Journal of Electronics and Information Science (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2023.080506
ISSN 2371-9524 Vol. 8 Num. 5

Population prediction in China based on maximum information coefficient and NAR-BP neural network

Fei Zeng*

School of Mathematics and Computer Science, Gannan Normal University, Ganzhou, 341000, China

**Corresponding author: 2570103360@qq.com*

Keywords: NAR-BP Dynamic Neural Network, Primary Constitution Analysis Model, GA-BP neural network, Census

Abstract: This paper studies the issue of national population census. Firstly, the paper collects census related data, establish a maximum information coefficient model, and preprocess the data. Then, it establishes a dynamic neural network prediction model based on NAR-BP to predict the total population of China in 2030. Furthermore, a PCA based NAR-BP dynamic neural network prediction model was established to predict the proportion of males and urban population in China by 2030. Finally, a neural network optimization model based on GA-BP was established to obtain the optimal search term. Based on the analysis of experimental results, it is proven that the frequency of chinese population census is appropriate to be a cycle of 10 years.

1. Introduction

Throughout the country, both domestically and internationally, changes in population size are regarded as important content and indicators to measure the development, prosperity, and social civilization level of a region[1]. The national census is organized by the state to conduct a comprehensive survey and registration of the existing population census sites in the country, household by person, in accordance with the law. The focus of the census is to grasp the changes in the existing population, gender ratios, and urban-rural population data in various regions, in order for the state to formulate next development policies. At present, the interval between national population censuses in China is about 10 years. From 1949 to 2021, China conducted 10 population censuses. It would be very meaningful to provide effective predictions of various census data through mathematical modeling.

In this paper, we tend to solve several essential problems. The total population of China in 2030, the proportion of males and the proportion of urban population in China in 2030 and the reasonable explanation for the frequency of the current census.

Firstly, consider using the maximum information coefficient to screen numerous factors, and then combine it with NAR-BP neural network[2] to establish a prediction model. Due to the mutual influence of various influencing factors, we consider using correlation statistical methods that require less data to explore the potential information and internal connections of the data, and obtain the correlation between population and various factors. We use principal component analysis to reduce

the dimensionality of the data, and then perform linear regression on the reduced dimensionality data to identify and rank the factors that directly affect the male population and urban population. Then the NAR-BP dynamic neural network model is used to predict the proportion of males and urban population in 2030[3-4].

2. Preliminary

2.1 Assumption

We do not consider relocating our population abroad, nor do we consider relocating foreigners to China.

2.2 NAR neural network

NAR neural network is a dynamic neural network model based on time series[5], where the inputs and outputs of the model are synthesized based on the dynamic results of the system before that time. This article selects 6 influencing factors through the MIC algorithm and uses the NAR neural network to predict the influencing factors.

2.3 Primary Component Analysis

The basic principle of principal component analysis (PCA)[6]: The principal component analysis method mainly concentrates information scattered on a set of variables onto certain comprehensive indicators, namely principal components. Each principal component is a linear combination of the original variables, with orthogonal relationships between the principal components, which can reduce the dimensionality of the multivariate time series, remove redundant information, reduce some noise contained in the multivariate time series, and reflect the correlation between different variables. When the sample data has a large number of dimensions and a complex structure, using principal component analysis can simplify the input samples, reduce training time, improve training efficiency, and achieve the goal of improving the generalization ability of the neural network, as shown in Table 1.

2.4 Symbols notation

Table 1: Neural network structure

Symbols	Notation
$B(n)$	xy upper limit of the number of grilles
I	I Maximum mutual information between two variables
n_x	x-axis number of grilles
n_y	y-axis number of grilles
xi	Population indicators
yi	Urban population indicators

3. Experiments

3.1 NAR-BP dynamic neural network prediction based on maximum information coefficient

This article collects the total population, per capita GDP, proportion of urban population, proportion of males, employment rate, birth rate, mortality rate, fertility rate, number of medical institutions, medical expenses, education expenses, number of marriages, age distribution between 0-

14 years old, 14-65 years old, and proportion of people aged 65 and above from 1978 to 2020, as shown in Table 2 and Table 3.

Table 2: Influencing of population (1)

Years	Total (10thousands)	Average-GDP (10thousands)	Ratio of Urban and Village	Male ration	Ration of employed	Birth rate	Death rate
1978	95616.5	3645.22	17.90	48.68	77.16	22.21	7.02
1979	96900.5	4062.58	18.62	48.68	77.11	21.34	7.02
1980	98123.5	4545.62	19.36	48.68	76.83	20.89	6.71
1981	99388.5	4891.56	20.12	48.68	76.60	20.89	6.63
1982	100863.0	5323.35	20.90	48.68	76.40	21.26	6.59
1983	102331.0	5962.65	21.55	48.68	76.15	21.90	6.58
1984	103682.5	7208.05	22.20	48.69	75.86	22.68	6.59
1985	105104.0	9016.04	22.87	48.69	75.60	23.40	6.61
1986	106679.0	10275.20	23.56	48.69	75.20	23.87	6.63
1987	107875.5	12058.60	24.26	48.69	74.82	23.95	6.66

Table 3: Influencing of population (2)

Year	Birth rate	Hospital	Curing payment	Educational fund	Marriage	0-14 years old	14-65 years old	65 years old
1978	2.94	169732	75.05	110.21	597.8	37.82	57.75	4.43
1979	2.75	176793	93.16	126.19	637.1	36.92	58.53	4.55
1980	2.61	180553	114.15	143.23	720.9	35.94	59.39	4.67
1981	2.55	190126	122.79	160.12	1041.7	34.79	60.39	4.82
1982	2.54	193438	137.61	177.53	836.9	33.69	61.34	4.97
1983	2.56	196017	155.24	207.42	765.4	32.63	62.27	5.10
1984	2.61	198256	180.88	242.07	784.8	31.63	63.16	5.21
1985	2.65	200866	226.83	279.00	831.3	30.74	63.95	5.31
1986	2.67	203139	274.72	315.90	884.0	30.14	64.47	5.39
1987	2.64	204960	293.93	379.58	926.7	29.66	64.89	5.45

$$u_k = \sum_{j=1}^{p} w_{kj} x_j, \tag{1}$$

$$v_k = net_k = u_k - \theta_k, \tag{2}$$

$$y_k = f(v_k), \tag{3}$$

3.2 Data preprocess

BP neural network is back propagating, mainly composed of three parts: input layer, middle layer and output layer. The number of nodes in the input and output layers is relatively easy to determine, but the determination of the number of nodes in the hidden layer is a very important and complex problem.

(2) Normalize the maximum mutual information scores found above and compile them into one.

The characteristic matrix AM of the rows and columns and the normalized score between 0 and 1.

$$A=\begin{bmatrix} a_{22} & a_{23} & \cdots & a_{2y} \\ a_{32} & a_{33} & \cdots & a_{3y} \\ \vdots & \vdots & \cdots & \vdots \\ a_{x2} & a_{x3} & \cdots & a_{xy} \end{bmatrix} \tag{4}$$

(3) Using, and the normalized score as the point coordinates in three-dimensional space, the total maximum mutual information scores can form a surface, and the highest point of the formed surface is the final MIC value. MIC does not rely on the distribution assumption of measurement data and can identify a wide range of associations compared to previous studies. Assuming a bivariate large dataset 2DR containing *n* samples, the MIC of the sum of two vectors is defined as follows.

$$MIC = \max_{xy \le B(n)} \left\{ \frac{I^*(D,x,y)}{\log \min\{x,y\}} \right\} \tag{5}$$

$$I(x,y) = H(x) + H(y) - H(xy) \tag{6}$$

Where *x*,*y* are the number of grids in the x-axis and y-axis zones, respectively.

Table 4: The related index of 13 influencing factors

Number	Name	MIC	Influencing factor	Rank
X1	Average GDP	0.79	Strong	7
X2	Ratio of Urban	0.96	Very strong	1
X3	Ratio of Male	0.53	General strong	12
X4	Ratio of employment	0.34	weak	13
X5	Ratio of Educated	0.93	Very strong	3
X6	Ratio of Death	0.57	General strong	11
X7	Ratio of Birth	0.90	Very strong	5
X8	Ratio of Marriage	0.62	Strong	10
X9	Ratio between 14-65	0.96	Very strong	1
X10	Ratio above 65	0.91	Very strong	4
X11	Institution of Medic	0.69	Strong	9
X12	Medic payment	0.82	Very strong	6
X13	Educational payment	0.77	Strong	8

This paper uses the MIC algorithm for data preprocessing, and sequentially obtains the correlation strength between the total population and 13 influencing factors such as per capita GDP, mortality rate, and birth rate. The results are shown in Table 4.

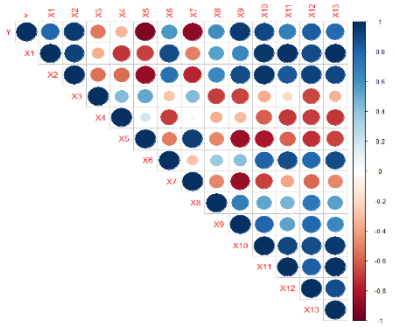

Figure 1: Thermodynamic diagram of 13 influencing factors related to intensity

In Figure 1, each column represents a sample, and each row represents a variable. The color represents the correlation strength between the population and each variable, indicating the difference in correlation strength between these screened variables and the population. At this point, the color

represents the size of the correlation coefficient. Therefore, from the graph, it can be seen that the variable itself has a correlation coefficient of 1 with itself, which is the darkest blue color. The closer the white color is, the weaker the correlation is. A blue (positive correlation) or red (negative correlation) color indicates a strong correlation.

Select highly correlated influencing factors from 13 factors, which are in order: proportion of labor force population, proportion of urban population, birth rate, proportion of aging population, fertility rate, and medical expenditure.

3.3 Experiments

The paper takes 6 main influencing factors as input nodes, inputs 10 hidden neuron numbers and 6 lagging orders, and selects the Levenberg Marquardt algorithm to train the NAR neural network. The training function is trainlm, the transfer function is tansig function, and the weight adaptive learning function is learngd function; Divide the data into training set 70%, validation set 15%, and testing set 15%. Train the NAR network, stop training when the sample mean square error increases, and calculate the prediction result; Subsequently, the trained NAR model is combined with a BP neural network to obtain the predicted values of impact factors and population from 2021 to 2030 through rolling grouping.

The population is showing a continuous growth trend, and by 2030, the total population of the country will reach 1488.9633 million people.

To verify the effectiveness of the PCA-NAR-BP model, the sample data from 1978 to 2010 was used as the training sample set to predict the proportion of males and urban population in China from 2011 to 2020. The average errors were found to be 0.0501126 and 0.21304, respectively, indicating a high feasibility of this combined model. Table 5 shows the results.

The model predicts the data from 2021 to 2030, and use the rolling grouping method to predict the male and urban population proportions of the following year using the 10 year prediction factors until 2030. Finally, the male and urban population proportions will be 51.38% and 66.33% respectively in 2030. The proportion of males will remain relatively stable over the next decade, but there will be a downward trend from the first census to 2030; The proportion of urban population is showing a gradual growth trend in the next decade. Due to the lack of early statistical data, the urban population has been increasing year by year from the third census to 2030.

Table 5: Female percentage from 2011-2020

Year	Actual percentage%	Prediction percentage%	error	Error rate%
2011	51.39	51.39466	0.004658	0.009063
2012	51.38	51.39686	0.016859	0.032812
2013	51.38	51.39742	0.017424	0.033911
2014	51.38	51.39869	0.018689	0.036375
2015	51.36	51.39911	0.039109	0.076147
2016	51.35	51.40025	0.050247	0.097852
2017	51.33	51.40061	0.070613	0.137566
2018	51.32	51.40085	0.080852	0.157546
2019	51.31	51.40118	0.091177	0.177698
2020	51.29	51.4015	0.111498	0.217388

4. Analysis

4.1 The suggestions for census

If the census frequency is once every 5 years, it is found that from 1975 to 1990. During the year, the fertility rate, the proportion of the working population, and the proportion of the aging population remained at normal levels, but they also consumed a large amount of manpower, material resources, and financial resources due to high-frequency surveys that did not identify problems.

If the frequency of the census is once every 15 or 20 years, from 1985 to 2005, the fertility rate, proportion of working population. The proportion of the aging population has exceeded the range value, with a sudden decrease in fertility rate to 16%, a sharp increase in the proportion of the working population to 70%, and an increase in the proportion of the aging population to 7.5%, all exceeding the range value. This has led to a deepening of aging, heavy family burden, and a decrease in the number of new forces. The inability to adjust policies in a timely manner has seriously affected the country's fertility level, population balance, and economic development.

The most suitable frequency for conducting a census is once every 10 years, which not only allows for proactive measures, but also targeted measures by the country. Developing population related strategies and policies to promote long-term balanced population development provides strong statistical information support, which to some extent saves a lot of manpower and material resources.

4.2 Sensitive analysis

Taking into account the impact of various factors on population change, a model suitable for predicting population change was established. The data was analyzed using the Maximum Information Coefficient (MIC) algorithm and the NAR-BP neural network to identify the main indicators that affect population change. It is possible to make more accurate predictions of abnormal population changes in the early stages and make corresponding policy adjustments based on the abnormal changes in corresponding indicators.

The method mentioned in this paper can rely on different data to establish different models, and can also integrate multiple data modeling. The model adopts various data analysis methods, such as the maximum information coefficient (MIC) algorithm, principal component analysis (PCA), and NAR-BP dynamic neural network model, which have good pertinence to the problem and are compared with other methods. This allows us to choose a model that is more closely related to population prediction based on the output results of the model, thereby improving the accuracy of the model.

5. Conclusion

This paper provides a reference basis for the current trends in population distribution, quantity, and structure, and provides a better grasp of the future. Not only can it perform good analysis on predictions, but it can also effectively solve similar evaluation and prediction problems. However, there are many dynamic factors that affect population growth predictions, and they cannot all be affected, so there is still some distance between the model and reality. Different models have high predictive power at corresponding time stages, but once they leave this time stage, the predictive power of the model will decline. In today's increasingly high demand for scientific and quantitative decision-making, our work is undoubtedly in line with the trend of the times and the development needs of the situation.

Acknowledgements

The authors gratefully acknowledge the financial support from xxx funds.

References

[1] Chang D, Gao D, Wang X, et al. Influence mechanisms of the National Pollution Source Census on public participation and environmental consciousness in China[J]. Journal of cleaner production, 2022.
[2] Liu Y, Zhang X, Yang C, et al. Accelerating HPCG on Tianhe-2: A hybrid CPU-MIC algorithm[C]. IEEE International Conference on Parallel & Distributed Systems. IEEE, 2015.
[3] Shuling C, Erbing L I, Liang C, et al. The time series prediction of tunnel surrounding rock deformation based on FA-NAR dynamic neural network [J]. Chinese Journal of Rock Mechanics and Engineering, 2019.
[4] Han J H, Huang D S, Lok T M, et al. A novel image retrieval system based on BP neural network[C]. Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on. IEEE, 2022.
[5] Donatelli R E, J. -A. P, Mathews S M, et al. Time series analysis[J]. American Journal of Orthodontics and Dentofacial Orthopedics, 2022(4):161.
[6] Wang Q, Xi H, Deng F, et al. Design and analysis of genetic algorithm and BP neural network based PID control for boost converter applied in renewable power generations [J]. IET renewable power generation, 2022(7):16.