Imperial College
London

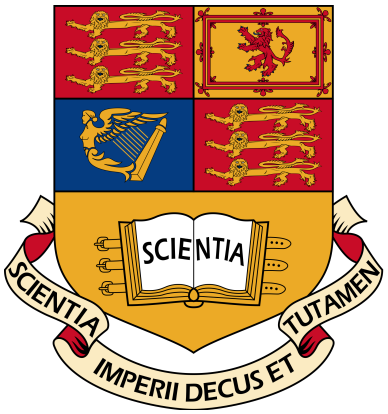
IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Research on the prediction of Chinese domestic migration rates and the Approximation for intention of migrants through Data Mining

Author:
Jiawei Li

Supervisor:
Chao Wu



Submitted in partial fulfillment of the requirements for the MSc degree in Type of Degree of Imperial College London

September 2016

Abstract

This study builds models for domestic migration in China. Using methodologies like random forest, artificial neural networks, maximal information coefficient, principle component analysis and K-means. Based on surveys conducted in each year for migrants in China, this paper employs data mining approach to (1) Building models for predictions over migration rates; (2) Specifying places in China that migrants are intending for traveling to.

The models in this paper achieves a high accuracy of predicting inter-provincial migration rates in China and considerable accurate over intentions of migrational families by providing a list of potential areas for their destinations. These models are built mostly by the sight of data but not theories in demography, which suggests the models built in this paper can be trained and tested in other related analysis without much prior knowledge automatically.

Acknowledgments

This research is supported by Dr Chao Wu who provided insight and expertise that to assist this research. Thanks for Imperial College, London for the research documents in this area that provided.

Thanks for Dr Yue Liu from Zhejiang University for assistance with methodologies in demography and Prof. Hong Mi for comments that greatly improved results in this research.

I shall also show gratitude to the my colleagues from Imperial College, London for sharing their advice of this research, although they might not totally agree with all the conclusions of this paper.

Contents

Chapter 1

Introduction

With the fast growing export industry, 140 million of Chinese citizens are under rural-to-urban migrants (10 per cent of the total population) in leaving their home provinces for seeking for better opportunities for jobs and better environment for families(1). According to the International Labor Organization, China is now the country with the largest number of domestic migration(33). This paper is to use data mine methods for analyzing this missive domestic migration.

1.1 Motivation

Various reasons such like economic situation in difference provinces, cultural differences, accents, distances from home and policies can cause them for migrations.

Traditional researches used to proposes theories that focus on the influence casted by only few variables to domestic migration. Most of these approaches proposed theories that successfully illustrate certain parameters and their contribution to Chinese internal migration but limits at the number of parameters of analysis.

In addition, traditional approaches require huge prior knowledge especially in the process of choosing parameters, the limitation of prior knowledge could directly cause the variable that not correlated to the targets to be chosen.

To overcome these limitations, a data mining approach for analyzing internal migration in China, this approach can select relevant variables and can discover patterns from them to extract migration information with understandable structures and without much prior knowledge should be used.

This paper is to build models that can semi-automatically analyzing internal migration with as little prior knowledge as possible with data mining methods.

1.2 Aim

To build models that can successfully cast predictions on Chinese internal migration, this paper should take consideration of the following facts that might interfere the process.

In spite of the difference of policies among provinces, the analysis of Chinese internal migration needs to take the impact of policies into consideration. Super large cities like Shanghai and Beijing are implementing policies for controlling the total amount of migrators each year by enhancing the requirement for migrators to be registered as local residences, which is also called the Hukou (Residence Registration) policy. Some cities, however, took actions in welcoming migrators from other provinces, especially migrators with high education background. Since Hukou policies play important role in migrations, an internal migration model in China need considering these policies. Understandingly, these policies are hard to be represented as numbers, which makes building models based on priori knowledge difficult in this area. To get rid of this problem, instead of using the data directly represent the policy impact, this paper utilizes economics and environmental data to indirectly represent it, which means, this paper is not using any direct policy data as priori knowledge for analyzing migrations but utilizing many other data that will be effected by the policy and, at the same time, highly related to migrations for modeling.

When considering the behaviour of immigration, the hidden residence as mentioned by Robert is effective for the statistics analysis. Many Chinese citizens migrate for 1-year or even short time jobs but not for residence, this causes the real situation of migrators are hard to be represented by the data of local changes in registered residence. In this paper, with the contribution of National Bureau of Statistics of China, employs several investigated data sets from 2009 to 2014. These data are conducted by thousands of face-to-face surveys for migrators between 2009 to 2014, which can more clearly present the true situations for individual migrator, especially for those without residence registrations.

In addition, the different cultures from various places inside China remain highly influent for migrations. A destination with a different culture, or with an accent that can not even be understandable for one migrator clearly cannot be attractable. The process of finding how different culture exact change the decision for one migrator could be unavailable since the description of one culture cannot be represented as numbers. In this case, the historical migration rates with distances between provinces can be used for representing the culture effect of one destination to migrators since with similar valuables in other attributes, the higher the migration rate is, the more similar culture they may share tends to.

This paper is to establish models for analyzing migration attentions between provinces and individual intention for migration without much priori knowledge. Rather than using some attributes that have been considered as related to migration by priori knowledge for further analyzing, this paper use various attributes and only select few from them for further analysis by calculating and comparing the maximal information coefficient scores before further classification algorithms and Artificial Neural Network (ANN) applied. Since one can select as many attributes as possible and the model will select potentially correlated features automatically, the process should build with little priori knowledge, which enables related problems in this area using the same process.

1.3 Contribution

This paper proposed approaches to make prediction on Chinese internal migration rates and the migrational intention of each migrator.

With the process in this paper, using statistics data of each region and interviewing data for migrators, parameters that can be selected from a larger set of potential parameters and predictions can be made based on values of these parameters. The results show good performance of the process.

Throughout the whole process, the amount of prior knowledge has been reduced as much as possible in order to achieve high accuracy. Due to the reduced prior knowledge, the process that built models for Chinese internal migration can also be employed into domestic migration in other countries, especially in other developing countries. This means, the approach this paper proposed is a common solution that could be used in other similar analysis.

Chapter 2

Background

The project focuses on the domestic migrations between provinces and major cities inside China. Data analyzed in this project mainly collected from the National Bureau of Statistics of the People's Republic of China, which includes the annual information of economy and the environment in each area of China and also few collections of individual data from 2009 to 2014 each records around 200,000 families, which migrates from one province to another. Papers about migration models, population growing models and related policies of the Chinese government are also used in this project.

2.1 Migration Models

Based on the migration data from each country, various migration models have been constructed. One of the famous ones is called Harris-Todaro model, which analyzes how the economic environments in each state influence the migrations and conducts into the conclusion that the government should control the number of migrators into cities since it would cause loss of jobs(10).

Another famous model is called Lee's Law, which divides factors into Pull Factors and Push Factors. The pull factors are positive factors that attract people from other areas into one certain area and the push factors, reversely, the push factors are reasons that push locals out of a certain area(11).

In 2006, a new theory named Bauder's regulation of labor markets has been published. It changes conventional view that market changes migrations into a new theory that migration regulates the labor market(12).

All the models above are originally built for migrations in countries other than China. Specifically, in China, the distinguished culture and policies could lead the domestic migration to differ from the rest of the world(30).

Typically, in the research for Chinese domestic migration rates, Jin Lu proposed a theory for that the inter provincial migration will largely influence the economy in different provinces and held the opinion that if the migration rate tends to 0, the growth of economy would be

stopped. In addition, he analyzed the influence cast by migration to cities and rural area and suggested that the migration will trend to be a dynamic balance finally(13).

Rather than focusing on inter provincial migration, Xiaoling Yuan uses the Shaanxi Province, which locates in the west part of China for researching and to analyze relationships between economy and migration rate inside one province. She proposes that the increasing inner provincial migration rate will balance the inner differences of economy among areas inside the province and there is a large correlation between economy, investment and migration(14).

In 2008, Zhaoyuan Xu analyzed the influence casted by domestic migration in China towards the differences in economy of each area. He proposed that even if the economy can be influenced by number of migrators floating in, the dynamics of capital still leads the difference of difference provinces remain stable(15). This conclusion supports theory that provided by PZ Duan, who shows the influence of migration to differences in growth of economy in different area is mighty between 1970 to 1980, however, after the 1990s, the influence is becoming increasingly weak(16).

2.2 Research Methods

As can be found in these models, most models require huge amount of prior knowledge and lack in consideration towards the influence of policies. However, in China, policies are changing quickly and some policies that reducing the floating of population will cast a huge influence on migration. This paper proposes models that only require little priori knowledge and uses data in economical and environmental fields that to represent the local situation including the policies.

In addition, most researches introduced above do not have a large detailed data set to support theories due to the difficulties in collecting and processing data at these time. This paper will focus on the analyzing of data without hypothesis and only get conclusions from the data mining results.

The choice initial parameters that for analyzing of domestic migration in China is influenced by the researches listed above. After selecting these parameters, this paper will illustrate a list of methods that could contribute to the building of the model.

The collections of Individual data, around 200,000 items of individual records which collected each year from 2009 to 2014 by National Bureau of Statistics is used in this project for analyzing people from each places and each income level from China. One item in this data set represents one family instead of one person and all items are collected averagely from each province. In each item, family income, expenses, health conditions and educational level of each family number are recorded. From this data set, general types of migrators inside China and be clustered for further analysis and correlations of each feature can be conducted using methods like the maximal information coefficient (MIC).

Another data set used is the provincial data from National Bureau of Statistics. The provincial data includes data about the general economic and environmental information about provinces in China. Some part of this data can be used together with the previous individual

data as the background data of provinces. Details of the data sets will be illustrated later.

One background knowledge that important to this paper is the Hukou System in China. Hukou system is a household registration system for citizens in mainland China and Taiwan. The household registration is the identification of a household for the resident area along with the recording of family information like parents and spouse.

In China today, although citizens do not need to get the registration or hukou for finding a job, the registered resident area still influence the citizen by the welfare offered by local government and in some mega cities, citizens without the local registration will be extra charged for buying houses. The hukou system actually reduces the floating of labours heavily and will influence the process of analysis in this paper .

Chapter 3

Design

Almost all meaningful analysis of internal migration in China begins with the analysis of the impact of variables, such as economic development, policies, health-care and insurance onto the migration figures(20). However, in traditional methods of demographic analysis, to propose assumptions for the impact from several variables to internal migration requires much priori knowledge and will be time-consuming for researchers for analyzing each variable themselves.

3.1 Parameters Selection

To improve the process of internal migration analysis, a more automatic method is necessary. This method should be with the ability for selecting variables from received provided data, which are most correlated to target variable relation types, which is migration rate or the destination for one individual migrator each year in this paper, and be possible to reduce dimensions if necessary, in case of large amount of variables or data set for further reduction and, finally, be a method for comparing various data mining algorithms for picking the best algorithm for a particular data set. This paper provides such an automatic process for migration analysis and use it for the Chinese internal migration research. The whole process contains the following particular algorithms.

Maximal information coefficient (MIC) is a measure of a wild range of associations both functional and not. A given internal migration dataset contains plenty of variables, which may contain important, undiscovered relationships between them. To compare correlations between pairs of data, a score for relationships between variables is necessary to be calculated and sorted for picking up important variables for further research. As a statistical method with the property of generality and equitability, MIC shows great performance in providing scores for correlations. As a general algorithm, MIC can capture a large range of relationships between variables based on the idea that if a concrete relationship exists between variables, then on the scatterplot, a grid of these two variables can be drawn that partitions the data to encapsulate that relationship (2).

The MIC score and strength relationships between variables can be shown as table 3.1.

Relationship	Random	Linear	Cubic	Exponential	Categorical	Parabolic	Sinusoidal
MIC	0.18	1	1	1	1	1	1

Table 3.1: MIC Scores

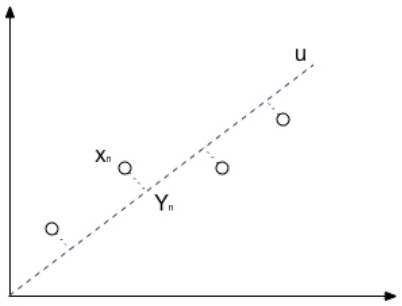


Figure 3.1: Principle Component Projection

A MIC score approximating to 1.00 denotes a very strong relationship between variable and a MIC score around 0.18 means the relationship is quite weak.

After MIC, if too many parameters have been selected, it is highly likely that they will be correlated and may lead the model over-fitting. In this case, a method that could select parameters from these pre-selected parameter should be used.

K-means can also be used for dimension reduction, it split features into groups and selects the most central feature in each group to represent the parameter in the whole group. It is employed in the process of predicting migration rates since the parameter that influences the migration rates is valuable to be known in order to analysis the importance of each parameter.

3.2 Principal Components Analysis

Principal component analysis (PCA) is a method for reducing dimensions for further analysis by converting a set of possible correlated data into a set of non-linearly correlated variables. The first few components calculated carries the larger variance (information). Then, the data will be projected into such components, which can maximise the mutual information between original and projected data(32). Taking a two-dimensional data for example, the direction u_1 is calculated to preserve the most information in the data set.

As Figure 3.1, the observations marked by X_n are projected onto a one-dimensional linear subspace with direction u, which can preserve largest possible information by one-dimensional subspace.

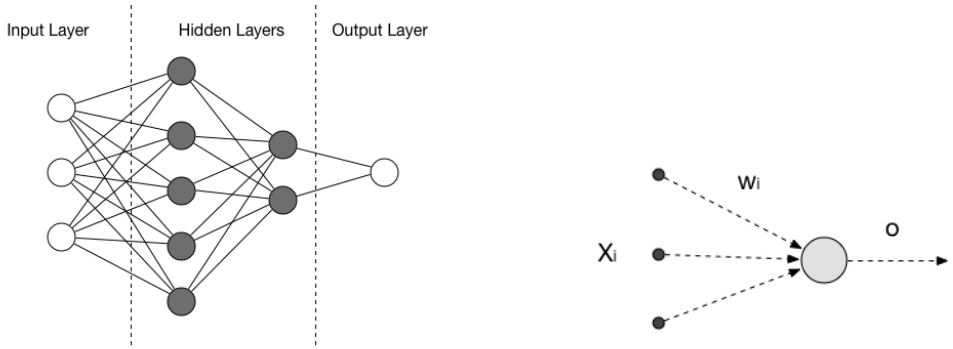


Figure 3.2: Neural Network and Perceptron

More formally, for a data matrix

$$X = [x1]...[xn]^T \tag{3.1}$$

Mean normalization should be done first for reducing the effects of different ranges of variables by computing the mean and replacing all data points xi with the following equation to result in a data covariance matrix $\bar{x}$.

$$\bar{x} = x_i - \mu \tag{3.2}$$

Then, calculating the eigenvectors, which represent the variance of each dimension calculated by PCA, and the corresponding eigenvalues of the data covariance matrix.

$$S = \frac{1}{N}\bar{x}^T\bar{x} \tag{3.3}$$

After obtaining all eigenvectors and eigenvalues, a decision of the total amount of dimensions M is required. To choose the principal subspace, simply choose the M largest eigenvalues as shown in the equation and their associating eigenvectors, which are the basis of the principal subspace.

$$U = [u_1, u_2, ..., u_M] \tag{3.4}$$

Finally, the projected vector can be calculated by

$$UU^T(x - \mu) + \mu \tag{3.5}$$

The Artificial Neural Networks (ANN) are composed of nodes or units connected by links and each link also has a numeric weight $w_(i, j)$ associated with it, which determines the strength and sign of the connection (3) as shown in Figure 3.2.

In this paper, only feed-forward networks have been deployed, which have connections only in one direction.

With perceptron training rule, a simplified one hidden layer neural network (perceptron) as shown in Figure 3.2 can be trained in following steps(23):

1. Set weights randomly;

2. With one example come through, the output of the neural network can be calculated with the activation function and the weight between x_i and the output layer w_i by:

$$o = \sigma \sum_{i=0}^{n} x_i w_i \tag{3.6}$$

3. When an example is misclassified, change the weights with the following function using the weight between x_i and the output layer W_i, the learning rateη and the training set with vectors x_i with the corresponding targets t_i:

$$W_i \leftarrow W_i + \eta(t_i - o)x_i \tag{3.7}$$

After MIC, K-means and PCA, parameters can be selected and dimensions can be reduced automatically, which can reduce the requirement of priori knowledge, which used to be largely required by normal migration analysis.

In order to test if the result of one algorithm is really better than others, student t-test has been used to determine whether or not two sets of result data is significantly different from each other.

There are two types of t-test, the one-sample t-test and paired-sample t-test. One-sample t-test designed for returning decision for the null hypothesis that all the data in data sets come from normal distribution shows significant difference before and after some process. Paired-sample t-test is for a that come from independent random samples.

In this paper, paired-sample t-test is used for determining the difference in predictions by calculating the significance in difference of results after different algorithms. The t value can be calculated as(19)

$$t = \frac{\overline{X}_D - \mu_0}{\frac{S_D}{\sqrt{n}}} \tag{3.8}$$

3.3 Layouts of Models

In this paper, the resulted models are conducted in the following process.

The model for predicting migration rates between places:

1. Collecting data upon migration rates and other related variables as many as possible;
2. Calculating MIC scores between each variable and migration rate. Sorting the results and choose M most correlated variables for further analysis;
3. using K-Means to reduce the M variables into N features.
4. Using the N variables in various classification algorithms like Bagging, Linear Regression and Multilayer Perceptron. Choose the algorithm that with the largest RMS score;

5. Using the model from step 3 for further prediction.

The model for predicting migration destination of individual migrators.

1. Collecting data for individual migrators and their family situations and using the average GDP per capita in the destinations to represent them;
2. Employing PCA in the data set to reduce dimensions;
3. Running Artificial Neural Network (ANN) with the data after PCA;
4. Predicting the destination of individual migrator by the nearest province or city with similar GDP per capita.

Chapter 4

Data Description

Most data used in this paper is conducted by National Bureau of Statistics of China, the official agency under the State Council of China for collecting data related to the economy, population and society of the China at global and local levels (4).

4.1 Data Sets

There are basically three parts of data used in this paper.

The first set is the one that describes local economy, environment and population. This data set contains more than 20 dimensions, which can be collected from the database in National Bureau of Statistics. This data set illustrates the general economical, environmental circumstances of each province and direct-controlled municipalities, which are cities with status equal to provinces in China, which is similar to federal district. All the data in this data set is presented by the year of collected and the province or direct-controlled municipality it belongs to, which means the analysis related to this data set can only be regulated to the migration status between provinces and direct-controlled municipalities but not specified cities or counties. All data from this data set is publicized by National Bureau of Statistics of China

The second one is the unchanging geological data of provinces and direct-controlled municipalities. The most important dimension is the distances between provinces, which records the distance between the capital cities of each province and direct-controlled municipalities.

The last one is the interview data conducted by the National Bureau of Statistics collected from thousands of migration families. This huge interviewing project was taken each year from 2009 to 2014. The samples in this data set are selected by local officers, who also worked closely with migrators for asking questions as documented in the questionnaires and recording their answers. The total number of samples from each province or direct-controlled municipality are kept equal, which means the analysis of this whole data can provide information towards the whole country without uncontrolled emphases on a certain geological area.

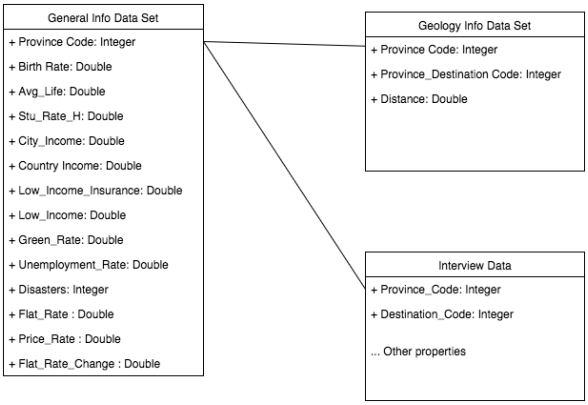


Figure 4.1: Data Sets

Since the questionnaires taken each year has slightly difference between each other, when selecting variables from the whole variable set for building models that not just suitable within one-year data, only common variables (data related to common questions) through out the 6 questionnaires within 5 years (2 different questionnaires are taken in 2010).

The form of data sets are shown in the following Figure 4.1. Due to the amount of parameters in Interview data, detailed columns of this data set is not shown below. Some of the columns of this data sets will be listed in the following chapters.

4.2 Missing Data

There are two types of missing data in this data sets. One is the missing data that caused by the design of questionnaires and some are caused by the methodology.

In spite of the detail on the questionnaires, more than 200 questions are listed. Some of them are optional questions that might not be required to be recorded according to prior answers, which means the amount of missing data is not ignorable in this data set. For this reason, the particular missing data problem in this data set is different from normal missing data problem.

Generally, missing data are caused by the misreporting in surveys but the missing data in this data set are data that should be missing according to the prior questions or the circumstances of the migrator's families such like the whether the migrator has a partner or not.

Since missing data under certain questions may carry some information, common methods like adding the mean of a certain column to the missing data can misrepresent the true information carried by missing data. In this paper, the missing data are added according to the questions of them, which will be further discussed later in this paper.

Another type of missing data in this data set is caused by the miss-collection of interview data. In the data sets for interview data later than 2011, there is no data about migrators that travel to many provinces. The data missing for certain provinces are listed below in Table 4.1.

Table 4.1: Number of provinces missed from interview data

Number of	2011	2012	2013	2014
Missing Provinces	1	12	23	20
Existing Provinces	30	19	8	11

As can be shown in the table, the data later than 2011 fails in containing data that collected from all provinces with migrators that travelled to.

For this reason, in the further process of prediction, the prediction of inner provincial migration rate for provinces and the prediction of inter provincial migration rate with the destination province that without any data collected locally directly should not proceed.

To prevent the miss-prediction and miss-training that caused by the migration rate 0.00 between some provinces that have no data collected. In the process of selecting data for building models, the migration rates with only the destination province that has been collected data from exists in the data sets. This means, there is no migration rate with 0.00 for these pairs with destinations that not been collected data from.

However, this problem shows little influence in inter provincial migration predictions, since the destination with missing data will be ignored and only counted provinces are used for inter provincial migration. This means, the numerator and denominator in the process of migration rates migration will not change.

Chapter 5

Prediction of Migration Rates

For Chinese internal migration prediction, there are 3 different kinds of migration rates, the overall migration rate, the inner provincial migration rate and the interprovincial migration rate(31).

The inner provincial migration rate is the rate for the number of migration families N_n that from one province divided by the population N_p, this can be described by the equation below

$$InnerRate = N_n/N_p \tag{5.1}$$

The interprovincial migration rate is the adverse term from the inner provincial migration rate. It measures the number of interprovincial migration family N_t divided by the population

$$InterRate = N_t/N_p \tag{5.2}$$

The term overall migration rate is the ratio between the amount of families come from their origin to their destination N_{od} divided by the total amount of migration families in this origin N_o

$$Rate = N_{od}/N_o \tag{5.3}$$

For feature research on estimation over these kinds of overall migration rate, a look at the differences in figures of interprovincial migration and inner-provincial migration would be helpful. For a quick analysis on these figures, this paper calculated the InnerRate N_n and InterRate N_t and draw these figures on the map of China for comparisons between figures in each province.

In the interview data sets, the comparison between interprovincial N_n and inner provincial migration families N_t in each province based on data collected in 2011 is be shown in Figure 5.1, which stands for inner provincial migration rates in 2011 and Figure 5.2 that represents the inter provincial migration rates in 2011.

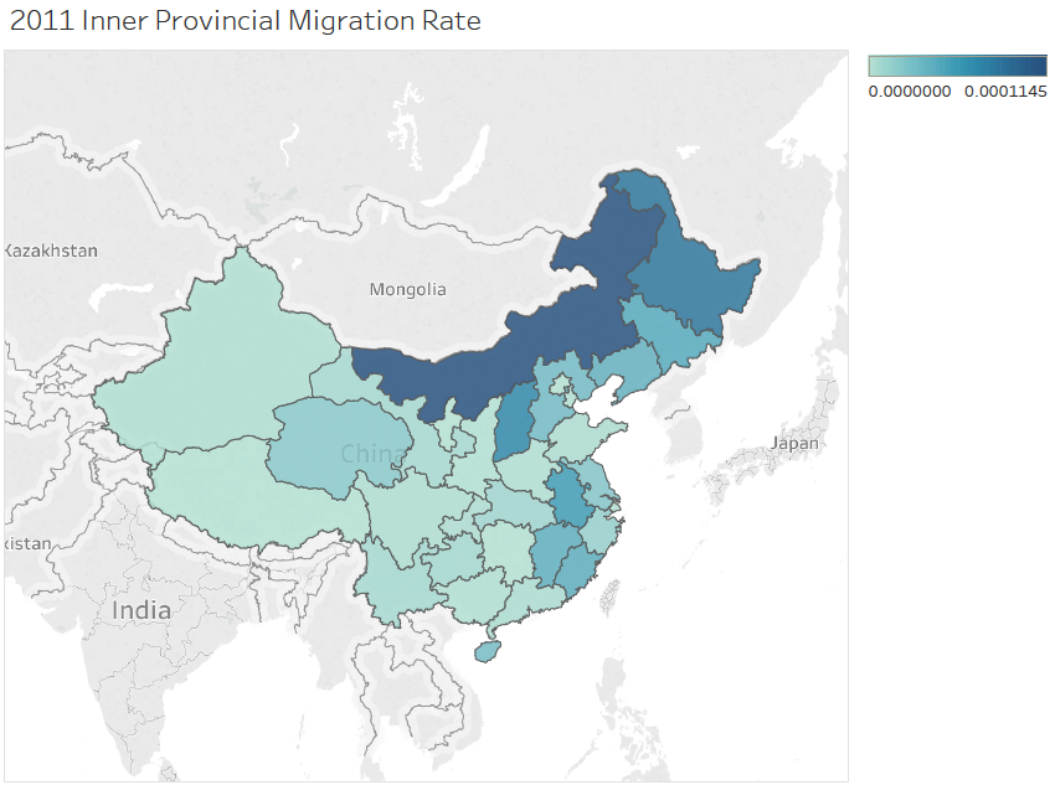


Figure 5.1: 2011 Inner Provincial Migration Rate

The InterRate and InnerRate are variant among provinces, among them, cities like Beijing and Shanghai hold the largest 'InterRate' while provinces like Hubei and Anhui have high 'InnerRate'.

This phenomenon can be caused by the differences within one province or direct-controlled municipality and with provinces nearby. For example, Zhejiang, a province that achieves relatively high InterRate and low InnerRate, has more healthy and positive economic environment than provinces nearby like Anhui, and considerable equal across the province(22). A place without much difference within the province and with relative apparent variance compared to places nearby intending to achieve a higher 'InterRate'.

Thus, the phenomenon on inner and inter provincial migration are different from each other and it may be helpful if they are analyzed separately. For confirming the assumption about the difference over inner and inter provincial migration rates, further maps towards the rates in different years are listed from Figure 5.3 to Figure 5.5.

It needs to be mentioned that data in these data sets do not include migrators that travel to all provinces in China. The provinces that drawn in these figures about inner provincial migration with data are provinces with inner provincial migration rate 0 but excludes Beijing, Tianjin, Shanghai and Chongqing.

From these graphs, the internal and inner provincial migration shows great difference between provinces but remain certain stability among years of collected. Thus, in the process below will be built on each data set for predictions.

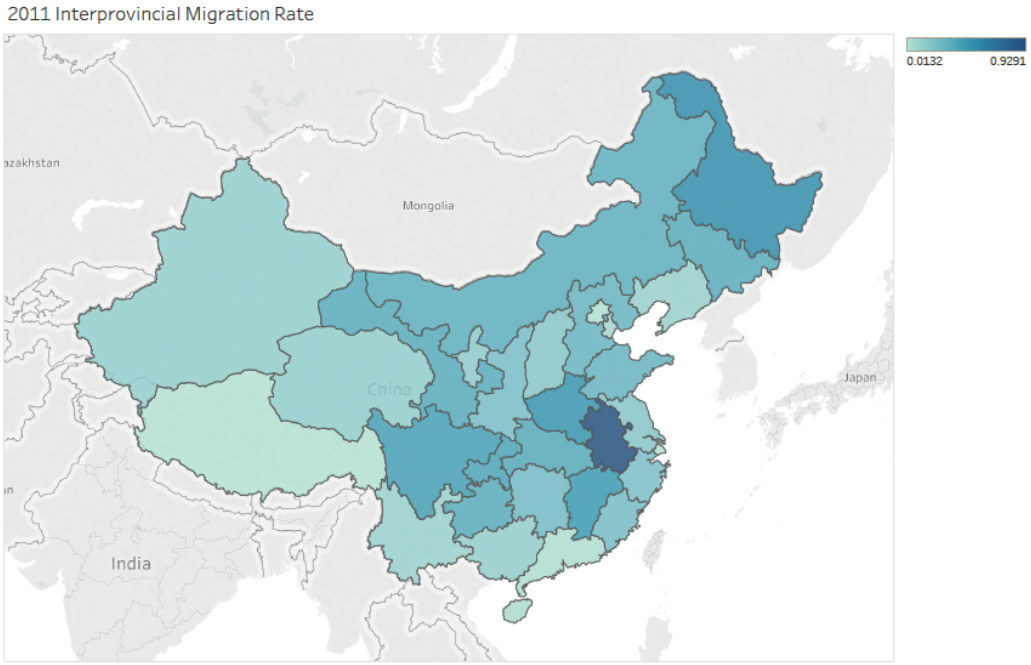


Figure 5.2: 2011 Inter Provincial Migration Rate

As listed before, the whole process of this chapter can be divided into four steps. Among them, the step of selecting parameters and the process of building models using selected parameters with the data sets are most important.

In the process of building models, this paper employs several different methods and test the accuracy for each before finally choosing the most fitted model. This choosing process can make the idea of this chapter be suitable not only for data mine for Chinese internal migration but also for any related project.

Typically, in this analysis process, each time for training models, an iteration number of 10 has been chosen while in each iteration. Before training, the data will be divided into 10 parts equally. In each iteration, 9 out of the 10 parts are selected to be the training set with

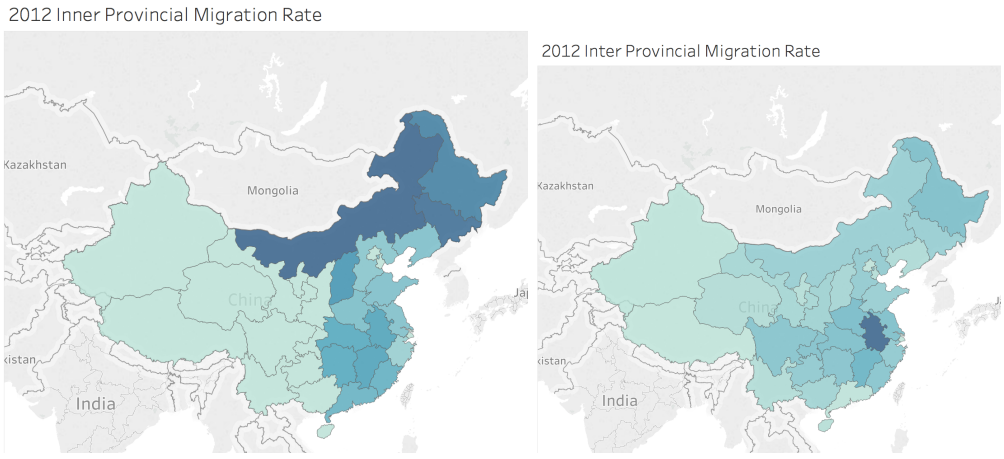


Figure 5.3: 2012 Provincial Migration Rate

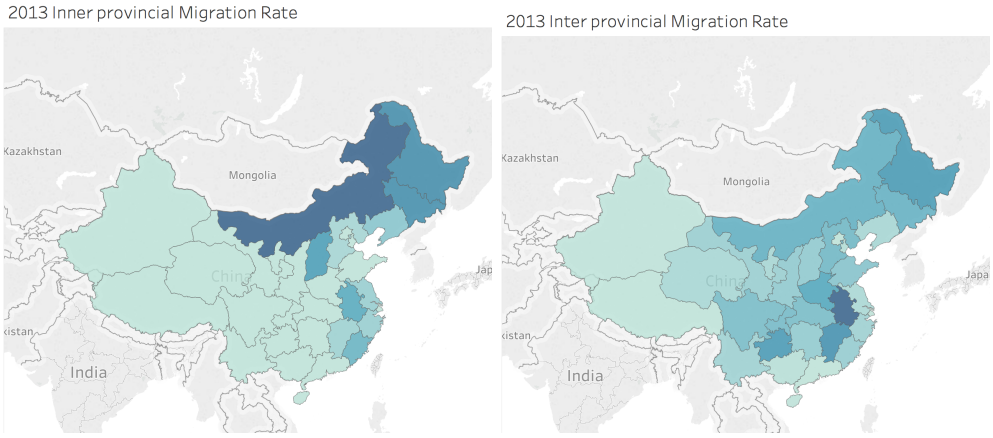


Figure 5.4: 2013 Provincial Migration Rate

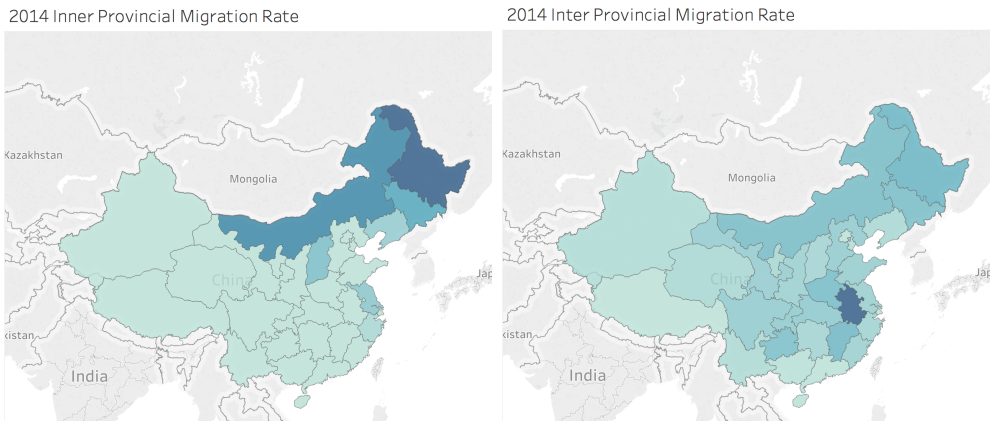


Figure 5.5: 2014 Provincial Migration Rate

the one part left to be the testing set.

This method forces the same training method to be trained and tested by 10 times rather than once and calculate the final accuracy by average accuracy among test sets as shown in Figure 5.6.

All the results for prediction in this chapter are trained and tested with this process.

5.1 Migration Data

There are mainly three parts of data are used for training and testing in this chapter: General Information Data Set, Geography Info Data Set and Interview Data set.

In this chapter, the data sets need to be connected to further analysis. The final processed data set should contain variables about the two provinces that migrators come from and go to with the corresponding migration rates and other general information data of the provinces.

The relationship between data sets is shown in Figure 5.7.

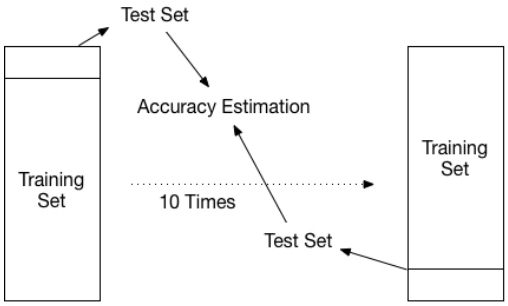


Figure 5.6: cross-validation

Unlike other data that can directly obtain from three data set provided, the migration rates between provinces have not been directly recorded. Unfortunately, there is no official record about the real amount of migrations between provinces. If there are records of migration rates in some provinces, the large amount of unregistered migrators (so-called hidden migrators) would lead the records inaccurate.

In this circumstance, the migration rates used in this paper are calculated by the interview data sets. Since data in these sets are real data about migrators no matter whether they have been registered or not, using this data set to calculated the migration rate can provide a more closely migration rate to the real figure. However, there would be one problem exists in this process. Since records this data set is only sampling from the whole migrators, errors might exist, especially to these provinces with small migration rate, the rate calculated might close to 0.

5.2 Selecting features

For predicting the future immigration rate, features must be selected to build the model. During the process of evaluating the relevance between feature and target variable, MIC scores are used to select the valuables that contribute more influence on immigration rate than others.

The initial features can be selected from data in various areas, which considered might be correlated with the variable to be predicted. With the increasing amount of features selected, the finally picked parameters will become worse based on the method of picking highest MIC scores. This means, for any related topic, researchers should find as many features as they can that might influence the result.

In this paper, the listing parameters are selected for MIC sorting with migration rate shown in Table 5.1

The MIC scores of all parameters in statistics data sets, to be simplified are shown in Table 5.2

Table 5.1: Parameters

orgGDP	The GDP per capita of the place immigrates from
Population	The population of the destiny
Birth_Rate	Birth Rate of the destiny
Avg_Life	Life expectancy
GDP_PC	GDP per capita of the destiny
Unemploy_Rate	Rate of unemployment
Flat_Rates	Rates of flats
Price_Rates	Rates of Price (last year as 100)
City_Income	Average annual income for citizens live in the city area
Country_income	Average annual income for citizens live in the rural area (Start publishing since 2013)
Green_Rate	Rate of Green land in city area (Start publishing since 2011)
Disasters	Amount of natural disasters happened in the area
Stu_Rate_H	Rate of citizens that obtains college certificates
Low_Income	Rate of citizens that are classified as low income
Low_Income_Insurance	Monthly insurance for low income citizens
Flat_Rates_Change	Change in average rates of flats

Table 5.2: MIC Scores

Parameters	MIC (strength)	Linear regression (p)
Population	0.67098	0.5396318
Birth_Rate	0.67098	-0.29075608
Avg_Life	0.67098	0.4073089
Flat_Rates	0.67098	0.53999525
City_Income	0.67098	0.5662188
Country_income	0.67098	0.5470502
Green_Rate	0.67071	0.30561876
GDP_PC	0.66989	0.48178324
Disasters	0.66989	-0.075478196
Stu_Rate_H	0.66989	0.3403138
Flat_Rates_Change	0.63526	-0.3043905
Low_Income_Insurance	0.62499	0.45970377
Low_Income	0.61332	-0.226193
Price_Rates	0.37312	-0.12579323
Unemploy_Rate	0.21025	-0.2366798
orgGDP	0.14091	-0.014871608

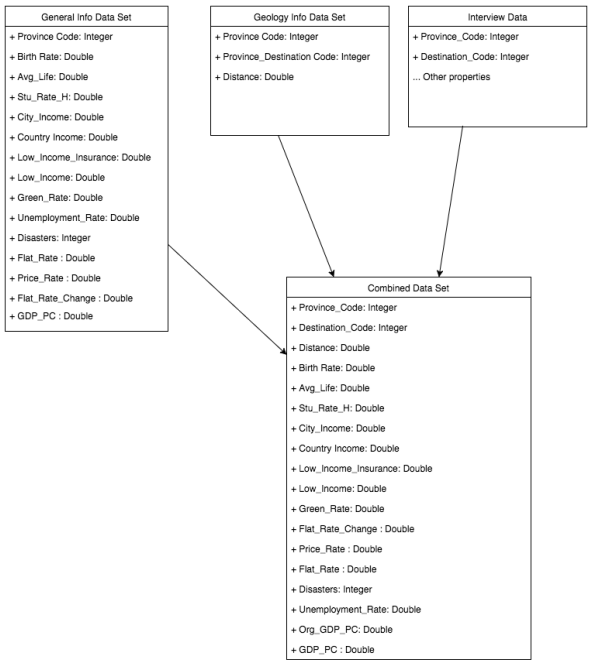


Figure 5.7: cross-validation

This table with the MIC scores of the data in 2012 shows the features selected are mostly relevant to migration rate of 2011. As can be shown in the table, $orgGDP$ and $Unemploy\_Rate$ obtain lower MIC scores than others, which illustrates they uncorrelated with migration rate in 2012. Particularly, in this paper, a threshold is set to be 0.5 in MIC scores.

In this case, parameters except $orgGDP$ and $Unemploy\_Rate$ have been deleted for further analysis.

In addition to these parameters, features from the geological data set such as 'Distances', 'areaX', 'areaY' are also be used.

For further illustration of MIC and relationships between variables. A strong relationship between the amount of migration from each place to Beijing and the economic condition in their hometown is shown in Figure 5.8, in which the red line indicates the Highest GDP per capita value minus GDP per capita value and the blue line indicates the number of migrators and the total number of migrators has been used, which equals to migration ratio multiplied by total migrators from Beijing.

As shown in Image 1, the provinces with fewer province codes follow the case that more rich (indicated by GDP per capita) one province is, more people will move out of the province, which, follows the migration model proposed by Harris and Todaro (6), which illustrates that the development economics is the important reason for rural-urban migration.

The list of Province names represented by Province Codes is shown in Table 5.3.

However, interestingly, provinces with large province code like 31 has very few people moves out while the GDP per capita in those provinces stays low. My assumption for this situation is that those provinces are located in far west part of China, which means their distance to

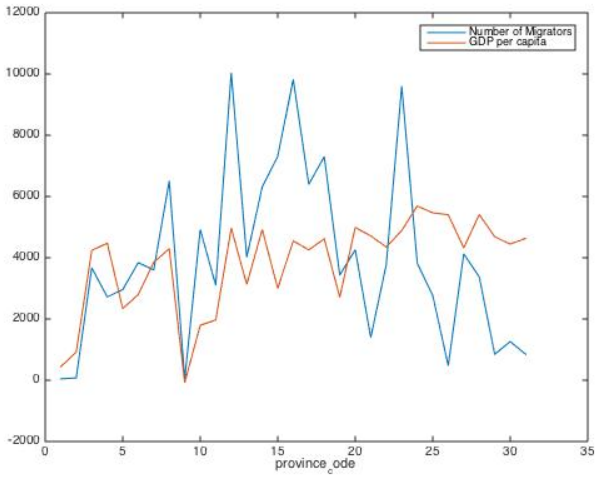


Figure 5.8: GDP Per Capita to Migration

Table 5.3: Province Codes

Code	Province	Code	Province	Code	Province
11	Beijing	42	Hubei	53	Yunnan
12	Tianjiang	43	Hunan	54	Xizang
13	Hebei	44	Guangdong	61	Shanxi
14	Shanxi	45	Guangxi	62	Gansu
15	Neimenggu	35	Fujian	63	Qinhai
21	Liaoning	36	Jiangxi	64	Ningxia
22	Jilin	37	Shandong	65	Xinjiang
23	Heilongjiang	41	Henan	53	Yunnan
31	Shanghai	46	Hainan	51	Sichuan
32	Jiangsu	50	Chongqing	52	Guizhou
33	Zhejiang	34	Anhui		

major cities with high GDP per capita such as Beijing and Shanghai are quite far. Since this, people may feel unfavourable for migrating to richer areas typically, in east coast of China.

With comparing the MIC score, the scores for pairs of variables in the data set are not high enough to judge one single variable could strongly influence the others. However, among those pairs, the most relevant one is the pairs between the location migrators see the doctors with other features like monthly income and even the amount of children.

The scores of the relationship between Income and Number of Children are shown in Table 5.4.

Against the idea from most of the papers that the number of children of one family will be influenced by the monthly income of the family, the MIC score is just around 0.00586, which is one out of then when compared with the MIC scores with locations for migrators to see doctors.

In addition, relationships between the amount of people leaving the provinces with their

Table 5.4: MIC Scores

Scores	Income-No. of Children
MIC	0.00586
MIC-P2	0.005460929
Linear	0.01997676

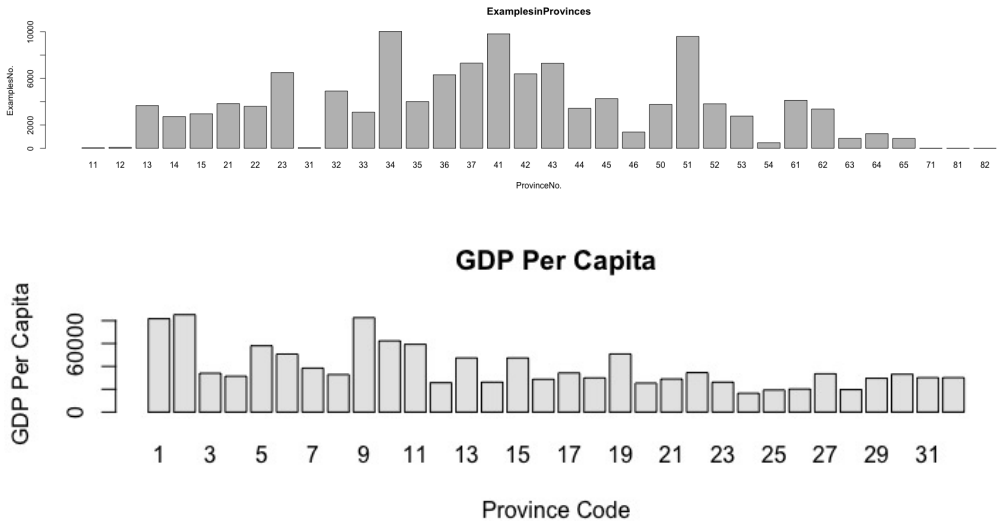


Figure 5.9: Number of Migrators and GDP Per Capita Recorded in Each Province

location also be founded. Some provinces like Anhui (which is the province with the largest number of citizens migrate out) and Sichuan (province with second largest number of migrators leaving the province), also their GDP per capita are not quite low, however, since they locate very close to major cities and very rich areas in China like Shanghai (highest GDP per capita in China) and Guangdong (largest GDP in total), citizens in those provinces seem more intended to move to the province next to them and earn higher salary. The population of migrators from each province show as Figure 5.9. (34: Anhui, 41: Sichuan)

For places like Hang Kong, Macau and Taiwan, since there is a border control between them and Mainland China, the migrators from those places are ignored in this project

In addition, taking $unemployment\_rate$, which has low MIC score for example. As can be judged from Figure 5.10, the relationship between the migration rate from Henan and unemployment rate in destinations are not quite related.

One more detail that should attract attention is that, in this paper, general data of 2011 is used to make prediction of migration rate of 2012 rather than 2011. This make sense in China due to the culture of Spring Festival (also named as Chinese New Year), which is the most important festival all year round.

Spring Festival is celebrated on the first day of the traditional lunisolar Chinese calendar each year, which is a date between January and February. In Spring Festival, all Chinese people are supposed to travel back home for a holiday for 7 to around 20 days. Most Chinese people, especially migrators are leaving their job in cities and only move to seek job opportunities

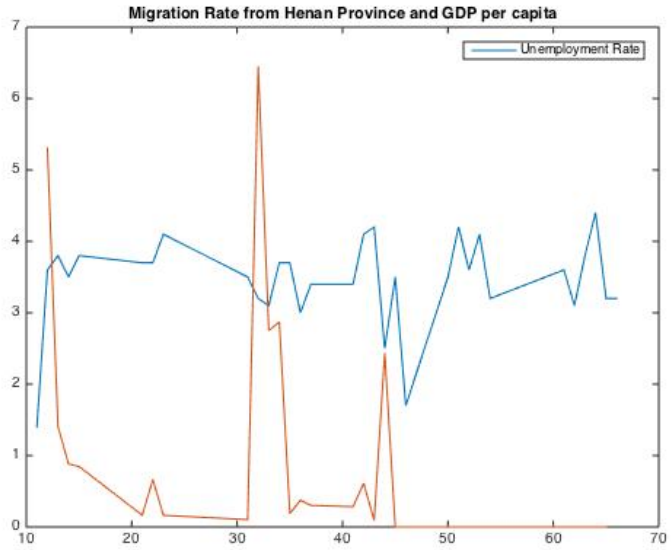


Figure 5.10: Parameters with Low MIC Score

after the holiday. Normally, most companies who employ those migrators from rural places are supposed to pay the salary to employees from rural places according to their contract, which marks an end the contract and possibly an end to one-time migration. After the Spring Festival, these people leave their hometown again and head to the same or different big cities from last year for jobs. This tradition makes the Spring Festival become an end of migration and a start of a brand new migration(29).

Considering the impact of Spring Festival, when most migrators decide their destination for the upcoming year, they will depend on the data of years before rather than years in the future. Thus, only the analysis of the data in past few years make sense to this project.

One more parameter that has be deleted from the pre-selected parameters table is the Flat_Rate_Change due to its unusual performance. This deletion requires prior knowledge as following.

Generally, the change of flat rates should be positively related to the migration rate. This is because, generally speaking, the increasing amount of population in one city will cause the increasing of flat rate due to the increasing of need. The increase rate of flat price and its relationship with immigration rate is not a positive relation as most academies expected. The MIC shows there is a certain relation between them with a correlation score as 0.63626. However, if the data is analyzed with linear model the score is -0.3043905, which denotes that, generally, there is a certain negative relation between. This might due to the influence of politics, since the government take actions for reducing the high flat rate in big cities for improving the living environment for locals(28).

Another reason for deleting this parameter is that, the upcoming years after 2012, the relationship between changes in flat rates and migration rate turns to be positive. This may also due to the unpredicted policy changes within the upcoming years. Thus, taking this parameter for further analysis will mislead the models and cause over fitting problem.

With the same MIC calculator, a data set with all statistics data from 2009 to 2013 is used to predict the immigration rate in 2014 between provinces. There relevance to immigration rate is as Table 5.5:

Interestingly, two phenomena can be found from this table. Parameters with similar meanings in different years mostly obtain similar ranks and for these parameters, in more recently they are collected, the more correlated they are to the migration rate.

After this selection process, 18 parameters have been preserved. However, using so many variables could cause overfitting for the final results.

For example, variables like 'City_Income' , 'Country_Income' and 'GDP_PC' all represents the local economical environment and seem to be correlated to each other. To show the correlation between variables, the Table 5.6 shows MIC score between variables and GDP_PC in the year of 2011

As can be seen from this table, most variables are highly correlated. With correlated variables, the model could be easily over-fitted by training on much data that not carries much difference between them. This will cause the model fitting the noise that exists in the data rather than learning the hidden information.

Thus, when researching social topics that include feathers highly rely on others, it is important for not to use too many data with similar meaning for building models.

To prove the influence from each algorithm towards the accuracy, using models that will be built in next section, the parameters that used for prediction are deleted on for each time and the accuracy that without that parameter using Random Forest are calculated as shown in Table 5.7. The first row describs errors for predictions over all parameters and next few rows are for errors in predictions after picking one parameter out one by one. All errors are represented in the form of RMS.

As can be shown in this table, the performance of the algorithms is robust. Even without key variable that may influence a lot to the performance, Bagging and Random Forest can still achieve in good results. This is because the selection of parameters and examples during the training process for building decision trees.

However, still as presented from this table, the parameter 'Distance' plays important rule in the accuracy of the model. This proves the importance of the geology information data set. The reason for this huge decline in accuracy is the distances between provinces can hardly be represented or abruptly measured by other parameters.

Unlike other parameters like GDP per capita, which can be roughly represented by parameters like $City\_Income$, the $Distance$ is irrelevant from other parameters. By declining one parameter like GDP per capita, the missing information from GDP per capita can be provided by related variables in certain level. However, deleting 'Distance' can make the permanent loss of geology information that could not be represented by others. This is the reason for the decline by reducing using Distance as one of the parameters.

To reduce the influence of this problem, a method that can select parameters from the whole parameter sets should be used. In this paper, K-means is used in reducing the number of

Table 5.5: MIC Scores

Rank	variables	MIC Score	Rank	variabes	MIC Score
1	Birth_Rate_13	0.67098	33	Low_Income_Insurance_13	0.62499
2	Birth_Rate_10	0.67098	34	Low_Income_Insurance_12	0.62499
3	Avg_Life_13	0.67098	35	Low_Income_Insurance_11	0.62499
4	Avg_Life_12	0.67098	36	Low_Income_Insurance_10	0.62499
5	Avg_Life_11	0.67098	37	Low_Income_Insurance_09	0.62499
6	Avg_Life_10	0.67098	38	Birth_Rate_13	0.62499
7	Avg_Life_09	0.67098	39	Birth_Rate_12	0.61807
8	GDP_PC_09	0.67098	40	Birth_Rate_11	0.61807
9	City_Income_13	0.67098	41	Green_Rate_09	0.61807
10	City_Income_12	0.67098	42	Green_Rate_11	0.61332
11	City_Income_11	0.67098	43	Low_Income_13	0.61332
12	City_Income_10	0.67098	44	Low_Income_12	0.61332
13	City_Income_09	0.67098	45	Low_Income_11	0.61332
14	Green_Rate_13	0.67011	46	Low_Income_10	0.61332
15	GDP_PC_13	0.66989	47	Low_Income_09	0.61332
16	GDP_PC_12	0.66989	48	Birth_Rate_09	0.58389
17	GDP_PC_11	0.66989	49	Green_Rate_10	0.56935
18	GDP_PC_10	0.66989	50	Green_Rate_12	0.56422
19	Stu_Rate_H_13	0.66989	51	Unemploy_Rate_12	0.52076
20	Stu_Rate_H_12	0.66989	52	Unemploy_Rate_11	0.52015
21	Stu_Rate_H_11	0.66989	53	Disasters_09	0.45272
22	Stu_Rate_H_10	0.66989	54	Unemploy_Rate_13	0.43731
23	Stu_Rate_H_09	0.66989	55	Disasters_10	0.42073
24	Population_13	0.66936	56	Price_Rates_09	0.41893
25	Population_12	0.66936	57	Disasters_12	0.3995
26	Population_11	0.66936	58	Price_Rates_13	0.39465
27	Population_10	0.66936	59	Disasters_11	0.38736
28	Population_09	0.66936	60	Price_Rates_09	0.37312
29	Country_income_13	0.66878	61	Price_Rates_12	0.32843
30	Country_income_12	0.66878	62	Unemploy_Rate_09	0.22774
31	Country_income_11	0.66878	63	Unemploy_Rate_10	0.21025
32	Country_income_10	0.66878	64	Price_Rates_11	0.20614

Table 5.6: MIC Scores

Variable	MIC	Variable	MIC
Disasters	0.9769869	Stu_Rate_H	0.42304355
Price_Rates	0.9338638	Population	0.3854314
Unemploy_Rate	0.9103233	Flat_Rates	0.3843422
Green_Rate	0.88238555	Avg_Life	0.37655085
Low_Income	0.83737725	Low_Income_Insurance	0.22350341
Flat_Rates_Change	0.78134805	City_Income	0.2158271
Birth_Rate	0.60136974	Country_income	0.15467566
amount	0.5621874	fromCode	0.001148369

Table 5.7: Decline Dimensions

Parameter	Bagging	Random Forest
Origin	0.0174	0.0163
Distance	0.02	0.0193
Population	0.0174	0.0164
Birth_Rate	0.0176	0.0162
Avg_Life	0.0175	0.0164
Area	0.0174	0.0169
Disasters	0.0174	0.0164
City_Income	0.0175	0.0164
Country_Income	0.0179	0.0163

parameters to get rid of over-fitting. This process will be illustrated in next few sections.

The further analysis in this chapter will be based on parameters that listed in this table.

5.3 Modelling

As mentioned before, 6 methods are used in accuracy comparison in this section, they are: Bagging, Multilayer Perceptron, Additive Regression, REP Tree, Decision Stump and Linear Regression.

5.3.1 Bagging

Bagging, short for Bootstrap Aggregating, is the method to generate several predictors from selecting data into different subsets randomly for several times and combine them into an aggregated predictor. To combine predictors, the aggregation calculates the mean over versions of predictors when making prediction.

Based on the idea that forming multiple versions of predictors by making bootstrap replicates of the whole learning set and using these as new learning sets. Bagging can give substantial gains in accuracy as proved by Leo Breiman in 1996 (5).

Based on the subsets generated in bagging aggregation, decision trees are built to classify the data and make predictions for further input.

Decision Tree (DTree) is a support tool using a dendritic graph or model of decisions and their possible consequences (7). It is a tree with leaf nodes, which has a class label, determined by majority vote of training examples reaching the leaf and internal nodes, which carry questions of features.

In this paper, C4.5 is used for building the trees after bagging aggregations. C4.5 is a method for choosing features in each internal nodes, which is similar to the commonly used ID3 method. The difference happens when deciding the chosen feathers in each inter node after

Table 5.8: Performance

Method	RMS	Mean absolute error
Bagging	0.0404	0.017

the calculation of information gain. Instead of using information gain for choosing parameters, the C4.5 uses an information ratio as quainter. The C4.5 defines split information with 'D' for total amount of examples in this node and D_i for the amount of examples with feature A valued as I and n stands for the total kinds of values in this feature (18) :

$$split\_info(A) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} log_2 \frac{|D_i|}{|D|} \tag{5.4}$$

After obtaining the split information, the information gain ratio can be calculated as:

$$infoGain\_ratio(A) = \frac{gain(A)}{split\_info(A)} \tag{5.5}$$

The value of gain(A) is calculated with the same way in ID3, which is:

$$gain(A) = info(D) - info_A(D) \tag{5.6}$$

The process of this algorithm can be shown in Figure 5.11

With the algorithms as mentioned before, the statistics data of 2010 is used below in building models for migration rate in 2011 to show the accuracy of each learning methods towards the data in name of overall migration rate.

The accuracy of each training method with the use of RMS(Root Mean Square) to comparing between different classifiers for predicting Immigration Rate is shown in Table 5.8:

Using this method for making prediction with statistics data of 2010 for migration to 2012 and compare them with the migration rates collected and calculated from the interview data set.

To indicate the errors, a bar plot with errors from predictions is shown in Figure 5.12

With errors as follow: Mean absolute error 0.0225 Root mean squared error 0.0604

As shown above, the Bagging aggregation with C4.5 makes prediction for 2 years with high precision with only around 0.05 in RMS. This method can provide predictions for future migration rates between provinces for policy makers.

To test the performance of this algorithm on $InnerRate$ and $InterRate$, subsets of migration data for inner provincial and inter provincial respectively are used in training and testing following the method above.

The results in Table 5.9 shows the RMS in the following sequence:

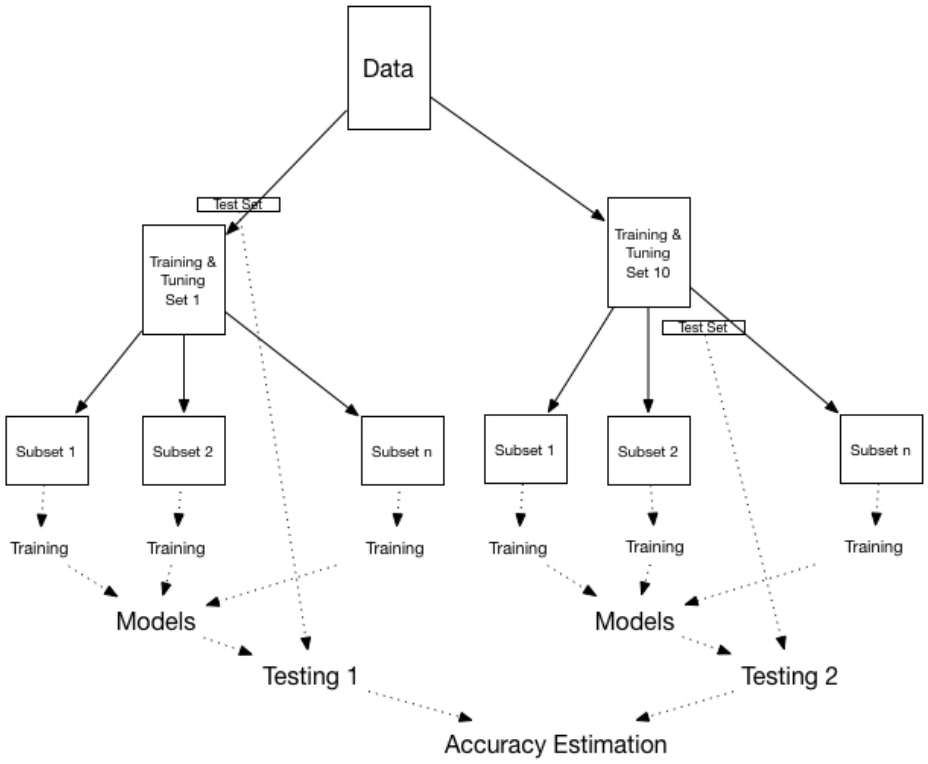


Figure 5.11: Process

Table 5.9: Inter Provincial Migration Rates of 2011 with Bagging

Method	2011Inter	2012Inter	2012InterAlone
Bagging	0.0174	0.032	0.0329

1. 2011 InterRate prediction using 10-folder cross-validation with statistics data on in 2010;
2. 2012 InterRate prediction using the same model built for 2011 InterRate prediction but with statistics data in 2011;
3. 2012 InterRate prediction using 10-folder cross-validation with statistics data on in 2011;

The result in Table 5.10 shows the RMS for $InnerRate$ in the following sequence:

1. 2011 InnerRate prediction using 10-folder cross-validation with statistics data on in 2010;
2. 2012 InnerRate prediction using the same model built for 2011 InnerRate prediction but with statistics data in 2011;

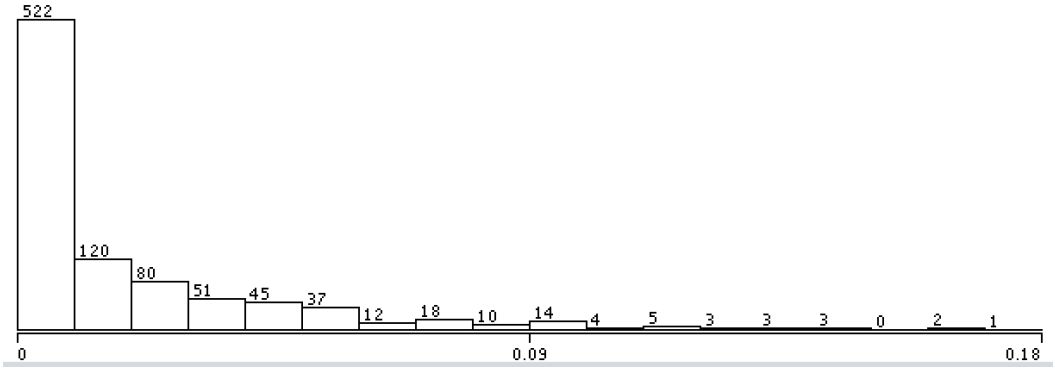


Figure 5.12: Error

Table 5.10: Inner Provincial Migration Rates of 2011 with Bagging

Method	2011Inner	2012Inner	2012InnerAlone
Bagging	0.2982	0.2609	0.2982

3. 2012 InnerRate prediction using 10-folder cross-validation with statistics data on in 2011;

To find the best method for building models for internal migration, several algorithms are also employed.

5.3.2 Linear Regression

Linear Regression is also selected as potential algorithm for building models for internal migration in China since most variables listed in the chapter for MIC calculation also obtains positive scores in linear regression. Especially, variables like city_income and population seem to have a relatively strong positive correlation with migration rates.

As a model that assumes all variables $\{x_1, x_2, ..., x_n\}$ follow linear relation towards the target variable y, linear regression can be expressed by the following equation with the denotes the small and acceptable noise.

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \tag{5.7}$$

More generally, for a set of samples with values $\{x_{1,i}, x_{2,i}, ..., x_{n,i}\}, i = 1, , p$ and target values $\{y_i\}, i = 1, , p$. The model takes the form

$$y = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_n x_{n,i} + \varepsilon_i, i = 1, , p \tag{5.8}$$

To be simplified, set $X = \{x_{1,i}, x_{2,i}, ..., x_{n,i}\}, i = 1, , p$ and $\{y_i\}, i = 1, , p$. Using ordinary least squares, which is a common estimator that reduce the total amount of squared

Table 5.11: Performance

Method	RMS	Mean absolute error
Linear Regression	0.045	0.0215

Table 5.12: Inter Provincial Migration Rates of 2011 with using Linear Regression

Method	2011Inter	2012Inter	2012InterAlone
Linear Regression	0.0204	0.0385	0.039

residuals, the equation for the estimated value of the parameter set β, which is short for $\{\beta_1, \beta_{2,\ldots}, \beta_n\}$ can be expressed as

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{5.9}$$

Employing linear regression to the 2010 statistics data and 2011 migration rates, the RMS and mean absolute error are shown in Table 5.11

Using this method for making prediction with statistics data of 2010 for migration to 2012 and compare them from the migration rates collected and calculated from the interview data set. The erros are:

Mean absolute error 0.0261 Root mean squared error 0.067

This is slightly worse than Bagging.

To test the performance of this algorithm on $InnerRate$ and $InterRate$, subsets of migration data for inner provincial and inter provincial respectively are used in training and testing following the method above.

The results in Table table: Inter Provincial Migration Rates of 2011 with Linear Regression shows the RMS in the following sequence:

1. 2011 InterRate prediction using 10-folder cross-validation with statistics data on in 2010 using Linear Regression;
2. 2012 InterRate prediction using the same model built for 2011 InterRate prediction but with statistics data in 2011 using Linear Regression;
3. 2012 InterRate prediction using 10-folder cross-validation with statistics data on in 2011 using Linear Regression;

The result in table 5.13 shows the RMS for $InnerRate$ in the following sequence:

1. 2011 InnerRate prediction using 10-folder cross-validation with statistics data on in 2010;
2. 2012 InnerRate prediction using the same model built for 2011 InnerRate prediction but with statistics data in 2011;

Table 5.13: Inner Provincial Migration Rates of 2011 with Linear Regression

Method	2011Inner	2012Inner	2012InnerAlone
Linear Regression	0.3681	0.2255	0.3681

3. 2012 InnerRate prediction using 10-folder cross-validation with statistics data on in 2011;

To find the best method for building models for internal migration, several algorithms are also employed.

5.3.3 Random Forest

Similar to bagging, Random Forest also uses the idea of splitting data sets for training. Moreover, Random Forest does also select feathers while building the trees for each subset.

Random forest was proposed by Breiman in 2001 as an extension for $bagging$. As described by Breiman, random forests are a group of decision trees such that each individual tree replies on the values of a random vector, which is sampled independently and with the same distribution for all trees in the forest (9).

Unlike the last algorithm, random forests change the way that decision trees are constructed. In standard decision trees, the node splitting methods use information of the node for finding the best splitting decision among all variables in the data set. However, in building decision forests, only the best feathers among the subset of predictors, which are randomly chosen over the nodes can be used for splitting the nodes. This strategy turns out to obtain a very well performance among many other classifiers as mentioned by Breiman in 2011(8).

The algorithms follow the steps below:

1. Using bagging for selecting subsets n_{tree} from the original data;
2. For each subset n_{tree} that generated by the bagging, generate a decision tree (using C4.5 in this paper). During the processing of generating, at each internal node, rather than choosing the best features among all predictors as described in Bagging, Random Forest selects a random sample m_{try} of the features and only use information gain ratio to select the best feature among them;
3. In the process of prediction after building the random forest, predict by aggregating the predictions made by the generated trees.

Specifically, in the following building process, the size of each bag and total number of iterations are all set to be 100.

The whole process is shown in Figure 5.13

As shown in the graph, if the subsets of features m_try are set to be the same as the whole feature set, random forests will be the same as bagging aggregating, which means that bagging

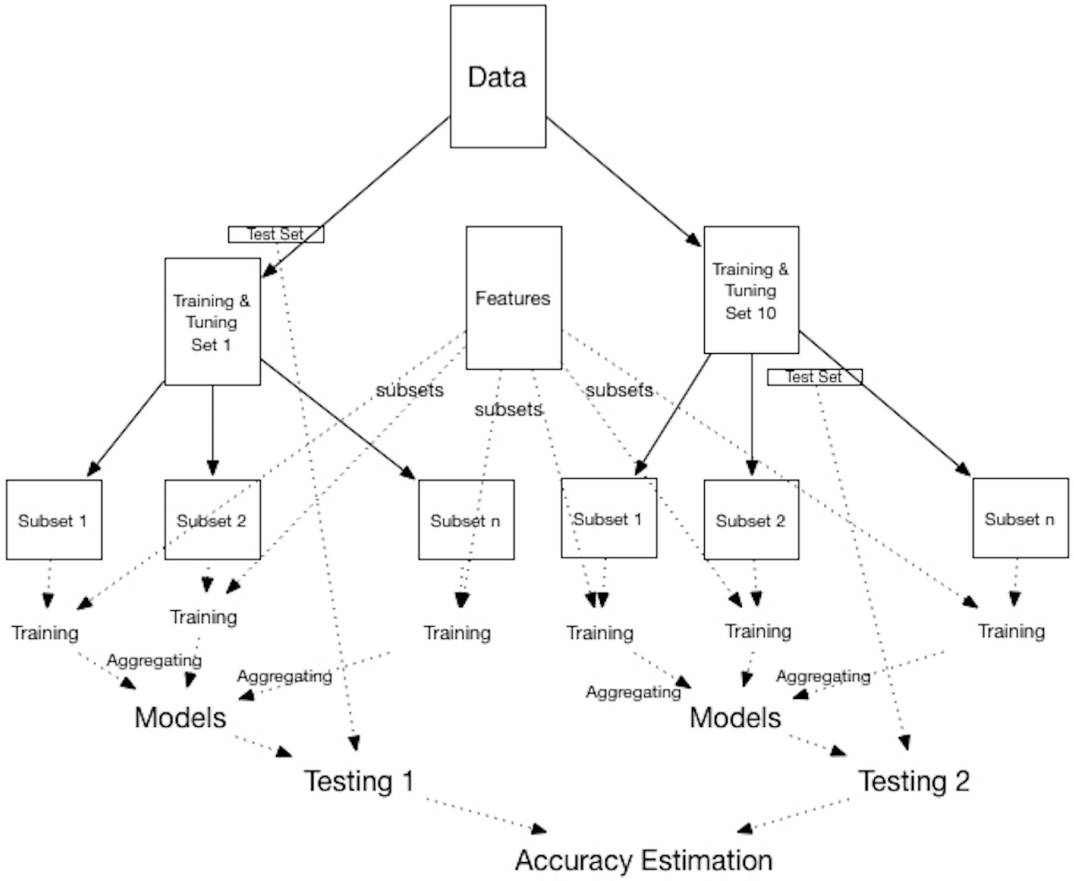


Figure 5.13: Process

Table 5.14: Performance

Method	RMS	Mean absolute error
Random Forests	0.0397	0.0169

can be taken as a special case of the algorithm of random forests.

After applying the algorithm of random forests on statistics data of 2010 and migration rates in 2011, the results are listed in table 6.6:

As shown in the table, in this data set, random forests have a better performance over bagging.

Following the steps of prior algorithms, the prediction accuracy of real data using the model just built based on statistics data of 2010 and migration rates in 2011 to predict the migration rates in 2012. The result is:

Mean absolute error 0.0201 Root mean squared error 0.0572

The result shows the random forests just built perform slightly better than bagging.

To test the performance of this algorithm on $InnerRate$ and $InterRate$, subsets of migration

Table 5.15: Inter Provincial Migration Rates of 2011 with using Random Forest

Method	2011Inter	2012Inter	2012InterAlone
Random Forest	0.0163	0.0305	0.0285

Table 5.16: Inner Provincial Migration Rates of 2011 with Random Forest

Method	2011Inner	2012Inner	2012InnerAlone
Random Forest	0.2851	0.3419	0.2851

data for inner provincial and inter provincial respectively are used in training and testing following the method above.

The results in Table 5.15 shows the RMS in the following sequence:

1. 2011 InterRate prediction using 10-folder cross-validation with statistics data on in 2010 using Random Forest;
2. 2012 InterRate prediction using the same model built for 2011 InterRate prediction but with statistics data in 2011 using Random Forest;
3. 2012 InterRate prediction using 10-folder cross-validation with statistics data on in 2011 using Random Forest;

The result in Table 5.16 shows the RMS for $InnerRate$ in the following sequence:

1. 2011 InnerRate prediction using 10-folder cross-validation with statistics data on in 2010;
2. 2012 InnerRate prediction using the same model built for 2011 InnerRate prediction but with statistics data in 2011;
3. 2012 InnerRate prediction using 10-folder cross-validation with statistics data on in 2011;

5.3.4 Neural Network

The feedforward neural network is a kind of neural network, which obtains no cycle among nodes.

In this chapter, a multi-layer perceptron is used for predicting the migration rates between provinces. A simple 2 layer perceptron can be shown as Figure 5.14.

As shown in the map, the number of nodes in each layer of the neural network need to be set specifically. Selecting the number of nodes in each hidden layer can directly interfere the accuracy of the whole model. However, up to now, there is no universal method for deciding the number of nodes in each layer before running the algorithm.

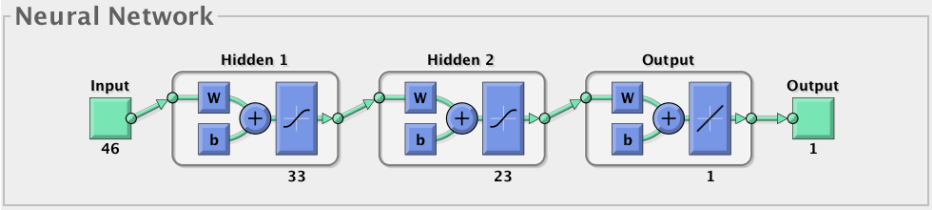


Figure 5.14: ANN

In this paper, a method of running the neural network with the combination of 1 to 40 nodes in hidden layers and evaluating the corresponding accuracy has ben employed in this paper to make judgment of the best number of nodes in each layer.

A library in Matlab is used for building neural networks in this chapter. With the use of this library, several different multilayer neural network algorithms have been employed. As to find the best algorithm fitted for Chinese internal migration, 4 neural network methods have been run one by one for each combination of 1 to 40 nodes in hidden layers.

To equally verified the accuracy for each iteration and each neural network method, since Matlab will automatically and randomly divide the data set into training set and test set before each iteration, which will definitely cause the difference in accuracy estimation even for the same method and number of nodes in each layer(25). This paper uses the whole data set as testing set, which controls the test set to be the same for all four different algorithms and for all iterations.

The four neural network algorithms are:

1. Gradient descent backpropagation (traingd) Parameter: learning rate;
2. Gradient descent with adaptive learning rate backpropagation (traingda) Parameters: learning rate and ratio increase/decrease learning rate;
3. Gradient descent with momentum backpropagation (traingdm) Parameters: learning rate, momentum constant.
4. Resilient backpropagation (trainrp) Parameters: Increment/Decrement to weight change.

To run four method in Matlab, simply set the network by following steps(27):

1. Set number of nodes in each layer: net = feedforwardnet([i,j]);
2. Set maximum training episodes by net.trainParam.epochs = 1000;
3. Configure the neural network by net = configure(net, Input, Output);
4. Train the network by net = trainrp(net, Input, Output);
5. Test the neural network by y = net(Input); p = perform(net,Output,y); with the whole data set

Table 5.17: Performance

Method	RMS	Nodes in 1st layer	Nodes in 2nd layer
Gradient descent backpropagation	0.07	1	2
Gradient descent with adaptive learning rate backpropagation	0.0583	7	17
Gradient descent with momentum backpropagation	0.007	7	1
Resilient backpropagation	0.045	10	2

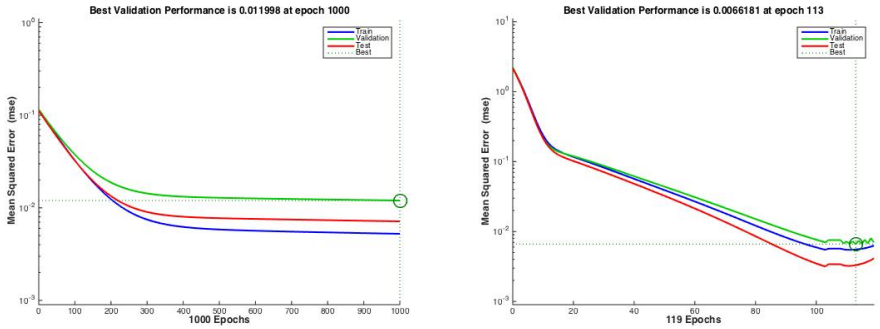


Figure 5.15: Gradient descent backpropagation (Left) and adaptive learning rate backpropagation (Right)

After setting maximum epochs as 1000 and training and testing all 1600 * 4 neural networks in Matlab, the result of using statistics data in 2010 and migration rate in 2011 is shown in Table 6.7:

With the regression from Figure 5.15 to Figure 5.16:

One point needed to be mentioned is that Gradient descent backpropagation and Gradient descent with momentum backpropagation has not reached the regression at 1000 epochs.

After using the 2010 data for 2012 migration rates, the result is:

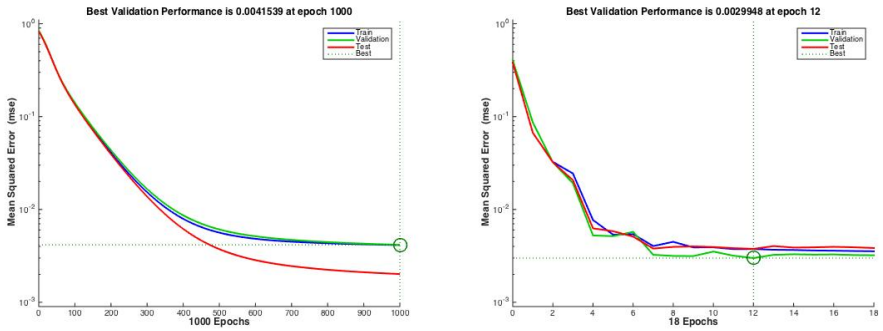


Figure 5.16: Gradient descent with momentum backpropagation (Left) and Resilient backpropagation (Right)

Table 5.18: Inter Provincial Migration Rates of 2011 with using Random Forest

Method	2011Inter	2012Inter	2012InterAlone
Neural Networks	0.0281	0.038	0.0385

Table 5.19: Inner Provincial Migration Rates of 2011 with Random Forest

Method	2011Inner	2012Inner	2012InnerAlone
Neural Networks	0.5195	0.0848	0.5195

RMS: 0.0624

To test the performance of this algorithm on $InnerRate$ and $InterRate$, subsets of migration data for inner provincial and inter provincial respectively are used in training and testing following the method above.

The results in Table 5.18 shows the RMS in the following sequence:

1. 2011 InterRate prediction using 10-folder cross-validation with statistics data on in 2010 using Neural Networks;
2. 2012 InterRate prediction using the same model built for 2011 InterRate prediction but with statistics data in 2011 using Neural Networks;
3. 2012 InterRate prediction using 10-folder cross-validation with statistics data on in 2011 using Neural Networks;

The result in Table 5.19 shows the RMS for $InnerRate$ in the following sequence:

1. 2011 InnerRate prediction using 10-folder cross-validation with statistics data on in 2010;
2. 2012 InnerRate prediction using the same model built for 2011 InnerRate prediction but with statistics data in 2011;
3. 2012 InnerRate prediction using 10-folder cross-validation with statistics data on in 2011;

5.4 Future Prediction

All the data mentioned in the prior chapter is the result of 1-year or 2-year prediction using only statistics data in one year. These results can reflect the accuracy and efficiency of each method used before. In this fast developing society of China, using more history data for future prediction or making predictions for future 5 more years become more meaningful.

Table 5.20: Performance with all history data

Method	RMS	Mean absolute error
Linear Regression	0.0587	0.0223
Random Forest	0.0491	0.0185
Neural Network	0.0684	0.0385
Bagging	0.0527	0.0197

Table 5.21: Performance with data in 2013

Method	RMS	Mean absolute error
Linear Regression	0.0604	0.0277
Random Forest	0.0491	0.0185
Neural Network	0.064	0.0383
Bagging	0.053	0.0198

As mentioned before in the chapter of calculating MIC scores, the potential variables for model building and predictions making has been posted. Based on the data within that table, a model that based on data in more than one years can be used in building models.

Since the data set mentioned before contains only data between 2009 to 2013. Using migration rate of 2014, which can be calculated by the interview data set of 2014, the comparison can be made by building models with same algorithms based on data from 2009 to 2013 and with the data of 2013 alone, to check the influence of the amount of history data.

After building the models with the methods mentioned above, the accuracy of each method can be calculated as Table 5.20

And with the data of 2013, the result is presented in Table 5.21

From these figures, predictions based on statistics data for more than one years can improve the performance of linear regression. However, the improvement on random forest and bagging are so slightly that can be ignored. Neither, for feedforward neural network, the data for more than one year cast negative effect on the precision of prediction as the shown in Figure 5.17.

This is an unusual phenomenon since most theories propose that in internal migration, more ages of statistics data employed in migration or relevant problem should improve the performance. More generously, most theories upon analysis on migration encourage researchers for applying more data into models to help the algorithms in learning potential information and relationships between variables(24). However, the data comparison shows that for some methods in this project, always employing a larger amount of data seems not a robust method for improving the performance.

There are two reasons for this problem.

1. Values of each variable scattered in each year for the same place show no distinct changes among years. As shown in the MIC scores, the relevance of most variables in various years in same area shows tiny differences between each other, which means, using the most recent value of the variable can about to represent the variable in each

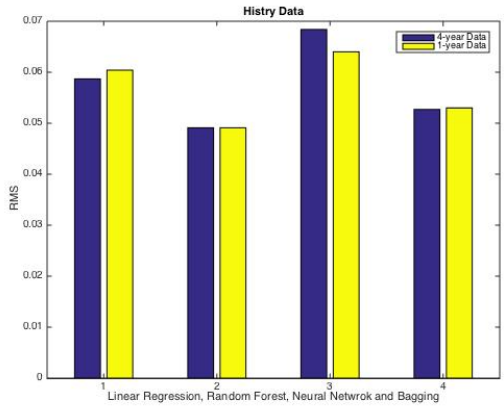


Figure 5.17: Performance Comparing

years. One more phenomenon to support this is that, for most variables, the most recent value (2013 in this case) obtains the highest MIC scores compared with the same variable in all other years, this illustrates that the most recent value be slightly more relevant to migration rate and values in other years may contain more irrelevant values that may interfere the result;

2. Over-fitting problem will occur when passing too much data(26).

For the second reason mentioned above, random forest and bagging contain ideology of preventing over-fitting by cutting of training set or cutting of features. Thus, for these methods, passing more features will be hard to interfere the accuracy to the already regressed model.

For the feedforward neural networks, adding data will cause an increase in validating accuracy, however, when testing with the test set although they might have already been over-fitting. Thus, when testing the model, the over-fitted model will behave worse than expected. For example, the Figure 5.18 shows an over-fitting neural network's training process, the best parameters for validating are not the same as for testing set.

As shown in the Figure 5.18 , the weights set at 69 epochs show the best among performance using validating data, however, while tested by testing data, it is slightly worse than the performance of the weight set around 63.

5.5 Evaluation

This chapter enhances four different kinds of algorithms and with four different propagation methods for feedforward artificial neural networks in the process of migration rate estimation. The prior process has built models based on several sets of statistics data in China with these algorithms and compared the performance among them.

After Predicting based on different number of variables, a set of 5 parameters can always achieve the best result.

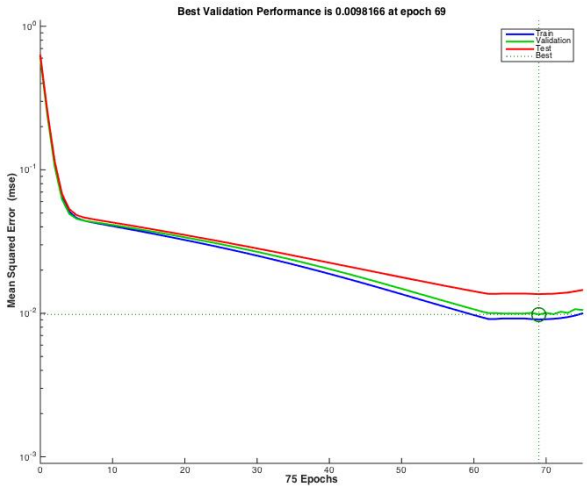


Figure 5.18: Overfitting Network

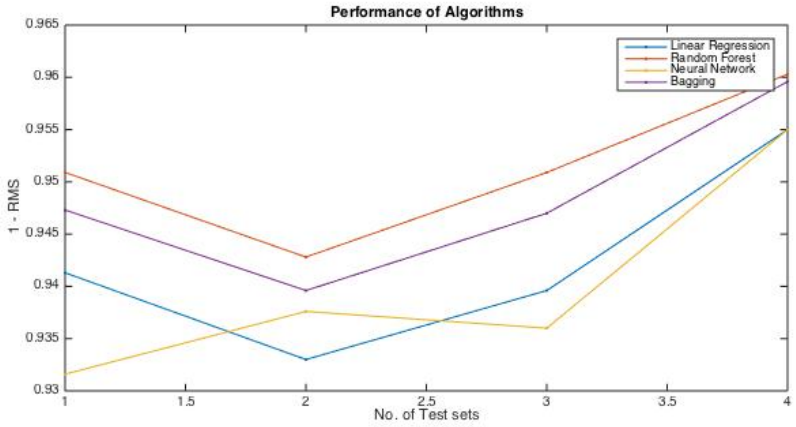


Figure 5.19: Performance of Algorithms

Based on the results shown in this chapter, random forest stays on top on the accuracy rank of all four kinds of algorithms, with slightly higher accuracy than Bagging. The average RMS is shown in Figure 5.19.

As shown in the graph, random forest obtains best accuracy can always perform better than other algorithms with bagging following it.

The general performance of these methods shows positive result towards the prediction of migration rates. As shown in 5.19, RMS for most algorithms are around 0.05.

Pacifically, for prediction the migration rates in China, random forest is the best in performance.

One point that should be mentioned here is whether or not values of all parameters should be selected for both origin and destinations (duel selection).

As mentioned before, the variables selected in this section are mostly used for training and

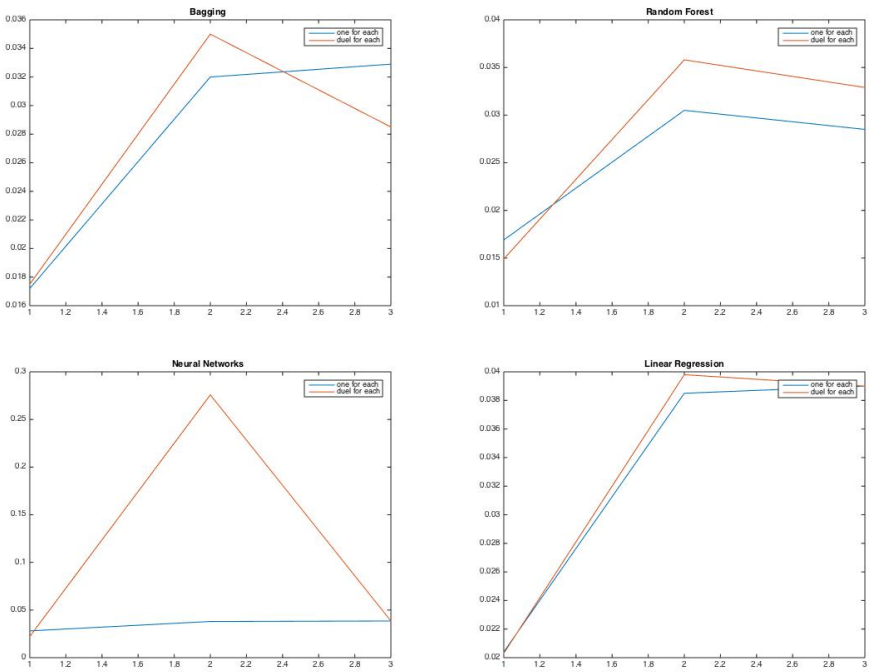


Figure 5.20: Performance on one or duel parameters

testing with both the value for either origin or value for destination. However, it is also possible that values of those parameters for both origins and destinations should be used for these models. This means, for one variable like $Price\_Rate$, two values are actually presented in the data sets, one for origin and one for destination. To test if this idea is meaningful in improving the accuracy, the comparisons are made shown in Figure 5.20 for the accuracy in prediction for 2011 and 2012 inter provincial migration data.

As can be shown Figure 5.20, the overall performance shows no apparent differences before and after changes. Taking values of parameters in both origin and destination seems to contribute little to the performance and would some times cast negative impact on it.

For this reason, parameters for destinations is not necessary be used for the models.

For the comparison of inter provincial and inner provincial migration rate prediction, as shown in Table 5.22, the accuracy of inner provincial migration rate prediction is obviously worse than the inter provincial prediction.

The difference between the accuracies are divergent, to make sure that the differences of these results are not caused by noise or chances, an extra t-test has been conducted into the results shown above.

Since the two sets of data record samples that not from the same group, a group t-test has been used.

The t-test shows the result reject the null hypothesis at the default 5% significance level with the p value of $2.3251 * 10^{-8}$.

Table 5.22: Inner and Inter Provincial Prediction Accuracy

Method	2011Inner	2012Inner	2012InnerAlone
Bagging	0.2982	0.2609	0.2982
Random Forest	0.2851	0.3419	0.2851
ANN	0.5195	0.0848	0.5195
Linear Regression	0.3681	0.2255	0.3681
Method	2011Inter	2012Inter	2012InterAlone
Bagging	0.0174	0.032	0.0329
Random Forest	0.0163	0.0305	0.0285
ANN	0.0281	0.038	0.0385
Linear Regression	0.0204	0.0385	0.039

As confirmed from the result of t-test, the performance of the model on inner and inter provincial migration rate shows to be divergent. This phenomenon can be caused by the following reasons.

1. The absence of parameters in statistics data sets that represent the economical and environmental situation within provinces. Most parameters chosen in this chapter are provincial data. However, one of the key reasons for inner migrations is related to differences among different areas inside provinces. For example, Within Jiangsu province, which shows high inner provincial migration rate ($2.188885 * 10^{05}$), the economic north part of the province (Subei) has very large difference from the south part (Sunan), which locates near Shanghai. Since the detailed economical and environmental data is not shown in the data sets, the analysis can hardly be accurate for inner provincial migration rates;
2. The strategy of data collection in the interview data sets. All inner provincial migration rates are calculated from this data sets. To make the rates accurate, the interviews should be conducted by randomly selecting samples from the whole country, which means, the number of samples in each province should be different from each other in spite of the migration families. However, the total number of samples collected in each province are equal, which makes the total number of inner migration samples fails in representing the total number of real inner provincial migrators.

Thus, using the data sets provided can hardly produce the accurate prediction on inner provincial migration.

To predict the inner provincial migration rates, two additional data should be provided

1. Detailed data about the economy and environment in different sub areas within provinces;
2. Interview data that conducted by the selecting samples randomly in each area based on its number of migrators.

Since the parameters used are not independent, the results would be over-fitting. To reduce the parameters, K-means is used to find the most representative parameters.

In case for migration predictions, given a set of parameters $(x_1, x_2, ..., x_n)$, K-means aim at finding k sets $S_1, S_2, ..., S_k$ that can partition all the parameters into so as to minimize the sum of squares as Equation 5.10

$$\sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu||^2 \tag{5.10}$$

Unlike common use of K-means for splitting sets of observations, since the range of values for each parameter influences the measurement of sum of squares, which is used in the process of grouping. Before using K-means, data under each parameter should be normalized to keep all data scattered with values in between 0-1.

After nomalization, k randomly generated sets can be generated and means of each group in each dimension is calculated $m_1, m_2, ..., m_k$. Using these sets for groups, the algorithm consists of the following two steps alternately.

1. Assign each parameter to the cluster according to the least within-cluster sum of squares. This means, the parameter will be partitioned in to set S_i if there is no S_j that achieve the situation as described in Equation 5.11

$$(x_p - m_i)^2 > (x_p - m_j)^2, 1 \leq j \leq k \tag{5.11}$$

2. Update the new means to be the centroids of the observations in the new clusters as Equation 5.12

$$m_i^{new} = \frac{1}{|S_i^{old}|} \sum_{x_j \in S_i^{old}} X_j \tag{5.12}$$

When the assignments no longer change, the algorithm will be converged. Then the parameters in each group that are mostly closed to the centre of the group are picked to represent all other variables in one group. This means, for a chosen parameter x_{picked} in group i, there is no other parameter x_p in this group that achievement the qualification as Equation 5.13

$$(x_{picked} - m_i)^2 > (x_p - m_i)^2 \tag{5.13}$$

This paper chooses the k value to be from 5 to 8 and find out the central variables. For variables in Inter Provincial Migration Rate, parameters are as described in Table 5.23

Using parameters that provided in Table 5.23, algorithms are employed for testing the performance using these parameters and to find the best size of sets for parameters.

The performance is shown in Table 5.24

For Inner Provincial Migration Rate, since variables like distances between provinces do not exist in this part, K-means should be calculated again. The parameters in Inner Provincial Migration Rate are described in Table 5.25

Table 5.23: K-means Centers

Size	4	5	6
1st	dis	dis	dis
2nd	areaX	areaX	areaX
3rd	City_Income	areaY	areaY
4th	Disasters	Country_income	Country_income
5th		Low_Income_Insurance	Stu_Rate_H
6th			Low_Income_Insurance
Size		7	8
1st		dis	dis
2nd		areaX	areaX
3rd		areaY	areaY
4th		Population	Population
5th		City_Income	City_Income
6th		Stu_Rate_H	Country_income
7th		Low_Income_Insurance	Stu_Rate_H
8th			Low_Income_Insurance

Table 5.24: Inter Provincial Migration Sets

Bagging				Random			
Size	2011Inter	2011-2012	2012	Size	2011Inter	2011-2012	2012
4	0.0193	0.0402	0.0336	4	0.0186	0.0399	0.0306
5	0.0184	0.0348	0.0336	5	0.0167	0.0312	0.0296
6	0.0182	0.0347	0.0337	6	0.0169	0.0314	0.0403
7	0.0185	0.0347	0.0337	7	0.0174	0.0334	0.0301
8	0.018	0.0353	0.0336	8	0.0175	0.0384	0.0295
Linear				ANN			
Size	2011Inter	2011-2012	2012	Size	2011Inter	2011-2012	2012
4	0.0213	0.0419	0.0379	4	0.0304	0.0462	0.0389
5	0.0206	0.039	0.0382	5	0.0341	0.0453	0.0383
6	0.0205	0.038	0.0382	6	0.035	0.0396	0.0403
7	0.0209	0.0393	0.0386	7	0.0299	0.0393	0.0395
8	0.0205	0.0353	0.0387	8	0.0331	0.0401	0.0402

Table 5.25: K-means Centers for Inner Provincial Migration

Size	4	5	6
1st	area	area	area
2nd	City_Income	City_Income	Birth_Rate
3rd	Disasters	Green_Rate	City_Income
4th	Low_Income	Disasters	Green_Rate
5th		Low_Income	Disasters
6th			Low_Income
Size		7	8
1st		area	area
2nd		Birth_Rate	Population
3rd		GDP_PC	Birth_Rate
4th		City_Income	City_Income
5th		Green_Rate	Country_income
6th		Disasters	Green_Rate
7th		Low_Income	Disasters
8th			Low_Income

The performance of Inner Provincial Migration Rate is shown in Table 5.26

In the process of choosing the best size of parameters in prediction, with the result of running same algorithms upon different sets of parameters, the average performance of each size of parameter set over all algorithms used is plotted as Figure 5.21.

In this figure, except ANN, all other algorithms achieve best performance with the size of parameter sets as 5. Thus, for the prediction of inter provincial migration rate, set of 5 parameters should be used.

For choosing the best size of parameters for inner provincial migration rate prediction, the result of performance upon different sets is shown in Figure 5.22

From this figure, 7 parameters seem to achieve the highest performance.

From the process of K-Means, the over-fitting problem caused by correlated parameters can be fixed. Comparing the new result with the performance before reducing parameters, the process of reducing parameters may reduce the performance slightly for some algorithms in some cases, but by using reduced set of parameters, the over-fitting problem that caused by multiple correlated variables.

Based on the results in this chapter, for inter provincial migration rates, the bust results are achieved using Bagging or Random Forest with parameters set of 5. For inner provincial migration rates, further data is required and Bagging or Random Forest can also achieve good performance.

The process of migration rate prediction thus can be listed below

1. Preparing, combining data sets with migration rates, economical and environmental variables in different social area. For best performance, migration rates data should be

Table 5.26: Inner Provincial Migration Sets

Bagging				Random			
Size	2011	2011-2012	2012	Size	2011	2011-2012	2012
4	0.2869	0.2618	0.2321	4	0.2851	0.3456	0.3028
5	0.2849	0.2578	0.2457	5	0.2831	0.3449	0.298
6	0.2943	0.2669	0.2601	6	0.2881	0.3492	0.3135
7	0.2935	0.239	0.212	7	0.2845	0.3389	0.3107
8	0.2953	0.2743	0.2536	8	0.2933	0.3383	0.321
Linear				ANN			
Size	2011	2011-2012	2012	Size	2011	2011-2012	2012
4	0.4731	0.2469	0.2327	4	0.413	0.2631	0.2367
5	0.4731	0.2458	0.2312	5	0.4777	0.2782	0.2476
6	0.3855	0.2742	0.2653	6	0.3645	0.3037	0.2849
7	0.3855	0.2742	0.2654	7	0.4092	0.2692	0.2529
8	0.3971	0.2742	0.2668	8	0.4441	0.438	0.395

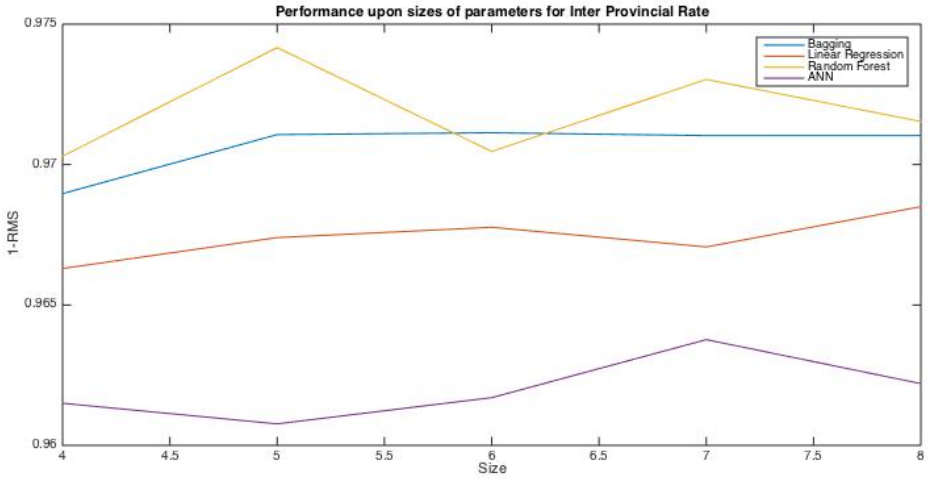


Figure 5.21: Performance upon sizes of parameters for Inter Provincial Rate

one year earlier than other statistics data;

2. Calculating and ranking MIC scores for each variable towards the target variable and selecting the variables based on their scores;
3. Using K-Means to find key variables;
4. Training and testing several training methods includes neural networks, linear regression, random forest and bagging for choosing the best method for prediction with one-year internal migration data;
5. Using the algorithm with the best result of internal migration and use the model for data in other years.

Using this process, a prediction of 2016 inter provincial migration rate can be made using the statics data in 2015 with 5 parameters mentioned above with Random Forest. The Figure

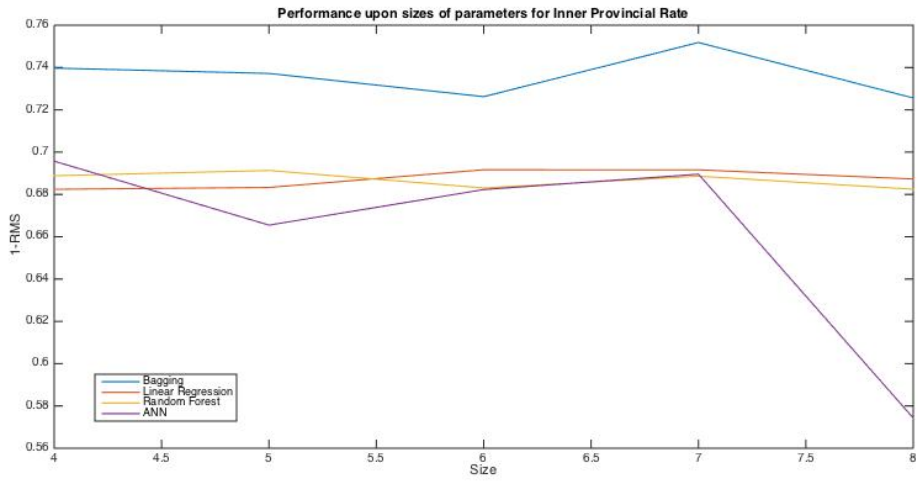


Figure 5.22: Performance upon sizes of parameters for Inner Provincial Rate

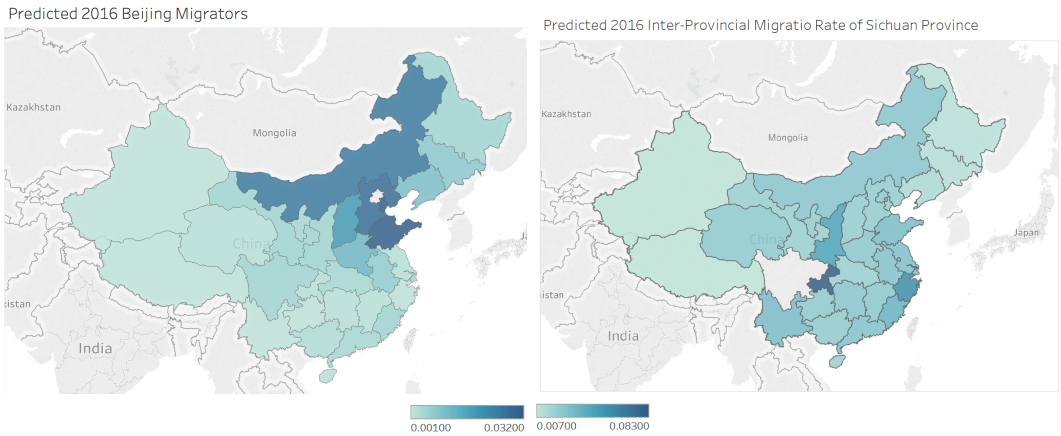


Figure 5.23: 2016 Predictions for Beijing (Left) and Anhui (Right) Migrators

5.23 shows the inter provincial migration rate from Beijing and the inter provincial migration rate from Sichuan Province.

To be specified, Beijing is a place with high GDP Per Capita, the citizens from Beijing keen to move forward to Tianjin, a city also with positive economic environment and Inner Mongolia, a province with abundant natural resources but not to nearby provinces like Liaoning, which is northeast next to Beijing and Henan, which is southwest next to Beijing.

However, as a province locates in south-west China, Sichuan province shows a different phenomenon from Beijing. Sichuan province does not hold a high GDP Per Capita and many provinces around Sichuan province do not hold high GDP Per Capita neither except Chongqing, which locates on the east of Sichuan Province. This is the reason why migration around Sichuan seem to be more evenly compared to Beijing except the migration rate to Chongqing. Also, situation for Sichuan province is less dependent on distances. For example, the migration rate to Zhejiang Province that locates in south east of China, which is over 1500 kilometers away is even higher than the rate to Yunnan, which is just next to Sichuan. This phenomenon shows that the migration rates do not overly dependent on one variable,

however, they are resulted from several economic and geographical features.

From the prediction for 2016 migration rate, a prediction for migration rate in this year can be conducted from data that from the last year. Using the same process, with the predicted data in future, migration rates in next few years can also be calculated.

Thus, a model that can select parameters from data sets using MIC and K-means and build models for making predictions automatically have been built.

Chapter 6

Individual Predictions

Chapter 5 illustrates a process that achieves a relatively high accuracy over Chinese internal migration rates. The prediction of internal migration rates contributes to the overall general inspect over situations about the migration data. However, figuring out what are the reason for leading the individual migrators for choosing their own destinations is vital for understanding what would be the key reason for their migrating and for analyzing their willing of migration based on individual data.

In this chapter, methods that for making predictions about the potential destinations of each family of migrator will be illustrated and tested. Since in the interview data, which is the only concrete source of data concerning the real situation of Chinese internal migration, all interviewers were recording the migrators along with the information of their families into one records, the data of an individual migrator in this chapter also carries the information about family members of that individual.

For making prediction towards the individuals, a further notice about the data set of the interview data should be further illustrated since unlike other data sets used in the prior chapter, these data sets contain problems like considerable large amount of missing data, nonnumeric values in variable and larger size of data.

Besides of theses problems in the data sets, the represent of destination in the data base will also cause big influences towards the accuracy due to all destinations are shown with province codes. This paper uses the GDP per capita of the destinations for representing the province codes with the reason that will be illustrated in the following chapter. This representation actually changes the aim of algorithms in this model from predicting the exact province to predicting the economic environment one actually would like to move to. After the prediction of GDP per capita, further analysis based on the geographic location for prediction of the real potential destinations can be made. Thus, the process can be described as follow

1. Data preparation for generating data sets that can be used for further analysis and replace the province code (both the origin and destination) with local GDP per capita;

2. Decreasing the dimensions of the data;

3. Training and testing the model, employing algorithms for predicting the GDP per capita that one migrator would prefer to migrate to;

4. If necessary, list provinces that locate near one's origin with the similar GDP per capita as the predicted GDP per capita of the destination.

Thus, the result of this model is either a predicted GDP per capita that one migrator likely to travel to or a list of potential provinces but not just one province for the migrator.

6.1 Data Preparation

The over 200 parameters are listed in the interview data set and the differences in questionnaires among years create difficulties for even just selecting variables that exists in every questionnaire. For example, on the questionnaire of 2013, there is a question for whoever or not the migrator holds a residence pass (question number: q202), which is a pass that provides migrators, who have not registered as a permanent residence in some cities with same welfare policy as local permanent residence. However, on questionnaires before 2013, no such question is listed on them since there was no such a policy for residence pass until 2013.

After deselecting questions that related that contains parameters that not exist in every interview data sets, those data with values of nonnumeric values should also be replicated or deleted.

The most important nonnumeric parameter in the questionnaires is about the place that the migrator is now living in. Although all questionnaires contain such a question but not all data sets record the place by its province codes. Some questionnaires like the one taken in 2012, record the Chinese characters rather than province codes.

For these records, a further action is needed to replace the characters with province code as used in other data sets. However, since the destination province is the target variable in the model of individual migrator prediction, simply just use province codes in the process of predicting will cause the increase of errors since codes fall in carrying much internal information or integrity for provinces considering that few provinces with vary different economic environment and geographic localization may have very close codes.

For this reason, GDP Per Capita is used in presenting different provinces. In the process of prediction, the Neural Network will output predicted GDP Per Capita of the destination that one family likely intended to travel to. If further prediction requires knowing exactly the place rather than the GDP Per Capita of the destination, a province that near the original province which obtains close GDP Per Capita of the prediction will be chosen to be the place one probably travels to.

After filtering the data set following steps above and deleting parameters that contains too much Null values, 43 feathers are used to represent the data. Most of them are from the interviews taken by Chinese government. Unfortunately, among the 43 feathers, 28 of them contains Null value. So, before further analysis, missing data must be imputation.

Table 6.1: Detail on Queationnaire

Parameter code	Number of missing data	Asking about
Q203	9531	The industry of one migrator
Q204	9531	The kind of job
Q206	9531	The role
Q207	9531	The average working days
Q208	9531	The average working hours

Table 6.2: Feature Value Accounting

Choices	Amount of selected	Total amount
Being employed	55968	55968
Lost the job	866	9531
Have not found a Job	2690	
Not willing to take a job	5777	
Retired	198	

Multiple methods have been created by researchers for the imputation of missing data like multiple imputation and partial imputation. However, most of them are designed for data that missing at random or even missing completely at random, which is not the case in this project (16).

As mentioned in prior chapters, most of them are caused by the designing of the questionnaire, which requiring some of people only take parts of questions. For example, in question marked by q204 of questionnaire taken in 2012 that asking for the type of job the migrator currently taking. Some migrators do not have a job while at the time of the interview and there is no choice for not having a job. Thus, such an answer will be recorded empty and finally be taken as a missing data in this project.

To back in this point of view, several parameters are listed below for valuing how much data are missing due the reason and how much are really missing randomly. The Table 6.1 below records the amount of missing records in parameters.

Table 6.2 shows records in q202, which asking whether the migrators have jobs:

As shown in the tables, a clear conclusion can be made that most the missing data in these parameters have no jobs during that time. Based on this conclusion, we can frankly extend the conclusion to all missing data in the interview data sets, since the data collected are in good condition, which has been shown by above tables since there is no even one mismatch between the amount of unemployed migrators and the amount of missing data in these parameters.

This paper proposes various default values for feathers valued as Null according to the questions themselves, which means different questions will have different default values.

Methods for Missing Data imputation:

1. For data missing in question q101a2 to q101a6, which requests information about whether immigrates have certain family members travelling with them, a method sim-

ply refill the blank with 0 has been used in this project.

2. For records under question q203, which asking about the area that immigrates working in, a number 15 denoting that the immigrate worked in another unmentioned area.

3. For records under question q204, which asking about the kind of job that immigrates working in, a record 80 denoting that the immigrate worked in another unmentioned area is used in filling the blanks.

4. For records under question q205, which asking about the kind of cooperation that immigrates working in, a record 12 denoting that the immigrate did not work in any kind of cooperation as mentioned under the question is used in filling the blanks.

5. For records under question q206, which asking about the role of the cooperation that immigrates working in, a record 12 denoting that the immigrate did not work as any kind of role as mentioned under the question is used in filling the blanks.

6. For records under question q207, which asking about the income last month for the immigrate, a record 0 denoting that the immigrate did not have any income is used in filling the blanks.

7. For records under question q208, which asking about the average working days for the immigrate last month, a record 0 denoting that the immigrate did not work last month is used in filling the blanks.

8. For records under question q209, which asking about the average working hours per day for the immigrate last month, a record 0 denoting that the immigrate did not work last month is used in filling the blanks.

9. For records under question q301, which asking about the time for the wedding of the immigrate, a record 201300 denoting that the immigrate has not married until 2012 is used in filling the blanks.

10. For records under question q304, which asking about the whether the family obtained the prof for one child family, a record 0 denoting that the immigrate has no child until 2012 is used in filling the blanks.

11. For records under question q306, q308, q309, q310, which ask about the whether the sexual life about migrators, a record 0 for q306 to q309 and a record 9 for q310 denoting that the immigrate has no sexual life during the last month for taking the questionnaire is used in filling the blanks.

12. For records under question q311, which asks the times of abortion made in recent 1 year, a record 0 is added the blanks.

13. For records under question q312, q313, q314, q315, q316, which ask if the migrator or his or her partner has signed or obtains some documents for their children in the name of birth control. A record 0 is added to fill the blanks, which indicates that the migrator is neither signed nor required to sign nor cannot remember because the migrator does not need to sign it.

After appending these data, the data sets can be fulfilled properly and will be ready for further analysis.

6.2 Methods

Unlike the prior process of building the model described, the MIC scores do not show much meaning in this section. After calculating the MIC scores, a very interesting point has been found that almost all parameters show very low correlation towards the destination of migrators.

The reason for this phenomenon should lay on the fact that single parameter of one record cannot determine the destination from a migrator. More generally, there is no common reason for every one in China that can determine the destination of one migrator.

According to the algorithm of MIC, only parameters, whose values as below that can be drawn in small grids on scatterplot can achieve high scores.

Unfortunately, since there is no dominant variable in this data set, sorting MIC scores seem to be meaningless and even misleading since almost all parameters show very low correlation to target variable on their own.

For this reason, the following process should proceed in analyzing all parameters in the data set, which makes the data sets relatively huge and high dimensional.

Since there is no direct relation between variables, algorithms like random forest, bagging and linear regression can hardly show high accuracy. For this reason, the following chapter focuses on using artificial neural network for making predictions.

As mentioned before, the most important parameters for feed forward artificial neural network lay on the number of nodes in each layer and the algorithm of propagating in each term.

In the process of training the model for predicting potential destinations for individuals, to find the number of nodes in each layers for best result, a training process, which running neural network for 40*40 times are used to iterate every possible combination of layers.

However, unlike the data sets used in the prediction of migration rate, the data set used in this chapter contains larger amount of samples and parameters. With 43 dimensions of data, running Neural Network through out all the data with dimensions from 1 to 40 for less than 3 layers and 4 propagating algorithms would be time-expensive(21).

For this reason, a method that can reduce the dimensions and keep the variance within parameters need to be used before running ANN in these datasets.

6.3 Dimension Reduction

This paper uses principle component analysis for the reducing in dimensions in the refined interviewing data sets.

As mentioned in prior chapters, PCA calculates the eigenvalues and eigenvectors of Equation 6.1 and ranking eigenvalues for finding the principle component with the largest possible

Table 6.3: EigenValues

Rank	Eigenvalues
1	10784886532
2	4863993192
3	161469356.2
4	9666079.312
5	559689.6398
6	252.5125925
7	110.7743358
8	40.65309036
9	29.88745188
10	10.61645235

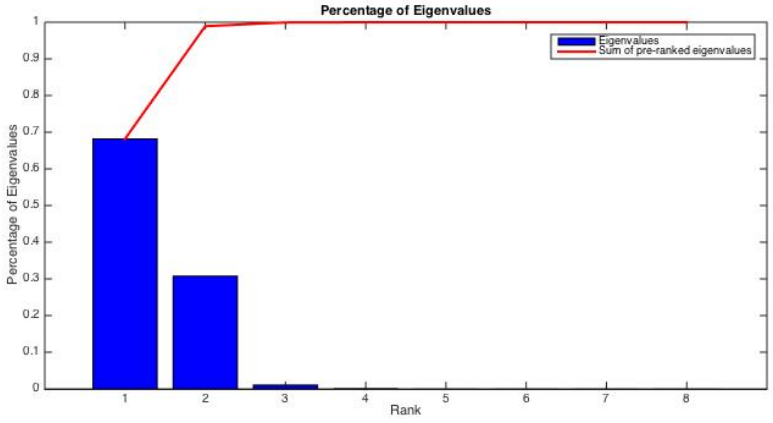


Figure 6.1: Percentage of Eigenvalues for each principle component

variance (17) .

$$S = \frac{1}{N}\bar{X}^T\bar{X} \tag{6.1}$$

After mean normalization and the calculation of eigenvalues and eigenvectors over the interviewing data set of 2012 after imputing missing data with the method above, 46 eigenvectors and their corresponded are obtained. Among them, the top ten eigenvalues are listed in Table 6.3:

The percentage of each eigenvalue towards the sum of eigenvalues are shown in Figure 6.1

As shown in Figurelfig:poefepc0 , the sum of the first 3 eigenvalues is nearly 100% of the sum of all 46 eigenvalues. For this data set, picking top 3 eigenvectors to generate a 3-dimensional data set based on the original data set is a good method for both reducing dimensions and improving efficiency.

However, whether pick 3 dimensions from data bases in all 4 others years is questionable. Thus, running the same PCA process on all other data sets from 2011 to 2014 in internal migration is necessary for testing if picking top 3 eigenvectors are always a good performance.

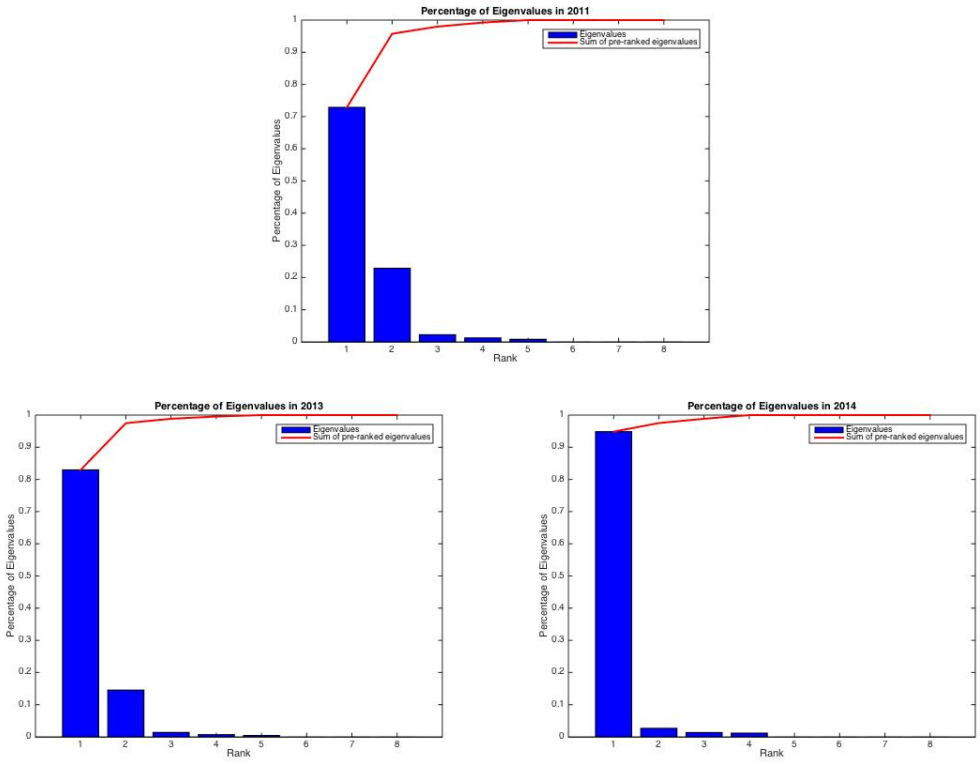


Figure 6.2: Percentage of Eigenvalues for each principle components

The Figure 6.2 shows the percentages of eigenvalues in each year.

After calculating the eigenvalues in the process of PCA over every data sets, the top 3 eigenvalues are found to cover around 99% of the sum of all eigenvalues. This proves that using the 3 principle components with the top 3 eigenvalues can cover most possible variance of the original 46-dimensional data set.

For this reason, in the process of building all neural networks in this chapter for interviewing data that collected in any year, only 3 principle components are selected and only 3-dimensional data sets are to be trained, validated and tested.

This process largely promotes the efficiency of building neural networks and also can decrease the influence of noise that may exist in components with low eigenvalues.

With 43 dimensions fewer, artificial neural networks can be trained with higher time efficiency and lower vibration.

6.4 Modeling

In the last section, PCA is introduced for de-dimensioning interview data into few 3-dimensional data sets. With these data sets, feed forward neural network can be efficiently employed over these data sets.

Table 6.4: Layout

Mean Squared Error	Number of Nodes in 1st Layer	Number of Nodes in 2nd Layer
$9.7*10^7$	27	28

Table 6.5: Accuracy with Imputed data

Mean Squared Error	Number of Nodes in 1st Layer	Number of Nodes in 2nd Layer
$9.0061*10^7$	19	23

As for running neural network, this time, the maximum number of nodes for each layer are also set to be 40 and the four algorithms used are:

1. Gradient descent backpropagation (traingd) Parameter: learning rate;
2. Gradient descent with adaptive learning rate backpropagation (traingda) Parameters: learning rate and ratio increase/decrease learning rate;
3. Gradient descent with momentum backpropagation (traingdm) Parameters: learning rate, momentum constant.
4. Resilient backpropagation (trainrp) Parameters: Increment/Decrement to weight change.

The Neural Network Toolbox in Matlab is used for building the neural networks.

First, a few tests are run to prove the presumption made above is correct.

To prove the specified imputation method for individual internal migration data can improve the accuracy of the model, a test is run by building the model based on the data of migrators to Beijing without any reduction in dimensions.

The training process runs all possible combination of number of nodes in layers and result in the following data:

The accuracy of the data that imputes 0 for missing data is shown in Table 6.4:

With the regression as Figure 6.3:

The accuracy with the data that imputed following steps in this paper is shown in Table 6.5

With the regression as shown in Figure 6.4:

As shown above, the steps enhanced in this paper successfully in reducing the error of predictions. Thus, in the following analysis, all data are imputed following the steps described in this paper.

Next, the effect of PCA is to be tested. To compare the performance that without PCA as described above, this section uses the same interviewing sub data set with only data recorded in 2012 of migrators travelling to Beijing.

After Picking the top 3 eigenvalues, the result is shown in Table 6.6:

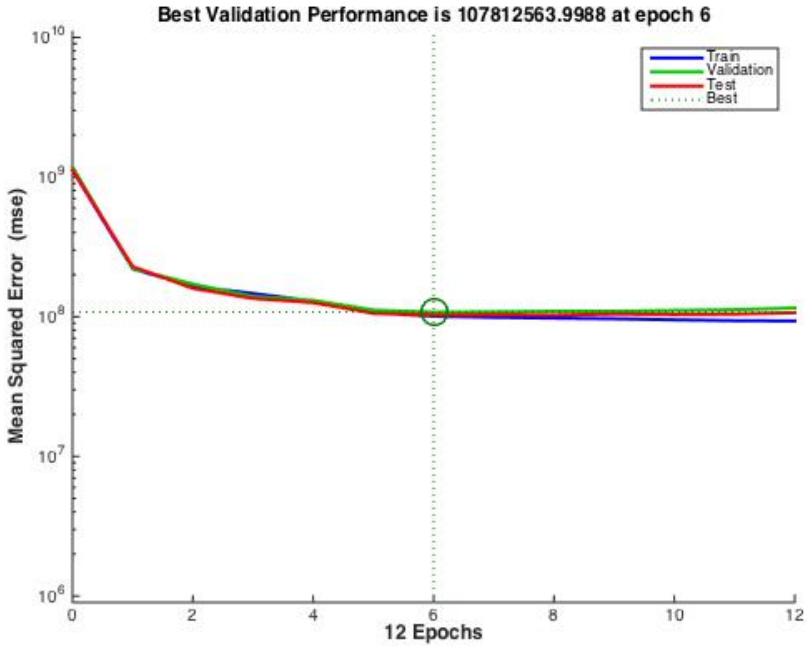


Figure 6.3: Regression with the Sub Data Set

Table 6.6: Performance

Propagation	MSE	Layer 1	Layer 2
Gradient descent backpropagation	1.27*10^8	3	2
Gradient descent with adaptive learning rate backpropagation	1.2578*10^8	2	39
Gradient descent with momentum backpropagation	1.3106*10^8	10	1
Resilient backpropagation	1.2471*10^8	39	16

The propagation algorithm with the least means squared error (MSE) is only about 15% lower than the result trained in the original data set. Thus, the data sets after principle component analysis still keep most variance within the data sets and shows good performance in comparison with the original data set without PCA.

All data sets used later are data sets after running PCA with the selection of top 3 principle components.

6.5 Evaluation

With the methods provided above, artificial neural networks are trained, validated and tested with the interview data in 2012. After iterated over 1600 different topology of neural networks and four propagation methods, the results are shown in Table 6.7:

The result shows the square errors are around 10^8 , which means real errors are around 10^4

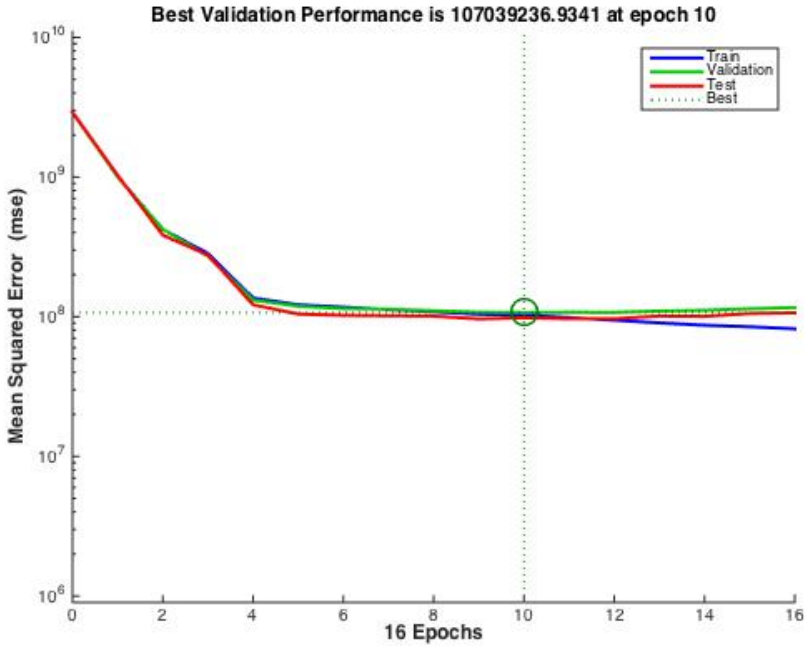


Figure 6.4: Regression with the Sub Data Set

Table 6.7: Performance

Propagation	MSE	Layer 1	Layer 2
Gradient descent backpropagation	1.4335*10^8	10	2
Gradient descent with adaptive learning rate backpropagation	1.2594*10^8	16	36
Gradient descent with momentum backpropagation	1.2719*10^8	1	9
Resilient backpropagation	1.2545*10^8	39	16

in Chinese Yuan, which can about to decide which province the immigrate about to travel to, since GDP Per Capita in provinces near to each other is mostly larger than 10^4 .

The scatter map of errors in predictions is shown in the Error Map in Figure 6.5.

In the error map, each dot represents the error of the prediction of one sample from the GDP per capita of on migrational family. As can be illustrated in the figure most prediction has error between $1 * 10^4$ to $1 * 10^{-4}$, this could narrow the selection of provinces into, mostly, less than 4 potential destinations.

As can be seen from the error map, almost all errors are quite small relative to the GDP per capita in each province.

This method does not specify the exact province that one migrator most intended for travelling to but describe the economic condition of the destination.

Actually, if using the same method to predict the exact province number of the destinations

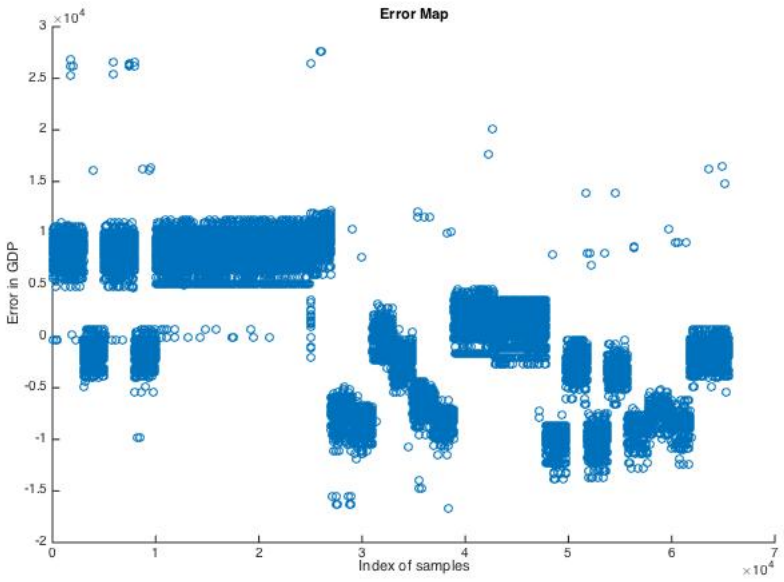


Figure 6.5: Error Map

Table 6.8: Performance for Predicting Province Code

Propagation	MSE	Layer 1	Layer 2
Gradient descent backpropagation	120.4155	2	2
Gradient descent with adaptive learning rate backpropagation	113.2838	3	5
Gradient descent with momentum backpropagation	113.5526	1	1
Resilient backpropagation	100.2972	3	8

will not be accurately working. The following process is to use 2012 migration data as mentioned above with the same algorithms for predicting the province code.

After training, validating and testing the interview data in 2012, the results are shown in Table 6.8.

As can be shown in the results, considering the range of province code (11- 65), the variance is considerable and not ideal. The Figure 6.6, which shows the predicted province codes and the real province code of each individual.

In this figure, almost all prediction seems to be none relative to the true values. Thus, using this method for predicting the province code directly is not possible.

However, the performance of casting predictions on the GDP per capita of the destinations remain positive in accuracy.

Compared to Figure 6.6, the scatter map that records the real and estimated data is shown in Figure 6.7

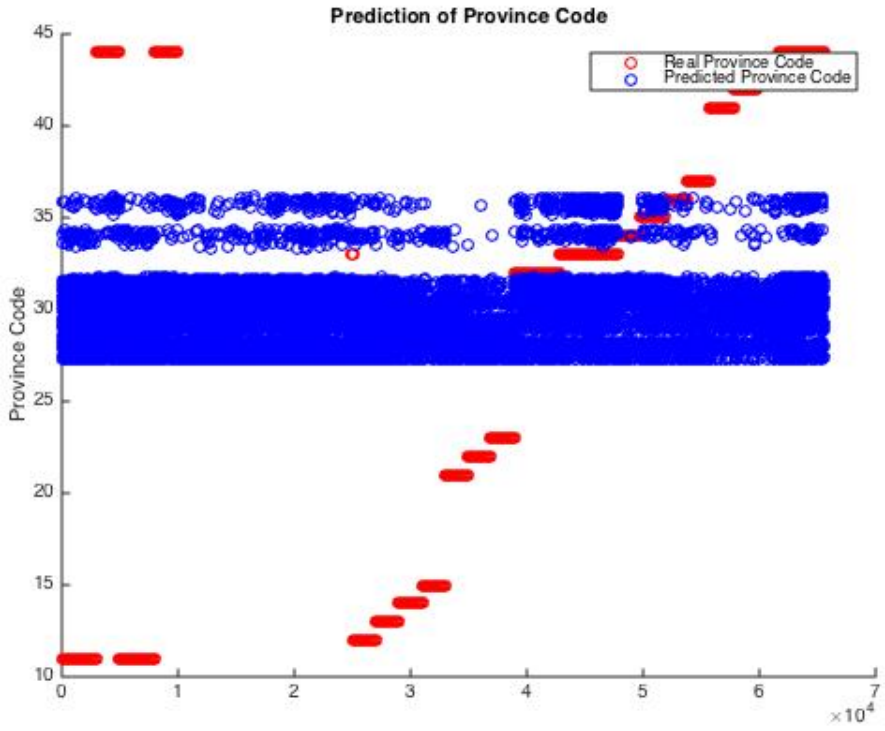


Figure 6.6: Province Code Prediction

In figure 6.7, the blue dots represent the real GDP per capita of destinations of migrators and the red dots represents the estimated dots. The scattered prediction applies relatively close to the destination's GDP per capita.

If the exact destinations of migrators are required for the analysis, the province code of the destinations can he predicted by the distances between origin and each province and the corresponded GDP per capita. The process is listed below:

1. Iterating all upper and bottom boundaries that classifying migrators to provinces with certain GDP Per Capita. In this paper, to compare the performance of prediction, 3 different lengths between two boundaries have been set as following:

   Narrow: 1500 CNY (Top to Bottom)

   Medium: 2000 CNY (Top to Bottom)

   Large: 2500 CNY (Top to Bottom)

2. Selecting provinces within certain boundaries into a list of potential choices.
3. Sorting the provinces with distances from origins and output the result. The first province within the set is the most potential one.

After this process, several provinces with the possibilities can be computed.

The Table 6.9 shows the accuracy of the prediction out of the total number of examples : 65499.

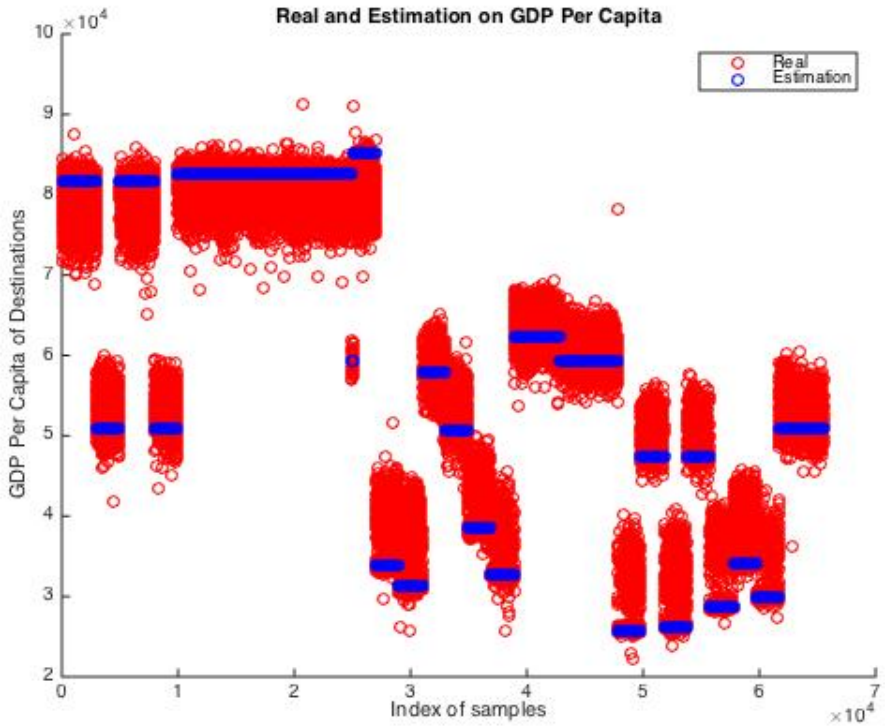


Figure 6.7: Error Map

Table 6.9: Correct Classifications

	Narrow	Medium	Large
Accuracy	29741	32704	37671
Up Bound	649	772	848
Bottom	-851	-1228	-1652

As shown in this table, with the increasing size of space between boundaries, the number of examples that have been correctly classified is increasing. The choice between Narrow, Medium and Large can be made for different aims of analysis. If researches require more accurate results, then larger space between boundaries would be ideal to choose. In contract, if the reduction of the size of potential choices of destinations is required, then, 'Narrow' should be a nice choice.

This chapter describes a process that can make prediction of destinations over individual migrators. The accuracy of result is not as good as the prior chapter for predicting the migration rate. This is because individual destinations are some how depend on information that can not be represented by provincial data. Some of the reasons for making choices might even be the place that once friends are migrating to or even by the impact from rumours. Thus, making predictions towards individual migrators is always be hard for data mining over general economical and environmental data.

For this reason, this chapter enhances methods that to predict the general economical environment of the destinations and make a list of potential destinations. This idea directly

increases the accuracy of predictions.

Chapter 7

Conclusion

This paper focuses on using several methods and data mine tools for the analysis on Chinese internal migration. In the process of analysis, this paper divides the process into two, one in the general view of the internal migration rate and one in the view of individuals.

The first one is to use several classification methods for the prediction of general national migration rates between provinces. This process is done by using several data mine methods including random forests, bagging aggregation with c4.5 decision tree, linear regression and feed forward artificial neural networks.

The migration rates in this process contain both inter-provincial and inner-provincial migration, which means the migration rates provided are between the province with all provinces including itself. One point needs to be mentioned here is that the word province in this paper also includes direct-controlled municipalities and autonomous municipalities but excludes Hong Kong, Macro and Taiwan since the visa or permits are required among these areas.

Another one is using artificial neural network to analysis the individual intention for the destination of migration. This includes the process of data preparation, principle component analysis and artificial neural networks modeling. Data used for this process must be imputed properly.

For the first process, MIC scores are first calculated for selecting parameters. Since the potential parameters especially in the area of economy and environment in each province are redundant when to be chosen, using MIC scores to deselect parameters that are not correlated to the migration rate shows a nice action in choosing parameters that can make the prediction be away from the vibration caused by over-fitting and only learning from the information that has enough correlation with migration rates.

After selecting parameters via MIC scores, K-means is used for reducing the size of parameters. For reducing the over-fitting problem, this paper uses K-means to split the features into few groups and choose the parameter that mostly close to the center within one group to represent the group. By employing the different sizes of parameters into the models, 5 parameters are found to achieve the best performance.

The built models are successfully in the prediction of migration rates between and inside the provinces in China. As mentioned before, the accuracy can be controlled around 0.015 in measured by RMS for inter-provincial migration rates and the best method for this process is always random forest for its best prediction accuracy and perfect low vibration among different data sets in this area with 5 parameters.

Then, a prediction upon not only one year but few years is introduced. With the same model that built in 2011, even using 2012 data as input, the prediction can be controlled around 0.03 in RMS by random forest. This shows the ability to use one model for prediction of migration rates for future years with the prediction of other economic data.

This paper also tries to use the history data that collected not only in one year before the potential data but also in several past years. However, this does not improve the accuracy as wished. Adversely, the increasing in strongly correlated data (parameters in the same filed but collected in different years) decrease the accuracy of some methods, and, as the best method found by this paper, random forests do not show much improvement after be trained and tested by these data. This paper analyzes the reasons should be the potential over-fitted training with too much correlated data and the increasing of noise with the historical data.

In addition, this paper also make prediction on 2016 migration rates based on statistical data in 2015 and analyzes the results based on the surrounded economical and environmental situation.

The second process is for the analysis on individual intention on the choosing of destinations for migration. This paper first introduces methods that for the imputation on data sets. Since most missing data (actually, all data that in the data set tested by this paper) are caused not randomly but by the design of the questionnaire, the missing data in certain parameters actually carry meaningful information such like the loss of jobs. For this reason, this paper imputes the data sets by analyzing the reason behind missing data in each parameter. After fulfilling the 23 dimensions of missing data each by its own meanings, this paper also tests the performance with such a method of imputation and without. With the tests on such imputations, a conclusion that this method will apparently increase the accuracy of our further prediction.

As the original representation of provinces, which is a simple list of codes from 11 to 85 hardly shows much difference between the statistics economical circumstances of on province, this paper proposes a method that to use the GDP per capita in representation of each province for the destination of migrators. This change shows in contribution in the increasing accuracy of prediction by artificial neural networks.

Since the data sets contain too many dimensions, which will obviously slow own the speed of training neural networks, principle component analysis is employed in these data sets by selecting the 3 principle components with top 3 eigenvalues in PCA. To make sure the major information is reserved, the data generated based on these components is used into building the artificial neural networks with Levenberg-Marquardt propagation along with the same data set but without PCA. Through the comparison between the accuracies, the fact that using PCA in these data sets do not cause the apparent lose of important information for making prediction. Thus, all data sets used later are processed by PCA for the reducing of dimensions.

After the reduction of dimensions, artificial neural networks are used in making predictions towards the GDP per capita of the destinations. After training 40*40 times for each propagation algorithms and totally 4 algorithms, results with considerable accuracy are generated and shown in prior chapter. The mean error of each prediction is around 10% of the true value of GDP per capita.

Finally, a further method that can predict the exact destination of one migrator rather than the GDP per capita is introduced by this paper by the selection of GDP per capita that near the predicted one and giving the possibility of each by comparing the distances between origin and destinations.

Based on the process before, a general method that for internal migration can be conducted. These process shows very low dependence on priori knowledge and can be used by any similar project with only a few changes in the process of collecting and preparing data. For example, data mine in international migration can also follow the steps introduced in this paper for the prediction of migration rates and individual intention over migration.

Chapter 8

Future Work

Further work can be down in several aspects.

Firstly, the collection of individual migration data does not contain enough data for prediction. The migration rate calculated by these data sets are common to be 0 between some provinces. But, actually, with such a huge population in each province, the migration rate can hardly be 0. Increasing the amount of interviews taken will benefit the accuracy of prediction.

Secondly, almost most candidates took the interviews as recorded in the interview data sets are labours but not students nor citizens in middle class, who are also a large portion in migrator. This problem can directly cost the prediction made in this paper shows more keen to the situation of labors but not the whole group of migrators.

In addition, values of some parameters in the interview data sets can not clearly represent carry the difference between the true information between them. For example, in the question q202 in the year 2012 as mentioned before, the value 2-5 stands for several situations of not having a job, which carries similar meaning (not employed) and only 1 stands for employed. This may easily cause neural networks miss-weighting it.

In spite of inner provincial migration rate prediction. The further detail data about economic and environment situation among cities and regions should be provided. With these data, using models built in this chapter, a more accurate model for inner provincial migration could be achieved.

During the process of individual predictions, the detailed information about the place of migrators including once friends and even rumours, which is hard to be represented by numerical data would increase the accuracy of predictions.

Bibliography

[1] Hu, Xiaojiang, Sarah Cook, and Miguel A. Salazar. ”Internal migration and health in China.” The Lancet 372.9651 (2008): 1717-1719. pages 1

[2] D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander. Detecting novel associations in large datasets.Science 334, 6062 (2011) pages 7

[3] Artificial Intelligence: A Modern Approach (3rd Edition) (11 December 2009) by Stuart Russell, Peter Norvig (p728) pages 9

[4] Yearbook, China Statistical. ”National Bureau of statistics of China.” China Statistical Yearbook (2011). Breiman, Leo. ”Bagging predictors.” Machine learning 24.2 (1996): 123-140. pages 12

[5] Harris, John R., and Michael P. Todaro. ”Migration, unemployment and development: a two-sector analysis.” The American economic review (1970): 126-142. pages 27

[6] Zhang, Yudong, et al. ”Binary PSO with mutation operator for feature selection using decision tree applied to spam detection.” Knowledge-Based Systems 64 (2014): 22-31. pages 21

[7] Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002, 2(3): 18-22. pages 27

[8] Breiman, L. Machine Learning (2001) 45: 5. doi:10.1023/A:1010933404324 pages 32

[9] Harris, John R. Todaro, Michael P. (1970), ”Migration, Unemployment and Development: A Two- Sector Analysis”, American Economic Review 60 (1): 126-142, JSTOR 1807860. pages 32

[10] Bauder, Harald. Labour Movement: How Migration Regulates Labour Markets. Oxford University Press, 1st edition, February 2006, English, 288 pages. pages 4

[11] Bauder, Harald. Labour Movement: How Migration Regulates pages 4

[12] LU J, ZHOU H. Economic Growth Effects of Population Migration of China’s Provinces: An Empirical Analysis of Endogenous Growth[J]. Population and Development, 2013, 5: 008. pages 4

[13] YUAN X, ZHANG B, HU D. The Effect of Population Migration on the Regional Disparity of Economic Growth - Take Shan Xi as an Example [J][J]. East China Economic Management, 2009, 9: 009. pages 5

[14] Zhaoyuan X, Shantong L. The effect of inter-regional migration on economic growth and regional disparity[J]. The Journal of Quantitative Technical Economics, 2008, 2: 38-52. pages 5

[15] Duan P Z, Liu C J. The effect of population mobility on the regional disparity of economic growth[J]. China Soft Science, 2005, 12: 99-110. pages 5

[16] Cleophas T J, Zwinderman A H. Missing data imputation[M]//Clinical Data Analysis on a Pocket Calculator. Springer International Publishing, 2016: 93-97. pages 5, 51

[17] Syms C. Principal components analysis[J]. 2008. pages 54

[18] Quinlan J R. Improved use of continuous attributes in C4. 5[J]. Journal of artificial intelligence research, 1996, 4: 77-90. pages 28

[19] O'Connor J J, Robertson E F. The MacTutor history of mathematics archive[J]. World Wide Web pagei http://www-history. mcs. st-and. ac. uk/¿(accessed April 22, 2004), 2007. pages 10

[20] Hu X, Cook S, Salazar M A. Internal migration and health in China[J]. The Lancet, 2008, 372(9651): 1717-1719. pages 7

[21] Krogh A, Vedelsby J. Neural network ensembles, cross-validation, and active learning[J]. Advances in neural information processing systems, 1995, 7: 231-238. pages 53

[22] ZHU H, WANG J, LI P, et al. Firm Migration of Clusters in East Coastal Areas of China: A Case Study of Lamp-making Clusters in Wenzhou, Zhejiang [J][J]. Progress in Geography, 2009, 3: 004. pages 16

[23] Widrow B, Lehr M A. 30 years of adaptive neural networks: perceptron, madaline, and backpropagation[J]. Proceedings of the IEEE, 1990, 78(9): 1415-1442. pages 9

[24] Chan K W. China: internal migration[J]. The encyclopedia of global human migration, 2013. pages 38

[25] Strik D P, Domnanovich A M, Zani L, et al. Prediction of trace compounds in biogas from anaerobic digestion using the MATLAB Neural Network Toolbox[J]. Environmental Modelling Software, 2005, 20(6): 803-810. pages 35

[26] Cawley G C, Talbot N L C. On over-fitting in model selection and subsequent selection bias in performance evaluation[J]. Journal of Machine Learning Research, 2010, 11(Jul): 2079-2107. pages 39

[27] Samanta B, Al-Balushi K R, Al-Araimi S A. Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection[J]. Engineering Applications of Artificial Intelligence, 2003, 16(7): 657-665. pages 35

[28] Just D R, Wansink B. The flat-rate pricing paradox: conflicting effects of 'all-you-can-eat' buffet pricing[J]. The Review of Economics and Statistics, 2011, 93(1): 193-200. pages 24

[29] Hare D. 'Push' versus 'pull' factors in migration outflows and returns: Determinants of migration status and spell duration among China's rural population[J]. The Journal of Development Studies, 1999, 35(3): 45-72. pages 24

[30] Chan K W, Zhang L. The hukou system and rural-urban migration in China: Processes and changes[J]. The China Quarterly, 1999, 160: 818-855. pages 4

[31] Zhang K H, Shunfeng S. Rural-urban migration and urbanization in China: Evidence from time-series and cross-section analyses[J]. China Economic Review, 2003, 14(4): 386-400. pages 15

[32] Vyas S, Kumaranayake L. Constructing socio-economic status indices: how to use principal components analysis[J]. Health policy and planning, 2006, 21(6): 459-468. pages 8

[33] International Labour Organization. Retrieved 2013-10-20. pages 1