An SVM-Based Classification Model for Migration Prediction of Beijing

Lan Zhang, Lina Luo, Lei Hu and Maohua Sun

***Abstract*—In this paper, a classification model for the migration of residents without local household registration in Beijing is established through the algorithm of Support Vector Machine (SVM) and the model is verified by using the migration data of Beijing, which is collected from various surveys. Our result shows that, compared to BP Neural Network and Logistic Regression, SVM performs better in terms of accuracy and generalization for these particular classification tasks. We identify ten classification features, which, we believe, are crucial as the determining factors to predict the migration trend in Beijing. These ten features include age, education, occupation, income, family status, housing status, leisure status, insurance status, temporary residence permit status and residence time. Our research shows that, taking into account the population demographic attributes and behavioral characteristics, our SVM classification model is able to predict the migration trend with a high accuracy rate. We believe that the results presented in this paper will provide valuable practical insights for various governmental departments of megacities in grasping the migration trend of different types of residents without local household registration, as well as in improving the residence policies, in order to encourage outward migration and tackle the issue of rapid population growth.**

***Index Terms*—megacity, residents without local household registration, migration, support vector machine.**

I. Introduction

In recent years, massive domestic migration has become a prominent phenomenon in China. In fact, according to “2017 Statistical Report on National Economic and Social Development” published by National Bureau of Statistics, by the end of 2017, the number of migrant population in China reached 244 million. Majority of these 244 million people originated from mid-western provinces of China and moved to eastern coastal provinces, most notably to China’s four megacities: Beijing, Shanghai, Guangzhou and Shenzhen. This migration created rapid population expansion and initiated the issue of urban diseases and social sustainability [1] to these megacities. Thus, the demand for an effective population control strategy is high, since authorities in said megacities have yet found such strategy to solve this ongoing problem. In the attempts of tackling this problem, these megacities have applied various regulations and restrictions, in the hope of tuning down the number of migrant residents. One particular example of strategies used is the forming of a satellite city, named Xiongan, in order to migrate some people out of Beijing [2]. The applications of these regulations and restrictions result in the initiation of people’s emigration from these megacities, and, for the first time in years, Shanghai’s population decreased by 150,000 people [3].

In this paper, we focus on addressing the questions of who, among the migrant residents, will stay, who will leave, and how to predict their migration. To clarify some concepts discussed later in this paper, we provide a few important definitions as follows.

Household registration/Hukou in China is a booklet issued by Public Security Bureau, and is used to officially identify a person as a resident of a specific area. Hukou determines where a person can claim his/her social welfare, e.g. health insurance, school allocation, etc. Such geographic registration system creates various inconveniences and limitations for people living away from the area his/her hukou is registered [4]. As we can see, hukou acts as a domestic passport and imposes restrictions on its holder’s migration.

Megacities [5] (as defined by The State Council of P. R. China in 2014) are the cities with a permanent population that exceeds 10 million in urban areas. Currently, the megacities in China include Beijing, Shanghai, Guangzhou and Shenzhen.

Migrant population refers to the population of people residing outside their hukou area in a long term.

Permanent residents are people who live within the area of their hukou.

Non-local residents are residents who are not permanent residents. *Long-term non-local residents* (LTNL residents) are non-local residents who have been living in a specific area for longer than six months. The ‘LTNL residents’ label was introduced by Beijing Statistics Bureau in order to study the issue of the rapid urbanization of megacities.

In short, the purpose of this paper is to predict and classify residents’ migration in Chinese megacities and analyse the types of residents that are likely to stay and those likely to leave. The conclusion presented in this paper provides great practical insights for governments in megacities, such as Beijing, Shanghai, etc., in tackling megacity urbanisation issues. By utilizing our model to analyse their current residents data, they will be able to tailor the policies imposed on various resident types, in order to encourage outward migration, hence, solving the rapid population growth issue.

II. Literature review

To date, majority of urban population classification study have mainly been focused on migration prediction, typologies of urban migration and preference of migrant population. We briefly mention these studies in the following.

A. Research on migration prediction

The gravity model, as shown in Figure 1, was proposed by Zipf [6] and originally used to predict migration. Assuming

Manuscript received October 11, 2019; revised August 18, 2020. This work is supported by the project of National Natural Science Foundation of China (61303018) and Research Project of CUEB Social Computation and Quality Development for Megacities.

Lan Zhang, Lina Luo, Lei Hu and Maohua Sun are with Department of Data Science and Big Data Technology, Capital University of Economics and Business, Beijing China

Corresponding author: Lina Luo, luolinatop123@163.com

that people migrate from Area A to Area B and the number of migrant people is C, Zipf believes that (1) C is positively correlated with the size of population of A and B, respectively; and (2) C is negatively correlated with the distance between A and B. The gravity model became a base model for many other approaches later.

In recent years, researchers, who study urban population classification, mainly focus on migration prediction, typologies of urban migration and preference of migrant population.

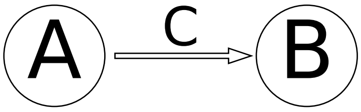

Fig. 1. The gravity model

According to a certain history of people's migration behavior, Rogers [7] uses Markov chain to predict migration behavior at a certain time in the future. Rogers's prediction takes age, education and occupation factors of the migrant population into consideration.

By taking into account of certain individuals' preference to work within the close proximity of their home, regardless of any better job opportunities, Simini proposed a radiation model in [8]. Based on Zipf's work, Simini's radiation model extended the gravity model by introducing a new and vital variable that represents the total population in a circular area, of which the center is area A and the radius is the distance between A and B. This extension provided better capabilities in capturing the mobility and migration patterns, and thus, improving the accuracy of migration prediction.

The rapid development of telecommunication and smartphone technologies, including GPS, Wi-Fi and Bluetooth, has enabled easy and precise personal location data collection. This further enables the analysis and prediction of migration patterns and paths based on these location data. Based on the historical behavior data of individuals and groups collected from mobile phone data of the downtown Boston-USA residents, Calabrese et al. [9] proposed a series of migration prediction models, which are proved to be highly accurate.

In considering the frequency of individual migration paths, Lu et al. [10] utilised entropy to measure the uncertainty of individual's mobility and found that the maximum theoretical prediction accuracy of individual migration paths is 88%. They further confirmed that this 88% accuracy can be approached in reality by applying Markov chain on the mobile phone data.

In addition, many researchers in the field of population dynamics have confirmed that the migration's spatial distribution is subject to the power-law distribution [11]–[13], and have established many models via various statistical methods.

B. Research on typologies of urban migration

The research on the typologies of urban migration utilizes various classification approaches of classification to investigate population migration. In the mid-1990s, Nabi [14] pointed out typological approaches should be applied to the field of migration, which turned to be a rather complex dynamic process. Existing literature surveys showed that most discussions on migration preference involve some demographic variables and three typical features; where people come from, where they move into, and the length of residence time. Studies found that, in addition to occupation, gender [15], [16], age [17]–[21], education [18]–[23] and marital status [16] can also significantly distinguish individuals' decision to migrate. Wan et al. [24] found that places, with more people moving in than leaving, are usually places where specific type of occupation is dominant for the local job market.

C. Research on the preference of migrant population

Migration preference is a subjective intention of whether non-local residents wish to reside permanently in their current location. Majority of existing work studied this migration tendency by considering influence factor analysis.

Ren [25] pointed out that the non-local residents in Shanghai have a "precipitation effect", i.e. the longer they live in Shanghai, the more likely they are going to stay permanently. Whereas Meng et al. [26] found that, due to how the availability and accessibility of social welfares are heavily depended on hokou, this system itself has great impacts on migration preference of certain non-local residents. Through the study of many LTNL residents and their residential preference, Zheng et al. [27] showed that these preferences are affected by many factors such as residential permits, length of residence period, family companionship, living standards and the availability of social security they can obtain. Cai et al. [28] found that income and occupation stabilities, property (or house) ownership and support payments for the parents are among the main factors that affect migration preference in Beijing.

In [29], Wang studied the migration preference of migrant population specifically in Xinjiang, Beijing and Guangdong, and found that community participation, local residents' acceptance and comparable well-being to their previous residence ave significant positive influences of the preference of staying in the new area. Furthermore, having analyzed the 2015 National Domestic Migrant Population Survey results specifically on Beijing's data, Yu et al. [30] found that education, length of residence time, the household size, monthly income, and housing expenditure have significant positive influences over the migration preference. Moreover, (1) LTNL residents, with ages between 30 and 40, have the strongest preference to stay; (2) the LTNL seniors, ages ranging from 60 to 70, whom are living with their children, also have strong tendency to stay; (3) individual, who is living and married to another LTNL resident, will have significantly stronger preference to stay; and (4) individual, whom are an employer or a freelancer, will have a very strong preference to stay.

Duan et al. [31], in order to establish various classifications of the migration population, created a system that quantifies economic situation, migration reasons, living conditions, and education level of migrants. After the division, They then studied the features of these individual classification accordingly. Zhao et al. [32], on the other hand, studied the migration preference by focusing on housing property ownership, housing space and housing quality.

In short, most of the existing research on megacities' migration prediction focused on big data to establish a model of people's mobility. However, there is only a few that focuses on the micro level, i.e. individuals' choices. There is yet any clear patterns and rules as to how an individual decides to migrate to a megacity. In this paper, we use individual's factual migration data to create a prediction model to study domestic migration to megacities in the micro level. We intend to utilize this model to provide a practical tool to gain a better understanding of this particular domestic migration in China.

III. Selection of resident classification features

Since the predication's accuracy of LTNL residents' migration in megacities greatly depends on the selection of resident classification features, it is crucial to decide which features are to be taken into account when creating the model. Insufficient number of features will decrease the prediction's accuracy significantly. In contrast, having too many features will decrease the model's practical value dramatically. Therefore, considering the results of various existing research and combining them with our reasoning, we have decided on using ten particular resident classification features that will discuss on the later part of this section.

Numerous studies have shown that, due to their non-local hukou's status, to be able to settle permanently in a megacity, LTNL residents must overcome the limited accessibility of various social welfares, as well as enduring the high living costs, in terms of both physically and psychologically. Diverse living conditions result in different capability levels for LTNL residents in overcoming and dealing with the said limitations and costs. These individual and unique features of living conditions are the crucial points for the residents in deciding to leave or stay in a megacity permanently. In summary, these features can be grouped into two main categories: LTNL residents' demographic attributes and behavioral characteristics.

A. Demographic attributes of LTNL residents

The following are the demographic attributes we have chosen to include in creating our prediction model:

1) Age: Studies showed strong correlation between age and migration tendency. For example, Li [33] pointed out a unique life cycle of many Chinese migrant workers, where they migrated to work in megacities at young age and returned home to farm, when they grow older. Furthermore, Tang et al. [34] found that the younger generations of migrant workers have stronger desires to work and settle in megacities rather than returning to their former homes.

2) Education: Education plays a pivotal role on employment possibilities and the adaptation to urban life in megacities, hence, it significantly affects the migration tendency. In our model we categorize education into eight different levels, starting from the lowest level of education to the highest as follows: elementary school, junior high school, high school, vocational high school, diploma, bachelor degree, master degree and doctoral degree.

3) Occupation: Occupation is a very important attribute to decide migration tendency, as it measures and affects each individual's income status. Taking the occupation characterization of Shanghai residents given by Qiu in [35] and combining it with the occupational characteristics of Beijing's LTNL residents, we categorize occupation into five classes as follows:

- The first class is mainly the leader types of occupation. This includes enterprise/company managers, private entrepreneurs, agents of the foreign firms, etc.;
- The second class is composed of general clerks or staff, such as bank staff, employees of foreign company/trading corporation, government employees, secretaries, etc.;
- The third class includes people of professionals such as lawyers, actors/actress, musicians, painters, journalists, scientific researchers, engineers, accountants, etc.;
- The fourth class includes commercial personals, e.g. brokers, enterprise/company employees, selfemployed/freelancers, salesmen/saleswomen, shop-assistants, etc.; and
- The fifth class includes farmers, workers and service providers, such as taxi drivers, nurses, mechanics, cleaners, waiter/waitress, couriers, etc.

4) Income: This is the most important feature to LTNL residents, as it directly decides their abilities to afford the cost of living in megacities. Having a decent or even substantial income will also enhance their confidence in settling down permanently.

B. Behavioral characteristic features of LTNL residents

Results of various studies have confirmed that marriage, occupation, living standards, leisure, friendships and health are the six vital fields of life in a megacity, and that there is a strong and positive correlation between the satisfaction level in those fields and the subjective feeling of happiness [36].

Here, the fields of occupation and living standards correspond to the aforementioned occupation and income, respectively, in Section III-A. As to friendship, many studies [26], [29], [37] pointed out that great social interactions between local residents positively affect the migration preference of LTNL residents. However, despite being unwelcomed by the locals, it is still possible for these residents to establish good social relationship and networks among themselves and live satisfactorily (Liu [38]). Thus, we decide to not regard friendship as a classification feature for our model. The behavioral characteristics that we take into account are as follows:

1) Family status: Family status refers to whether an LTNL resident is migrating with his/her family and is directly related to marriage, one of the six vital fields mentioned above. In recent years, taking the family along gradually become the main pattern of migration [39], [40]. Agreeing with the papers [39], [40], we believe that, due to Chinese traditional and family-oriented culture [41], those, who are migrating with their main family members, have stronger tendency to settle in megacities, as opposed to those whose main family members residing in their former home. Here, main family members refer to the spouse, children, parents and parents-in-law. We categorize family status as: living alone, living with main family members, living with other relatives, and others.

2) Housing status: As one of the measurement of living standards, housing is one of the essential requirements for

LTNL residents to work and live in cities. Zhao et al. [32] confirmed the significance of housing status related to migration tendency. Furthermore, Chen et at. [42] showed there exists a direct relationship between housing and residential satisfaction. While in [43], Wang et al. pointed out how housing wealth inequality between urban and rural areas has continued to grow unbalanced. Thus, housing status becomes a factor that cannot be ignored by migrants. We categorize housing status as living in own house, living in relatives' or friends' house, living in dormitory provided by companies or organizations, renting a house alone and renting a house with others.

3) Leisure status: Leisure status refers to the activities LTNL residents mainly do outside working hours. Leisure is an important part of daily life and it can reflect the quality of life of LTNL residents in a specific way [44]. The paper [34] further pointed out that, the current of migrant residents have lower leisure level, in terms of time and expenses, compared to locals. This affects their perception of satisfaction in life and shows how leisure time and space usage differs between various social groups (Whyte [45]). Further research [46] showed that the gap of leisure levels between these groups in Chinese cities are increasing rapidly and leisure status has become one of the deciding factors of migration tendency. Taking the characterization of residents' leisure status in Shanghai proposed in [35] and combining it with our own research data of Beijing's residents, we categorize this status into five types: leisure, fitness, learning, entertainment and no leisure.

4) Insurance status: Health, as one of the six vital fields mentioned earlier, relates to pension and medical insurance and has significant impact on an LTNL resident's life satisfaction [34]. Yu et al. [30] pointed out that LTNL residents with medical insurance are more likely to settle in Beijing. The categorization of the pension insurance status is as follows: no pension insurance, pension insurance in Beijing, pension insurance in former residence area and others. The categorization of the health insurance status is as follows: no medical insurance, medical insurance in Beijing, urban medical in former residence area and rural medical insurance in former residence area.

5) Temporary Residence Permit (TRP) status: TRP status refers to a migrant's possession of a temporary residence permit. Currently, some megacities in China have established and applied a TRP system, of which LTNL residents can apply to the local government and be granted with a TRP, given that they meet all the requirements. The requirements, in general, include a legal and stable occupation as well as a long-term housing status, as minimum. Processing a TRP gives LTNL residents' children guarantied opportunities to study within the city, as well as the possibilities of obtaining the local hukou in some megacities. This increases the likelihood of LTNL residents with TRP to eventually settle down in megacities. We categorized TRP status as no Beijing TRP and Beijing TRP.

6) Residence time: Residence time refers to the length of time a migrant resides in Beijing; from moving into Beijing to moving out, or until the time he/she was surveyed. Some studies [26], [29], [37] confided that the longer the residence time, the more likely an LTNL resident stays permanently. Consequently, we chose residence time as one of the features considered. In our model, the residence time is measured in the number of months.

C. Summary of chosen features

Following the analysis given above, we chose ten particular features to be used in our model in order to classify the residents' migration in megacities, as summarised and shown in Table I.

TABLE I
FEATURES FOR CLASSIFYING RESIDENT MIGRATION OF MEGACITIES

Source	Chosen Features
Demographic attributes	Age
	Education
	Occupation
	Income
Behavioral characteristics	Family status
	Housing status
	Leisure status
	Insurance status
	Temporary residence permit status
	Residence time

IV. RESEARCH METHODS

A. Support Vector Machine (SVM)

The classification of residents' migration is a binary-classification problem, i.e. they either settle down permanently in a megacity or leave for other places. To solve such problem, Support Vector Machines (SVM) is a very effective approach to use. SVM was first proposed by Vapnik [47], in mid-1990s, for pattern classification and non-linear regression. This approach has been successfully used in many research and industry fields, in recent years to solve various classification problems, such as intrusion detection for wireless sensor networks [48], power-line degradation detection [49], email author identification [50], stock investment classification [51], financial performance classification for enterprises [52], rainstorm/haze classification [53] and text classification [54]. In particular, SVM was utilised to study migration in Africa and produced a convincing result [55]. In this paper, we also choose to adopt SVM model to classify the residents' migration in Chinese megacities, since the classification problem we are dealing here is considered to be a complex non-linear classification problem.

In general, SVM algorithm works as follows: it maps the original inputs, which are often not linearly separable in their space, into a higher-dimensional feature space using a non-linear function, known as kernel function; then it searches for the optimal linear classification surface, also called hyperplane, in the new space. If such hyperplane exists, it defines an SVM classifier. Further details, as written in [56], are explained below.

Let the dataset $T = (x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$ be the training dataset containing N points in some feature space, where $x_i (x_i \in X = \mathbb{R}^n)$ is the i_{th} feature vector and $y_i (y_i \in Y = \{+1, -1\})$ is the class (or label) of which x_i belongs to. When $y_i = +1$, it is called the positive of x_i, similarly, $y_i = -1$ is the negative of x_i. We call (x_i, y_i) a sample. Then, the SVM algorithm goes as follows.

(1) With the chosen kernel K and the penalty parameter C, construct and solve the optimization problem which is defined by:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{N} \alpha_i$$

$$s.t. \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C, i = 1, 2, \cdots N$$

Find the optimal solution: $\alpha^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_N^*)^T$.

There is a wide selection of kernel functions that can be used in SVM. These kernel functions include polynomial kernel, RBF (Gaussian) kernel, and string kernel. In this paper, we use RBF kernel, which is defined as:

$$K(x, z) = exp\left(-\frac{||x - z||^2}{2\sigma^2}\right).$$

.

(2) Choose an element α_j^* of α^* such that $0 < \alpha_j^* < C$, and compute:

$$b^* = y_i - \sum_{i=1}^{N} y_i \alpha_i^* K(x_i, x_j).$$

.

(3) Construct the separable hyperplane $w^* \cdot x + b^* = 0$ to further obtain the decision function:

$$f(x) = sign((\sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i)) + b^*)$$

.

B. SVM classification model for residents' migration of Beijing

We establish the SVM classification model for the migration of LTNL residents in megacities in three steps:

(1) Obtain the migration data of megacities' residents, and use the ten resident classification features, mentioned in Section III-C, as our feature space for learning.

(2) Construct the classification model by using SVM algorithm. Looking closely, the classification problem of the residents' migration in megacities can be translated as a problem to divide two types of labelled points with different attributes in high-dimensional space. Suppose that each sample is viewed as a p-dimensional feature vector in the feature space, the number of LTNL residents who leave Beijing is m and the number of LTNL residents who still stay in Beijing is n. Then the feature space is a p-dimensional Euclidean space, in which there exist: (a) m points $x_i(x_{i1}, x_{i2}, \ldots, x_{ip})$, where $i = 1, 2, \ldots, m$, labelled with the "Left" tag; and (b) n points $y_j(y_{j1}, y_{j2}, \ldots, y_{jp})$, where $j = 1, 2, \ldots, n$, labelled with the "Staying" tag. The purpose of constructing this SVM classification model is to minimize the generalization errors and to maximize the margin to separate two sets.

Through experiments using various methods, We have confirmed that SVM is the most suitable method to solve our classification problem. We will present the results of our various experiments in Section V.

(3) Verify the performance of our model with the testing dataset. We run the training dataset and the testing dataset to test our model. The testing results for both data sets showed very good classification performance. To ensure, the stability of our model performance, experiments for each dataset are conducted for a number of times.

V. Research result

A. Data collection

All the data used for experiments presented in this paper were collected through surveys, which made specifically to obtain information relevant to the resident classification features we selected. We hired a leading market research company [1] to conduct the data collection process. This company currently is the largest online investigation, examination and voting platform in China.

Online data was collected from April 10 to May 15, 2018. In total, we received 1712 answered questionnaires; 1360 online and 352 paper-based, Out of which, 1661 were valid, that is 1324 online and 327 paper-based answers, giving us a 97% validity. Within the valid samples, we had 882 samples for "Left" and 779 samples for "Staying". Note that "Left" samples are data related to the LTNL residents, whom have left Beijing at the time the survey was taken, and "Staying" samples are data related to those, who were still staying in Beijing.

We have utilised these survey to investigate the migration trends of LTNL residents in Beijing and collected the following statistics:

- The percentage ratio of the respondents based on gender is 59.0% and 41.0%, that is 980 males and 681 females, respectively, as shown in Figure 2.
- The youngest respondent is 14 years old and the oldest one is 63 years old. The median of the respondents' ages is 31 and the statistics of these ages are shown in Figure 3.
- The lowest monthly income is zero, the highest is CNY300,000 with the median of the monthly income is CNY6,000. Further statistics of monthly income data is given in Figure 4.
- As shown in Figure 5, for respondents' highest education level, 5.5% respondents is elementary school, 11.0% junior high school, 9.7% high school, 5.6% vocational high school, 23.1% diploma, 37.3% bachelor degree, 6.7% master degree, and 1.1% doctoral degree. The specific numbers are 92, 183, 161, 93, 384, 619,111 and 18, respectively.
- Shown further in Figure 6, for occupations, 9.5% respondents are leaders, 10.8% are general clerks or staff, 15.2% are professionals, 25.8% are commercial personnel, 38.7% are workers, farmers or workers who provide service. The numbers for each class are 157, 179, 252, 429, and 644, respectively.

In recent surveys published by [39], [57], all the survey respondents, i.e., LTNL residents in Beijing were either doing business or providing business. Whereas in our data, there are only 64.5% people with such occupations among all the

[1] Since its launch in 2006, the company has sent out more than 9.48 million questionnaires and collected more than 5.11 million answered questionnaires. Moreover, their clients comprises of more than 90% of universities and research institutes in China. We can believe that the data collected by this company has a high credibility.

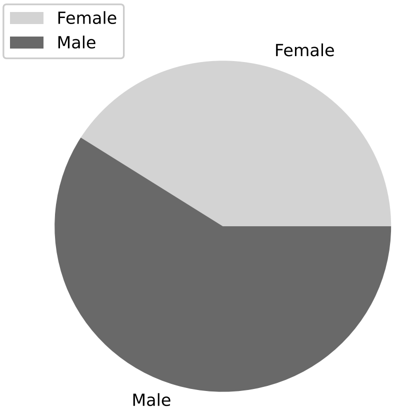

Fig. 2. The Statistics of Gender

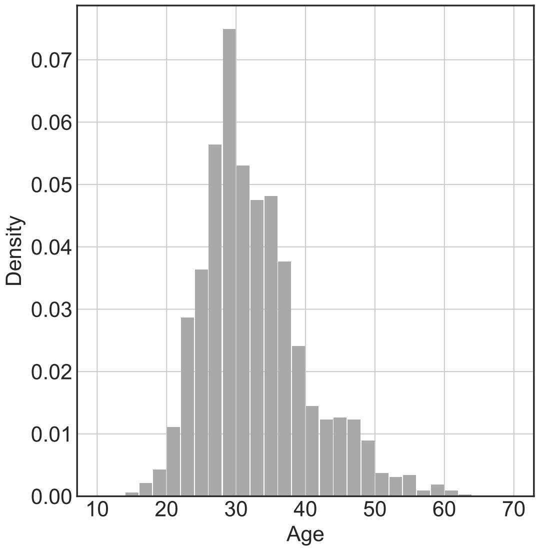

Fig. 3. The Statistics of Age

respondents. Thus, it can be deduced that these survey results have a better and more universal representativeness.

B. Data analysis

Let S be the set of all valid samples. Thus, based on the size of our sample, the cardinality of S is 1661, i.e., $|S| = 1661$. Randomly select 150 “Left” samples and 150 “Staying” samples from S, and use them to form a new set S_{test}. The set S_{test} represents the testing set that will be used later to verify our SVM model, and $|S_{test}| = 300$.

Let training set T be the set difference between set S and set S_{test} that is $T = S \backslash S_{test}$, then $|T| = 1661 - 300 = 1361$.

Let set $U = \{u | u \in T \text{ and } u \text{ is a “Left” sample}\}$ and set $V = \{v | v \in T \text{ and } v \text{ is a “Staying” sample}\}$. Then, we randomly select 50 samples from U and another 50 samples from V to combine them into another testing set S_{test2}.

Based on the definitions above, we have obtained three sets that we need to train and test our model: (1) training set T, (2) testing set S_{test} that will be used to verify our SVM model later, and (3) testing set S_{test2} which is to be used in the training process.

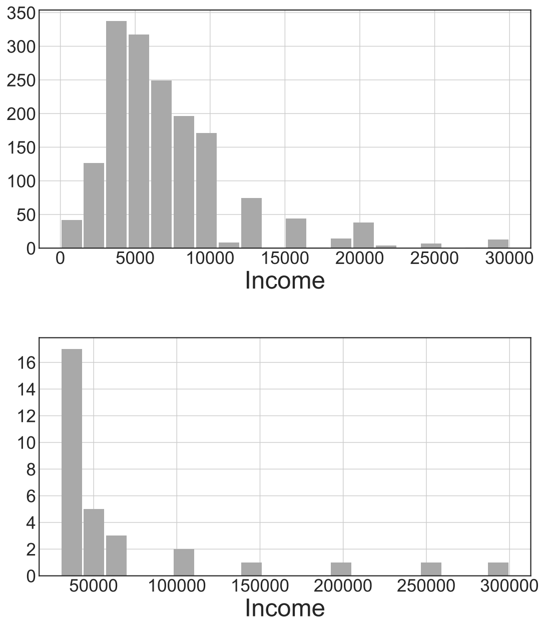

Fig. 4. The Statistics of Monthly Income

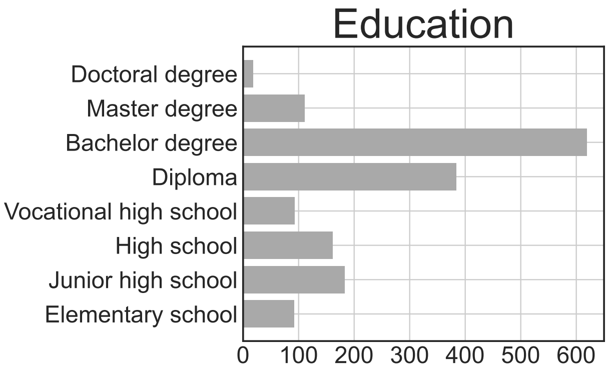

Fig. 5. The Statistics of Highest Education Level

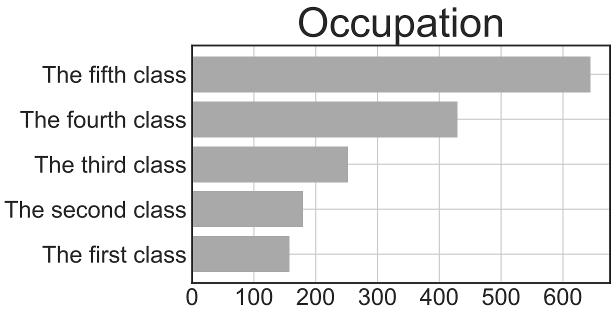

Fig. 6. The Statistics of Occupation

For the SVM configuration, we use RBF (Gaussian) kernel function and set $\sigma = 1$. Using the programming language R, in a CentOS 7.0 Linux server, we train SVM model using set T and test the trained model using set S_{test} and set S_{test2}.

In order to evaluate our SVM model, we use the same training and testing sets to train and test the BP Neural Network model and the Logistic Regression model. We then compared all the results together.

We repeated the experiment 5 times, meaning randomly forms the three set T, S_{test}, and S_{test2} through the way we described earlier 5 times. The 5 experiment results of the

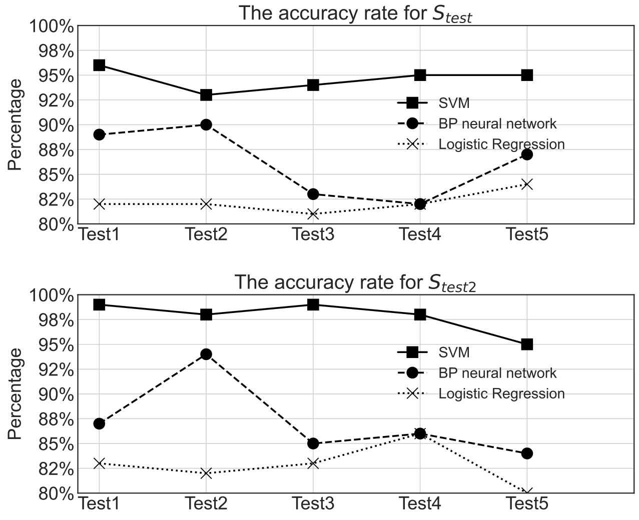

Fig. 7. The 5 experiment results of SVM model,BP Neural Network model and Logistic Regression model

SVM model, BP Neural Network model and the Logistic Regression model are shown in Figure 7. The averages of 5 times of classification accuracy rate are shown in Table II.

Based on the empirical results we obtained and the comparison between SVM model, BP Neural Network model and Logistic Regression model, it is very noticeable that SVM Classification Model possesses excellent abilities in classification and generalization. Furthermore, our SVM model has a better stability in general.

TABLE II
THE AVERAGES OF 5 TIMES OF THE CLASSIFICATION ACCURACY RATE OF SVM MODEL, BP NEURAL NETWORK MODEL AND LOGISTIC REGRESSION MODEL

	SVM	**BP neural network**	**Logistic regression**
The average of the accuracy rate for S_{test2}	97.9%	87.1%	83.2%
The average of the accuracy rate for S_{test}	94.8%	86.0%	82.0%

VI. CONCLUSION

In this paper, to predict LTNL residents' migration in megacities, we established an SVM classification model and used factual survey data to verify the model. Through our experiments, we were able to make the following deductions: Firstly, we found that (a) the SVM classification model we established has a great performance, (b) the classification and generalization of SVM model are better than those of BP Neural Network and Logistic Regression models, and (c) SVM models produce more stable predictions in general.

Secondly, the ten resident classification features: age, education, occupation, income, family status, housing status, leisure status, insurance status, Temporary Residence Permit (TRP) status, and residence time, highly reflect the tendency of LTNL residents' migration in megacities. Hence, creating the SVM model using the data of these particular features plays a pivotal role for the high accuracy rates we achieved.

Lastly, the results presented in this paper have important practical value for various governmental departments in megacities to (a) grasp the migration tendency of different categories of LTNL residents, (b) scientifically control megacities' urbanization, and (c) effectively moderate public services and policies for LTNL residents.

On the conclusions and results we presented on this paper, there are a couple of points to be noted:

- Although most of the megacities in China share similar typical characteristics, every city always has its own unique features. Thus, the application of our approach for other cities should be verified using the local data; and
- In this paper, we adopted the non-probability data sampling and the quantity of our samples is relatively small. We aim to address and improve these issues in our future work.

REFERENCES

[1] J. Jover and I. Diaz-Parra, "Who is the city for? Overtourism, lifestyle migration and social sustainability," *Tourism Geographies*, no. 1, pp. 1–24, 2020.

[2] H. Lin, "The shrinking of Beijing and the rising of Xiong'an: Optimize population migration in terms of transport service," *Discrete Dynamics in Nature and Society*, vol. 2020, pp. 1–6, 2020.

[3] 21st Century Business Reports, http://finance.sina.com.cn/roll/2017-03-02/doc-ifxpvutf3830596.shtml.

[4] G. Fields and Y. Song, "Modeling migration barriers in a two-sector framework: A welfare analysis of the hukou reform in China," *Economic Modelling*, vol. 84, pp. 293–301, 2020.

[5] The State Council of P. R. China, http://www.gov.cn/zhengce/content/2014-11/20/content_9225.htm.

[6] G. K. Zipf, "The P1P2/D hypothesis: On the intercity movement of persons," *American Sociological Review*, vol. 11, no. 6, pp. 677–686, 1946.

[7] T. W. Rogers, "Migration prediction on the basis of prior migratory behavior: A methodological note," *International Migration*, vol. 7, no. 1-2, pp. 13–19, 2010.

[8] F. Simini, M. C. Gonzalez, A. Maritan, and A.-L. Barabasi, "A universal model for mobility and migration patterns," *Nature*, vol. 484, pp. 96–100, 2012.

[9] F. Calabrese, G. D. Lorenzo, and C. Ratti, "Human mobility prediction based on individual and collective geographical preferences," in *International IEEE Conference on Intelligent Transportation Systems*, 2010, pp. 312–317.

[10] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific Reports*, vol. 3, no. 10, p. 2923, 2013.

[11] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.

[12] M. C. Gonzlez, C. A. Hidalgo, and A.-L. Barabsi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[13] F. Luo, G. Cao, K. Mulligan, and X. Li, "Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago," *Applied Geography*, vol. 70, pp. 11–25, 2015.

[14] A. K. Nabi and P. Krishnan, "Methodology for population studies and development." *New Delhi India Sage Publications*, pp. 82–121, 1993.

[15] S. R. Curran and A. C. Saguy, "Migration and cultural change: A role for gender and social networks?" *Journal of International Womens Studies*, vol. 2, no. 3, pp. 54–77, 2001.

[16] C. Duan, "Individual level determinants of interprovincial migration in China. On the effects of time sequence in migration studies," *Population Research*, vol. 24, no. 4, 2000.

[17] M. A. Maria, "International migration: A panel data analysis of the determinants of bilateral flows," *Journal of Population Economics*, vol. 23, no. 4, pp. 1249–1274, 2010.

[18] Y. Chen and S. S. Rosenthal, "Local amenities and life-cycle migration: Do people move for jobs or fun?" *Journal of Urban Economics*, vol. 64, no. 3, pp. 519–537, 2008.

[19] H. Zhou, "Capital form and national policy and inter-provincial migration," *Chinese Journal of Population Science*, vol. 31, no. 5, pp. 42–51, 2007.

[20] J. Tang and Z. Ma, "The selectivity of population migration in China: Analysis based on five data," *Population Research*, vol. 31, no. 5, pp. 42–51, 2007.

[21] Y. Yuan, H. Wang, and F. Chen, "Analysis on factors affecting decision-making of migrant population in large cities of China," *China Population Today*, pp. 36–36, 2011.

[22] M. David and R. Hillel, "Self-selection patterns in Mexico-U.S. migration: The role of migration networks," *Review of Economics & Statistics*, vol. 92, no. 4, pp. 811–821, 2010.

[23] J. Grogger and G. H. Hanson, "Income maximization and the selection and sorting of international migrants," *Journal of Development Economics*, vol. 95, no. 1, pp. 42–57, 2011.

[24] W. Fang and X. Zhou, "An analysis of immigration differences on population migration network in the perspective of occupation types," *Population & Economics*, pp. 21–29, 2015.

[25] Y. Reng, ""Step-by-Step Precipitation" and "Residency Decision Residing" — Analysis of Shanghai's immigrant population residency model," *Shanghai Journal of Economics*, vol. 3, pp. 67–72, 2006.

[26] Z. Meng and R. Wu, "A research on living intentions of floating population," *Population and Development*, vol. 17, no. 3, pp. 11–18, 2011.

[27] J. Zheng and C. Zhu, "The competitiveness of China's service trade and trade liberalization in pilot free trade zone," *Shanghai Journal of Economics*, pp. 122–129, 2014.

[28] X. Cai, J. Zhu, and J. Zhong, "Research on the willingness to move and the social psychological influence of residents in Beijing," *Beijing: Social Science Literature Publishing House*, pp. 73–86, 2016.

[29] P. G. Wang, "The study of inter-provincial floating population's long-term residence intention in Xinjiang from the perspective of social integration theory," *Population & Development*, vol. 21, no. 2, pp. 66–71, 2015.

[30] Y. Yu, C. Liu, and G. Li, "Study on the willingness to settle and the influencing factors of the floating population in the capital — Also on the mode of population control in the capital," *Modernization of Management*, vol. 1, pp. 49–52, 2017.

[31] L. Duan and L. Duan, "Population stratification based on AHP and sample survey data," *Mathematics in Practice & Theory*, vol. 41, no. 22, pp. 247–256, 2011.

[32] Y. Zhao and Z. Men, "Floating population: Social stratification and housing quality — Evidence from the six census data of Changning district in Shanghai," *Population & Development*, vol. 18, no. 5, pp. 59–66, 2012.

[33] Q. Li, "Analysis of the factors affecting the thrust and pulling forces of urban and rural floating population in China," *Social Sciences In China*, vol. 1, no. 151, pp. 125–136, 2003.

[34] S. Tang and J. Feng, "Disparities of life satisfaction within the new generation of rural migrants: A case study in Jiangsu province," *Population & Economics*, pp. 52–61, 2016.

[35] L. Qiu, "Professional status: An indicator of social stratification: A study of Shanghai's social structure and social stratification," *Sociological Research*, vol. 3, pp. 18–33, 2001.

[36] H. Bruce, V. Ruut, and W. Alex, "Top-down versus bottom-up theories of subjective well-being," *Social Indicators Research*, vol. 24, no. 1, pp. 81–100, 1991.

[37] Z. X. Wei, "A region-specific comparative study of factors influencing the residing preference among migrant population in different areas: Based on the dynamic monitoring & survey data on the migrant population in five cities of China," *Population & Economics*, vol. 4, pp. 12–20, 2013.

[38] N. Liu, "Integrate into urban society or adapt to urban life — A study on rural migrants in Beijing," *Social Sciences of Beijing*, vol. 7, pp. 61–67, 2015.

[39] Z. Zhai, C. Duan, and Q. Bi, "The floating population in Beijing: An update," *Population Research*, vol. 31, no. 2, pp. 30–40, 2007.

[40] Y. N. Sheng, "The determinants of whole family migration and migration behaviors decision," *Population Journal*, vol. 36, no. 3, pp. 71–84, 2014.

[41] W. U. Ye-Miao, "Notion support and life anticipation about the peasant workers' floating," *Journal of Southwest China Normal University*, vol. 30, no. 6, pp. 65–71, 2004.

[42] Y. Chen, Y. Dang, and G. Dong, "An investigation of migrants' residential satisfaction in Beijing," *Urban Studies*, vol. 57, no. 3, SI, pp. 563–582, 2020.

[43] Y. Wang, Y. Li, Y. Huang, C. Yi, and J. Ren, "Housing wealth inequality in China: An urban-rural comparison," *Cities*, vol. 96, 2020.

[44] Q. Meng and G. Qiao, "A research on floating population's life quality in the view of leisure in metropolis," *Urban Studies*, vol. 17, no. 5, pp. 4–7, 2010.

[45] P. W. Robert, "Constructed leisure space," *Annals of Tourism Research*, vol. 28, no. 3, pp. 581–596, 2001.

[46] X. Fen, "The dilemma and outlet of the protection of floating population rights in megacities — Taking Beijing as an Example," *Exploration and Free Views*, vol. 1, pp. 26–28, 2014.

[47] V. N. Vapnik, *The nature of statistical learning theory*. New York: Springer-Verlag, 1995.

[48] Y. Maleh and A. Ezzati, "Lightweight intrusion detection scheme for wireless sensor networks," *IAENG International Journal of Computer Science*, vol. 42, no. 4, pp. 347–354, 2015.

[49] U. Iruansi, J. R. Tapamo, and I. E. Davidson, "Classification of power-line insulator condition using local binary patterns with support vector machines," *IAENG International Journal of Computer Science*, vol. 46, no. 2, pp. 300–310, 2019.

[50] Q. Zheng, X. Tian, M. Yang, and H. Su, "The email author identification system based on support vector machine (SVM) and analytic hierarchy process (AHP)," *IAENG International journal of computer Science*, vol. 46, no. 2, pp. 178–191, 2019.

[51] L. I. Yun-Fei and X. F. Hui, "The classification model for stock investment value based on SVM," *China Soft Science*, vol. 1, pp. 135–140–66, 2008.

[52] Y. Jiang and C. Xu, "Analysis of classification model of companies' financial performance based on integrated support vector machine," *Chinese Journal of Management Science*, vol. 17, no. 2, pp. 42–51, 2009.

[53] W. Fan, P. Wang, Y. Yuan, and H. Y. Sun, "Heavy rain/hail classification model based on SVM classification credibility," *Journal of Beijing University of Technology*, vol. 41, no. 3, pp. 361–365, 2015.

[54] Z. Huang and F. Tang, "Performance and choice of Chinese text classification models in different situations," *Journal of Hunan University (Natural Sciences)*, vol. 46, no. 7, pp. 144–153, 2016.

[55] D. Dietler, A. Farnham, K. de Hoogh, and M. S. Winkler, "Quantification of annual settlement growth in rural mining areas using machine learning," *Remote Sensing*, vol. 12, no. 2, pp. 235–, 2020.

[56] H. Li, *Statistical learning method*. Beijing: Tsinghua University Press, 2012.

[57] X. Li, "Analysis of the social class structure of permanent resident population and its countermeasures — Taking Beijing as an Example," *Theory Monthly*, vol. 4, pp. 128–132, 2016.