For office use only	Team Control Number	For office use only
T1 _____	**1912947**	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	**D**	F4 _____

2019
MCM/ICM
Summary Sheet

This is title

Summary

There are many kinds of language with complex geographic distribution in the world. Transnational corporations all desires to maximize profits. Therefore, it is necessary to predict the number of users and the geographic distribution of different languages in the future, in which many experts and scholars make a deeper research.

Keywords: K-Means; K-Nearest Neighbor; Personality Compatibility Matching Perfect Dating Partner Online Recommendation System

Contents

1 Introduction

1.1 Background

There are 6,000 to 7,000 languages being used all over the world currently, but around half of the world's population are using the most important 15 languages. With the development of globalization, the whole world has changed deeply. The impact of this trend does not only exist in economy and society, but also in language and culture. Nowadays, with the continuous improvement of transportation and communication, most people are able to speak the second language besides the mother language, and the second language plays an important role in traveling abroad and international trade.

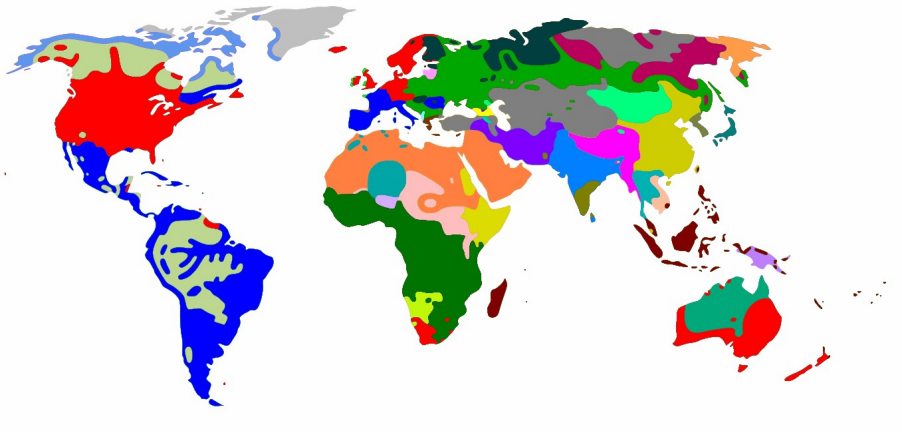


Figure 1: The Distribution of Various Language

The current geographical distribution of different languages in the world is shown in the figure 1. When considering the total number of language speakers, native speakers, the second language speakers and even the third language speakers should be taken into consideration. The language distribution is complex and diverse, so the statistics on the total number of languages are a complicated job. At present, many experts and scholars have conducted research in this area to explore whether the trend of language distribution tends to be simplification.

1.2 Restatement of the Problem

For the increase of the total number of language speakers, we divided them into native speakers and second language speakers for model building. For the geographical distribution of language speakers, we should not only consider the population increase in each country after 5 years, but also the number of speakers who moving to other countries. Analyzing of the various factors of the problem, we consider the problem how to select the site of the International Office in each

country as a risk-based site selection analysis: This problem can be divided into four parts:

- Establish an increase model of the total number of language speakers changed with time to predict the change in the total number of language speakers in the next 50 years.
- Establish linguistic geography distribution model changed with time and predict the change of the geographical distribution of the language speakers in the next 50 years.
- According to the geographical distribution of language speakers, select the international offices to maximize the profits of the office.
- Consider changes in the global communication methods and determine whether the number of international offices can be reduced.

1.3 Our Work

In order to make the prediction more accurate, our work is divided into the following four aspects:

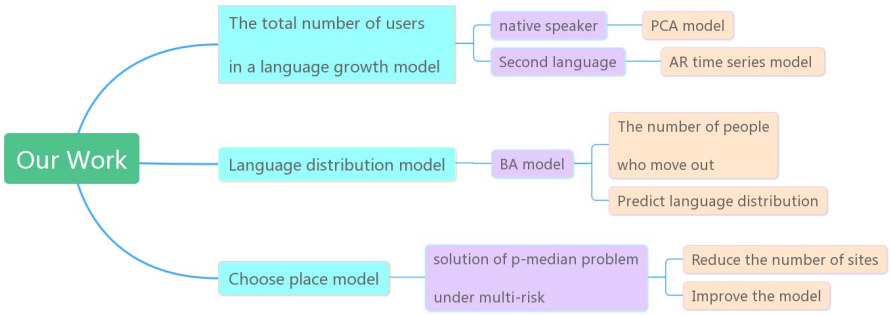


Figure 2: Flow Chart

1) To establish an increase model of the total number of language speakers, we should first obtain the number of native speakers in the future by predicting models of the population in different countries in the world. Secondly, we will determine the number of users who mainly use a certain language as a second language by determining the corresponding weights through principal component analysis .Finally, we will get the total of the two parts, and get the trend of the number of people who speak a certain language.

2) Secondly, we use the scale-free network model to calculate the country's total value, according to the data function fitting. The FMR model is used to represent the point-to-point population mobility as the ratio of the nationalą́s repulsive force to the attractive force of other countries .And we predict the migration pattern of the global population.

3) Thirdly, we adopt the solution of p-median problem under multi-risk. We convert the multi-objective problem and the line-by-line weight problem into single-objective problem and then make fuzzy decision. Finally, we have found the location of six new international offices and give advises to clients.

4) Finally, we submit a 2-page memo to summarize our findings and recommendations.

2 Terminology Explained in Model

- Nodes centrality degree: The number of edges connected to that node.
- L1 country: Countries referred to the country which has Russian as native language.
- L2 country: Countries referred to the country which has Russian as second language.
- L1 numbers: The numbers of native speakers of a certain language.
- L2 numbers: The numbers of second language speakers of a certain language.

3 Assumptions

1. We assume that each country uses only the third and fourth languages with very few people, which has little effect on native speakers and second language speakers;

2. We do not consider small probability events in our model, assuming no random events such as world war and economic crisis.

3. Assuming that each country moves in and out of balance, that is, the total population does not change;

4. Assuming that the total number of immigrants to a country speaks the official language of the country and does not consider the issue of the offspring after the move;

5. Assume that all immigrants are legal immigrants;

6. Assuming that national policy does not change significantly over time;

7. Assumptions data are all true and reliable;

4 Symbols and Definitions

Table 1: Symbols and Definitions

Symbols	Meanings
Y_t	A country's L1 number
$\Delta Y(t)$	the number of population growth one year
$W(t)$	A country's L2 speakers
$S(t)$	The population of the country
O_i	The total degrees of country i
$F_i(t)$	The thrust of country i
$N_j(t)$	The attractive force of country i
$N`_i(t)$	The tensile force of country i
$Q_i(t)$	The total number of people removed from country i
R	The total expected value of a local profit

5 Simulation Model

5.1 Justification of Our Approach

For this question, the total population of a given language A is divided into two parts by us, which including those countries whose native language is A and those countries whose second(or 3rd,etc) language is A. Therefore, in order to find the accurate total numbers of speakers of language A, we should both find the two parts of the numbers of language A.

For the first part of the population, we investigate the population of a country where a language A is the mother tongue. However, not all people speak the native language. Therefore, we remove immigrants and people with a low level of literacy. As for population growth, we use the time series forecast model to make a prediction about the world's population.

For the second part of the population, we use PCA model to find the proportion of the numbers of speakers who have language A as a second language in a country(L2 country). According to the background of the title, we choose four factors to measure a country's level of learning a second language, which are cultural soft power, social pressure, immigration, the degree of opening up. Considering that if a government encourage people to learn second language, the trend of school learning is bound to upwards. So we regard the two factors ąőlanguage A promoted by the governmentąŕ and ąőthe language used in schoolsąŕ as one factor which is quantized by Primary Education Enrollment. We can use the per capita GDP as a measure of social pressure, because social pressure is inversely proportional to the level of consumption. The immigration is measured by net migration in a country. The degree of opening up is measured by the volume of export trade and the volume of import trade.

After confirming the factors, we use Principal Component Analysis(PCA) to analyze and get the corresponding weight of each factor, and find the relation between the final scores and the proportion of the numbers of speakers in L2 country. Finally, we sum up all the numbers of speakers in different country and compare the calculated data with the number of second-language speakers given in the question. Even though we did not count the third and fourth languages and ignored a few countries with small population, the results were slightly different but basically feasible.

5.1.1 Justification of our approach

- **Why do we build the population growth model?**
  If we want to model the distribution of language speakers in the future, building the population growth model is necessary because the numbers of different language speakers is changing over time. If we don't consider this, we can not do the next step and the result must be unreliable.
- **Why do we divide each language into different country?**
  Because different country has different population growth rule and different L2 country has different proportion of L2 speaker. If we easily add up the native language and the second language and make a simple calculation of population growth model, the error of result is very big.
- **Why do we divide each language into two parts(L1 country, L2 country)?**
  It is obviously that L1 country and L2 country have different rule about calculating the proportion of the speakers of a particular language. So what we need do is to find the population of every country which speaks a particular language, and find the proportion of this language in these countries.

5.1.2 Basic Model

1) AR Time Series Model

In order to get a model which have the ability to predict, we build an AR Time Series Model. Autoregressive model is:

$$\Delta \mathrm{Y_t}{=}\alpha_0{+}\sum_{\mathrm{i}=1}^{n}\alpha_i \Delta Y_{i-1}+\mu_i$$

where α_0. is constant, α_i is coefficient of the model, μ_i is White Noise Sequence.

As an analogy, the model of population growth is:

$$Y_t = Y_{t-1} + \Delta Y_t \tag{1}$$

where Y_t is the population this year, ΔY_{t-1} is the number of population growth, $\{\Delta Y_t\}$is first-order difference sequence of the population.

Based on this, the numbers of speakers of a particular language in L1 country are:

$$Y = (1 - \beta - \lambda)Y_t$$

Where β is the proportion of immigrant this year,λ is the proportion of people with a low level of literacy. $n$is the number of L1 country.

2) PCA Model

Second language, as a supporting language, can be obtained in different environments. Due to various influences such as immigration, national policies and cultural groups, people who speak a second language will change over time. Therefore, we selected four indicators of cultural soft strength, social pressure, immigration and the degree of opening up. The principal component analysis (PCA) model was used to derive the weight of each dependent variable and calculate the number of people using the second language in the country.

$$\begin{cases} F_1 = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + a_{14}X_4 \\ F_2 = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + a_{24}X_4 \\ F_3 = a_{31}X_1 + a_{32}X_2 + a_{33}X_3 + a_{34}X_4 \\ F_4 = a_{41}X_1 + a_{42}X_2 + a_{43}X_3 + a_{44}X_4 \\ F_5 = a_{51}X_1 + a_{52}X_2 + a_{53}X_3 + a_{54}X_4 \end{cases}$$

Where $X_i\,(i = 1, 2, 3, 4, 5)$ respectively represent the Secondary school enrolment rate, per capita GDP, net immigration, export trade and import trade.

Set $\widetilde{a}_{ij}$ as the value of index variable j of model i.

$$\widetilde{a}_{ij} = \frac{a_{ij} - \mu_j}{s_j}, (i = 1, ..., 5, j = 1, ..., 5)$$

$$\widetilde{x}_j = \frac{x_j - \mu_j}{s_j}, (j = 1, ..., 5)$$

Where $\mu_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}, S_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - \mu_i)^2}$ Then we calculate the correlation coefficient matrix $R = (r_{ij})_{5*5}$

$$r_{ij} = \sum\nolimits_{k=1}^{n} \frac{\widetilde{a}_{ki} * \widetilde{a}_{kj}}{n-1}, (i, j = 1, ..., 5)$$

Then, we calculate the eigenvalues λ_j and the corresponding eigenvectors of R, to form the five new index variables. On the basis of the eigenvalues, we can get the contribution rate b and the accumulative contributions rate $\alpha$of the parameters.

$$b_j = \frac{\lambda_j}{\sum_{k=1}^{m}\lambda_k}, (j = 1, ..., 5)$$

$$\alpha_p = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=5}^{p} \lambda_k}$$

In descending order:$Z = \alpha_1 F_1 + \alpha_2 F_2 + ... + \alpha_i F_i$

We assume that there is a relation between the final scores and the proportion of the numbers of speakers in L2 country:

$$W(t) = \gamma Z(t)$$

Based on the available data, we find this weight and make a test which tell us it is feasible.

3) Total Numbers of Speakers Total numbers of language speakers of a certain language is:

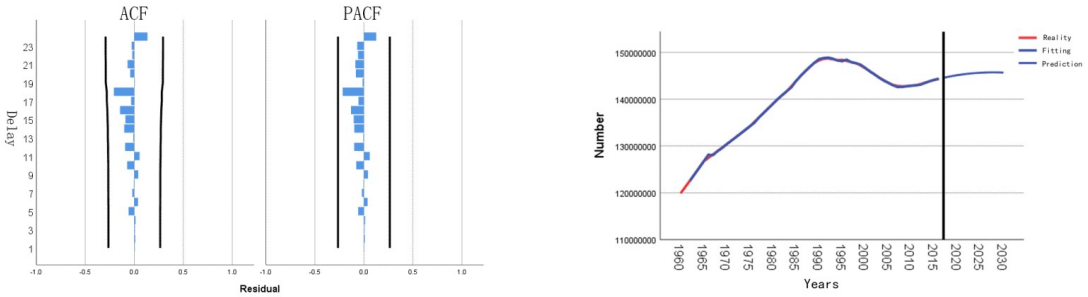
$$S(t) = Y(t) + W(t)$$

5.1.3 Solution and Result

In this problem we have to model the distribution of various language speakers over time. But due to the wide distribution of languages, we focus on the analysis of the four representative languages which are English, Russian, Chinese, Japanese. We use our model to predict their distribution ten years later.

For the growth of native languages speakers in these four languages, here is an example of analysis in Russian.

Figure.3a shows the autocorrelation and partial autocorrelation coefficients of Russia's population growth function, and Figure.3b shows the population trend ten years after the prediction of Russia.



(a) ACF and PACF (b) Change in Russia's Population

Figure 3: The Growth Model of Russia's population

As can be seen from Figure.3, Russia's total population is 145 million after 10 years. As for β, λ parameter in Equation.1, we find the data from 1960 to 2016 in World bank.

We choose four languages and make a prediction about their L1 speakers number. Result is shown in Table.2

Table 2: The Change of L1 Speakers in Four Countries

	Now	In ten years
Russian	153	162
English	371	405
Chinese	897	955
Jpanese	128	124

As for the number of L2 speakers, we almost find all the certain language's L2 country and use our model to calculate the number of their L2 speakers in 2017, as is shown in Figure.4. Comparing the result with the actual data in 2017, we find it that our model can basically figure out the number of L2 speakers.

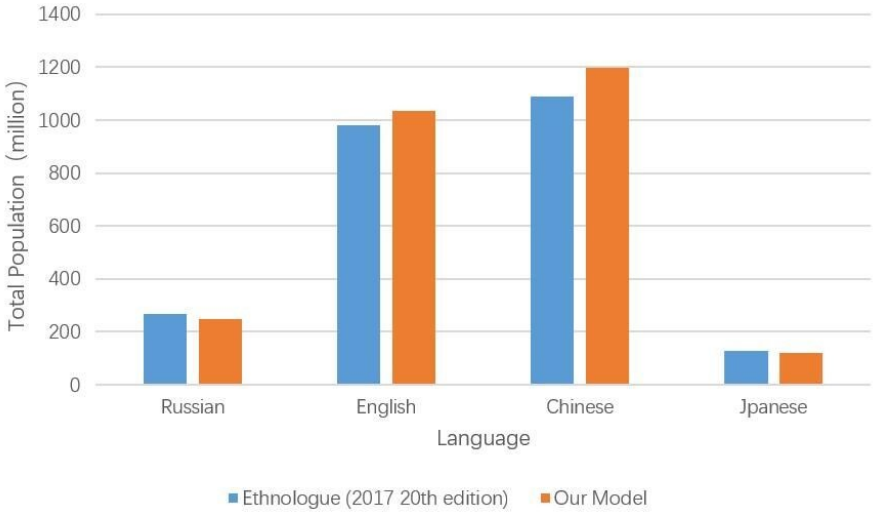


Figure 4: Model Test

Finally, we predict the total languages speakers of the four languages we choose.

Table 3: The Total Numbers of Four Languages

	Native Language		Second Language		Total	
	Now	In Ten Years	Now	In Ten Years	Now	In Ten Years
Russian	153	162	113	119	266	281
English	371	405	611	666	982	1071
Chinese	897	955	193	205	1090	1160
Jpanese	128	124	1	1.2	129	125.2

It can be seen from Table.3 that the total languages speaker change a little after 10 years.

5.1.4 Result and Analysis

The data given by the title shows that at present, the top five languages in all languages are Chinese, English, Hindi, Spanish and Arabic. After consulting the literature, we consider these countries' current total population, level of economic development, degree of opening to the outside world and other factors. We think the rankings of the top five will not change and we finally decide to rank the 6th to the 16th Language the total number of languages 50 years after. According to the first question predictive model, the results obtained are in Table.4.

Table 4: The Predictions of 6th-16th Language's Total Numbers of Speakers

	L1		L2		Total	
	Now	In 50 Years	Now	In 50 Years	Now	In 50 Years
Malay	77	107	204	271	281	378
Bengali	242	340	19	22	261	362
Russian	153	183	113	158	267	341
Portuguese	218	297	11	16	229	313
French	76	101	153	203	229	304
Hausa	85	132	65	93	150	225
Punjabi	148	192			148	192
German	76	106	52	69	129	175
Persian	60	84	61	85	121	169
Japanese	128	164	1	1.3	129	165.3
Swahili	16	22	91	131	107	153

We can compare this result with the ranking of Ethnologue in the previous years, and find that the top ten rankings of the total number of languages in previous years are also basically stable, which shows that the ranking of the total number of language users has not changed. The condition is related to many factors.

1. The country corresponding to the top-ten language can be divided into two parts. Some countries have a small population but have developed economically. Some countries have an underdeveloped economy but have a lager population base.

- As for the first kind of countries like America, they have strong comprehensive national strength so their native language has become world official language, the other country should learn their language to carry out business and other activities.
- As for the second kind of countries like China, their people have superiority, so the number of native language is hard to be exceeded by other countries.

2. Since the countries ranked after the tenth in the rankings are not economy powerhouse and have not a large population base, growth in both native speakers and second language speakers is not high.

3. We have ignored the small probability events, such as national split, war, invasion by other countries and other factors, so population model is the ideal model, so there will be such a result.

6 Model II

With a total population of 224 countries and regions in the world, some countries have a small population and a few other countries has very few immigrants, the impact on the first 26 languages is very small. Therefore, we only discuss forty countries with large populations and high immigration rates including China, India, the United States, Indonesia, Brazil, Pakistan, Nigeria, Bangladesh, Russia, Japan, Mexico, Ethiopia, the Philippines, Vietnam, Egypt, Germany, Iran, Turkey, the Democratic Republic of the Congo, Thailand, France, Great Britain, Italy, South Africa, Myanmar, Tanzania, South Korea, Columbus, Spain, Kenya, Argentina, Ukraine, Uganda, Sudan, Algeria, Poland, Canada, Iraq, Morocco, Afghanistan. These 40 countries account for 85% of the world's total population, while the remaining 15% are in more than 180 countries. Each country has a small population and a small impact on the entire population. Therefore, it is reasonable to do so.

For this question, we use the one-year world population migration data we found(cite), using a BA network model to calculate the total degree O_i of 40 countries. Based on the rank of O_i, we stratify the 40 countries into four layers. We calculate the Data function parameter by using the data we have found in World Bank. Finally we get the four function of the total immigrant population of each country

6.1 Prediction Model

1) BA Model

We select the top 10 countries in these 40 countries by the total degree. Then we use giving weights method to assign them 10 to 1, to indicate the score in this weighting network. The country with ordinate as the outflow of countries, the abscissa for the inflow country, enter the top 10 of the relationship data.

Nodes centrality degree is expressed as the number of edges connected to that node. In this paper, for a country which has indegree and outdegree, can be simplified as directed weighted network.

Definition T_{ij} is the coefficient of the population that flows from country i to

country j, T_{ji} indicates the opposite vector value.

$$R_i = \sum_j T_{ji}$$

$$C_i = \sum_j T_{ij}$$

,

where R_i indicate the indegree of country i and C_i indicate the outdegree of country i, There are directions in the network, and there will be no equal flow of population between countries, therefore $T_{ij} \neq T_{ji}, R_i \neq C_i$.

In order to study the connection strength and flow strength of nodes in a network, we define $O_i = R_i + C_i$ as the total degree of a node which is regarded as the degree of the node being in the center of the network.

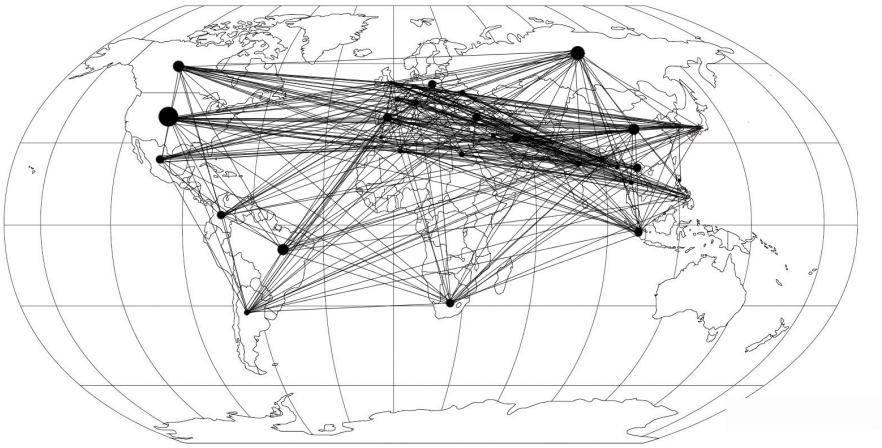


Figure 5: Global Population Mobility Network Space Pattern

According to the rankings of all national nodes O_i, it is found that the inter-country population mobility network shows a clear hierarchical level. In order to enhance the homogeneity of countries in the same level and the differences between different levels, we divided the 40 countries into four levels using the natural breakpoint classification method as is shown in Table.5.

Table 5: The Different Layer of The Forty Countries

Hierarchy	Country				
The First Network Layer	America Germany	Mexico China	Russian	Ukraine	India
The Second Network Layer	Bangladesh Italy	Pakistan Philippines	U.K Iran	France	Canada
The Third Network Layer	Turkey Morocco	Spain Japan	Afghanistan Viet Nam	Algeria Korea	Poland
The Forth Network Layer	Brazil South Africa Indonesia Ethiopia	Colombia Nigeria Sudan	Argentina Thailand Egypt	Iraq Tanzania Kenya	Congo Myanmar Uganda

6.2 Applications of Our Models

7 Strengths and Weaknesses

7.1 Strengths

-
-
-

7.2 Weaknesses

-
-
-

References

[1]

[2]

[3]

[4]

Appendices

Appendix A First appendix

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

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

Input matlab source:

```
function [t,seat,aisle]=OI6Sim(n,target,seated)
pab=rand(1,n);
for i=1:n
    if pab(i)<0.4
       aisleTime(i)=0;
    else
        aisleTime(i)=trirnd(3.2,7.1,38.7);
    end
end
```

Appendix B Second appendix

some more text **Input C++ source:**

```
//============================================================================
// Name        : Sudoku.cpp
// Author      : wzlf11
// Version     : a.0
// Copyright   : Your copyright notice
// Description : Sudoku in C++.
//============================================================================

#include <iostream>
#include <cstdlib>
#include <ctime>

using namespace std;
```

```
int table[9][9];

int main() {

    for(int i = 0; i < 9; i++){
        table[0][i] = i + 1;
    }

    srand((unsigned int)time(NULL));

    shuffle((int *)&table[0], 9);

    while(!put_line(1))
    {
        shuffle((int *)&table[0], 9);
    }

    for(int x = 0; x < 9; x++){
        for(int y = 0; y < 9; y++){
            cout << table[x][y] << " ";
        }

        cout << endl;
    }

    return 0;
}
```