For office use only	Team Control Number	For office use only
T1 _____	**78951**	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	**B**	F4 _____

2018
MCM/ICM
Summary Sheet

A prediction of the world's language

Summary

In this paper, we model the distribution of various language over time and predict what will happen to the numbers of native speakers and total language speakers in the next 50 years based on our model. In addition, we come to a conclusion that geographic distributions of these languages will not change over this same period of time, given the global populations and human migration patterns predicted for the next 50 years.

In our model, multi-variable linear regression method and cluster method are the main structure, which is a combined system. For every language, one factor is a variable, whose predicted value is the input of the system, because the solution of the value is not linear. We find the distribution of a language always focuses on some developed country or district and spread around. In other words, they distribute as clusters. As a result, to model the distribution is to model the size of clusters. So that we can predict the number of speakers as well, that's the total size of clusters.

According to our model, we predict that any of the languages in the current top-ten lists will not be replaced by another language. Because the weight **w** in the economy is relatively big so that we can conclude that the existence of a language mainly depends on the economy. And we find that the centers of the top-ten lists are developed countries. Thus we make this prediction. And the geographic distributions of these languages will not change over this same period of time. However, the total speakers of some languages such as English, Chinese will increase.

Finally, we use our model to analyze the situation for a company and write a memo of our results and recommendations in the end.

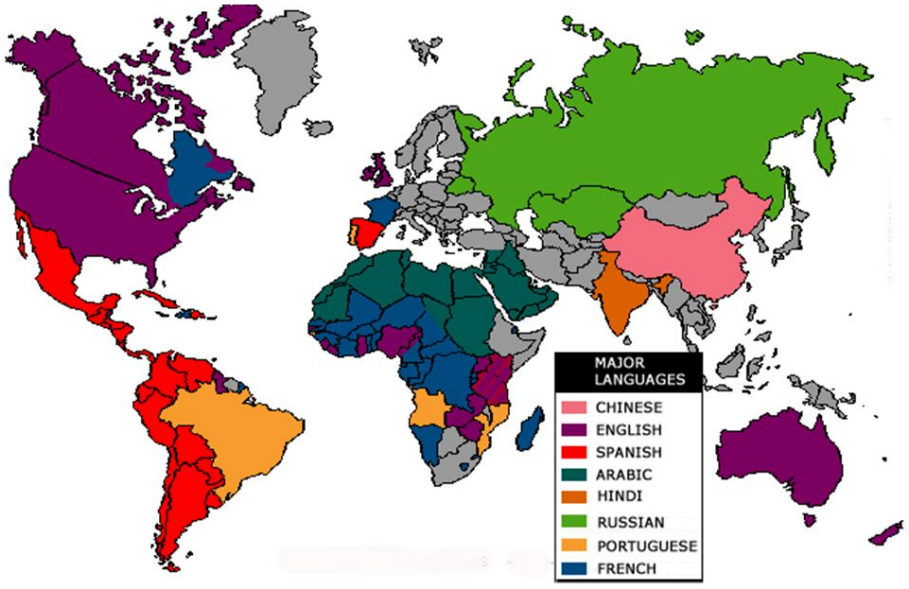
Keywords: Multi-variable linear regression; Cluster method; Combined system

Contents

1 Introduction

1.1 Background

In a multilingual and globalized society, language is the prerequisite for us to communicate with others and to participate in the social, cultural and economic activities.Although there are currently about 6900 languages spoken on Earth, only about half the world's population claim one of the following ten languages as a native language: Mandarin (incl. Standard Chinese), Spanish, English, Hindi, Arabic, Bengali, Portuguese, Russian, Punjabi, and Japanese. In addition, much of the world's population also speaks a second language. The total number of speakers of a language may increase or decrease over time due to many factors.
A person's native language may continue to be used as the daily life and working language within one's own nation. One or several foreign languages may be used as an international auxiliary language in one's native language for trading activities with other people.



1.2 Our Work

We focus on modelling the speakers number of various languages. And the use the cluster method to predict the distribution. We use multi-variable linear regression method to model the prediction. However, the choice of variable is a problem. So we use the PCA model to handle it. In addition, we also add regularization item to solve the possible exist over-fitting and use kernel trick to solve

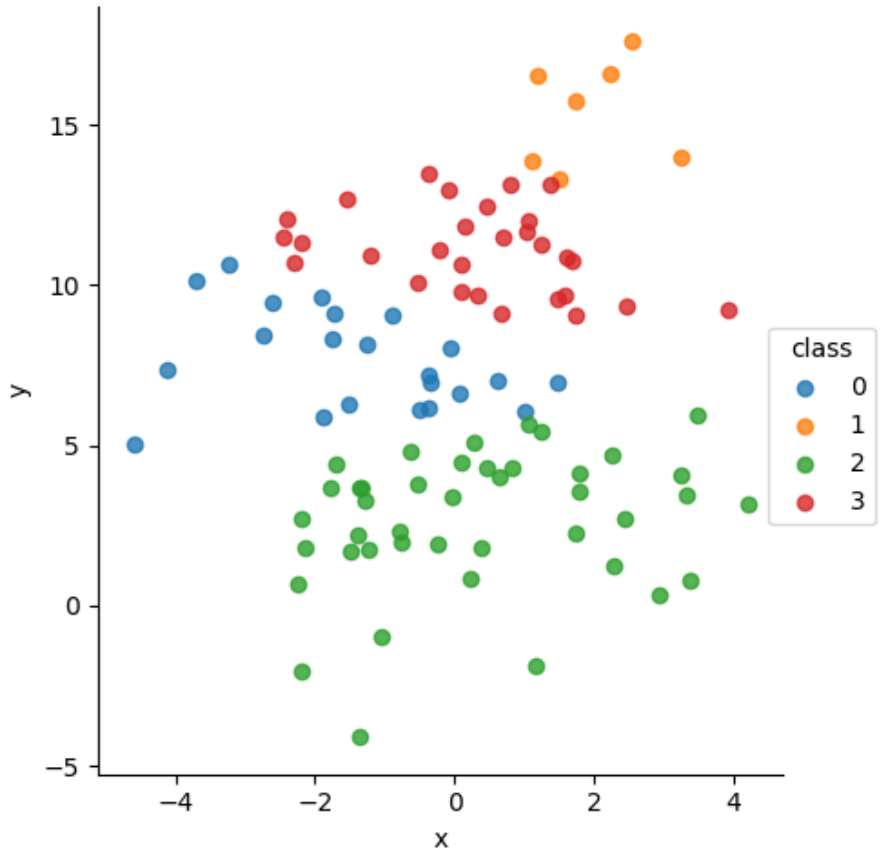


Figure 1: Cluster Method

the possible non-linear data.

1.3 Assumptions

In this paper, our model is based on some identified valid assumptions.

- The very small clusters are ignored because we think they will not have too many effects on our results.
- The changes of rate of natural increase and Engel's Coefficient are also not in our consideration because we do not think they will have too many effects on your results.
- We simplify the model, on one hand, we ignore the unpredictable factors and we just choose some important factors we think. Because these factors have a great weight in the prediction.
- The cost of building the office is not in our consideration.

2 Analysis of the Problem

2.1 The abstract of the problem

Assume that we have a prediction system that its input is language name and its output is the number of its speakers as the following figure. Our task is to

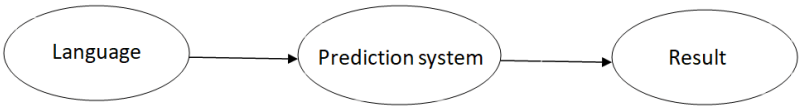


Figure 2: Abstract Model

build the prediction system. Now that we can predict the number of its number of speakers, we can use clustering methods to predict its distribution.

2.2 Structure

The whole structure of our model is the multi-variable linear regression analysis. And its input is the effect factors' values, which is predicted from former data, and we call it a sub-model, which can be solved by formula or some statistical learning methods like least square method. The nature that decides the distribution of language is the move of people and communication. Here we only consider the factors from economy, migration, culture and rate of natural increase.

3 Model Design

3.1 Notation

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \cdots & \vdots & \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{bmatrix}$$

X means test input vector. $x_{i,j}$ means the i-th sample j-th feature.

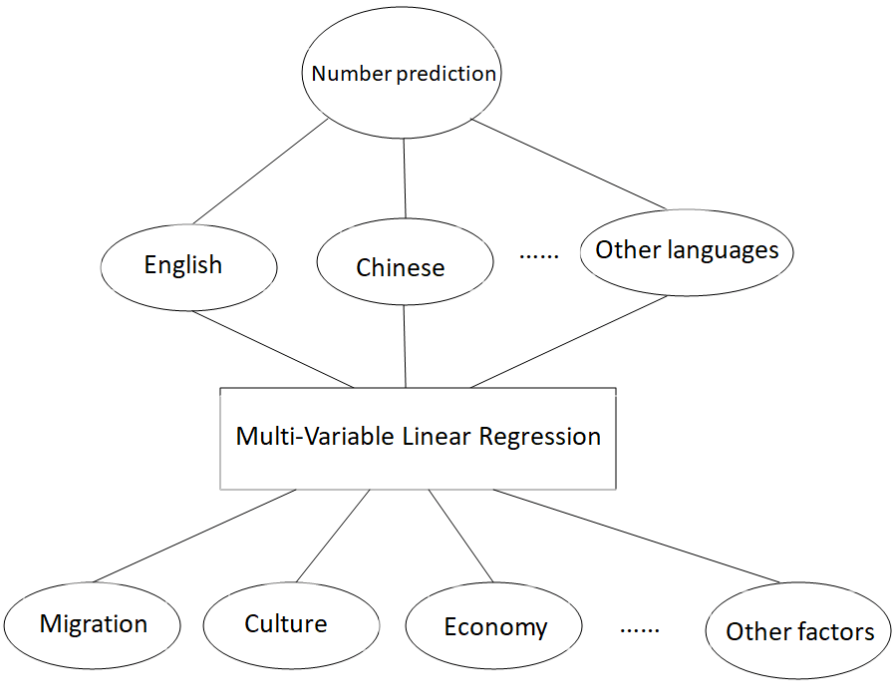


Figure 3: Model Structure

Notation	Representation
X	Input Vector
R	Real space
R^n	n-dimensional real space
var	Variance
I	Identity matrix
ϕ	Reflect function
K	Kernel function
$\|\|A\|\|_2$	L2 norm
t	Time
W	Variable coefficient

Table 1: Notation List

3.2 Multi-variable linear regression analysis

3.2.1 Normal form

We define the output is

$$y = W^T x \tag{1}$$

where $W \in R^n$ is a parameter vector. n denotes the dimension size of input data $X^{(train)}$

The loss function is

$$MSE_{test} = \frac{1}{m}\sum_{i}(y_{predict} - y_{test})_i^2. \tag{2}$$

Also we can write it in another form.

$$MSE_{test} = \frac{1}{m}||y_{predict} - y_{test}||_2^2 \tag{3}$$

And the target is to minimize the loss function.
Objective function:

$$\min_{W,b} MSE_{train} \tag{4}$$

By taking the derivation of MSE_{test} and set it zero, we can simply get the optimal answer.

$$\nabla_w MSE_{train} = 0 \tag{5}$$

We can solve the W:

$$W = (X^{(train)T}X^{(train)})^{-1}X^{(train)T}y^{(train)} \tag{6}$$

$X^{(train)}$ consists of the sub-models' answers. In addition, the initial value of the W is random.
W denotes $[Migration, Culture, Economy, Rateofnaturalincrease, \cdots, others]$

3.2.2 Regularization

Taking over-fitting into consideration to help us predict more accurately, we add a regularization item as follow:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}(y^{predict} - y^{test})^2 + \frac{\lambda}{2}||w||_2^2 \tag{7}$$

It equals that:

$$J(W) = \frac{1}{2}\sum_{i=1}^{n}(y^{predict} - y^{test})^2 + \frac{\lambda}{2}W^TW \tag{8}$$

By taking the derivation of x, we can solve the W:

$$W = (\lambda I + W^TW)^{-1}W^Ty^{train} \tag{9}$$

where I denotes the identity matrix.

3.2.3 Kernel Trick

Since we may not make sure the data can be divide linearly, so we also try the kernel trick. That's not so complicated. Here we give the function form.

$$K(x_i,) = \phi(x_i)K(x_i, x_j) = \phi(x_i)\phi(x_j) \tag{10}$$

That means we reflect the data into a higher dimension, which is a Hilbert space. And we choose RBF Function for K, which is called Gaussian kernel function:

$$K(x,z) = exp(-\frac{||x-z||^2}{2\sigma^2}) \tag{11}$$

Thus, we can rewrite the loss function:

$$J(W) = \frac{1}{2}\sum_{i=1}^{n}(y^{predict} - W^T\phi(X))^2 + \frac{\lambda}{2}W^TW \tag{12}$$

And its solution is:

$$W = (\lambda I + \phi^T\phi)^{-1}\phi^T y^{train} \tag{13}$$

About the proof of positive definite kernel function, we will not give, Gaussian kernel function is a frequently-used kernel function.

3.3 Sub-model

3.3.1 Rate of natural increase

We can approximately think the rate of natural increase of next 50 years in a country is η, which is a constant. So we can get the function to calculate the future people number of the country:

$$N_t = N_0 * (1+\eta)^t \tag{14}$$

N_t denotes the people number of the t year from now. N_0 denotes the current people number.

3.3.2 Economy

We choose GDP and Engel's coefficient. Because Engel's coefficient will not change too much, it will not affect too much to our model, hence we can consider it as a constant. As for GDP, we combine the data collected from Internet with the result we predict to analyse.
On the supply sight, we have:

$$y_t = A_t K_t^{\alpha} L_t^{1-\alpha} \tag{15}$$

where y_t denotes t-th year's GDP, A_t denotes the technology of t-th year , k_t denotes the capital of t-th year, L_t denotes the labor of t-th year, α is a capital output elastic coefficient. On the demand sight, we have:

$$y_t = C_t + I_t + G_t + (x_t - M_t) \tag{16}$$

where y_t denotes t-th year's GDP, C_t denotes the consumption, I_t denotes the investment, G_t denotes the investment of government, X_t denotes the export, M_t denotes the import.

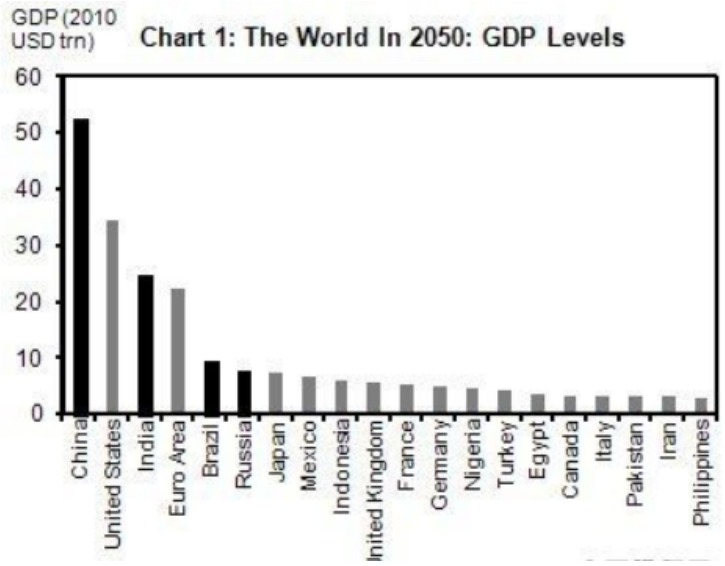


Figure 4: GDP Levels

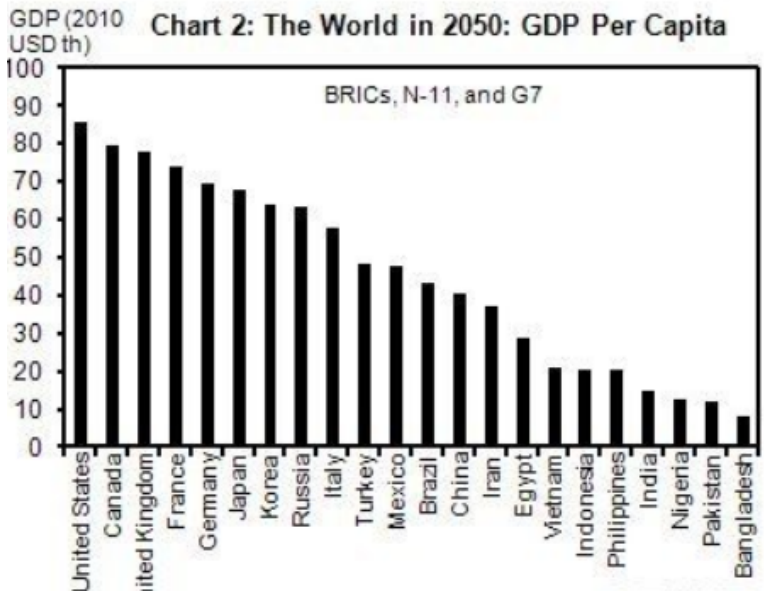


Figure 5: GDP Per capita

3.3.3 Migration

As for migration, this is very difficult to predict and it is very complicated. So we take the following formula to help us complete our model.

$$ER(0) = \int_0^T [P_t(t)P_2(t)\gamma_d(t) - P_3(t)\gamma_0]e^{-rt}dt - C(0) \tag{17}$$

ER(0) is the pure profit before migration, t is the time variable, $t \in [0, T]$. According to the analysis by the relative researcher, the pure profit will be affected by the seven basic factors. Three factors of the first group come from the target

country, which include:

- Inbound residence control in destination country.(P_1)
- The chance of finding a job in the target country.(P_2)
- The possible salary.(Y_d)

The two factors of the middle group come from home country, which include:

- The probability of being hire staying at home.(P_3)
- The possible salary.(Y_o)

The factors of the third group is the possible future currency change.
The last item is the cost of migration, which won't affect our model too much, hence we can ignore it. P_1 is also unpredictable, which we also ignore by setting it a constant p. So the formula we will use is:

$$ER(0) = \int_0^T [pP_2(t)\gamma_d(t) - P_3(t)\gamma_0]e^{-rt}dt \tag{18}$$

Also the $P_2(t), p_3(t), \gamma_d(t)$ can be predicted by fitting the former data. Hence we can solve the formula:

$$ER(0) = pp_2'\gamma_d'T + \frac{1}{\gamma_0}p_3'(1 - e^{rT}) \tag{19}$$

p_2', p_3', γ_d' are the value predicted by fitting. Take China's rate of unemployment as an example as follow:

4 PCA Model

4.1 Normalization

To handle the data conveniently, we had better normalize the data using the next method.

$$x_{i,j} = \frac{x_{i,j} - \bar{x_j}}{s_j} \tag{20}$$

$\bar{x_j}, s_j$ denote the mean of data and the standard deviation partly
Thus, the data value will be reflected into the interval [0, 1] According to the additive property of our variables, we can approximately consider the data satisfy Gaussian Distribution for central limit theorem. So the rationality of our model depends on the data subject to the dependency among the variable and satisfying the same distribution. Thus we should validate the model and we adopt

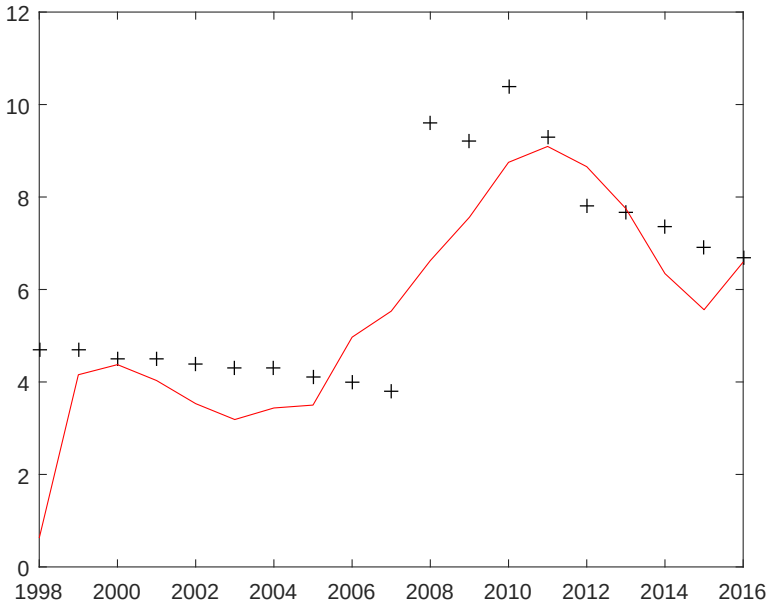


Figure 6: China's rate of unemployment

PCA to finish it. To make the data as disperse as possible, that's

$$\begin{aligned} max(var(w_1x_1 + w_2x_2 + \cdots + w_nx_n)) \\ s.t. \quad ||W||_2^2 = 1 \end{aligned} \tag{21}$$

$$\begin{cases} Z_1 = w_{1,1}x_1 + w_{1,2}x_2 + \cdots + w_{1,p}x_p \\ Z_2 = w_{2,1}x_1 + w_{2,2}x_2 + \cdots + w_{2,p}x_p \\ \cdots\cdots \\ Z_p = w_{p,1}x_1 + w_{p,2}x_2 + \cdots + w_{p,p}x_p \end{cases} \tag{22}$$

Z_i denotes the i-th principle component, the formula satisfies $max(Var(Z_i))$ for any pair of (i, j) and

$$\begin{aligned} (w_{i,1}, w_{i,2}, \cdots, w_{i,p}) \perp (w_{j,1}, w_{j,2}, \cdots, w_{j,p}) \\ w_{i,1}^2 + w_{i,2}^2 + \cdots + w_{i,p}^2 = 1 \end{aligned} \tag{23}$$

Thus we can solve the p principle components.

4.2 Solve the equation

$$M_2^* = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2 = \frac{1}{n}\sum_{i=1}^{n} z_i^2 = \frac{1}{n}X^TX \tag{24}$$

To maximum M_2^*, we can get the eigenvalue $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ and their corresponding orthonormalization eigenvector by using eigen decomposition. So the

value that maximum M_2^* can be fetch at the $l = \eta_1$, and the max value is λ_1, when the first rank component is $z_1 = X * \eta_1$. We can get the k principle component for the same method.

$$z_k = X * \eta_k, k = 1, 2, \cdots, p \tag{25}$$

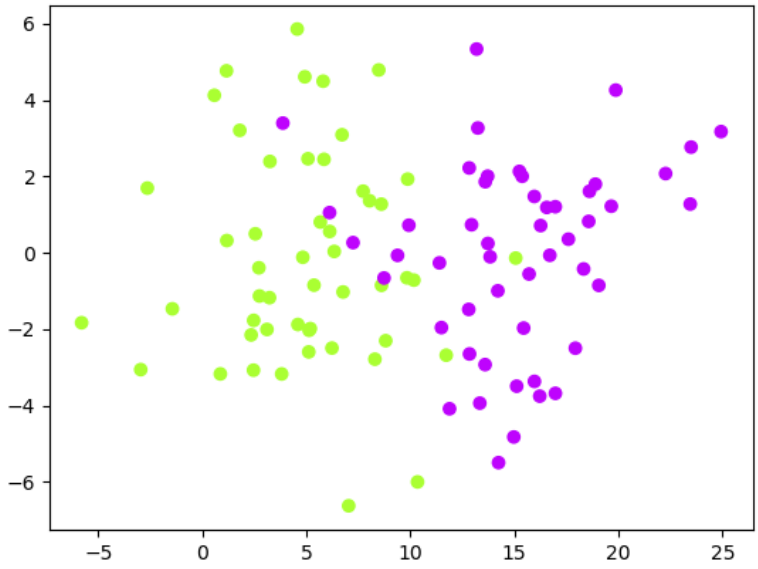


Figure 7: First two dimension

5 Office Location

Without considering the price of building offices, service company should choose the developed high-quality life districts, which is promising. According to the data and our analysis, we think that only rich people will be more willing to spend their money on service. Thus, the keys of choosing the office places are rich and many people. Our method is as follow based on GDP and Engel's coefficient.

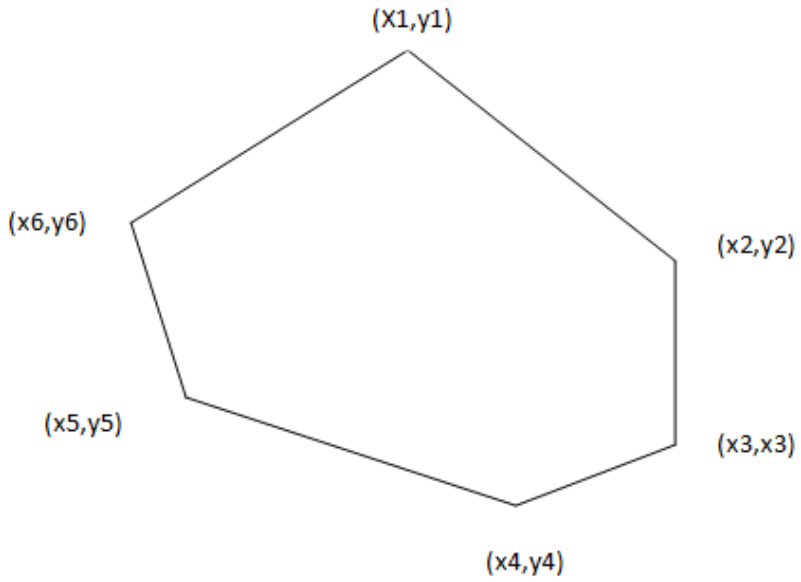
5.1 Choose method

- sort the country according to GDP, then choose the districts from small to big. Of course not all districts meet the demands. The choosing condition will be listed as follow.

- In the chosen districts, find the place where has the biggest people density, then spread around it until the density decreases to $\alpha\rho$. α is an adjustive variable. Here we set it 0.8.
- Connect the bound points and make it a polygon, the chosen point is the center of the polygon (x_g, y_g) using the following formula:

$$\begin{cases} x_g = \sum_{i=1}^{n} x_i \\ y_g = \sum_{i=1}^{n} x_i \end{cases} \tag{26}$$

where n denote the number of bound points.



5.2 Choose condition

- They can speak English.
- The distances between two points mutually are not too big.
- Political status.

5.3 Short term and long term

So the six office locations are Moscow, Paris, GuangZhou, London, Tokyo, Sao Paulo.

In short term, we can make it as above. However, in long term, the GDP we used is supposed to be the expectation of the GPD, which is a predicted value.

6 Evaluate Model

6.1 The measurement of fitting degree

$$\begin{aligned} R^2 &= \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} \\ &= 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} \end{aligned} \tag{27}$$

where:

$$\begin{aligned} \sum (y - \hat{y})^2 = \sum y^2 - (W_0 \sum y + W_1 \sum x_1 y + W_2 \sum x_2 y + \cdots + \sum W_d y) \\ \sum (y - \bar{y})^2 = \sum y^2 - \frac{1}{n} (\sum y)^2 \end{aligned} \tag{28}$$

The greater the R^2 is, the greater the degree of fitting the regression side to the sample data points, and the closer the relationship between the independent variables and the dependent variables is.

6.2 Estimate the standard error

$$\begin{aligned} S_y &= \sqrt{\frac{\sum (y - \hat{y})^2}{n - d - 1}} \\ v_d &= \frac{S_y}{y} \end{aligned} \tag{29}$$

The standard error is estimated, that is, the standard error between the real value of the dependent variable y and the estimated value $\hat{y}$ of the regression equation. The smaller the estimated standard error is, the more fitting the regression equation is.

6.3 The saliency test of the regression function

The saliency test of the regression function, that is, to test the saliency of the whole regression function, or to evaluate the close relationship between the linear relation of all independent variables and the dependent variables. F test can be

used often, and the formula of F statistics is as follows:

$$\begin{aligned} F &= \frac{\frac{\sum(\hat{y}-\bar{y})^2}{d}}{\frac{\sum(y-\hat{y})^2}{n-d-1}} \\ &= \frac{\frac{R^2}{d}}{\frac{1-R^2}{n-d-1}} \end{aligned} \tag{30}$$

According to the given significant level a, (d, n-d-1) is used to check the F distribution table, and the corresponding critical value F_a is obtained. If F > F_a, the regression equation has significant significance, and the regression effect is significant. F < F_a, the regression equation has no significant significance, and the regression effect is not significant.

7 Conclusions

7.1 Part I

According to our model and analysis, we draw a conclusion that any of the languages in the current top-ten lists will not be replaced by another language. In our consideration and model, we think that economy is the core and has the great weight, the top-ten listed languages' cores are developed countries or districts, in other words, it's economy maintains the development of language. What's more, the geographic distributions of these languages will not change too much over this same period of time.

7.2 Part II

The six office locations are Moscow, Paris, GuangZhou, London, Tokyo, Sao Paulo. In short term, we can make it as above. However, in long term, the GDP we used is supposed to be the expectation of the GPD, which is a predicted value. They can communicate in English. Because according our census and model prediction, English will be the language that has the most people to learn in the next 30 years. I suggest that the company still open six international offices. Because the sizes of clusters will not change a lot. What's more, as the second language is developing and developing, we think the value of service will become more and more important, especially the developed countries or districts. In a word, it all depends on the development of economy and culture.

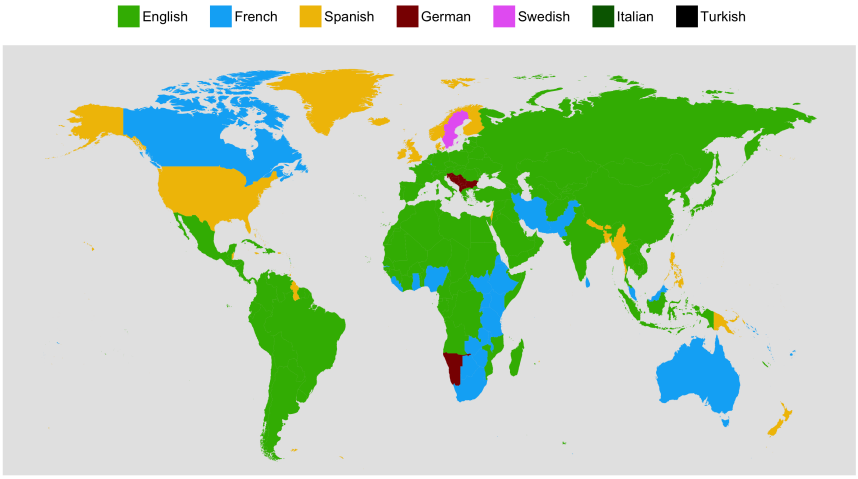


Figure 8: The most welcome learning language

8 Strengths and weaknesses

8.1 Strengths

- **Applies widely and good flexibility**
  This multi-variable linear variable model can be easily modified to accommodate to various situation. Our model can find the optimum solution under various kinds of conditions and we consider many situations.
- **Easy to solve and result is a global optimum solution** Of course, we can handle the solution form in our paper. The proof has been given above.

8.2 Weaknesses

- **Inaccuracy** There are too many unpredictable factors, such as policy, nature disasters. In addition, lack of data is another reason. Thus, it may not predict well. This is just a estimation method.
- **Simplifying assumptions** If some clusters are not so big, we may ignore them, however, if too many such clusters, the answer will be affected.
- **Too many parameters** For some method such as PCA, we must adjust the parameter p manually.

References

[1] JOURNAL OF XIAMEN UNIVERSITY(Arts Social Sciences)

[2] http://finance.sina.com.cn/worldmac/compare.shtml?indicator=NE.CON.PETC.KD.ZG&nation=CN&type=0

[3] Lane, J. (2017). The 10 Most Spoken Languages in the World. Babbel Magazine. Retrieved from https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world

[4] Noack, R. and Gamio, L. (April 23, 2015). The WorldâĂŹs Languages in 7 Maps and Charts. The Washington Post. Retrieved from https://www.washingtonpost.com/news/worldviews/wp/2015/04/23/the-worlds-languages-in-7-maps-and-charts/?utm_term=.a993dc2a15cb

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

[6] http://dspace.xmu.edu.cn/bitstream/handle/2288/101034/%E5%BD%93%E4%BB%A3%E8%A5%BF%E6%96%B9%E5%9B%BD%E9%99%85%E7%A7%BB%E6%B0%91%E7%90%86%E8%AE%BA%E5%86%8D%E6%8E%A2%E8%AE%A8.pdf?sequence=1&isAllowed=y

[7] Abstract of speakers' strength of languages and mother tongues âĂŞ 2000, Census of India, 2001

[8] 2."Summary by language size". Ethnologue. Retrieved 2016-04-06.

[9] Crystal, David (March 2008). "Two thousand million?". English Today. doi:10.1017/S0266078408000023.

[10] "Hausa speakers in Nigeria now 120mâĂŞ Communique - Vanguard News". vanguardngr.com. Retrieved 2017-04-26.

[11] "World Arabic Language Day | United Nations Educational, Scientific and Cultural Organization". www.unesco.org. Retrieved 2017-07-14.

Appendices

Appendix A First appendix

Here are simulation programmes we used in our model as follow.

some more text **Input Python source:**

```
import tensorflow as tf
import math
from sklearn import datasets
from sklearn.manifold import TSNE
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import cm
from mpl_toolkits.mplot3d import Axes3D
import seaborn as sns

data, label = loadDataSet()

class TF_PCA:

    def __init__(self, data, target=None, dtype=tf.float32):
        self.data = data
        self.target = target
        self.dtype = dtype

        self.graph = None
        self.X = None
        self.u = None
        self.singular_values = None
        self.sigma = None

    def fit(self):
        self.graph = tf.Graph()
        with self.graph.as_default():
            self.X = tf.placeholder(self.dtype, shape=self.data.shape)

            # Perform SVD
            singular_values, u, _ = tf.svd(self.X)

            # Create sigma matrix
            sigma = tf.diag(singular_values)

        with tf.Session(graph=self.graph) as session:
            self.u, self.singular_values, self.sigma =
            session.run([u, singular_values, sigma],
                        feed_dict={self.X: self.data})

    def reduce(self, n_dimensions=None, keep_info=None):
        if keep_info:
            # Normalize singular values
```

```
            normalized_singular_values =
            self.singular_values / sum(self.singular_values)

            # Create the aggregated ladder of kept information per dimension
            ladder = np.cumsum(normalized_singular_values)

            # Get the first index which is above the given information threshold
            index = next(idx for idx, value in enumerate(ladder)
                         if value >= keep_info) + 1
            n_dimensions = index

        with self.graph.as_default():
            # Cut out the relevant part from sigma
            sigma = tf.slice(self.sigma, [0, 0],
                             [self.data.shape[1], n_dimensions])

            # PCA
            pca = tf.matmul(self.u, sigma)

        with tf.Session(graph=self.graph) as session:
            return session.run(pca, feed_dict={self.X: self.data})


tf_pca = TF_PCA(np.mat(data), np.mat(label))
tf_pca.fit()
pca = tf_pca.reduce(keep_info=0.9)  # Results in two dimensions

color_mapping = {0: sns.xkcd_rgb['bright purple'], 1:
                 sns.xkcd_rgb['lime'], 2: sns.xkcd_rgb['ochre']}
colors = list(map(lambda x: color_mapping[x], label))

plt.scatter(pca[:, 0], pca[:, 1], c=colors)
plt.show()
```

To: Chief Operating Officer of the service company
From: MCM Team 78951
Date: February 12, 2018
Subject: International Offices

Office Location

The six office locations are Moscow, Paris, GuangZhou, London, Tokyo, Sao Paulo. In short term, we can make it as above. However, in long term, the GDP we used is supposed to be the expectation of the GPD, which is a predicted value. They can communicate in English. Because according our census and model prediction, English will be the language that has the most people to learn in the next 30 years. According our model, the predict result is as follow:

Mainly considering from the economy, we choose the six country for more

Country	People Number
India	165655
China	130372
United States	43901
Brazil	26069
Russia	10919
Japan	9367
German	7361
France	6977
English	6398

Table 2: Notation List

profit. There are some reasons as following table.

- These are developed districts or countries, which have developed economy.
- There are many people, in other words, potential consumers.
- Some may not very big like Paris, however, we choose it because it can lead the surrounding districts and European Union.

Number of offices

I suggest that the company still open six international offices. Because the sizes of clusters will not change a lot. What's more, as the second language is developing and developing, we think the value of service will become more and more important, especially the developed countries or districts. In a word, it all depends on the development of economy and culture.

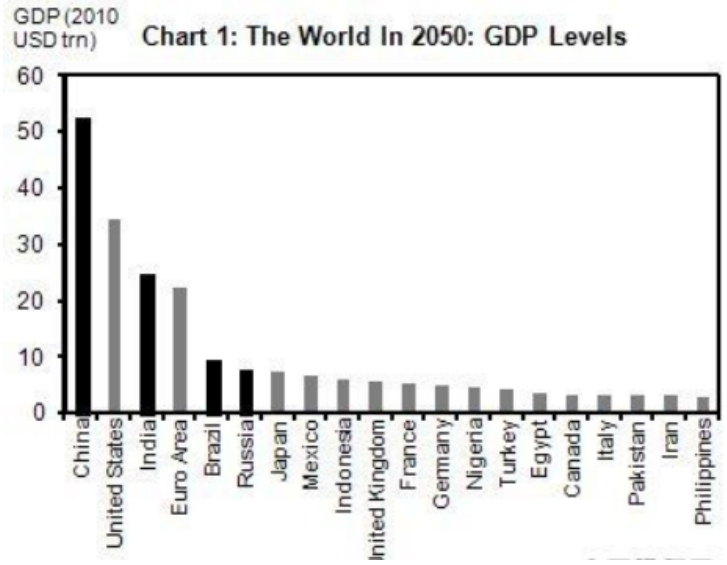


Figure 9: GDP Levels

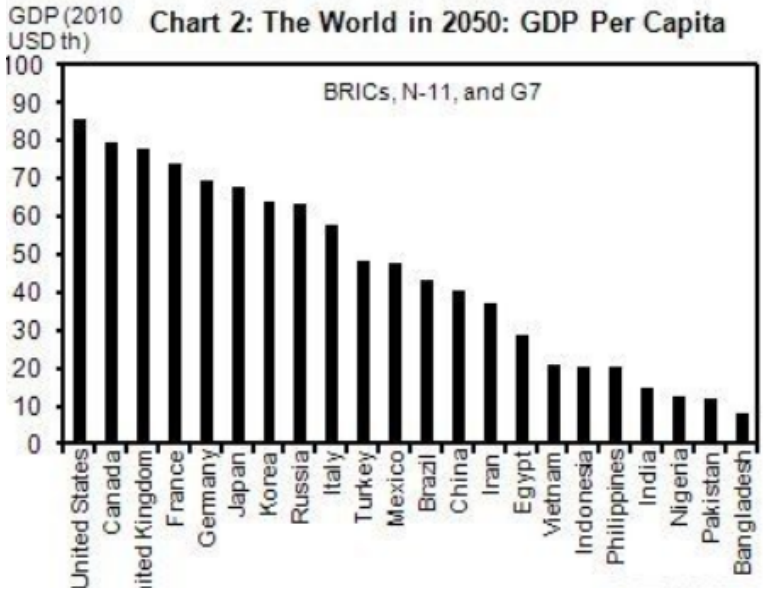


Figure 10: GDP Per capita