chinese

The graphs show that the introduction of random error doesn't change the time paths of L1 speakers and total speakers. For different variances, the fluctuation is acceptable. At the forecast point (50,994), the L1 speakers' stochastic volatility does not exceed 40 and the stochastic volatility of the total speakers does not exceed 80 at the point (50,1337). Then the prediction is still applicable. Therefore, the model is robust over random fluctuations.

4 Model 2 - Geographic distribution

4.1 Migration Patterns – Markov Process

4.1.1 Assumptions and Variables

a. Assumptions

- (1) Population migration is based on the state and we do not consider domestic population movements.
- (2) People in economically underdeveloped areas will migrate to the economically developed areas, resulting in stable trend of migration and fixed probability of transfer.
- (3) The probability of immigration in each country is not related to the time stage, but relating to countries.
- (4) The growth rates of population in different countries are not consistent. For different countries, the population of each period changes according to the growth rate of each country.
- (5) The migration of immigrants from various countries can be completed at the same time and can be considered as one-off completed by the end of the year.

b. Variables

VARIABLES	DEFINITION
	Population of i^{th} country in t^{th} year
	Vector made up of elements N
	The Growth Rate of population of i^{th} country in t^{th} year
	Vector made up of elements + 1
P	Vector made up of elements One-step transition probabilities matrix
	The number of migrants from i^{th} country to j^{th} country in 2015
	Matrix formed by 0
	Steady-state probability of i^{th} country
I, J	Steady-state probability vector Country's serial number

Table 13: Variables for Model 2

4.1.2 Model Settings

The transition process is a discrete-time stochastic process. The occurrence of immigrants satisfies Markov conditions. Regard to the conditional probability $P(N_{n+1} = j | \mathcal{F}_n)$, it satisfies the following expression.

$$P(N_{n+1} = j | N_0 = i_0, N_1 = i_1, \dots, N_{n-1} = i_{n-1}, N_n = i_n) = P(N_{n+1} = j | N_n = i_n) \quad (26)$$

This means that when the status of the immigration process at time n is known, the status of immigration process after time n has nothing to do with the status before n , that is, no post-validity. We define

$p_{i,j} = P(N_{n+1} = j | N_n = i_n)$, then the entire migration of the process $\{N_n\}$ is determined by the $p_{i,j}$ and the initial distribution of N_0 . As we know from assumption (3), $p_{i,j}$ is only related to country i, j , but has nothing to do with n , then Markov chain is time-aligned Markov chain. Then write the $p_{i,j}$ in matrix form:

$$P = (p_{i,j}) \begin{pmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,226} \\ p_{2,1} & p_{2,2} & \dots & p_{2,226} \\ \vdots & \vdots & \ddots & \vdots \\ p_{226,1} & p_{226,2} & \dots & p_{226,226} \end{pmatrix} \quad (27)$$

Because the transition probability is positive, and residents will certainly either stay in their own country or move to other countries in the next period, thus, the matrix has the following properties.

$$(1) p_{i,j} > 0, i, j = 1, 2, \dots, 226$$

$$(2) \sum_{j=1}^{226} p_{i,j} = 1, \forall i = 1, 2, \dots, 226$$

Thus, according to the Chapman-Kolmogorov Equation, we have n -step transition probability matrix:

$$P^{(n)} = P \cdot P^{(n-1)} = \dots = P^n = \begin{pmatrix} p_{1,1}^{(n)} & p_{1,2}^{(n)} & \dots & p_{1,226}^{(n)} \\ p_{2,1}^{(n)} & p_{2,2}^{(n)} & \dots & p_{2,226}^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ p_{226,1}^{(n)} & p_{226,2}^{(n)} & \dots & p_{226,226}^{(n)} \end{pmatrix} \quad (28)$$

Therefore, the number of residents in each country changes as follows.

In the 0th period, the vector $N_0 = (N_{1,0}, N_{2,0}, N_{3,0}, \dots, N_{226,0})$ of resident numbers of each country is calculated by the weighted moving average of the historical natural population growth rate as a natural vector of growth rates $g_0 = (g_{1,0}, g_{2,0}, g_{3,0}, \dots, g_{226,0})^T$, which will converge to 0 according to a certain rate in the next 50 years. Then we have the vector $G_0 = (G_{1,0}, G_{2,0}, G_{3,0}, \dots, G_{226,0})^T$ of population growth.

Define the rule of operator \otimes is that $\begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \otimes \begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix} = \begin{pmatrix} a_1 * b_1 & a_2 * b_2 \\ a_3 * b_3 & a_4 * b_4 \end{pmatrix}$, which has the same function to multidimensional matrix.

In the 1st period, $N_1 = N_0 * P \otimes G_1$.

Namely, $N_1 = (N_{1,1}, N_{2,1}, N_{3,1}, \dots, N_{226,1})$

$$= (N_{1,0}, N_{2,0}, N_{3,0}, \dots, N_{226,0}) * \begin{pmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,226} \\ p_{2,1} & p_{2,2} & \dots & p_{2,226} \\ \vdots & \vdots & \ddots & \vdots \\ p_{226,1} & p_{226,2} & \dots & p_{226,226} \end{pmatrix} \otimes (G_{1,0}, G_{2,0}, G_{3,0}, \dots, G_{226,0})^T \quad (29)$$

$$\text{In the 2nd period, } N_2 = N_1 * P \otimes G_2 = N_0 * P^{(2)} \otimes G_1 \otimes G_2 = N_0 * P^2 \otimes G_1 \otimes G_2 \quad (30)$$

$$\text{Thus, in the } n^{\text{th}} \text{ period, } N_n = N_{n-1} * P \otimes G_n = N_0 * P^{(n)} \otimes \prod_{i=1}^n G_i = N_0 * P^n \otimes \prod_{i=1}^n G_i \quad (31)$$

Here, $\prod_{i=1}^n G_i = G_1 \otimes G_2 \otimes \dots \otimes G_n$.

In the meanwhile, as we know, $\{N_n\}$ is ergodic, which means that this Markov chain tends to steady state. So, there is a unique asymptotically stable equilibrium $\pi = (\pi_1, \pi_2, \pi_3, \dots, \pi_{226})$, $\pi_j = \lim_{n \rightarrow \infty} (p_{i,j})^{(n)}$. And it satisfies the following equations.

$$\begin{cases} \pi P = \pi \\ \sum_{i=1}^{226} \pi_i = 1 \end{cases} \quad (32)$$

Over time, the proportion of the population in each region will approach a more stable level. According to the model of population growth, after 50 years, the population will gradually become the maximum sustainable value in the future and fluctuate around the maximum value, with the natural growth rate converging to zero. Population in all regions fluctuates around its maximal value.

4.1.3 Solution and Visualization

$$\text{As known, } \begin{pmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,226} \\ p_{2,1} & p_{2,2} & \dots & p_{2,226} \\ \vdots & \vdots & \ddots & \vdots \\ p_{226,1} & p_{226,2} & \dots & p_{226,226} \end{pmatrix} \otimes \begin{pmatrix} N_{1,0} & N_{1,0} & \dots & N_{1,0} \\ N_{2,0} & N_{2,0} & \dots & N_{2,0} \\ \vdots & \vdots & \ddots & \vdots \\ N_{226,0} & N_{226,0} & \dots & N_{226,0} \end{pmatrix}$$

$$= \begin{pmatrix} M_{1,1} & M_{1,2} & \cdots & M_{1,226} \\ M_{2,1} & M_{2,2} & \cdots & M_{2,226} \\ \vdots & \vdots & \ddots & \vdots \\ M_{226,1} & M_{226,2} & \cdots & M_{226,226} \end{pmatrix}, \text{ namely, } P \otimes N_0 = M_0. \text{ (Details in Appendix 9)} \quad (33)$$

($M_{i,j}$ is the number of migrants from i^{th} country to j^{th} country in 2015, $i,j=1,2,\dots,6$)

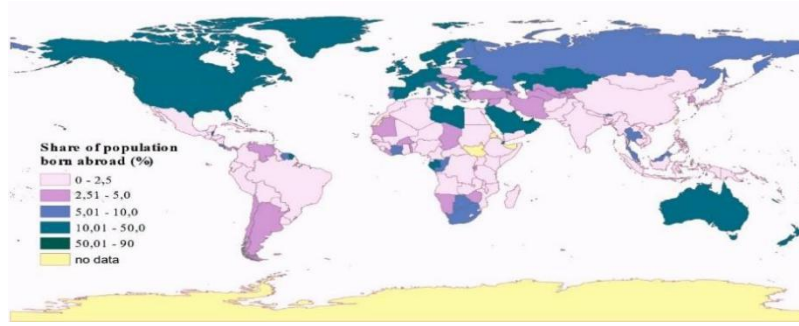


Figure 15: Share of population born abroad of Each Country (Visualized Moving-in Matrix M_0)

The matrix P of 226×226 is deduced from (33), and Appendix 10 shows the partial calculation results of P . And we get the natural growth rate of population vector G by moving average method.

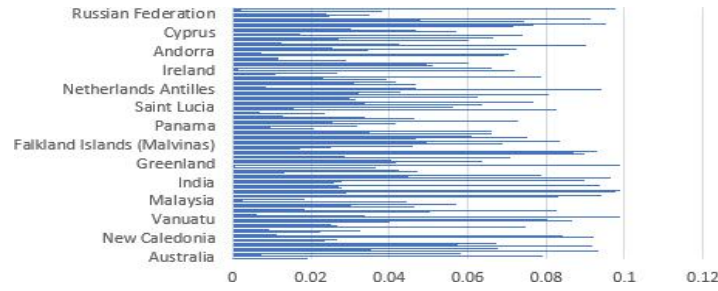


Figure 16: Visualize G by The Mobile Average Data of Population Naturally Growth Rate (2005-2016 Partially)

After obtaining P and G , according to the Markov transfer model, we can calculate the population quantity vector N_t , which can be seen in Appendix 11 for details.

4.2 Language Distributions

4.2.1 Prediction

We calculate the distribution of the number of Mandarin Chinese in the world, for example, and compare the situations in different periods.

For every population vector, there is a matrix of languages distribution of all countries, H , which indicates the ratio of the number of people in a given language in each country to the total population of the country, namely,

$$H = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1(26)} \\ h_{21} & h_{22} & \cdots & h_{2(26)} \\ \vdots & \vdots & \ddots & \vdots \\ h_{226(1)} & h_{226(2)} & \cdots & h_{226(26)} \end{pmatrix} \quad (34)$$

h_{ij} denotes the ratio of the number of people who have a demand of the language in the i^{th} country, $i=1,2,\dots,226$ and $j=1,2,\dots,26$ (for the 26 given languages in the problem). We can infer the distribution of all languages in the t^{th} period by formula (35).

$$\text{diag}(N_t) * H \quad (35)$$

$\text{diag}(N_t)$ indicates the diagonal matrix of diagonal elements N_t .

In the resultant matrix, every column represents the distribution of total usage of a language in 226 countries. Combining them with the latitude and longitude of the corresponding capitals, we get 226 scattered points in the 3D real number domain, and use Lagrange interpolation method to smooth the space between the scattered points to obtain the spatial distribution curve of the language in the world.

The data for each element in matrix H are mainly derived from the world language distribution data for the 2017 EPS database disclosure but for each of the post-2017 calculations, we manually adjust the H matrix manually so that it is generally consistent with the Law of population migration. Based on equation (35), we extract and visualize Mandarin Chinese.

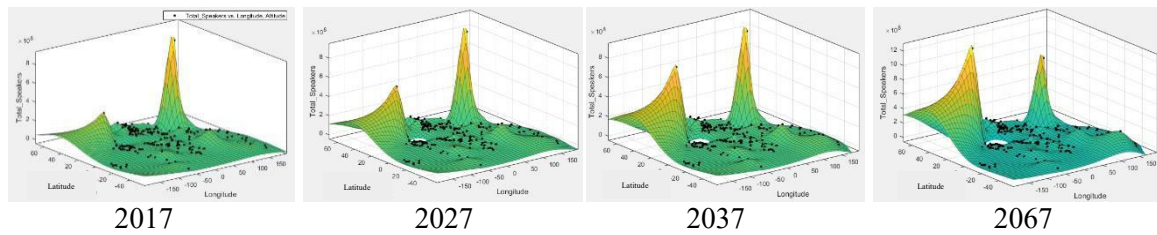


Figure 17: Geographic Distribution of Mandarin Chinese in The Next 50 Years

4.2.2 Conclusion

According to the Markov model of world immigrants, we simulate the geographical distribution of Mandarin Chinese total speakers in 4 periods. Two peaks were estimated, one in China in East Asia and one in North America. The peak in East Asia has basically stabilized at an unchanged high point. The peak in North America has become more pronounced. 50 years later, East Asia will become where Mandarin Chinese is most widely used. The reason for the change is that the population growth in China has entered a slowdown phase. The increase of native speakers of Mandarin Chinese is not obvious. With the influx of a large number of Chinese immigrants into North America and the promotion of Chinese influence, the number of L2 speakers will obviously rise, changing the long-term structure of Mandarin Chinese speakers. Due to the above results of the differential equation model, the increment of Mandarin Chinese's non-native speakers accounts for 71.4% of the increment of total speakers, which is 3 times of the native speakers' increment and is consistent with the result of Markov Model. The geographical distributions of other languages are detailed in Appendix 12.

According to the result, there is a tendency that the geographical distribution of all languages will be mixed up in the next 50 years and the non-native language will be more abundant. English is the most stable and equal, which will still be the most important language in the world. Non-native languages, speakers, for example Spanish and German, will rise. And the use of French will increase significantly.

4.3 Comparison of Model 1&2

In this part, we will compare the results of Markov model and long-term differential model to ensure the model is applicable.

According to formula (35), we get the distributions of each language in each country. To sum up all columns of matrix $\text{diag}(N_t) \cdot H$, we will get the number of total speakers in t^{th} period, namely the predicted number of each language in t^{th} period. Let $t=50$, the result is,

Languages	Predictions by Markov Model	Predictions by Differential Model	Difference
Mandarin Chinese	1502	1337	165
English	1488	1551	-63
Hindustani	645	651	-6
Spanish	798	763	35
Arabic	416	450	-34
Malay	295	237	58
Russian	306	379	-73
Bengali	133	160	-27
French	368	274	94
Portuguese	270	262	8

Table 14: Difference Between Predictions of 2 Models

An analysis of variance on the 2 sets of predicted data shows that 94% of the probabilities have no significant difference. Therefore, we think the prediction results of the two models are the same, which mutually prove the feasibility. The analysis of variance is detailed in Appendix 14.

5 Model 3 – Offices Location Decisions

5.1 Model in Short-term and Long-term - Cluster Analysis & MODM

5.1.1 Assumption

Except the previous assumptions, we assume the company prefers the capital of each country for their new offices.

5.1.2 The Settings and Solutions of Model

The company hopes to promote the internationalization, so we choose 6 cities from the capitals of the residual 224 countries except China and the United States. We construct a matrix L_t , which represents the number of each language in each country (including L1 and L2) in t^{th} period.

$$L_t = \text{diag}(N_t)H \quad (36)$$

$L_t(i,j)$ denotes the total number of the j^{th} language speakers of the i^{th} country. The meaning of N_t and H is the same of the meaning in Markov model (formula 11). As the employees have a demand of English at least, we prefer the number of English speakers firstly. In matrix L_t , we extract the vector of the total number of English speakers in each country, and then introduce the vector of the ratio of English speakers to the population of the country, which is denoted by $E_{t(224 \times 1)}$.

Then we analyze the elements of the vector by cluster, so there are k_1 categories $G_{11}, G_{12}, \dots, G_{1k_1}$. Considering the economic, culture factors, we hope to invest in the developed countries as possible. We use the number of net immigrants in each country MI_t and the per capita GDP vector GDP_t of each country as a measure of their level of development. MI_t can be calculated by summing up the columns of Markov Model. Then we throw a cluster analysis to get k_2 categories $G_{21}, G_{22}, \dots, G_{2k_2}$ and k_3 categories $G_{31}, G_{32}, \dots, G_{3k_3}$. Finally, we introduce the Chinese speakers to facilitate the communication with the office in Shanghai. We extract the ratio vector C_t of the number of Chinese speakers and cluster analyze to get k_4 categories $G_{41}, G_{42}, \dots, G_{4k_4}$.

Firstly, here comes the short-term model. Let $t=10$, we estimate E_{10}, MI_{10}, C_{10} by Markov Model. According to the error correction equation in the short-term difference model and the influence of GDP per capita on the other hand, the effect of the factors such as GDP per capita is approximately stable. Therefore, we use the GDP per capita data in this model as the estimation of actual GDP per capita in the short term. The result of systematic cluster analysis as following,

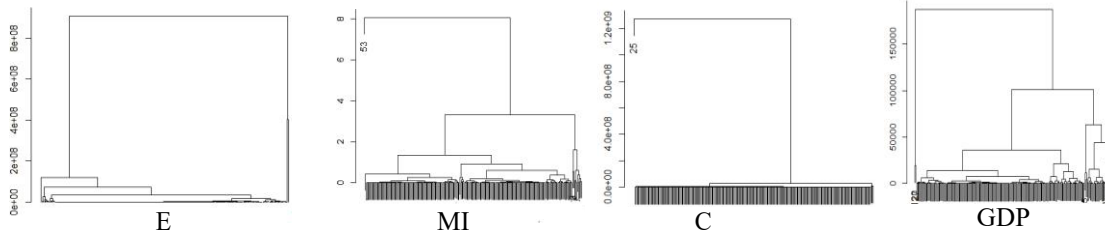


Figure 18: Cluster Dendrograms for 4 Vectors of Variables when $t=5$

According to the result of clustering, the number of 4 categories is determined respectively, $k_1=2$, $k_2=3$, $k_3=4$, $k_4=2$. Countries with more English speakers are G_{11} , and G_{12} otherwise. In accordance with the number of attracting immigrants from more to less in the order, divide countries into G_{21}, G_{22} and G_{23} . As sorting GDP per capita in descending order, divide countries into $G_{31}, G_{32}, G_{33}, G_{34}$. Countries with more Chinese users are defined as G_{41} , otherwise G_{42} . Part of the country classification details is in Appendix 15.

The Multiple Objective Decision Making (MODM) of Fuzzy Evaluation is used to determine the location of offices. The English speaker proportion, the number of attracting immigrants, GDP per capita and the Chinese proportion are taken as the indicators respectively corresponding to the index function (37) -(40). Because the company has established offices in the United States and China, we consider English and Chinese into the decision-making system.

$$EI = \begin{cases} 1, G1 = 1 \\ 0, G1 = 2 \end{cases} \quad (37)$$

$$MI = \begin{cases} 2, G2 = 1 \\ 1, G2 = 2 \\ 0, G2 = 3 \end{cases} \quad (38)$$

$$GDPI = \begin{cases} 3, G3 = 1 \\ 2, G3 = 2 \\ 1, G3 = 3 \\ 0, G3 = 4 \end{cases} \quad (39)$$

$$CI = \begin{cases} 1, G4 = 1 \\ 0, G4 = 2 \end{cases} \quad (40)$$

Establish membership function based on the index (37) - (40).

$$\mu_{EI}(x) = x \quad (41)$$

$$\mu_{MII}(x) = x/2 \quad (42)$$

$$\mu_{GDPI}(x) = x/3 \quad (43)$$

$$\mu_{CI}(x) = x \quad (44)$$

Determine the membership degree according to the affiliation, and get the fuzzy relation matrix

$$R = \begin{pmatrix} r_{EI1} & r_{EI2} & \dots & r_{EI224} \\ r_{CI1} & r_{CI2} & \dots & r_{CI224} \\ r_{MII1} & r_{MII2} & \dots & r_{MII224} \\ r_{GDPI1} & r_{GDPI2} & \dots & r_{GDPI224} \end{pmatrix} \quad (45)$$

The element r represents the score of each country in each indicator. According to the demand, English is a necessary language demand for the new office members, and secondly, the regional economy is taken into account. The Chinese index is considered as supplement, so the English language index is given the highest weight. Therefore, the weight vector is $A = (0.4, 0.25, 0.25, 0.1)$ (46)

The comprehensive evaluation matrix of each country is $B = AR$ (47)

Take the largest six elements in B and its corresponding country (Details of B vectors will be seen in Appendix 16) to get the short-term model results.

Using the same method for **long-term** analysis, where t is set to be 50, and four indicators as well as countries' scores are obtained by combining the long-term differential model and the Markov model prediction result in Model 2 to obtain the top six countries.

5.2 Comparison

The result of the short-term model is as follows.

Country	Score	Capital City	English Speaker Proportion	GDP per Capital	Chinese Speaker Proportion
The U.K	0.88	London	0.91	45673	0.06
Canada	0.84	Ottawa	0.85	51962	0.17
Singapore	0.83	Singapore City	0.83	54776	0.76
Australia	0.75	Canberra	0.9	67488	0.21
Germany	0.74	Berlin	0.82	45088	0.04
New Zealand	0.71	Wellington	0.91	40652	0.15

Table 15: Short-term results

To summarize, in the short term, it is advisable to locate the new offices in London, Ottawa, Singapore City, Canberra, Berlin and Wellington. We set the office languages in English, French, English or Chinese, and English, German, English, respectively, in accordance with the highest proportion of languages in each country.

According to the prediction result of long-term model, the top 6 countries are shown in the table 16.

Country	Score	Capital City	English Speaker Proportion	Chinese Speaker Proportion
Singapore	0.91	Singapore City	0.93	0.81
The U.K.	0.88	London	0.9	0.05
Australia	0.82	Canberra	0.85	0.17
France	0.78	Paris	0.79	0.18
Canada	0.75	Ottawa	0.84	0.18
Malaysia	0.66	Kuala Lumpur	0.58	0.78

Table 16: Long-term results

There are a few differences in the choices of different phases. In the long run, it is advisable to locate the new offices in Singapore City, London, Canberra, Paris, Ottawa, Kuala Lumpur. The language suggested to use in each office are Chines, English, English, French, French, Malay or Chinese, respectively.

5.3 Resource-saving Suggestions – MINE Model

Except for the original assumptions, one more assumption is added:

Global communication is between New York, Tokyo, Hong Kong and London, four major economic centers.

Based on the above model, we have set up eight international offices. We are considering reducing some of our new offices to reduce costs. New offices are expected to spread around the world as far as possible so as to improve coverage, because the coverage radius of offices that are too close from each other will easily overlap, resulting in a waste of resources. We identify overly dense offices by the Maximum Information-based Nonparametric Exploration proposed by M Filosi (2014).

First, the latitude and longitude coordinate plane is divided into blocks centered on New York, Tokyo,

Hong Kong and London. The eight sample points are distributed in four blocks according to the coordinates. The grid formed by the four blocks is denoted as G , and the probability density function of G is defined as

$$p(x, y) = \begin{cases} \frac{\text{Number_of_Samples_in_block1}}{8}, (x, y) \in \text{block1} \\ \frac{\text{Number_of_Samples_in_block2}}{8}, (x, y) \in \text{block2} \\ \frac{\text{Number_of_Samples_in_block3}}{8}, (x, y) \in \text{block3} \\ \frac{\text{Number_of_Samples_in_block4}}{8}, (x, y) \in \text{block4} \end{cases} \quad (48)$$

Define
$$I(G) = \int_r \int_x p(x, y) \ln \left(\frac{p(x, y)}{p(x)p(y)} \right) dx dy \quad (49)$$

Construct characteristic matrix $\{m_{xy}\}$, where
$$m_{xy} = \frac{\max\{I_G\}}{\ln \min\{x, y\}} \quad (50)$$

Max required taking over all possible grids G of coordinate (x, y) . Then define maximum information coefficient (MIC) as
$$\text{MIC} = \max\{m_{xy}\} \quad (51)$$

And max need to take over all possible (x, y) .

MIC value range $(0, 1)$. The closer to 1, the more concentrated the sample point distribution. The closer to 0, the more dispersed the sample points. For the short-term eight sample points, MINE model is as following.

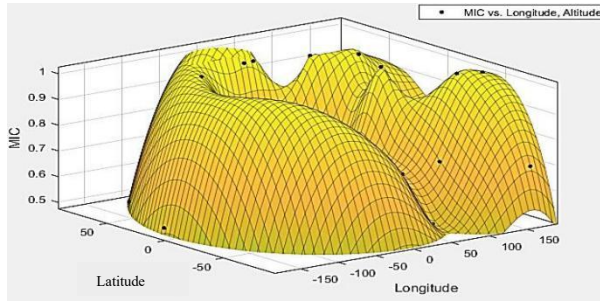


Figure 19: Distribution of MIC for Short-term Samples

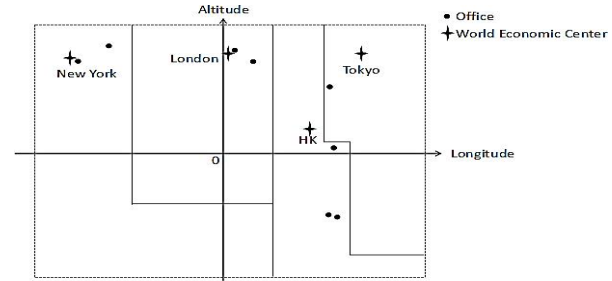


Figure 20: Geographic Distribution of 8 Offices and Economic Centers

From the results, we know that the MIC value approaches 1 in most regions and less than 0.5 in only a few regions, which indicates that the distribution of sample points is too concentrated. Considering the score of each country in question A and the distance between offices, it is suggested Reduce the office in Willington, Berlin, resulting in four new offices for model 2: London, Singapore City, Ottawa and Canberra.

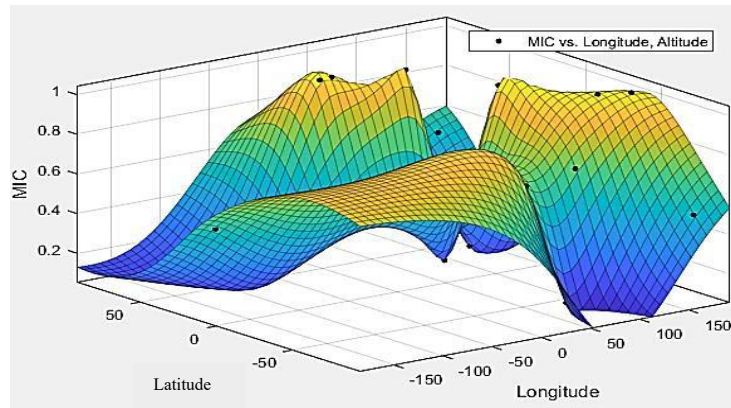


Figure 21: Distribution of MIC after Improving

The result shows that the removal of the 2 redundant offices helps relaxing the over-intensity, which enhance the economic proficiency.

Similarly, the improvement result of the **long-term** model shows the offices are redundant in Kuala Lumpur and Ottawa. So, we recommend building 4 new offices in Singapore, London, Canberra and Paris.

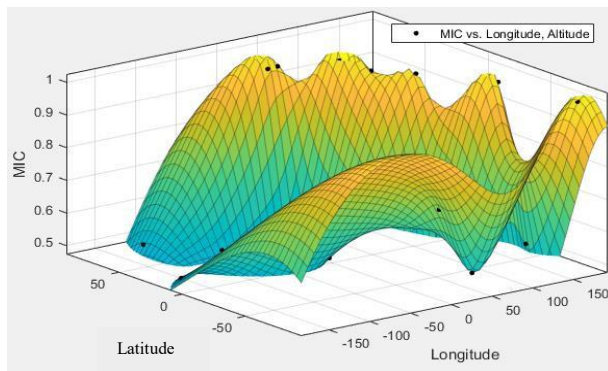


Figure 22: Distribution of MIC for Original Plan

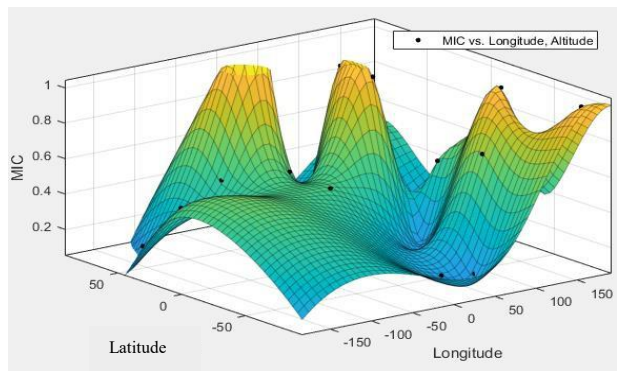


Figure 23: Distribution of MIC for New Plan in Long-term

6 Model Assessments

6.1 Strengths

- The time series model represents the influence of the historical information of each variable on the situation of the system and entirely describes the dynamic law of the change of number of speakers.
- The ECM model combines the horizontal and the differential variables organically, makes full use of the information given by the two, and combines the long-term and short-term effects between the variables to enhance the ability of model interpretation and prediction. When the variables are non-stationary, there is a co-integration relationship among the variables in the ECM model. Since neither the error term nor the differential variables are stable, neither false regression nor long-term information will be lost.
- For the description of geographical location, we use cluster analysis, simplify the 2-D plane into coordinate points and enhance visualization effects as well as simplify the calculation process.
- The Markov Chain describes the process of random migration well, which is derived from the dynamic prediction in accordance with the reality.
- The MODM (Multiple objective decision making) theory reflect the relationship between each single-objective optimal solution and multi-objective satisfactory solution, can consider the satisfaction of different nature and conflicting multiple goals, which may optimize all goals as possible and provide a new approach to optimize the multi-objective systematic problems.
- The introduction of MINE model promotes the analyzing efficiency, the revising the quantifying method and makes the result intuitive.

6.2 Weaknesses

- There is two main drawbacks in Monte Carlo Simulation. If a random number which may not be random as expected have to be put in the particular part of system, some subtle non-random patterns will appear. And there would be deviate from the authentic simulation results. And Monte Carlo simulation with static characteristics has trouble to complete long-term dynamic simulation.
- The division of 4 economic centers is rough and simple. And the division of grid shapes is subjective.

6.3 Improvements

- Introduce the Markov process in the random process into the Monte Carlo simulation to achieve dynamic Monte Carlo simulation. That is to say, a Monte Carlo-Markov chain (MCMC) is constructed. By sampling repeatedly, a Markov chain with the same distribution of system probabilities is established to obtain the state samples of the system. The dynamic simulation of changing the sampling distribution as the simulation progresses makes up for the shortcomings that the traditional Monte Carlo integrals can only be modeled statically.
- Additional information is needed. For example, we divide the world in a more specific way to get the economic zones, combining the additional information such as the culture, economic structure, the international trade and the financial markets, which contributes to promote the accuracy of the division of grid and quantify the international coverage of offices in a more specific way.

Memo

To: Chief Operating Officer of the service company
From: Team 74316
Date: 12 February
Subject: Results and Recommendations on language distribution and offices decision

We are writing to report our key results and Recommendations in these four days.

Results

1. Languages spread in 2 ways, the variety of the number of native speakers and the non-native speakers. In the short-term, the distribution of the former is mainly reflected in the temporal sequence correlation. The temporal sequence of most L1 speakers of most languages is a non-stationary unit root process, which increases as the local population increases except Russian. The temporal distribution of the latter is related to all kinds of influences and factors. We think that economic, social, welfare and cultural factors have a positive impact on the number of non-native speakers and the total number of speakers.
2. In the long-term, the variety of L1 speakers is consistent with the natural increment of logistic, which usually is a stable system. L2 speakers are subject to economic, cultural and other external factors, which can be simplified into an unchanged parameter. The total number is equal to the sum of L1 and L2 speakers, which is a non-stationary dynamic system. The number of most native speakers increases as the population increases; however, the temporal path of non-native speakers is affected by the language influences.

Languages that are influential such as Mandarin Chinese and English increases because the non-native speakers increase and the number of total speakers will increase in 50 years. Languages that is little influential will slowly increase as the population increase, and the number of total speakers is not significant even negative.

3. Russian and Malay are at risk of dropping out of top 10. The reason is that Russian native speakers experienced negative growth and Malay speakers grew negatively and had few native speakers, which are replaced by Hausa with a fast-growing mother tongue and German with wide spread speakers.
4. The geographical distribution of languages presents a trend of mutual integration. For example, the increase of Mandarin Chinese non-native speakers in Europe exceeds the increase of native speakers in China. The number of non-native speakers in the world maintained a steady growth. The distribution of Spanish and Portuguese tends to be even. The number of French non-native speakers in Europe is growing rapidly.

Recommendations

1. In the short term (about 10 years), it is recommended that the company set up international offices in London, Ottawa, Canberra, Singapore, Berlin and Willington. These cities have performed best in comprehensive English usage, Chinese usage, and economic development, conducting to the company's international development. And the office languages are suggested to be English, French, English, English or Chinese, German, English, respectively.

In the long term (over the next 50 years), the proposed locations are Singapore, London, Canberra, Paris, Ottawa, and Kuala Lumpur. It is recommended that the office languages be English or Chinese, English, English, French, French, Malay or Chinese.

2. If six new offices were not necessarily required, it would be enough to suggest the establishment of four new offices. Because too many offices would lead to a over dense geographical distribution, overlapping coverage, and increase marginal costs as well as reduce marginal benefits, resulting in waste of resources. With the MINE model, we identify grids that are too densely distributed and eliminate duplicated offices, retaining the highest level of comprehensive evaluation and coverage. We suggest that opening new offices as shown below.

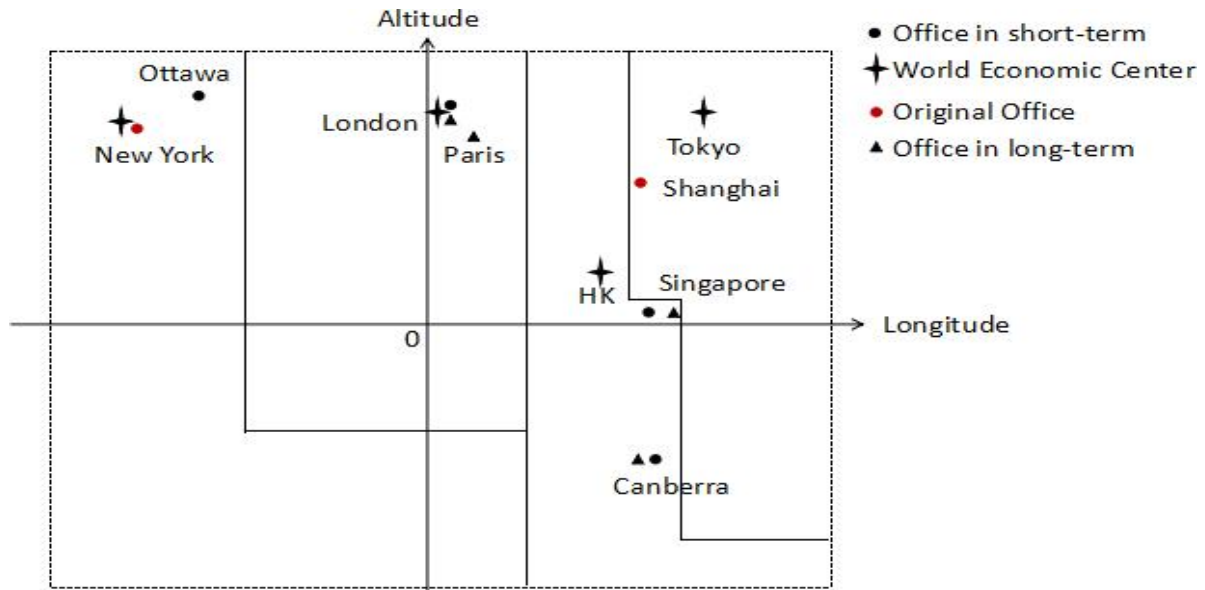


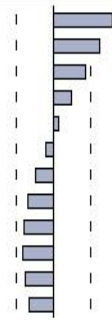
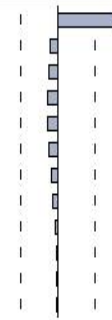
Figure: Appropriate Locations of International Offices According to New Plan

Work cited

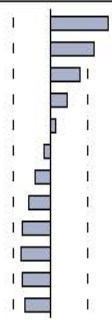
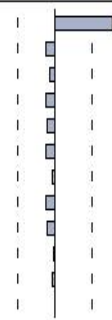
- [1] Bouallègue, Zied Ben. “Calibrated Short-Range Ensemble Precipitation Forecasts Using Extended Logistic Regression with Interaction Terms” *Weather and Forecasting*; Boston Vol. 28, Iss. 2, (Apr 2013): 515-524.
- [2] Cuevas, Ángel; Pérez -Quirós, Gabriel; Quilis, Enrique M. *Revista de Economía Aplicada*. “Integrated Model of Short-term Forecasting of the Spanish Economy (Mpire Model)” Zaragoza Vol. 25, Iss. 74, (Fall 2017): 5-25.
- [3] Daniel Rudolf. Error bounds for computing the expectation by Markov chain Monte Carlo[J]. *Monte Carlo Methods and Applications*, 2010,16(3-4).
- [4] Filosi M, Visintainer R, Albanese D. minerva: Maximal Information-Based Nonparametric Exploration Rpackage for Variable Analysis[J]. 2014.
- [5] Haikun Qi, Feng Huang, Hongmei Zhou, Huijun Chen. Sequential combination of k - t principle component analysis (PCA) and partial parallel imaging: k - t PCA GROWL[J]. *Magnetic Resonance in Medicine*, 2017, 77(3).
- [6] Kehrer Johannes, Hauser Helwig. Visualization and Visual Analysis of Multi-faceted Scientific Data: A Survey[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2012.
- [7] McGhan, Anna C; Lerman, Dorothea C. “AN ASSESSMENT OF ERROR-CORRECTION PROCEDURES FOR LEARNERS WITH AUTISM” *Journal of Applied Behavior Analysis*; Malden Vol. 46, Iss. 3, (Fall 2013): 626-39.
- [8] Paulo Fernandez, Sandra Mourato, Madalena Moreira, Luísa Pereira. A new approach for computing a flood vulnerability index using cluster analysis[J]. *Physics and Chemistry of the Earth*, 2016, 94.
- [9] R. Kruse, M. Steinbrecher. Visual data analysis with computational intelligence methods[J]. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 2010, 58(3).
- [10] Tracey, Brendan; Wolpert, David; Alonso, Juan J. “Using Supervised Learning to Improve Monte Carlo Integral Estimation” *American Institute of Aeronautics and Astronautics. AIAA Journal*; Virginia Vol. 51, Iss. 8, (Aug 2013): 2015-2023.

Appendix

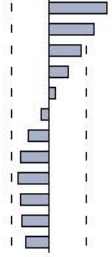
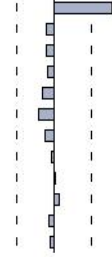
Appendix 1

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 0.821	0.821	12.278	0.000
		2 0.642	-0.100	20.351	0.000
		3 0.454	-0.135	24.724	0.000
		4 0.260	-0.146	26.297	0.000
		5 0.067	-0.154	26.411	0.000
		6 -0.110	-0.127	26.757	0.000
		7 -0.253	-0.083	28.803	0.000
		8 -0.357	-0.067	33.456	0.000
		9 -0.418	-0.045	40.879	0.000
		10 -0.430	-0.013	50.326	0.000
		11 -0.404	-0.013	60.720	0.000
		12 -0.349	-0.020	71.078	0.000

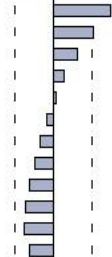
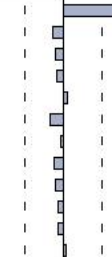
Autocorrelation and Partial Correlation of Hindustani

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 0.807	0.807	11.852	0.001
		2 0.607	-0.126	19.075	0.000
		3 0.423	-0.076	22.884	0.000
		4 0.240	-0.130	24.221	0.000
		5 0.068	-0.114	24.338	0.000
		6 -0.090	-0.121	24.570	0.000
		7 -0.207	-0.049	25.930	0.001
		8 -0.308	-0.122	29.391	0.000
		9 -0.390	-0.108	35.847	0.000
		10 -0.417	-0.013	44.695	0.000
		11 -0.410	-0.042	55.387	0.000
		12 -0.367	-0.004	66.810	0.000

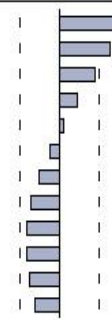
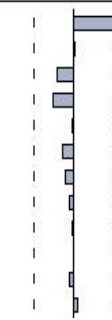
Autocorrelation and Partial Correlation of Arabic

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 0.816	0.816	12.127	0.000
		2 0.629	-0.110	19.886	0.000
		3 0.445	-0.104	24.104	0.000
		4 0.273	-0.094	25.828	0.000
		5 0.089	-0.171	26.028	0.000
		6 -0.114	-0.227	26.397	0.000
		7 -0.284	-0.117	28.960	0.000
		8 -0.390	-0.026	34.492	0.000
		9 -0.430	0.021	42.332	0.000
		10 -0.406	0.076	50.723	0.000
		11 -0.378	-0.070	59.839	0.000
		12 -0.336	-0.065	69.430	0.000

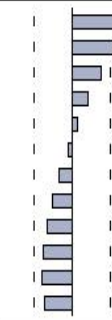
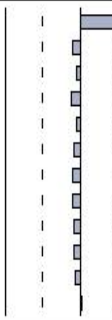
Autocorrelation and Partial Correlation of Russian

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 0.773	0.773	10.873	0.001
		2 0.537	-0.149	16.526	0.000
		3 0.319	-0.108	18.686	0.000
		4 0.132	-0.087	19.092	0.001
		5 0.031	0.052	19.116	0.002
		6 -0.092	-0.187	19.358	0.004
		7 -0.175	-0.033	20.334	0.005
		8 -0.256	-0.128	22.722	0.004
		9 -0.332	-0.110	27.417	0.001
		10 -0.370	-0.081	34.385	0.000
		11 -0.388	-0.073	44.005	0.000
		12 -0.335	0.038	53.537	0.000

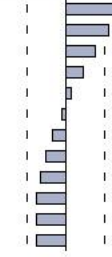
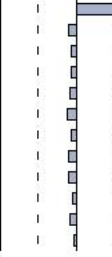
Autocorrelation and Partial Correlation of Portuguese

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 0.812	0.812	11.999	0.001
		2 0.663	0.011	20.613	0.000
		3 0.471	-0.205	25.325	0.000
		4 0.238	-0.278	26.638	0.000
		5 0.063	-0.023	26.740	0.000
		6 -0.127	-0.155	27.195	0.000
		7 -0.276	-0.098	29.618	0.000
		8 -0.375	-0.054	34.741	0.000
		9 -0.433	-0.017	42.717	0.000
		10 -0.432	0.003	52.213	0.000
		11 -0.408	-0.054	62.801	0.000
		12 -0.327	0.053	71.893	0.000

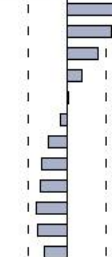

Autocorrelation and Partial Correlation of Spanish

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 0.785	0.785	11.232	0.001
		2 0.577	-0.103	17.769	0.000
		3 0.395	-0.065	21.080	0.000
		4 0.215	-0.124	22.149	0.000
		5 0.067	-0.060	22.262	0.000
		6 -0.061	-0.087	22.365	0.001
		7 -0.174	-0.101	23.326	0.001
		8 -0.271	-0.107	25.998	0.001
		9 -0.345	-0.092	31.054	0.000
		10 -0.395	-0.090	39.028	0.000
		11 -0.415	-0.066	50.013	0.000
		12 -0.381	0.013	62.361	0.000

Autocorrelation and Partial Correlation of Malay

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 0.782	0.782	11.146	0.001
		2 0.574	-0.099	17.599	0.000
		3 0.388	-0.071	20.803	0.000
		4 0.224	-0.079	21.963	0.000
		5 0.077	-0.086	22.115	0.000
		6 -0.063	-0.121	22.229	0.001
		7 -0.175	-0.073	23.200	0.002
		8 -0.271	-0.107	25.867	0.001
		9 -0.349	-0.106	31.039	0.000
		10 -0.388	-0.052	38.707	0.000
		11 -0.406	-0.082	49.216	0.000
		12 -0.386	-0.027	61.897	0.000

Autocorrelation and Partial Correlation of Bengali

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 0.764	0.764	10.634	0.001
		2 0.584	0.000	17.323	0.000
		3 0.416	-0.072	21.005	0.000
		4 0.190	-0.254	21.841	0.000
		5 0.022	-0.066	21.854	0.001
		6 -0.091	-0.012	22.090	0.001
		7 -0.248	-0.207	24.047	0.001
		8 -0.352	-0.108	28.555	0.000
		9 -0.363	0.052	34.155	0.000
		10 -0.410	-0.135	42.741	0.000
		11 -0.397	-0.037	52.808	0.000
		12 -0.298	0.087	60.333	0.000

Autocorrelation and Partial Correlation of FRE

Appendix 2

Data for the Short-term Model of Non-native Speaker in Model 1

	Language	Typical Country(Region)	GDP(per capita)	Crop Yield(per capita)	Average Years of School	Gini Coefficient	Gross National Happiness Index
2008	Madarin Chinese	China	2214	413	7.62	0.66	9424
	English	The U.S. , The UK	43063	8547	13.41	0.37	9524
	Spanish	Spain , Mexico	10891	2126	11.57	0.45	9527
	Arabic	Saudi Arabia	8766	1670	10.09	0.47	9523
	Russian	Russian Federation	4102	773	11.21	0.49	9494
	French	France , Canada	52478	10491	11.73	0.34	9555
2009	Madarin Chinese	China	3642	638	7.81	0.63	9554
	English	The U.S. , The UK	44385	8787	13.73	0.32	9467
	Spanish	Spain , Mexico	12462	2481	11.52	0.39	9541
	Arabic	Saudi Arabia	9525	1810	10.62	0.47	9455
	Russian	Russian Federation	5431	1007	11.02	0.42	9601
	French	France , Canada	61545	12271	12.63	0.39	9582
2010	Madarin Chinese	China	4361	811	7.93	0.61	9532
	English	The U.S. , The UK	47251	9440	14.12	0.38	9547
	Spanish	Spain , Mexico	14555	2852	11.12	0.47	9591
	Arabic	Saudi Arabia	10640	2075	10.14	0.57	9582
	Russian	Russian Federation	6665	1295	11.11	0.48	9446
	French	France , Canada	74389	14788	12.84	0.36	9471
2011	Madarin Chinese	China	7210	1391	8.01	0.55	9570
	English	The U.S. , The UK	49987	9992	14.56	0.33	9433
	Spanish	Spain , Mexico	16714	3273	11.6	0.35	9499
	Arabic	Saudi Arabia	10840	2093	10.25	0.42	9559
	Russian	Russian Federation	7426	1402	11.37	0.36	9589
	French	France , Canada	79654	15899	11.87	0.27	9587
2012	Madarin Chinese	China	7467	1399	8.09	0.49	9528
	English	The U.S. , The UK	52839	10518	14.98	0.34	9569
	Spanish	Spain , Mexico	18291	3523	11.99	0.34	9521
	Arabic	Saudi Arabia	11215	2228	10.44	0.34	9442
	Russian	Russian Federation	7864	1491	11.2	0.39	9578
	French	France , Canada	86455	17223	11.22	0.23	9618
2013	Madarin Chinese	China	7683	1466	8.15	0.44	9455
	English	The U.S. , The UK	56763	11309	15.25	0.31	9580
	Spanish	Spain , Mexico	19675	3954	11.16	0.34	9422
	Arabic	Saudi Arabia	12033	2374	10.17	0.4	9563
	Russian	Russian Federation	8193	1626	11.17	0.34	9470
	French	France , Canada	89543	17882	11.91	0.22	9609
2014	Madarin Chinese	China	8566	1698	8.22	0.41	9598
	English	The U.S. , The UK	58992	11723	15.75	0.28	9565
	Spanish	Spain , Mexico	21054	4129	11.08	0.34	9508
	Arabic	Saudi Arabia	12095	2347	10.17	0.39	9483
	Russian	Russian Federation	8542	1680	11.22	0.38	9422
	French	France , Canada	95356	19080	11.66	0.29	9499
2015	Madarin Chinese	China	8970	1715	8.26	0.42	9425
	English	The U.S. , The UK	61068	12190	15.77	0.22	9495
	Spanish	Spain , Mexico	23555	4693	11.85	0.25	9556
	Arabic	Saudi Arabia	13254	2580	10.23	0.27	9426
	Russian	Russian Federation	8861	1723	11.07	0.26	9487
	French	France , Canada	105955	21151	12.42	0.25	9462
2016	Madarin Chinese	China	9485	1893	8.32	0.38	9575
	English	The U.S. , The UK	64250	12792	15.91	0.29	9506
	Spanish	Spain , Mexico	25098	4957	11.87	0.39	9583
	Arabic	Saudi Arabia	13896	2766	10.69	0.44	9563
	Russian	Russian Federation	9088	1775	11.05	0.43	9615
	French	France , Canada	11543	2218	12.38	0.33	9453

Data Source: The Work Bank Database

	Language	Typical Country/Region	Number of Migrants	Labor Productivity	Consumer Price Index	Income of Tourism(per capita)	Amount of Transation softwares
2008	Madarin Chinese	China	575	5555	1.24	355	13
	English	The U.S , The UK	430	5708	1.13	301	985
	Spanish	Spain , Mexico	475	5915	1.27	355	90
	Arabic	Saudi Arabia	655	5806	1.29	489	25
	Russian	Russian Federation	491	5755	1.6	497	10
	French	France , Canada	677	5847	1.43	431	78
2009	Madarin Chinese	China	449	5548	1.93	319	18
	English	The U.S , The UK	547	5045	1.21	355	1033
	Spanish	Spain , Mexico	514	5052	1.13	357	85
	Arabic	Saudi Arabia	527	5876	1.9	435	92
	Russian	Russian Federation	454	5977	1.07	389	95
	French	France , Canada	585	5589	1.13	423	57
2010	Madarin Chinese	China	541	5453	1.45	483	21
	English	The U.S , The UK	511	5448	1.39	408	1048
	Spanish	Spain , Mexico	495	5032	1.31	494	7
	Arabic	Saudi Arabia	448	5634	1.45	458	42
	Russian	Russian Federation	512	5085	1.5	347	17
	French	France , Canada	599	5605	1.42	350	62
2011	Madarin Chinese	China	643	5444	1.08	459	25
	English	The U.S , The UK	450	5249	1.19	358	1095
	Spanish	Spain , Mexico	421	5245	1.92	385	56
	Arabic	Saudi Arabia	453	5850	1.87	452	78
	Russian	Russian Federation	612	5510	1.15	325	19
	French	France , Canada	682	5354	1.04	475	65
2012	Madarin Chinese	China	700	5042	1.92	344	33
	English	The U.S , The UK	645	5309	1.48	355	1126
	Spanish	Spain , Mexico	458	5924	1.52	313	48
	Arabic	Saudi Arabia	497	5655	1.98	335	59
	Russian	Russian Federation	529	5900	1.65	405	57
	French	France , Canada	510	5395	1.74	311	99
2013	Madarin Chinese	China	512	5718	1.5	452	39
	English	The U.S , The UK	637	5458	1.42	343	1134
	Spanish	Spain , Mexico	551	5045	1.93	474	61
	Arabic	Saudi Arabia	682	5570	1.35	382	23
	Russian	Russian Federation	484	5777	1.74	397	61
	French	France , Canada	560	5187	1.08	305	58
2014	Madarin Chinese	China	451	5378	1.82	359	42
	English	The U.S , The UK	677	5115	1.78	350	1154
	Spanish	Spain , Mexico	475	5792	1.98	318	12
	Arabic	Saudi Arabia	439	5202	1.28	338	65
	Russian	Russian Federation	525	5074	1.88	383	27
	French	France , Canada	607	5879	1.83	424	49
2015	Madarin Chinese	China	450	5245	1.82	443	47
	English	The U.S , The UK	698	5356	1.15	453	1157
	Spanish	Spain , Mexico	644	5885	1.27	332	75
	Arabic	Saudi Arabia	638	5489	1.53	340	73
	Russian	Russian Federation	453	5014	1.07	443	25
	French	France , Canada	633	5405	1.88	309	25
2016	Madarin Chinese	China	485	5358	1.69	404	52
	English	The U.S , The UK	484	5885	1.34	308	1171
	Spanish	Spain , Mexico	480	5425	1.59	458	14
	Arabic	Saudi Arabia	683	5552	1.35	325	72
	Russian	Russian Federation	522	5123	1.13	425	33
	French	France , Canada	628	5441	1.23	481	80

Table2: Formed Panel by factors influencing the number of non-native speakers in2008-2016

(Notes: The units of the ten variables are \$, \$, year, -, -, 10,000 people, \$10,000, %, \$, unit)

Data Source: The Work Bank Database

Appendix 3

Threshold Tables for Cointegration Test of EG Method

Engel-Granger cointegration critical values

	No constant or trend		Constant but no trend			Constant and trend		
	5%	10%	1%	5%	10%	1%	5%	10%
β_0			-3.9001	-3.3377	-3.0462	-4.3266	-3.7809	-3.4959
β_1			-10.534	-5.967	-4.069	-12.531	-9.421	-7.203
β_2			-30.03	-8.98	-5.73	-34.03	-15.06	-4.01
T								
25			-4.37	-3.59	-3.22	-5.00	-4.18	-3.79
50			-4.12	-3.46	-3.13	-4.65	-3.98	-3.64
75			-4.05	-3.42	-3.10	-4.54	-3.91	-3.59
100			-4.01	-3.40	-3.09	-4.49	-3.88	-3.57
200			-3.95	-3.37	-3.07	-4.41	-3.83	-3.53
∞			-3.90	-3.34	-3.05	-4.33	-3.78	-3.50

Appendix 4

Error Correction Terms of Each Language

	Mandarin Chinese	English	Spanish	Arabic	Russian	French
ECM2008	2.7	4.9	-2	-2.6	-0.9	-4.9
ECM2009	3	3.6	1.8	-2.4	0.5	0.4
ECM2010	-0.7	-3.9	1.3	-0.5	1.7	2.2
ECM2011	3	-2.8	3.3	-3	-1	3.5
ECM2012	-3.3	1.7	4.8	-4.2	-5	0.6
ECM2013	-4.5	3.9	-2.5	3.1	1.7	-4.3
ECM2014	-4.3	-3.8	-0.3	-2	-4	-2.7
ECM2015	2.8	5	2.1	0.1	1.6	2.4

Appendix 5

Solution of Prediction model

The predicted one-step result of formula (3) is $N_t(1) = \phi N_t$. The prediction error is $N_{t+1} - N_t(1) = \varepsilon_{t+1}$

and the error variance is σ^2

The predicted two-step result is $N_t(2) = \phi N_t(1)$. The prediction error is $N_{t+2} - N_t(2) = \phi \varepsilon_{t+1} + \varepsilon_{t+2}$

and the error variance is $(1 + \phi^2 + \phi^4)\sigma^2$

The predicted three-step result is $N_t(3) = \phi N_t(2)$. The prediction error is

$N_{t+3} - N_t(3) = \phi (\phi \varepsilon_{t+1} + \varepsilon_{t+2}) + \varepsilon_{t+3}$, and the error variance is $(1 + \phi^2 + \phi^4)\sigma^2$

Due to the Induction, the error variance of one-step prediction is

$$(1 + \phi^2 + \phi^4 + \dots + \phi^{2l})\sigma^2 = \frac{1 - \phi^{2l+2}}{1 - \phi^2}\sigma^2$$

When $\phi > 1$, the curve shows an exponential growth trend.

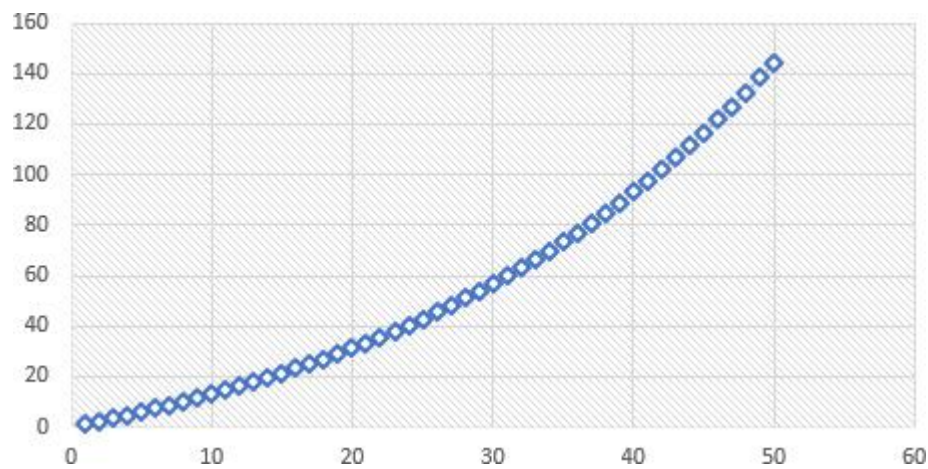


Figure 5: Analysis Simulation for Prediction Error of Marinda Chinese (Variance of Error Versus Prediction Step-length=1:50)

Appendix 6

Dickey-Fuller Test for Unit Root of Each Language

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-3.01506	0.0013	1	6
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-0.98766	0.1617	1	6
ADF - Fisher Chi-square	4.99241	0.0824	1	6
PP - Fisher Chi-square	12.3488	0.0021	1	7

Mandarin Chinese

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-2.13839	0.0162	1	7
Breitung t-stat	-0.55180	0.2905	1	6
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-0.12399	0.4507	1	7
ADF - Fisher Chi-square	2.62124	0.2697	1	7
PP - Fisher Chi-square	3.67956	0.1589	1	7

English

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-3.30737	0.0005	1	7
Breitung t-stat	-1.27200	0.1017	1	6
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-0.36145	0.3589	1	7
ADF - Fisher Chi-square	3.99704	0.1355	1	7
PP - Fisher Chi-square	7.96535	0.0186	1	7

Spanish

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-2.71761	0.0033	1	7
Breitung t-stat	-1.67909	0.0466	1	6
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-0.82871	0.2036	1	7
ADF - Fisher Chi-square	6.60071	0.0369	1	7
PP - Fisher Chi-square	12.6712	0.0018	1	7

Arabia

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-2.53077	0.0057	1	7
Breitung t-stat	-0.63201	0.2637	1	6
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-0.17826	0.4293	1	7
ADF - Fisher Chi-square	2.88458	0.2364	1	7
PP - Fisher Chi-square	5.80578	0.0549	1	7

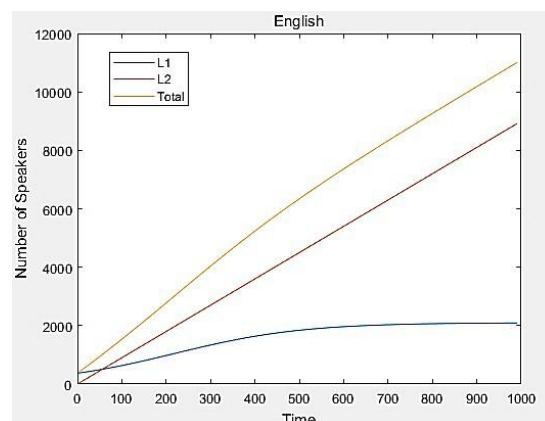
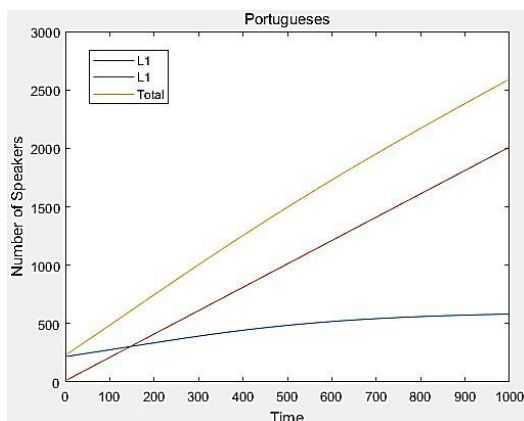
Russian

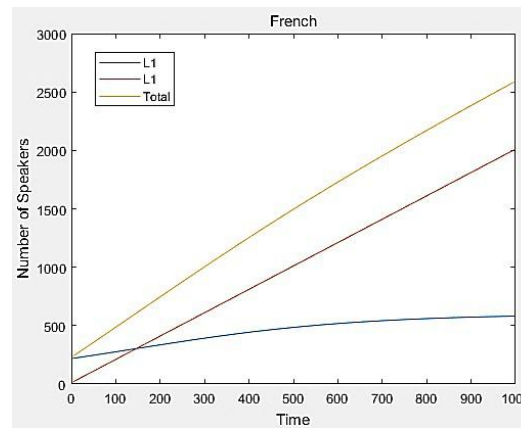
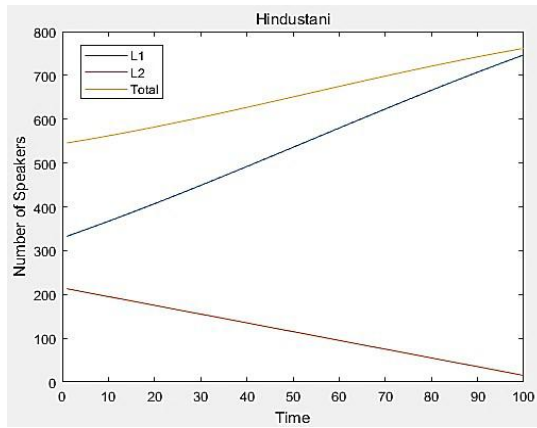
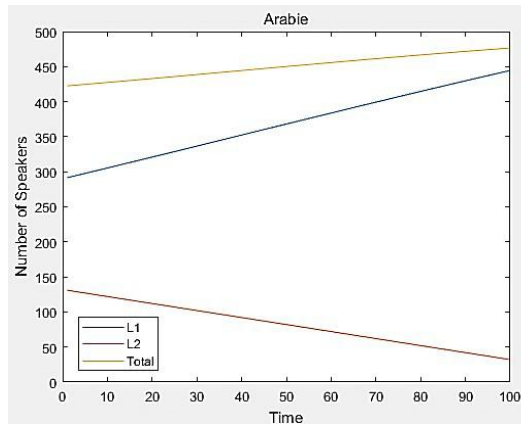
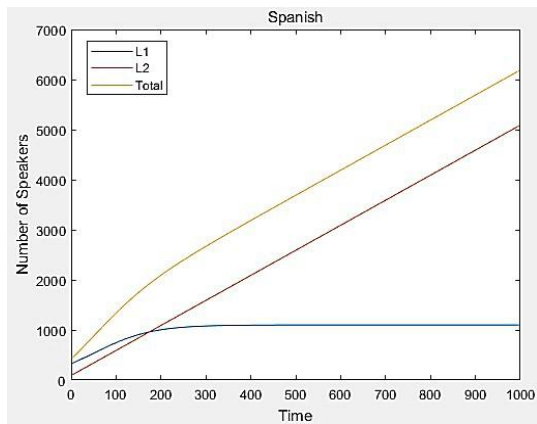
Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-1.79931	0.0360	1	7
Breitung t-stat	-0.40939	0.3411	1	6
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	0.03607	0.5144	1	7
ADF - Fisher Chi-square	1.77352	0.4120	1	7
PP - Fisher Chi-square	1.67350	0.4331	1	7

French

Appendix 7

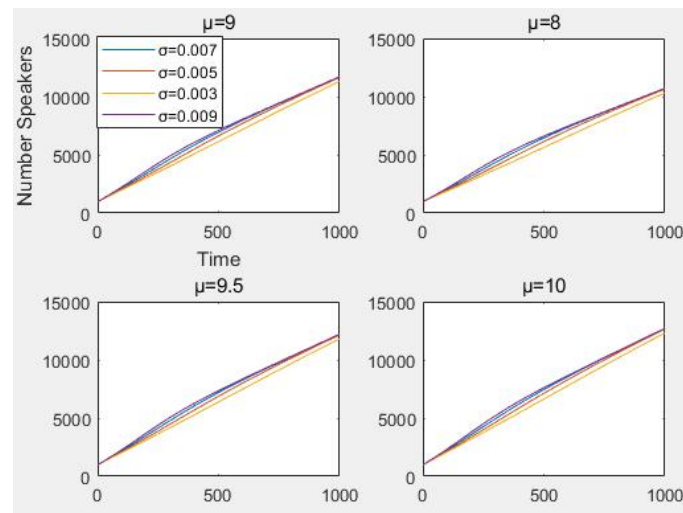
Number of Speakers Versus Time for Each Language



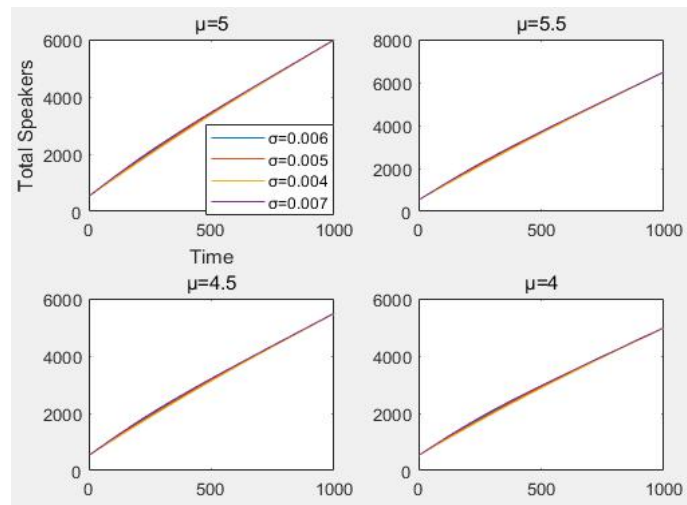


Appendix 8

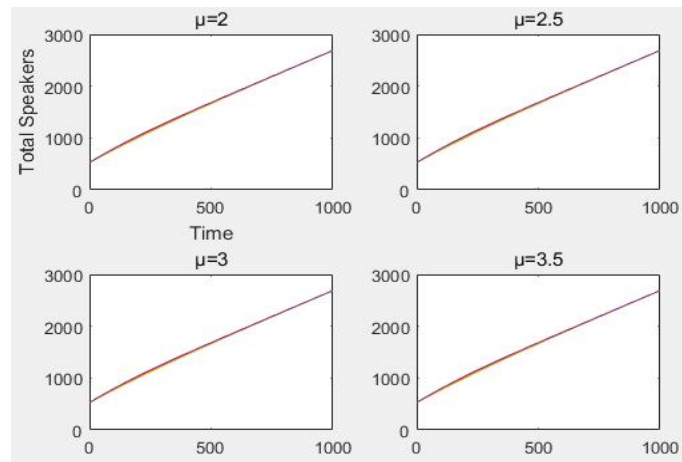
Sensitivity Analysis Results of Each Language



English



Spanish



Portuguese

Appendix 9

Part of Matrix Formed by the Number of Migrant in Different Countries

Origin Countries	State Code	AUS	NZL	ASM	COK	FJI	PYF	GUM	KIR	MHL	FSM	NRU	NCL	NFK	MNP
Australia	AUS	0	56173	6	129	1521	14	5	9	141	0	67	23	776	0
New Zealand	NZL	355765	0	230	1028	1410	18	6	43	1	0	327	110	481	0
American Samoa	ASM	154	402	0	31	142	1	3	135	0	0	3	1	0	6
Cook Islands	COK	4742	15224	20	0	62	1	5	47	0	0	18	6	1	4
Fiji	FJI	44261	25733	153	114	0	5	33	9	1	0	67	2534	2	24
French Polynesia	PYF	347	465	3	3	638	0	1	584	0	0	3	1085	0	1
Guam	GUM	50	33	71	68	512	3	0	0	1	0	2	1	0	1590
Kiribati	KIR	407	507	3	3	265	0	1	0	0	0	2	1	0	1
Marshall Islands	MHL	18	16	9	8	167	0	251	0	0	0	1	0	0	132
Micronesia, Federate	FSM	10	29	20	18	392	1	6983	1	0	0	9	3	0	2197
Nauru	NRU	465	219	1	1	36	0	0	0	0	0	0	0	0	0
New Caledonia	NCL	1074	171	2	1	39	180	0	0	0	0	1	0	0	0
Norfolk Island	NFK	199	117	0	0	0	0	0	0	0	0	0	0	0	0
Northern Mariana Isla	MNP	7	4	8	8	96	0	2183	0	0	0	2	1	0	0
Niue	NIU	494	5328	6	5	7	0	1	5	0	0	5	2	0	1
Palau	PLW	18	10	9	9	58	0	1334	0	0	0	2	1	0	1308
Papua New Guinea	PNG	23616	1153	45	40	873	2	11	5	0	0	37	14358	1	8
Samoa	WSM	13254	47126	17712	84	585	4	24	420	1	0	72	24	2	17
Solomon Islands	SLB	1326	507	4	3	73	0	1	0	0	0	3	1323	0	1
Tokelau	TKL	262	1662	27	2	4	0	1	5	0	0	2	1	0	0
Tonga	TON	7693	18058	1146	40	331	2	12	240	0	0	25	8	1	8
Tuvalu	TUV	97	1017	1	1	30	0	0	26	0	0	1	0	0	0
Vanuatu	VUT	898	276	4	3	35	0	1	0	0	0	2	471	0	1
Wallis and Futuna	WLF	18	16	6	5	47	2402	2	35	0	0	6	3728	0	1
China	CHN	142780	38951	167	11	218	193	2711	24	31	120	182	80	6	15583
Hong Kong	HKG	67122	11301	7	1	27	24	25	9	6	15	71	26	2	22
Japan	JPN	25471	8622	4	2	34	30	2454	4	75	18	35	14	1	892
Korea, Republic of	KOR	38900	17934	183	3	54	48	3250	7	13	30	56	24	2	1797
Taiwan	TWN	22418	12486	187	1	20	18	18	4	5	11	32	12	1	16
Macau	MAC	1948	249	0	0	4	4	4	0	1	2	2	1	0	3
Mongolia	MNG	125	15	0	0	0	0	0	0	0	0	0	0	0	0
Korea, Democratic P	PRK	55	12	1	1	17	15	16	0	4	9	0	2	0	13
Indonesia	IDN	47158	3792	6	3	63	56	57	6	15	34	48	1050	2	49
Malaysia	MYS	78858	11460	8	1	26	23	23	11	6	14	82	30	3	20
Philippines	PHL	103942	10135	664	6	126	111	32625	20	6	69	150	61	5	15822
Singapore	SGP	33485	3909	3	0	9	8	9	4	2	5	34	12	1	7
Thailand	THA	23600	5154	3	1	29	26	26	3	7	16	26	11	1	23
Viet Nam	VNM	154831	3946	204	4	76	67	68	19	17	41	144	592	5	59
Brunei Darussalam	BRN	2069	246	0	0	1	1	1	0	0	0	2	1	0	0
Cambodia	KHM	22979	4770	3	1	12	10	10	3	3	6	25	9	1	9
Lao People's Democr	LAO	9565	1017	1	1	13	12	12	1	3	7	10	4	0	10
Virgin Islands, British	VGB	7	9	0	0	0	0	0	0	0	0	0	0	0	0
Austria	AUT	19313	1202	2	4	3	15	5	5	4	0	19	125	1	1
Belgium	BEL	4900	515	0	4	3	15	5	4	4	0	5	118	0	1
Denmark	DNK	9024	1435	1	2	2	8	3	3	2	0	10	64	0	1
Finland	FIN	8258	373	1	3	3	11	4	3	3	0	8	93	0	1
France	FRA	17272	1629	3	14	13	23183	20	16	15	0	39	457	1	4
Germany	DEU	108220	8398	9	26	32	134	48	41	35	0	107	1111	3	11
United Kingdom	GBR	1036245	218410	99	110	884	140	50	177	37	0	1137	1505	84	11
Greece	GRC	116431	945	9	7	7	30	11	20	8	0	106	276	3	2
Ireland	IRL	50235	6730	4	8	7	31	11	13	8	0	52	266	2	2
Italy	ITA	218718	1455	17	26	25	106	38	48	28	0	200	919	6	8
Luxembourg	LUX	141	30	0	0	0	1	1	0	0	0	0	12	0	0
Netherlands	NLD	83324	22242	8	6	6	25	9	18	7	0	96	236	3	2
Portugal	PRT	15441	148	1	15	15	63	22	15	16	0	15	506	0	5
Spain	ESP	12662	344	1	11	10	44	16	10	11	0	12	353	0	3
Sweden	SWE	6818	961	1	2	2	9	3	3	2	0	7	78	0	1
Switzerland	CHE	10753	2765	1	3	3	14	5	4	4	0	12	113	0	1
Iceland	ISL	463	84	0	0	0	1	0	0	0	0	1	9	0	0
Liechtenstein	LIE	11	0	0	0	0	0	0	0	0	0	0	1	0	0
Norway	NOR	4324	466	0	1	1	6	2	2	2	0	4	48	0	0
Andorra	AND	11	0	0	0	0	0	0	0	0	0	0	2	0	0
Bosnia and Herzegov	BIH	23848	464	2	10	10	41	15	11	11	0	22	338	1	3
Faroe Islands	FRO	24	9	0	0	0	0	0	0	0	0	0	0	0	0
Gibraltar	GIB	416	60	0	0	0	1	0	0	0	0	0	5	0	0
Macedonia, the form	MKD	43527	592	3	2	2	8	3	7	2	0	40	79	1	1
Monaco	MCO	40	0	0	0	0	1	0	0	0	0	0	5	0	0
San Marino	SMR	4	0	0	0	0	0	0	0	0	0	0	2	0	0
Serbia and Monteneg	SCG	55365	2622	5	13	12	53	19	18	14	0	53	443	2	4
Albania	ALB	1451	66	0	6	6	25	9	5	7	0	2	201	0	2
Bulgaria	BGR	2571	507	0	7	6	28	10	6	7	0	3	222	0	2
Croatia	HRV	51909	2282	4	5	4	19	7	10	5	0	49	170	2	1
Cyprus	CYP	19482	300	2	0	6	6	6	2	1	3	18	7	1	5
Czech Republic	CZE	6973	664	1	3	3	12	4	3	3	0	7	98	0	1
Hungary	HUN	22752	989	2	3	3	13	5	6	3	0	22	114	1	1
Malta	MLT	46998	366	4	1	1	4	1	6	1	0	43	44	1	0
Poland	POL	58110	1946	5	16	16	66	24	21	17	0	55	550	2	5
Romania	ROM	12821	922	1	8	8	34	12	9	9	0	13	274	0	3
Slovakia	SVK	2984	140	0	4	4	16	6	4	4	0	3	131	0	1
Slovenia	SVN	6685	174	1	1	1	4	1	2	1	0	6	30	0	0
Estonia	EST	2389	100	0	1	1	6	2	1	2	0	2	47	0	0
Latvia	LVA	6688	247	1	2	2	7	3	2	2	0	6	61	0	1
Lithuania	LTU	3687	109	0	3	2	10	4	3	3	0	4	85	0	1
Republic of Moldova	MDA	15024	2061	2	104	00	416	140	00	100	0	21	2227	1	22

Data Source: The Global Migrant Origin Database

Appendix 10

Part of One-Step Transition Probability Matrix

Origin Countries	State Code	AUS	NZL	ASM	COK	FJI	PYF	GUM	KIR	MHL	FSM	NRU
Australia	AUS	0.98162	0.00236	2.52E-07	5.42E-06	6.39E-05	5.88E-07	2.10E-07	3.78E-07	5.93E-06		0.282E-06
New Zealand	NZL	0.07741	0.88498	5.00E-05	0.00022	0.00031	3.92E-06	1.31E-06	9.36E-06	2.18E-07		0.712E-05
American Samoa	ASM	0.00277	0.00724	0.26653	0.00056	0.00256	1.80E-05	5.40E-05	0.00243	0		0.540E-05
Cook Islands	COK	0.0335	0.60062	0.0009	0.24441	0.0028	4.52E-05	0.00023	0.00212	0		0.00081
Fiji	FJI	0.04961	0.02884	0.00017	0.00013	0.83963	5.60E-06	3.70E-05	1.01E-05	1.12E-06		0.751E-05
French Polynesia	PYF	0.00125	0.00167	1.08E-05	1.08E-05	0.0023	0.98736	3.60E-06	0.0021	0		0.108E-05
Guam	GUM	0.00031	0.0002	0.00044	0.00042	0.00316	1.85E-05	0.4464		6.18E-06		0.124E-05
Kiribati	KIR	0.00362	0.00451	2.67E-05	2.67E-05	0.00236		8.90E-06	0.96731	0		0.178E-05
Marshall Islands	MHL	0.00034	0.0003	0.00017	0.00015	0.00315		0.00474		0.78318		0.189E-05
Micronesia, Federated States of	FSM	9.58E-05	0.00028	0.00019	0.00017	0.00375	9.58E-06	0.06687	9.58E-06	0	0.76433	8.62E-05
Nauru	NRU	0.03727	0.01756	8.02E-05	8.02E-05	0.00289		0	0	0		0.91776
New Caledonia	NCL	0.00393	0.00063	7.33E-06	3.66E-06	0.00014	0.00066		0	0		0.366E-06
Norfolk Island	NFK	0.09352	0.05498	0	0	0	0		0	0		0
Northern Mariana Islands	MNP	0.00013	7.30E-05	0.00015	0.00015	0.00175		0.03982		0		0.365E-05
Niue	NIU	0.10984	0.25019	0.00458	0.00381	0.00534		0.00076	0.00381	0		0.00381
Palau	PLW	0.00085	0.00047	0.00042	0.00042	0.00272		0.06266		0		0.939E-05
Papua New Guinea	PNG	0.00298	0.00015	5.68E-06	5.05E-06	0.00011	2.53E-07	1.39E-06	6.31E-07	0		0.467E-06
Samoa	WSM	0.0684	0.24322	0.09141	0.00043	0.00302	2.06E-05	0.00012	0.00217	5.16E-06		0.00037
Solomon Islands	SLB	0.00226	0.00086	6.81E-06	5.11E-06	0.00012		0.170E-06		0		0.511E-06
Tokelau	TKL	0.1467	0.37066	0.01512	0.00112	0.00224		0.00056	0.0028	0		0.00112
Tonga	TON	0.07233	0.16978	0.01077	0.00038	0.00311	1.88E-05	0.00011	0.00226	0		0.00024
Tuvalu	TUV	0.00882	0.09245	9.09E-05	9.09E-05	0.00273		0	0.00236	0		0.909E-05
Vanuatu	VUT	0.00339	0.00104	1.51E-05	1.13E-05	0.00013		0.378E-06		0		0.756E-06
Wallis and Futuna	WLF	0.00116	0.00103	0.00039	0.00032	0.00304	0.15517	0.00013	0.00226	0		0.00039
China	CHN	0.0001	2.84E-05	1.22E-07	8.02E-09	1.59E-07	1.41E-07	1.98E-06	1.75E-08	2.26E-08	8.75E-08	1.33E-07
Hong Kong	HKG	0.00919	0.00155	9.58E-07	1.37E-07	3.70E-06	3.29E-06	3.42E-06	1.23E-06	8.21E-07	2.05E-06	9.72E-06
Japan	JPN	0.0002	6.78E-05	3.15E-08	1.57E-08	2.67E-07	2.36E-07	1.93E-05	3.15E-08	5.90E-07	1.42E-07	2.75E-07
Korea, Republic of	KOR	0.00076	0.00035	3.59E-06	5.88E-08	1.06E-06	9.41E-07	6.37E-05	1.37E-07	2.55E-07	5.88E-07	1.10E-06
Taiwan	TWN	0.00095	0.00053	7.96E-06	4.26E-08	8.51E-07	7.66E-07	7.66E-07	1.70E-07	2.13E-07	4.68E-07	1.36E-06
Macau	MAC	0.00324	0.00041	0	0	0.666E-06	6.66E-06	6.66E-06		0.166E-06	3.33E-06	3.33E-06
Mongolia	MNG	4.20E-05	5.04E-06	0	0	0	0	0	0	0	0	0
Korea, Democratic People's Republic	PRK	2.18E-06	4.75E-07	3.96E-08	3.96E-08	6.73E-07	5.94E-07	6.34E-07		0.158E-07	3.57E-07	0
Indonesia	IDN	0.00018	1.47E-05	2.32E-08	1.16E-08	2.44E-07	2.17E-07	2.21E-07	2.32E-08	5.81E-08	1.32E-07	1.86E-07
Malaysia	MYS	0.00257	0.00037	2.60E-07	3.25E-08	8.46E-07	7.49E-07	7.49E-07	3.58E-07	1.95E-07	4.56E-07	2.67E-06
Philippines	PHL	0.00102	9.96E-05	6.53E-06	5.90E-08	1.24E-06	1.09E-06	0.00032	1.97E-07	5.90E-08	6.78E-07	1.47E-06
Singapore	SGP	0.00605	0.00071	5.42E-07		0.163E-06	1.45E-06	1.63E-06	7.23E-07	3.61E-07	9.03E-07	6.14E-06
Thailand	THA	0.00034	7.51E-05	4.37E-08	1.46E-08	4.22E-07	3.79E-07	3.79E-07	4.37E-08	1.02E-07	2.33E-07	3.79E-07
Viet Nam	VNM	0.00169	4.30E-05	2.22E-06	4.36E-08	8.29E-07	7.31E-07	7.41E-07	2.07E-07	1.85E-07	4.47E-07	1.57E-06
Brunei Darussalam	BRN	0.00496	0.00059	0	0	2.39E-06	2.39E-06	2.39E-06		0	0	0.479E-06
Cambodia	KHM	0.00148	0.00031	1.93E-07	6.44E-08	7.73E-07	6.44E-07	6.44E-07	1.93E-07	1.93E-07	3.87E-07	1.61E-06
Lao People's Democratic Republic	LAO	0.00144	0.00015	1.50E-07	1.50E-07	1.95E-06	1.80E-06	1.80E-06	1.50E-07	4.50E-07	1.05E-06	1.50E-06
Myanmar	MMR	0.00021	1.34E-05	1.91E-08	1.91E-08	1.91E-07	1.72E-07	1.72E-07	1.91E-08	3.82E-08	9.54E-08	2.10E-07
Timor Leste	TLS	0.00757	2.42E-05	8.06E-07		0.806E-07	8.06E-07	8.06E-07	8.06E-07		0	0.725E-06
Bangladesh	BGD	5.63E-05	7.36E-06	4.34E-08	6.20E-08	3.07E-06	1.09E-06	1.11E-06	6.20E-09	2.85E-07	6.70E-07	6.82E-08
India	IND	7.29E-05	1.60E-05	1.45E-08	1.15E-08	2.89E-06	2.05E-07	2.10E-07	1.07E-08	5.35E-08	1.28E-07	8.40E-08
Sri Lanka	LKA	0.00255	0.00029	2.86E-07	9.54E-08	1.48E-06	1.29E-06	1.34E-06	3.34E-07	3.34E-07	8.11E-07	2.58E-06
Afghanistan	AFG	0.00033	2.19E-05	1.19E-07	1.19E-07	2.58E-06	2.28E-06	2.31E-06	2.96E-08	5.93E-07	1.39E-06	3.26E-07
Bhutan	BTN	8.13E-05	1.14E-05	0	0	0	0	0	0	0	0	0
Maldives	MDV	0.00043	8.80E-05	0	0	0	0	0	0	0	0	0
Nepal	NPL	9.16E-05	1.20E-05	3.49E-08	6.98E-08	1.05E-06	9.42E-07	9.42E-07		0.244E-07	5.93E-07	1.05E-07
Pakistan	PAK	6.29E-05	6.98E-06	2.64E-08	3.17E-08	2.62E-06	5.65E-07	5.86E-07	1.06E-08	1.48E-07	3.59E-07	6.86E-08
Canada	CAN	0.00076	0.00022	8.37E-08	1.39E-06	1.90E-06	1.23E-06	4.46E-07	1.12E-07	3.35E-07	2.20E-06	8.93E-07
United States of America	USA	0.00017	4.16E-05	1.12E-05	2.40E-07	3.49E-07	2.24E-07	5.95E-05	3.43E-08	9.35E-09	3.63E-06	2.68E-07
Mexico	MEX	9.17E-06	1.93E-06	0	0	0.434E-06	2.81E-06	1.00E-06		0.739E-07	5.02E-06	2.38E-08
Bermuda	BMU	0.0056	0.00253	0	0	0.153E-05	1.53E-05		0	0	0.153E-05	0
Greenland	GRL	0.00073	5.35E-05	0	0	0	0	0	0	0	0	0
Saint Pierre and Miquelon	SPM	0	0	0	0	0	0	0	0	0	0	0
Colombia	COL	8.98E-05	4.83E-06	0	0	0.124E-07	1.16E-06	4.15E-07	2.07E-08	3.11E-07		0.829E-08
Peru	PER	0.00018	1.32E-05	0	0	0.956E-08	7.97E-07	2.87E-07	3.19E-08	1.91E-07		0.159E-07
Venezuela	VEN	3.56E-05	3.40E-06	0	0	0.321E-08	3.85E-07	1.28E-07		0.963E-08		0.321E-08
Bolivia	BOL	6.13E-05	9.60E-06	0	0	0.186E-07	1.21E-06	4.66E-07		0.280E-07		0.932E-08
Ecuador	ECU	8.21E-05	3.59E-06	0	0	0.186E-07	1.42E-06	4.96E-07		0.372E-07		0.619E-08
Argentina	ARG	0.00025	8.98E-06	2.30E-08		0.461E-08	4.84E-07	1.61E-07	2.30E-08	1.15E-07		0.230E-07
Brazil	BRA	2.29E-05	3.23E-06	0	0	0.194E-08	1.46E-07	5.34E-08	4.86E-09	3.88E-08		0.243E-08

Appendix 11

Part of Population of countries in decades

Origin Countries	N1	N10	N20	N50
Australia	27810051	67567820	113896811	248354923
New Zealand	4852473	6502407	8488640	14451496
American Samoa	39649	12593	15949	24041
Cook Islands	5978	7039	14646	57465
Fiji	770635	258521	158226	202442
French Polynesia	308266	630757	1063082	2439136
Guam	147281	151672	171295	226775
Kiribati	113097	119197	129428	208355
Marshall Islands	43169	11430	8072	9086
Micronesia, Federated States of	83335	21596	18480	26129
Nauru	16732	65269	137126	432209
New Caledonia	317744	750983	1297600	3307790
Norfolk Island	3218	12034	20613	42908
Northern Mariana Islands	88798	199043	214060	223765
Niue	203	115	108	108
Palau	14926	11560	12267	14367
Papua New Guinea	8056356	9777543	11618479	17847944
Samoa	94677	10424	13653	22505
Solomon Islands	598668	732139	863487	1233140
Tokelau	479	128	170	395
Tonga	57711	2842	2504	2357
Tuvalu	9641	3663	2403	3309
Vanuatu	267503	308122	347144	475133
Wallis and Futuna	10177	3101	2590	1986
China	1373328546	1395031690	1397583980	1342538404
Hong Kong	9346775	21305096	26553158	29040544
Japan	127404190	128781763	129779753	129998799
Korea, Republic of	49898660	41536594	33699857	18904722
Taiwan	24468537	38481381	52992777	89584611
Macau	515348	231378	209418	253002
Mongolia	3027950	3640373	4206935	5436547
Korea, Democratic People's Repu	24919516	22358728	19538354	12649865
Indonesia	259663204	279058836	291211385	298995502
Malaysia	32155860	46765899	61206350	92150294
Philippines	101190568	100509247	98928813	101535666
Singapore	5895828	9621373	13419532	21807341

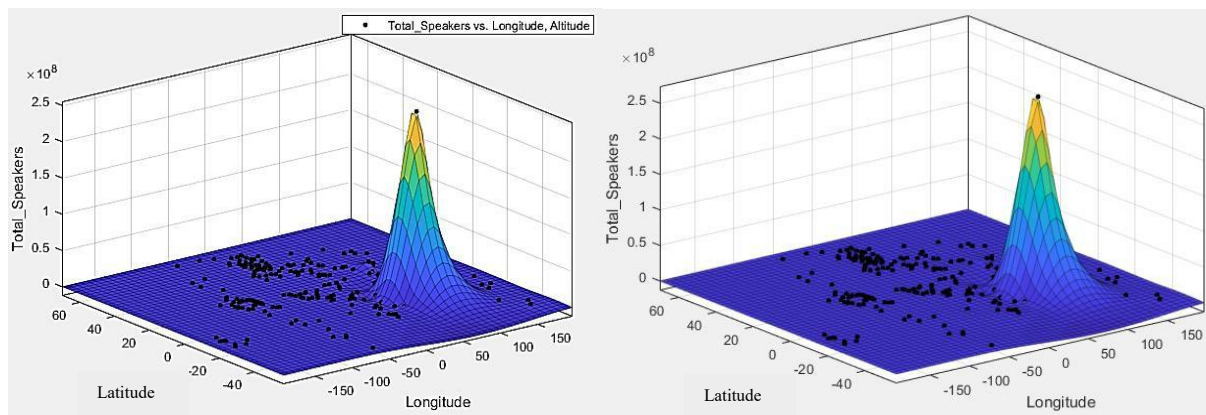
Part of Latitude and Longitude of Coordinates of Various National Centers

Origin Countries	Latitude	Longitude	Origin countries	Latitude	Longitude	Origin Countries	Latitude	Longitude	Origin Co	Altitude	Longitude	Origin Countries	Latitude	Longitude
Australia	35°15'S	149°08'E	Singapore	1°3'N	103.8°E	Denmark	55°41'N	12°34'E	Finland	60°15'N	25°03'E	Estonia	59°22'N	24°48'E
New Zealand	41°19'S	174°46'E	Thailand	13°45'N	100°35'E	Suriname	05°50'N	55°10'W	France	48°50'N	02°20'E	Latvia	56°53'N	24°08'E
American Samoa	14°16'S	170°43'W	Viet Nam	21°05'N	105°55'E	Belize	17°18'N	88°30'W	Germany	52°30'N	13°25'E	Lithuania	54°38'N	25°19'E
Cook Islands	21°14'S	159°46'W	Brunei Darussalam	04°52'N	115°00'E	Costa Rica	09°55'N	84°02'W	United Kingdom	55°15'N	3°W	Russian Federation	55°45'N	37°35'E
Fiji	18°06'S	178°30'E	Cambodia	11°33'N	104°55'E	El Salvador	13°40'N	89°10'W	Greece	37°58'N	23°46'E	Armenia	40°10'N	44°31'E
French Polynesia	17°32'S	149°34'W	Laos People's Democr	17°38'N	102°36'E	Guatemala	14°40'N	90°22'W	Ireland	53°21'N	06°15'W	Azerbaijan	40°29'N	49°56'E
Guam	13°30'N	144°48'E	Myanmar	16°45'N	96°20'E	Honduras	14°05'N	87°14'W	Italy	41°54'N	12°29'E	Belarus	53°52'N	27°30'E
Kiribati	01°30'N	173°00'E	Timor Leste	8°34'S	125°34'E	Nicaragua	12°06'N	86°20'W	Luxembourg	49°37'N	06°09'E	Georgia	41°43'N	44°50'E
Marshall Islands	9°N	168°E	Bangladesh	23°43'N	90°26'E	Panama	09°00'N	79°25'W	Netherlands	52°23'N	04°54'E	Kazakhstan	51°10'N	71°30'E
Micronesia, Federated States of	6°55'N	158°15'E	India	28°37'N	77°13'E	Antigua & Barbuda	17°03'N	61°48'W	Portugal	38°42'N	09°10'W	Kyrgyzstan	42°54'N	74°48'E
Nauru	0°32'S	166°56'E	Sri Lanka	7°N	81°E	Bahamas	25°05'N	77°20'W	Spain	40°25'N	03°45'W	Moldova, Republic of	47°02'N	28°50'E
New Caledonia	22°17'S	166°30'E	Afghanistan	34°28'N	69°11'E	Barbados	13°05'N	59°30'W	Sweden	59°20'N	18°03'E	Tajikistan	38°33'N	68°48'E
Norfolk Island	45°20'S	168°43'E	Bhutan	27°31'N	89°45'E	Dominica	15°20'N	61°24'W	Switzerland	46°57'N	07°28'E	Turkmenistan	38°00'N	57°50'E
Northern Mariana Islands	15°12'N	145°45'E	Maldives	04°00'N	73°28'E	Dominican Republic	19°N	70°40'W	Iceland	64°10'N	21°57'W	Ukraine	50°30'N	30°28'E
Niue	19°03'S	169°55'W	Nepal	27°45'N	85°20'E	Grenada	12°07'N	61°40'W	Liechtenstein	47°08'N	09°31'E	Uzbekistan	41°20'N	69°10'E
Palau	07°20'N	134°28'E	Pakistan	33°40'N	73°10'E	Haiti	18°40'N	72°20'W	Norway	59°55'N	10°45'E	Turkey	39°57'N	32°54'E
Papua New Guinea	09°24'S	147°08'E	Canada	45°27'N	75°42'W	Jamaica	18°00'N	76°50'W	Andorra	42°31'N	01°32'E	Bahrain	26°10'N	50°30'E
Samoa	13°50'S	171°50'W	United States of America	39°91'N	77°02'W	Puerto Rico	18°28'N	66°07'W	Bosnia and Herzegovina	43°52'N	18°25'E	Iran, Islamic Republic of	32°20'N	53°30'E
Solomon Islands	09°27'S	159°57'E	Mexico	19°10'N	99°10'W	Saint Kitts and Nevis	17°17'N	62°43'W	Faroe Islands	62°05'N	06°56'W	Iraq	33°00'N	44°00'E
Tokelau	9°10'S	171°50'W	Bermuda	32°20'N	64°45'W	Saint Lucia	14°02'N	60°58'W	Gibraltar	36°8'N	5°21'W	Israel	31°47'N	35°12'E
Tonga	21°10'S	174°00'W	Greenland	64°10'N	51°35'W	Saint Vincent and the Grenadines	13°10'N	61°10'W	Macedonia	41°6'N	21°E	Jordan	31°57'N	35°52'E
Tuvalu	08°31'S	179°13'E	Saint Pierre and Miquelon	46°46'N	56°12'W	Trinidad and Tobago	11°25'N	60°55'W	Monaco	43°44'N	7°25'E	Kuwait	29°30'N	48°00'E
Vanuatu	17°45'S	168°18'E	Colombia	04°34'N	74°00'W	Virgin Islands, U.S.	18°21'N	64°56'W	San Marino	43°55'N	12°30'E	Lebanon	33°53'N	35°31'E
Wallis and Futuna	13°18'S	176°12'W	Peru	12°05'N	77°00'W	Anguilla	18°23'N	63°05'W	Serbia and Montenegro	44°49'N	20°28'E	Palestine	31°53'N	35°12'E
China	39°55'N	116°20'E	Venezuela	10°30'N	66°55'W	Aruba	12°32'N	70°02'W	Albania	41°18'N	19°49'E	Oman	23°37'N	58°36'E
Hong Kong	22°38'N	114°2'E	Bolivia	16°20'S	68°10'W	Cayman Islands	19°20'N	81°24'W	Bulgaria	42°45'N	23°20'E	Qatar	25°15'N	51°35'E
Japan	35°N	136°E	Ecuador	00°15'S	78°35'W	Cuba	23°08'N	82°22'W	Croatia	45°50'N	15°58'E	Saudi Arabia	24°41'N	46°42'E
Korea, Republic of	37°31'N	126°58'E	Argentina	36°30'S	60°00'W	Guadeloupe	16°00'N	61°44'W	Cyprus	35°10'N	33°25'E	Yriani Arab Republic	33°30'N	36°18'E
Taiwan	25°02'N	121°38'E	Brazil	15°47'S	47°55'W	Martinique	14°36'N	61°02'W	Czech Rep	50°05'N	14°22'E	United Arab Emirates	24°28'N	54°22'E
Macau	22°10'N	113°33'E	Chile	33°24'S	70°40'W	Montserrat	16°45'N	62°12'E	Hungary	47°29'N	19°05'E	Yemen	15°N	48°E
Mongolia	47°55'N	106°53'E	Uruguay	34°50'S	56°11'W	Netherlands Antille	12°05'N	69°00'W	Malta	35°54'N	14°31'E	Morocco	32°N	6°W
Korea, Democratic People's Republic of	40°N	127°E	Falkland Islands (Malvinas)	51°45'S	59°51'W	Turks and Caicos	21°45'N	71°35'W	Poland	52°13'N	21°00'E	Tunisia	36°50'N	10°11'E
Indonesia	06°09'S	106°49'E	French Guiana	05°05'N	52°18'W	Virgin Islands, British	18°30'N	64°30'W	Romania	44°27'N	26°10'E	Algeria	36°42'N	03°08'E
Malaysia	03°09'N	101°41'E	Guyana	06°50'N	58°12'W	Austria	48°12'N	16°22'E	Slovakia	48°10'N	17°07'E	Egypt	30°01'N	31°14'E
Philippines	14°40'N	121°03'E	Paraguay	25°10'S	57°30'W	Belgium	50°51'N	04°21'E	Slovenia	46°04'N	14°33'E	Libyan Arab Jamahiri	32°49'N	13°07'E

Data Source: The Wikipedia

Appendix 13

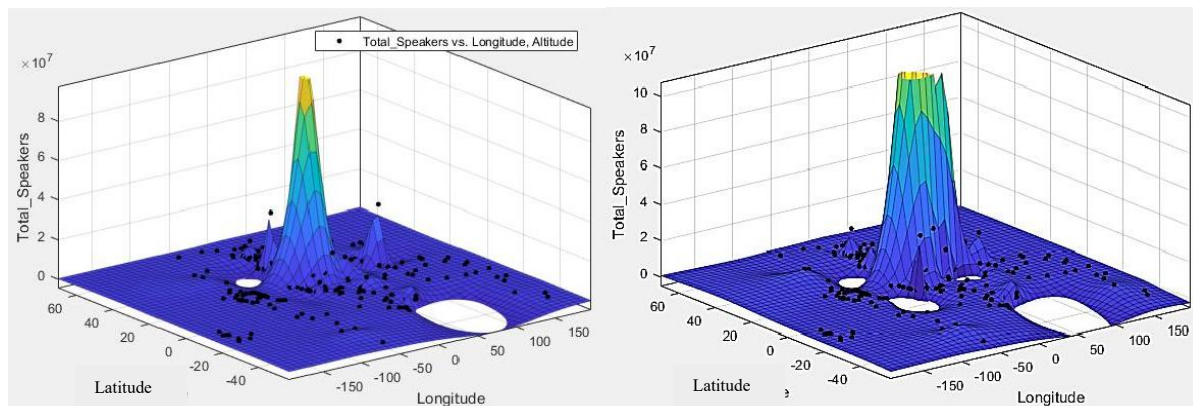
3D Image of the Number of Particular Language Speakers – Geographic Distribution



2017

2067

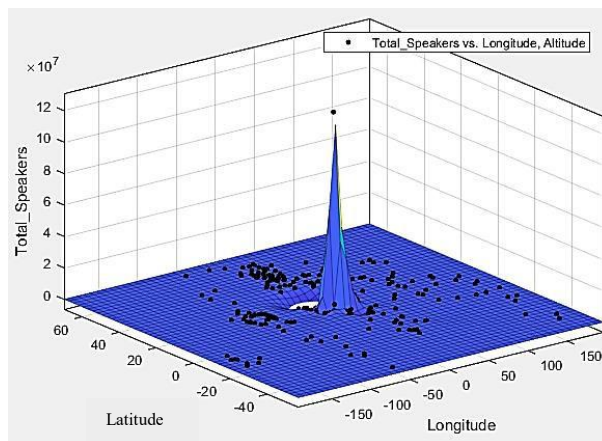
English



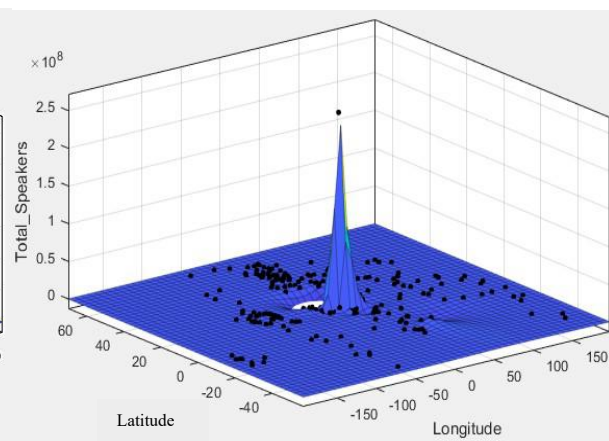
2017

2067

Spanish

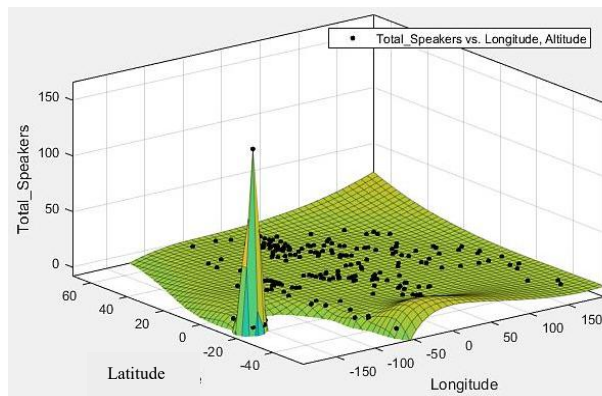


2017

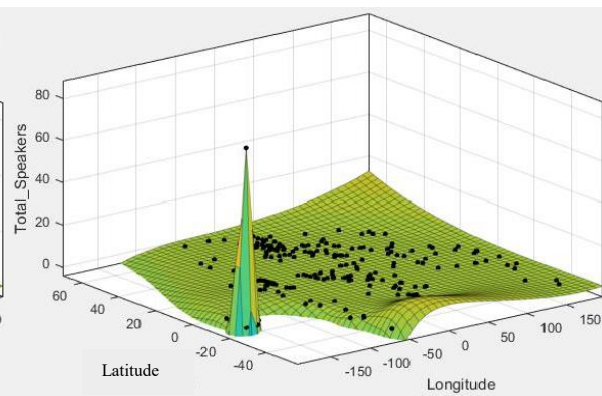


2067

French



2017



2067

German

Appendix 14

Variance Analysis Result

Source	SS	df	MS	F	Prob>F
Columns	1232.45	1	1232.5	0.01	0.9437
Error	4326183.3	18	240343.5		
Total	4327415.75	19			

Appendix 15

Score for Model 3 of Each Countries

Country	G1	G2	G3	G4
Australia	1	1	2	1
New Zealand	1	2	2	1
American Samoa	1	2	2	1
Cook Islands	2	1	3	2
Fiji	1	1	1	2
French Polynesia	2	2	2	2
Guam	1	2	1	2
Kiribati	1	2	3	1
Marshall Islands	1	3	3	1
Micronesia, Federated States of	2	1	3	2
Nauru	1	2	4	1
New Caledonia	2	3	4	1
Norfolk Island	1	3	1	1
Northern Mariana Islands	1	2	1	2
Niue	2	1	2	2
Palau	1	3	1	1
Papua New Guinea	1	1	3	1
Samoa	2	3	3	1
Solomon Islands	2	2	2	1
Tokelau	1	2	2	1
Tonga	1	2	4	2
Tuvalu	1	3	2	2
Vanuatu	1	3	3	1
Wallis and Futuna	1	2	2	1
China	2	2	1	1
Hong Kong	2	2	4	1
Japan	1	2	2	2
Korea, Republic of	2	3	3	1
Taiwan	2	2	4	2
Macau	1	2	3	1
Mongolia	2	1	1	1
Korea, Democratic People's Republic of	1	2	1	1
Indonesia	2	1	4	2
Malaysia	2	2	1	1
Philippines	1	2	3	1
Singapore	2	3	3	1
Thailand	2	2	3	2
Viet Nam	1	1	4	2
Brunei Darussalam	2	2	2	2

Appendix 16

National Comprehensive Evaluation Vector

Australia	New Zealand	American Samoa	Cook Islands	Fiji
0.6	0.69	0.76	0.51	0.76
Guam	Kiribati	Marshall Islands	Micronesia, Federated States of	Nauru
0.53	0.76	0.04	0.47	0.15
Norfolk Island	Northern Mariana Islands	Niue	Palau	Papua New Guinea
0.61	0.23	0.72	0.21	0.32
Solomon Islands	Tokelau	Tonga	Tuvalu	Vanuatu
0.54	0.25	0.95	0.57	0.55
Singapore	Hong Kong	Japan	Korea, Republic of	Taiwan
0.15	0.97	0.77	0.14	0.99
Mongolia	India	Indonesia	Malaysia	Philippines
0.46	0.39	0.52	0.76	0.63
Thailand	Viet Nam	Brunei Darussalam	Cambodia	Bangladesh
0.85	0.34	0.87	0.46	0.69