Population Monitoring and Forecasting Model Based on XG-Boost

Shihan Ma[1,a], Qiang Li[1,2,b,*]

[1]College of Engineering, Hebei Normal University, Shijiazhuang, 050010, China
[2]Hebei Provincial Key Laboratory of Information Fusion and Intelligent Control, Shijiazhuang, China
[a]2369485185@qq.com, [b]liqiang01@hebtu.edu.cn
[*]Corresponding author

Abstract: *The expansion of urban areas and the rise in population mobility have led to significant alterations in population structure and urban configuration, presenting substantial problems for monitoring population dynamics. Due of its immediacy and reliability, big data, including mobile phone signalling and geographic position, offers possibility for population dynamic monitoring and precise management. The study investigated the population size and age structure in Hebei province utilizing large data from Hebei Unicom mobile phone signalling and Baidu Huiyan, employing complex network analysis to create and examine population migration. Based on data from the sixth and seventh population censuses in Hebei Province, a monitoring model for population size, birth rate, and transient population has been established, employing the population development equation as the foundational model and the XG-Boost machine learning algorithm for identification. The model is ultimately adjusted by cellular signalling data and validated using census data. The result shows that the population growth rate is slowing down, aging intensification, and a spatial distribution of population movement characterized by a "multi-core" structure in Hebei Province. And the population monitoring model developed by cellular phone signals and other large datasets demonstrates practicality and accuracy, efficiently predicting population size and structure.*

Keywords: *Population dynamics monitoring, XG-Boost model, cellular signalling data, population forecasting*

1. Introduction

The advancement of urbanization in China has resulted in a significant population surge, leading to the ongoing expansion of metropolitan areas and enhanced interregional connectivity. The significant mobility and variable behavior of the inhabitants have substantially altered the basic urban configuration.[1] Monitoring and predicting population dynamics using big data is essential for addressing the aforementioned issues. Data analysis can address issues such as intra-regional population flow, migration, social integration of transient populations, urban-rural income disparity, medical insurance accessibility, and fertility intentions, while also facilitating a deeper examination of the laws and characteristics governing the comprehensive level of population development within the region [2-4]. It offers extensive foundational data support for demographic shifts, economic development, resource management, and environmental conservation in the region.

While the national population census is the most precise approach for assessing population figures, it fails to provide timely data for management due to its extended survey duration, infrequent occurrence, delayed outcomes, and significant impact of sampling variables. Simultaneously, statistical population data delineated by administrative units possesses a considerable spatial scale and necessitates substantial human and material resources. The increasing number of suburban units and urban social issues are emerging, leading to a growing necessity for a nuanced representation of population distribution. Cellular signaling data constitutes a novel category of extensive data source. Its distinctive application benefits, including real-time capabilities, comprehensiveness, and extensive coverage of temporal and spatial travel, serve as a crucial foundation for addressing the deficiencies of the existing demographic statistics system.

Big data serves as a crucial foundation for decision-making, service provision, evaluation, and forecasting. Utilizing Tencent's location big data from 2016 and 2018, Li Tianzi and Lu Mingjun [5] employed social network analysis and the QAP regression model to investigate the geographical structural characteristics and determinants of China's population mobility network. Additionally, Song Junru [6] examined the inter-provincial and intra-provincial flow characteristics of Jilin utilizing geographic information and population census data from Baidu Huiyan. Following the advent of advanced algorithms like deep learning, contemporary research employs the cohort factor technique and CA-Markov model to simulate and forecast future population size and land use patterns in the region, while estimating population size in conjunction with POI spatial big data. He Yanhu [7] et al. developed an advanced regional spatial distribution simulation model for forecasting future population size in the Pearl River Delta. Niu Xinyi [8] utilized mobile locating big data to replicate the population size of Wuhan, enabling more precise monitoring of population fluctuations over a brief duration.

Hebei Province is situated within the Beijing-Tianjin-Hebei metropolitan area, characterized by significant demographic heterogeneity and movement. The prior mechanical governance paradigm is insufficient to address current requirements, necessitating the promotion of standardized processes and methodologies. Utilizing cellular signaling, location and other extensive mobile communication datasets, this research aims to cohesively amalgamate the disparate service requirements of the transient population, establish a population monitoring framework, and subsequently employ big data technology to enhance the precision of governance processes. This approach will transition population monitoring from a conventional model that segregates components such as individuals, locations, events, and objects to a model characterized by the profound integration of all elements.

This paper is based on cellular signaling and Baidu migration big data, adopts a discrete population development equation as the basic model, and integrates the sixth and seventh Hebei population census data, uses the limit gradient boosting tree model (XG-Boost) to evaluate the complicated parameters of birth, death, mobility, and public service accessibility within a population, determines the transfer population rate matrix in the population development equation, and conducts a comprehensive evaluation and error analysis on the accuracy of the model using Hebei census data. To facilitate the modeling and monitoring of population dynamics, work-residential separation, traffic and commuting in Hebei Province, as well as the prediction of population size and structure. Figure 1 shows the population movement monitoring program in Hebei Province.

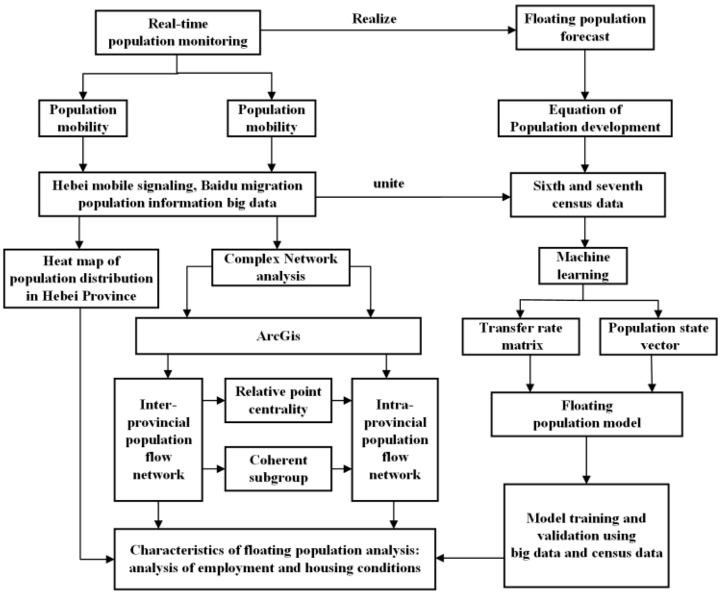
Real-time population monitoring
Realize
Floating population forecast
Population mobility
Population mobility
Equation of Population development
Hebei mobile signaling, Baidu migration population information big data
unite
Sixth and seventh census data
Heat map of population distribution in Hebei Province
Complex Network analysis
Machine learning
ArcGis
Transfer rate matrix
Population state vector
Inter-provincial population flow network
Relative point centrality
Coherent subgroup
Intra-provincial population flow network
Floating population model
Model training and validation using big data and census data
Characteristics of floating population analysis: analysis of employment and housing conditions

Figure 1: Hebei Province population flow monitoring plan design.

2. Data Sources and Processing Methods

2.1. Data Sources and Statistical Criteria

2.1.1. Data source

This topic utilizes signaling data from the Hebei mobile phone communication network, including position signaling MC and MME port data. The MC port data refers to the locational information of 2G/3G mobile devices. When the population's location changes within the designated area, or when communication activities such as calls and text messages take place, a location data signal will be generated, and information will be collected every 45 minutes while the mobile phone remains entirely immobile. An MC data signal is produced for an average duration of approximately 20 minutes daily. MME port data comprises 4G mobile phone location information, created when individuals enter and exit a base station's coverage area.

2.1.2. Data cycle

The figures for 2020 encompass the entire year, while the statistics for 2021 cover the period from January to December of that year. When the phone is entirely immobile, it produces data at five-minute intervals. Omit IoT card data utilizing the roster of IoT card numbers recorded in the billing accounting system. Utilize voice call CDRs to eliminate numbers that have not received voice calls for a duration of six months. The 13-digit numbers commencing with 106 and 144 are omitted.

2.1.3. Geographical definition

11 prefecture-level cities (Shijiazhuang, Tangshan, Qinhuangdao, Handan, Xingtai, Baoding, Zhangjiakou, Chengde, Cangzhou, Langfang, Hengshui) and 2 county-level cities (Dingzhou, Xindi) in Hebei Province. Xiongan New Area comprises Xiongxian County, Rongcheng County, Anxin County, Gaoyang County, Longhua Township, Renqiu City, Qijianfang Township, Gou Gezhuang Town, and Maozhou Town.

2.1.4. Statistical methodologies for demographic indicators

- Stable population: A stable day is defined as the occurrence of over 10 hours in the region on the same day, and it is classified as a stable population for that month if the number of stable days in a year exceeds one-half.

- Population dimensions (gender, age, residential registration location): stable population-associated identification number, gender identification bit to differentiate between male and female; Stabilize the population linked to the ID number and compute the age based on the birth date. Children aged 0-17 were used for precipitation analysis

- Migration population: Current year intra-provincial stability minus previous year intra-provincial stability, with the current year serving as the inter-provincial migration reference. In this context, Hebei users consider their current year roaming location as their stability for the current year.

2.2. Analysis of population size and population structure data

In Figure 2, the stable population of Hebei Province and prefecture-level cities in 2020 amounted to 75,996,000, and the stable population in 2021 amounted to 74,697,000, a year-on-year decrease of 1.71%. It can be seen that the population size of cities in Hebei Province does not change much, and there is a positive correlation between population size and urban location, economic level and traffic conditions.

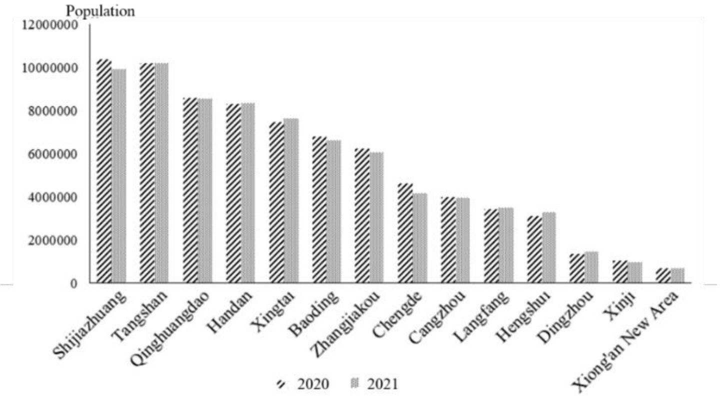

Figure 2: Stable population by region in Hebei Province in 2020.

From the perspective of gender (see Figure 3), the stable population of the province is more male than female, accounting for about 53% of males and 47% of females. The ratio of male to female in each city is basically in line with that of the whole province.

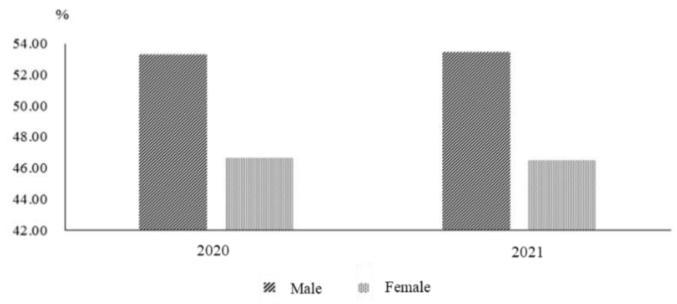

Figure 3: Gender distribution of population in Hebei Province in 2020.

According to the data of Table 1. The male population of all age groups in the stable population of Hebei Province in 2020-2021 is more than the female population. The proportion of children aged 0-14 in the total population and the size of the working-age population aged 15-64 has decrease, but the number of elderly people aged 65 and above has increased from the perspective of age structure.

Table 1: Population age changes in Hebei Province from 2020 to 2021.

Age (years)	2020		2021	
	Male	Female	Male	Female
0-14	3651269	3195741	3611076	3150506
15-64	33473097	30284423	32647247	29596098
65+	3403706	1987594	3748281	2152624

Figure 4 to 5 illustrates the demographic age distribution of Hebei Province for the years 2020 and 2021.The population pyramid has a steady aging trend, characterized by a reduction at the base. But the working-age demographic remains predominant, particularly among individuals aged 30 to 59 years.

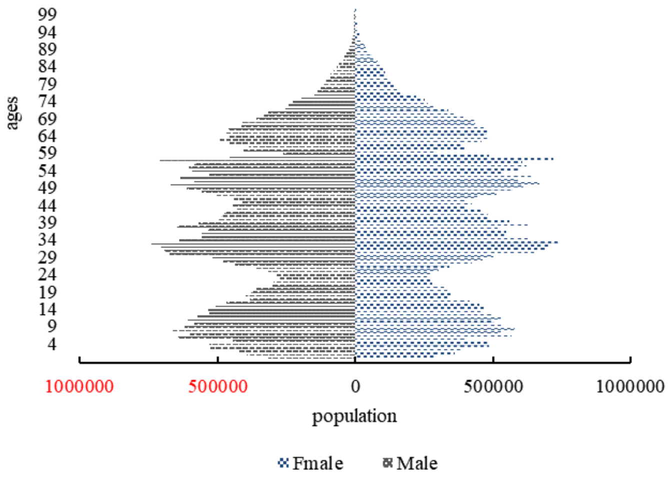

Figure 4: Age pyramid of the population in Hebei Province for the year 2020.

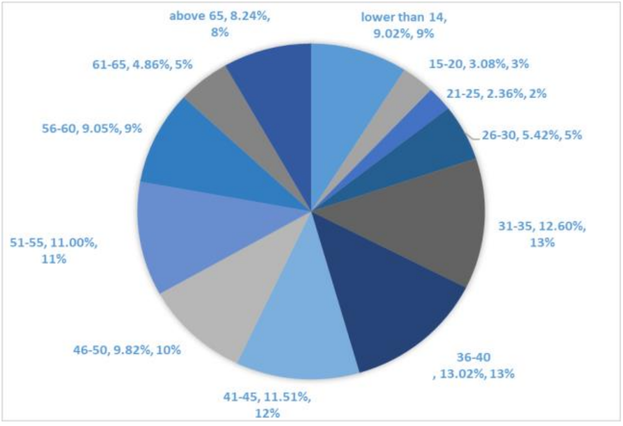

Figure 5: Age pyramid of the population in Hebei Province for the year 2021.

Sources of data: The population migration data is sourced from the 2020 Hebei Unicom cellular signaling data and Baidu Huiyan big data, while the GNP data for each region is derived from the economic census conducted by the National Bureau of Statistics.

2.3. Data processing method

2.3.1. Hebei Province retained training data of population model by gender and age

Based on the data of the sixth population census in 2010 and the seventh population census in 2020, the total population of Hebei Province by sex and age and the death rate of each age in the 10 years from 2010 to 2020 are estimated. The total population x^* is taken as input data. The annual death population by gender and age is used as output data y^* as to form the training data set of XG-Boost model.

2.3.2. Training data of population transfer rate matrix

Then, on the basis of the data of 2010-2020 in 2.3.1, the remaining population of the first year follows:

$$ {}^a h_i^t = {}^a x_i^t - {}^a d_i^t \tag{1} $$

${}^a x_i^t$ is the number of people whose age is i in year t, ${}^a d_i^t$ is the number of deaths whose age is i in year t, where $a = m$ represents males and $a = f$ represents females; By calculating the retained population ${}^a h_{i+1}^{t+1}$ with age $i+1$ in year $t+1$. According to the above method, the transferred population ${}^a w_i^t$ with age i in year t can be obtained. After collation, a dataset of migrant population by sex and age is available for the period 2010-2020.

3. Research method

3.1. Intricate network analysis

The complex network analysis method emerged from the stochastic network theory introduced by Paul Erdos and Alfred Renyi [13] is employed in the examination of population flow size and geographical organization. In the population flow network, the direction of population flow follows the connections between nodes, comprising the origin and destination of population flow, which are characterized as the degree of entry and the degree of departure in complex networks. The entry degree signifies the node's appeal, while the exit degree indicates the node's control authority in the network.

3.1.1. Node centrality

In a directed network, a node's centrality can be categorized into inflow and outflow degree. Inflow degree refers to the intensity of incoming flow to the node, while outflow degree pertains to the intensity of outgoing flow from the node [10]. The centrality of a node just accounts for those directly linked to it, which is a crucial metric for assessing the robustness of its external connections within the network.

Let X_{j-i} represent the inner degree of city node i, and X_{i-j} denote the outward degree of city node i. The overall degree value of city node i is given by

$$ A_i = \sum_j X_{j-i} + \sum_j X_{i-j} \tag{2} $$

A_i denotes the weighted total degree value of node i, reflecting the centrality degree of urban nodes and the overall external contact strength. Given that the population flow network is a directed multivalued network, where $X_{i-j} \neq X_{j-i}$, the idea of centrality T_i is introduced to further ascertain the population dispersion capacity of urban nodes.

$$ T_i = A_i / \sum_{i=1}^{n} A_i (i = 1, 2, ..., n) \tag{3} $$

T_i denotes the centrality of city i, and n signifies the number of city nodes. The cities in Hebei

Province were categorized according to their degree of centrality using ArcGIS (Arc Geographic Information System).

3.1.2. Cluster Analysis

Cluster analysis enables the identification of group quantities inside the population flow network of Hebei Province and elucidates the interrelations among these groups to characterize the network's local aggregation [11]. The clustering coefficient C_i of city node i is defined as:

$$C_i = \frac{2}{k_i(k_i-1)}\sum_{j\neq k} a_{ij}a_{ik}a_{jk} \tag{4}$$

a_{ij}, a_{ik}, a_{jk} denote the relational connections among nodes i, j, and k. If two nodes are directly linked, the value is 1. Otherwise, it is 0. k_i denotes the quantity of edges associated with this node.

3.2. Equation of Population Dynamics

3.2.1. Fundamental forecasting model

The gender-discrete population development equation serves as the foundational model for analyzing population size, structure, and migration trends, while machine learning techniques are employed to ascertain the transfer equation for forecasting alterations in total population, gender composition, migration scale and direction, as well as the total number of births. The equation for population development can be articulated as:

$$\begin{aligned} &X^s(t+1) = H^s(t)X^s(t) + [1\,0\cdots 0]'\,\eta_{0A}(t)\,y^s(0)_t + W^s(t)X^s(t) \\ &y^s(0)_t = \beta(t)S^s(t)F^T(t)X^f(t) \end{aligned} \tag{5}$$

3.2.2. Configuration of model parameters

- Demographic retention matrix

The statistical data from the 7th population Census in Hebei Province provided the total population by age group and sex, as well as the mortality figures for each age group and sex, enabling the calculation of the natural survival rate $h_i(i=1,2,\ldots,18)$ for each age group. So, the Leslie matrix is introduced.

$$H^s(t) = \begin{bmatrix} 0 & \cdots & \cdots & 0 & 0 \\ h_1 & 0 & \cdots & 0 & 0 \\ \vdots & h_2 & \cdots & \vdots & 0 \\ 0 & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & h_{i-1} & h_i \end{bmatrix} \tag{6}$$

The natural survival rate h_i for each age group can be determined using the following formula:

$$h_i = 1 - d_i = 1 - \frac{D_i}{X_i} \tag{7}$$

where, d_i represents the mortality rate of each age cohort, D_i denotes the number of fatalities within each age cohort, and X_i signifies the total population of each age cohort. Following data mining, the matrix for population retention rates can be derived:

$$H^s(t) = \begin{bmatrix} 0 & \cdots & \cdots & 0 & 0 \\ 1-d_1 & 0 & \cdots & 0 & 0 \\ \vdots & 1-d_2 & \cdots & \vdots & 0 \\ 0 & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1-d_{i-1} & 1-d_i \end{bmatrix} \tag{8}$$

- Age-specific rate vector of the transient population

Utilizing the total population and mortality statistics by age and sex from 2010 to 2020, the floating population by age and sex from 2010 to 2019 may be derived through cross-comparison in alternate years. The population movement matrix is presented below:

$$W^s(t) = \begin{bmatrix} w_1 & \cdots & \cdots & 0 & 0 \\ 0 & w_2 & \cdots & 0 & 0 \\ \vdots & 0 & \cdots & \vdots & 0 \\ 0 & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & w_i \end{bmatrix} \tag{9}$$

where $w_i = \frac{m_i}{x_i} = \frac{hx_i^{(1)} - hx_i^{(0)}}{x_i}$, m_i is age-specific transfer population, $hx_i^{(0)}$ is the retained population of each age in the first year, $hx_i^{(1)}$ is the retained population of each age in the subsequent year, and x_i is the total population of each age in the current year, dx_i is deaths by age. Including:

$$hx_i = x_i - dx_i \tag{10}$$

3.2.3. Machine Learning – XG-Boost Model

XG-Boost is an efficient implementation of gradient boosting [9], utilizing either a linear classifier or a tree as its base learner. Considering a training dataset $D = \{(x_i, y_i)\}(i = 1, 2, \cdots, N, x_i \in R^M, y_i \in R)$ including N samples and M features. Upon training the XG-Boost algorithm, an ensemble model can be derived by using K Classification and Regression Tree (CART) decision tree functions:

$$y_i^* = \sum_{k=1}^{K} f_k(x_i) \qquad f_k \in F \tag{11}$$

Where represents the output of the XG-Boost model; $F = \{f(x) = w_{q(x)}\}(q : R^M \rightarrow T, w \epsilon R^T)$ represents the set of CART trees, while q denotes the tree's structure during the construction of the tree model. A feature segmentation point is designated as leaf node j, T denotes the total number of leaf nodes, and w_j signifies the weight of leaf nodes, indicating the extent to which the node contributes to the overall tree. f_k is associated with a certain tree structure q and the weight vector w of the leaf nodes. The significance of y_i^* is to associate the sample data with the respective leaf node of each decision tree. Subsequently, aggregate the weights of K leaf nodes associated with the sample.

To quantify the discrepancy between the output and actual value of the training model, the loss function is articulated as follows:

$$L = \sum\nolimits_{i=1}^{N} l(y_i^*, yi) + \sum\nolimits_{k=1}^{K} \Omega(f_k) \tag{12}$$

Here, $l(y_i^*, yi)$ denotes the training loss function, with the logarithmic loss function and mean square error loss function employed to quantify the discrepancy between the output of the training model and the actual value. $\Omega(f_k)$ is the standard term employed to regulate model complexity, mitigate overfitting, and improve the model's generalization capacity. Its representation is as follows:

$$\Omega(f) = \gamma T + 0.5\lambda \sum_{j=1}^{T} w_j^2 \tag{13}$$

The initial term γT is employed to limit the number of nodes in the central portion of the tree model, hence preventing an excessively intricate tree structure. The second term regulates the weight distribution of leaf nodes. Typically, λ is set to 1, while γ is modified as required.

The procedure can be outlined as: When the model's accuracy falls below a predetermined threshold, the quantity of tree models K is no longer augmented, resulting in the final model $y_i^* = \sum_{k=1}^{K} f_k\left(x_i\right)$; The gradient lifting procedure is employed to minimize the loss function value of the decision tree, hence generating a new decision tree. The cessation of tree splitting can be regulated by two parameters: when the maximum depth of the tree $d_{\max}$ attains a predetermined value, or all leaf node splitting strategies fail to further reduce the loss function, resulting in the termination of tree splitting. The optimal weight vector w associated with the tree structure q can now be computed, yielding a new tree function f.

3.2.4. Queue Element Method

The population is segmented into various cohorts based on age and gender [12], while population change is analyzed through birth, death, and migration. Cohort is a group of individuals experiencing a demographic event simultaneously. This birth cohort expands throughout time and with age, diminishing due to mortality, augmenting and contracting via emigration. Upon reaching childbearing age, these individuals will generate a new demographic cohort through births to ensure continued population growth. Annually, the population across all age groups can be regarded as a distinct birth cohort. Population forecasts are often derived from the intrinsic principles of population dynamics, utilizing death, and migration rates in relation to the current population size and composition.

4. Identification, verification and application of population monitoring model

4.1. Maintained demographic models categorized by gender and age

The XG-Boost model serves as the training and forecasting tool for population data. The population development model stipulates that the input parameters of the training model consist of the population count categorized by sex and age, while the output parameters represent the mortality count of the population, also categorized by sex and age and the expected outcomes were consistent with the 2020 census data in Figure 6-1 to 6-2.

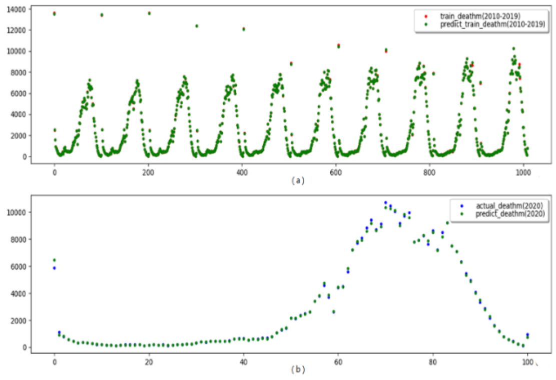

Figure 6-1: Training process and prediction results of death male population by age group.

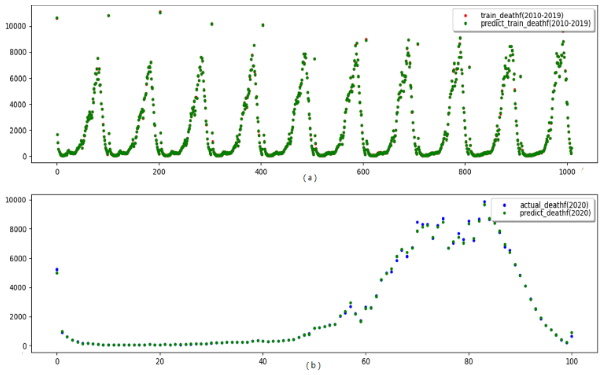

Figure 6-2: Training process and prediction results of death female population by age group.

4.2. Migration model specific to gender and age

Simultaneously, for population migration data. The output parameters comprise the population flow figures also categorized by gender and age when the input parameters of the training model consist of the total population segmented by gender and age.

Figures 6-3 and 6-4 illustrate the results of training(a) and forecasting(b) migration data by gender and age, which has a high prediction accuracy for the retention population.

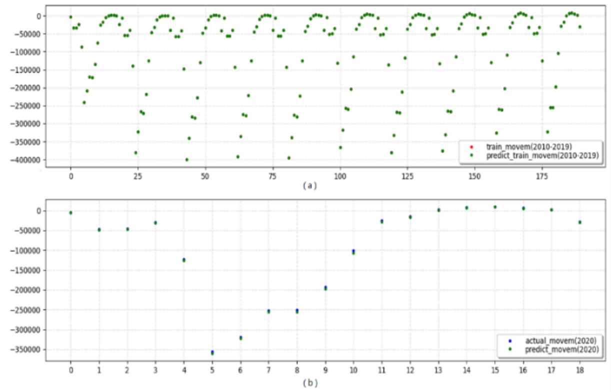

Figure 6-3: Training and forecasting male migration data by age.

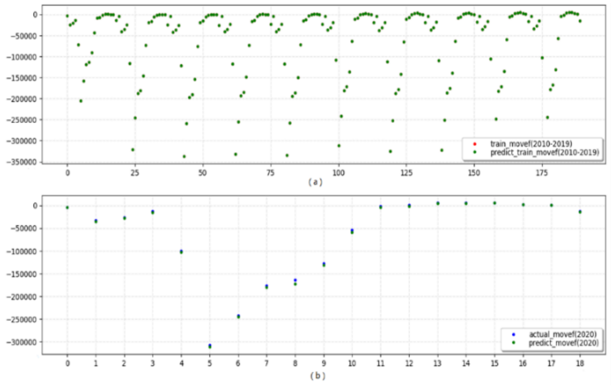

Figure 6-4: Training and forecasting of female migration data by age.

4.3. Revision of the model

The updated population monitoring model utilizes signaling data from Hebei Unicom mobile phones for the period of January to December 2021 and total population data by gender and age from the seventh population census of Hebei Province least squares method to establish the relationship between the two datasets to revise the population monitoring model.

Linear fitting of the data reveals that the link between mobile phone signaling data and population census adheres to the following model:

$$y_p = 1.18 y_b \tag{14}$$

Where y_p represents the population census data and y_b denotes the mobile phone signaling data.

4.4. Application of the Model

The updated model forecasts the demographic changes over the next five years (see Figure 7):

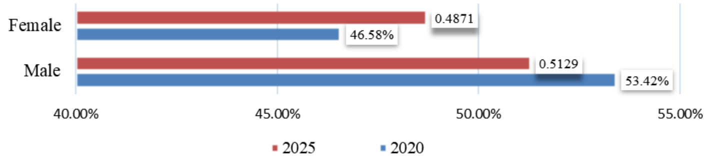

Figure 7: The gender ratio of population in Hebei province in 2020 and 2025.

From a gender standpoint, the province's constant population comprises of 53% males and 47% females. The forecast for 2025 suggests a male-to-female ratio of 0.5129:0.4871.

Forecasting the male and female demographic composition of Hebei Province for the forthcoming five years allows policymakers to foresee and mitigate prospective gender disparities. A distorted gender ratio may influence marriage patterns, fertility rates, and the general demographic stability of the area.

Secondly, comprehending the future gender demographic composition facilitates the strategic planning and distribution of public resources, including healthcare, education, and social services, thereby addressing the distinct requirements of both genders. The identified patterns, including the declining percentage of children and the working-age population alongside the rising number of old individuals, underscore the necessity for focused measures to address these demographic changes.

5. Conclusions

This research examined the demographics size and age composition of Hebei Province. A monitoring model for population size, birth rate, and migratory population was developed using complex network analysis, population dynamics equations, and machine learning methodologies. It takes into account the statistical mechanisms inherent in population change and also avoids, to a certain extent, prediction errors caused by pre-determined parameters.

In particular, it has certain advantages in analyzing the population structure by age and gender and predicting the migration of the population. With the development of mobile communication data, the model has a good application prospect for the training and prediction of large-scale data.

The analysis of population dynamics via big data transforms abstract data points into meaningful application value. It enables the compilation of information on fertility, infant mortality, and maternal mortality, allowing for the evaluation of age-specific fertility trends and the real-time monitoring of essential human attributes, which is helpful for government to implement effective policies that attract population influx.

Acknowledgement

This research was funded by Soft Science Research Project of Innovation Ability Improvement Plan in Hebei Province (Grant number: 23556103D.)

References

[1] Smith, S.K., Morrison, P.A. (2005). In: Poston, D.L., Micklin, M. (eds.) Small-Area and Business Demography, pp. 761–785. Springer, Boston, MA.

[2] Tayman, J. (2011). Assessing uncertainty in small area forecasts: state of the practice and implementation strategy. Popul. Res. Policy Rev.30, 781–800.

[3] Diamond, I., Tesfaghiorghis, H., Joshi, H. (1990). The uses and users of population projections in Australia. Aust Popul Assoc 7,151–170.

[4] Wei, H., Chen, Z., & Zhang, G. (2020). Innovation in demographic methods with the help of big data from operators. China Statistics, (05), 56-57.

[5] Li, T., & Lu, M. (n.d.) (2022). Characteristics and influencing factors of China's population mobility network: An analysis based on Tencent location big data. Contemporary Economic Management, 1-12.

[6] Song, J. (2021). Analysis of population mobility characteristics in Jilin Province: Based on Baidu Huiyan big data. In 2020/2021 China Urban Planning Annual Conference and 2021 China Urban Planning Academic Season (pp. 504-514).

[7] He, Y., Gong, Z., & Lin, K. (n.d.) (2022). Simulation study on the fine spatial distribution of future

population in the region based on geographic big data and multi-source information fusion: Taking the Pearl River Delta as an example. Geographical Science, 1-10.
[8] Niu, X., Lin, S., Qin, S., et al. (2020). Technical approaches to urban population size monitoring supported by mobile positioning big data. Contemporary Architecture, (12), 39-43.
[9] Chen, T., & Guestrin, C. (2016). XG-Boost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
[10] Singh, A., Singh, R. R., & Iyengar, S. R. S. (2020). Node-weighted centrality: a new way of centrality hybridization Computational Social Networks, 7, Article number: 6.
[11] Scitovski, R., Sabo, K., Martínez-Álvarez, F., & Ungar, Š. (2021). Cluster Analysis and Applications. Springer.
[12] Glynn, P. W. (2022). Queueing theory: past, present, and future. Queueing Systems, 100, 169-171
[13] Erdős, P., & Rényi, A. (1960). On the Evolution of Random Graphs. Publication of the Mathematical Institute of the Hungarian Academy of Sciences, 5, 17-61.