

Research on China's Healthcare Development Issues Based on the Grey Forecasting Model

Qing Xu^{1,a,*}, Yiyao Liu^{1,b}, Yiming Zhang^{1,c}

¹Northeast Agricultural University, Xiangfang District, Harbin City, Heilongjiang Province, China
a. 3023155527@qq.com, b. 2936017543@qq.com, c. 2860205406@qq.com

**corresponding author*

Abstract: This paper focuses on four major issues and provides in-depth analysis. For issue one, the three-spline interpolation method is used to fill missing data, and the IQR method and box plot are employed to detect and remove outliers. The growth trend of medical institutions in different regions is analyzed. For issue two, principal component analysis (PCA) is used to reduce the dimensionality of healthcare resource allocation indicators, with the applicability of the analysis determined by the KMO and Bartlett tests. The relationship between resource allocation and regions is analyzed using the variance explanation table and Spearman correlation coefficient, while Moran's I index reveals spatial distribution characteristics. For issue three, the ARIMA model is used to predict the number of medical institutions in various regions over the next five years, with data stationarity tests and parameter optimization conducted. The model's predictive ability is assessed through MSE. For issue four, the Pearson correlation coefficient is used to analyze key factors influencing the number of medical institutions, such as population aging and GDP, and the grey forecasting model is applied to predict future trends. The study shows that population aging and regional development imbalance will continue to affect the number of medical institutions, and economic development is highly correlated with healthcare levels. The paper proposes strategies to optimize the healthcare system, expand medical insurance coverage, and promote technological innovation, emphasizing the importance of addressing public health systems and the aging challenge.

Keywords: Moran's I index, Spearman correlation coefficient, Pearson correlation coefficient, Grey forecasting model

1. Problem Restatement

1.1. Problem Background

Healthcare is a crucial guarantee for social stability and development. It has a significant impact on promoting social stability, economic development, and cultural and technological advancements, and is also an important indicator of national development level and comprehensive strength. China's healthcare sector has achieved remarkable progress. However, healthcare resource distribution in China remains uneven, with high-quality medical resources concentrated in developed regions. Therefore, it is essential to understand the healthcare development levels in various regions, rationally

allocate healthcare resources, and improve the healthcare system to promote social fairness and stability, as well as to improve public health levels and quality of life.

1.2. Unresolved Issues

Issue 1: The annual growth rate of medical institutions in each province needs to be calculated, and the growth rates compared across different provinces to analyze regional trends.

Issue 2: Data collection is needed to explore the causes of disparities in the allocation of healthcare resources across provinces. This includes analyzing the changes and disparities in the number of medical institutions in each province of China and examining the underlying causes and their practical significance.

Issue 3: Based on data from 1990 to 2023, predictions should be made for the number of medical institutions in each region over the next five years (2024-2028). The model should be evaluated, and the impact of uncertainties on future trends should be discussed.

Issue 4: Based on current healthcare resource allocation, population structure, and social development trends, the key factors affecting the number of medical institutions should be analyzed. Additionally, a macro prediction should be made for the number of medical institutions in China over the next ten years (2024-2033). The results of these predictions should provide insights for future healthcare system construction and formulating strategies to address potential challenges.

2. Problem Analysis

Issue 1: After performing preliminary analysis on the data, it was found that some data points were randomly missing, and incomplete observational data may cause systematic differences compared to fully observed data. Therefore, the first step is to handle the missing data. The preferred method for filling the missing values is flexible and efficient. The cubic spline interpolation method strikes a reasonable balance between flexibility and computational speed. Compared to higher-order splines, cubic splines require less computation, are easier to store, and are more stable. In contrast to quadratic splines, cubic spline interpolation is more flexible when simulating arbitrary shapes[1]. Therefore, it was decided to use the cubic spline interpolation method to fill the missing values. The processed dataset still contains noise and outliers, which need to be addressed next. After initial detection, it was found that most of the data indicators do not follow a normal distribution. The IQR method, based on actual data, does not impose any distributional requirements on the data and provides a straightforward representation of the data's inherent distribution. On the other hand, the Boxplot method identifies outliers based on quartiles and interquartile range. Quartiles have certain robustness: outliers can become more distant without significantly affecting the quartiles.

Issue 2: Before proceeding with mathematical modeling, it is necessary to conduct exploratory analysis to preliminarily understand the differences in healthcare resource allocation across regions. The following steps can be taken: First, descriptive statistics should be calculated, specifically the basic statistical measures (such as mean, median, variance, and standard deviation) for the number of healthcare institutions in each region, in order to observe overall trends and distribution across different regions. Next, a heatmap can be plotted to display the spatial distribution of healthcare resource allocation levels across regions, which visually reflects the national disparities. Given that there are multiple variables and correlations among them, principal component analysis (PCA) is performed to reduce the dimensionality. Correlation analysis is then conducted to explore possible influencing factors of healthcare resource distribution and identify primary and secondary factors[2]. Afterward, the relationships between the variables should be analyzed. Next, a normality test is conducted on the data. If the results indicate a normal distribution, the Pearson correlation coefficient is used; otherwise, the Spearman correlation coefficient is applied. Moran's I index is used to analyze

the relationship between regions and healthcare resource allocation. The global Moran's I index is used to determine if there is spatial autocorrelation at the national level. If the global Moran's I index shows significance, a further investigation using the local Moran's I index can be conducted to explore more detailed relationships[3].

Issue 3: Based on data from the past 33 years, the number of healthcare institutions in various regions will be predicted for the next five years, and the accuracy of the model's predictions, as well as the results and their impact on future healthcare resource allocation, will be evaluated. The change in the number of healthcare institutions is reflected through dynamic sequences, and since the goal is to predict the future, the ARIMA model is used for analysis[4].

First, a stationarity test is conducted on the original dataset. If the series is non-stationary, differencing methods are applied to test for stationarity. The ARIMA model is composed of an AR(p) model, a differencing model, and an MA(q) model, and it is necessary to determine the order of the three parameters. The appropriate p and q orders are determined by calculating the correlation coefficients and partial correlation coefficients. The appropriate differencing order (d) is identified through the ADF test. Using the rule of minimizing AIC, the best AR model is identified using SPSSPRO. The MSE (Mean Squared Error) is highly sensitive to prediction errors; even small errors significantly affect the MSE value, which makes it an effective measure of model predictive capability. A smaller MSE value indicates higher model accuracy. If the evaluation metrics show a very high degree of fit and prediction accuracy, this suggests that the model effectively captures patterns and trends in the training data[4].

Issue 4: The number of healthcare institutions is influenced by various factors. A normality test is conducted on the dataset; if the results indicate a normal distribution, the Pearson correlation coefficient is used for correlation analysis; otherwise, the Spearman correlation coefficient is applied. Next, a grey forecasting model is used to predict future data. Before establishing the GM(1,1) grey forecasting model, a level comparison test is conducted on the time series[5]. If the series passes the level comparison test, it indicates that the sequence is suitable for constructing the grey model. If the sequence does not pass, a "translation transformation" is applied to the sequence to ensure that the new sequence meets the level comparison test. The GM model formula is determined by calculating the development coefficient $\hat{a}$ and the grey action amount $\hat{b}$, followed by prediction calculations.

3. Model Assumptions and Symbol Definitions

3.1. Basic Assumptions of the Model

- (1) It is assumed that the statistical data collected from the surveyed websites conforms to the principles of scientific research, ensuring the rationality and validity of the data.
- (2) It is assumed that there is no large-scale migration of populations.
- (3) It is assumed that there are no significant changes in healthcare policies.
- (4) It is assumed that there will be no major disasters or outbreaks of new pandemics in China's regions over the next ten years.

3.2. Symbol Definitions

Table 1: Symbol Definitions

Symbol	Meaning	Symbol	Meaning
n	Sample size	x_1	Number of healthcare institutions per 10,000 people
Gi	Growth rate	x_2	Number of healthcare institution beds per 1,000 people

Table 1: (continued).

$S(n)$	Interpolation function	x_3	Medical technology expenditure
x_j	Data point	x_4	Government financial expenditure
λ	Eigenvalue	x_5	Resident population size
r_i	Eigenvector	x_6	Proportion of aging population
$\hat{Y}_t^{(i)}$	Predicted value for region i	x_7	Gross Domestic Product (GDP)
W_{ij}	Spatial weight	x_8	GDP per capita
$\varepsilon_t^{(i)}$	White noise sequence	x_9	Number of healthcare technology professionals per 1,000 people

4. Model Establishment and Solution

4.1. Model Establishment and Solution for Issue 1

First, a preliminary analysis of the data was conducted, revealing that some data points were missing randomly. The incomplete observational data would likely cause systematic differences when compared with the fully observed data. If such data were analyzed directly, there is a high likelihood of drawing erroneous conclusions. Therefore, the missing data was filled. After filling, the dataset still contained noise and outliers, which required further treatment. The next step was to eliminate the outliers and smooth the noisy data to obtain high-quality data and more accurate results.

4.1.1. Data Cleaning

(1) Handling Missing Values

Since the number of missing values is relatively small, the gaps between the data points are relatively narrow, and the data is a time series with a generally linear trend, the "spline" interpolation method from MATLAB (1D interpolation) was used to perform cubic spline interpolation. This method constructs a cubic polynomial for interpolation[1]. The results were rounded to integers to match the actual number of healthcare institutions, thus improving the accuracy and precision of the imputed data.

(2) Handling Outliers

After testing, it was found that most of the data indicators did not follow a normal distribution. The IQR method was used to detect outliers, as it is based on actual data without imposing any distributional requirements. The Boxplot method, based on quartiles and interquartile range (IQR), was employed to identify outliers. Quartiles are robust, meaning that outliers cannot significantly affect the quartile values. Therefore, the Boxplot method is more objective and superior in detecting outliers. Using Boxplot for visualization, points that exceeded the upper quartile plus 1.5 times the IQR or fell below the lower quartile minus 1.5 times the IQR were considered outliers. These outliers were then treated as indicators that align with the underlying development patterns of the data.

4.1.2. Cubic Spline Interpolation

(1) Constructing the Spline Interpolation Function

For the data points $\{x_1, x_2, \dots, x_n\}$ in the interval $[a, b]$, with corresponding function values $\{y_1, y_2, \dots, y_n\}$, the function is expressed as $S(x)$:

$$S_j(x) = a_j x^3 + b_j x^2 + c_j x + d_j, (j = 1, 2, \dots, n - 1) \quad (1)$$

$S(x)$ satisfies the following conditions:

$$\begin{cases} S(x_j) = y_j \\ S(x_j - 0) = S(x_j + 0) \\ S'(x_j - 0) = S'(x_j + 0) \\ S''(x_j - 0) = S''(x_j + 0) \end{cases}, j = 1, 2, \dots, n-1 \quad (2)$$

The second derivative of the spline at the data points is given by $S''(x_j) = g(x_j)$ where $h_j = x_{j+1} - x_j$.

$$S''(x) = g_j \frac{x_{j+1}-x}{h_j} + g_{j+1} \frac{x-x_j}{h_j}, \forall x \in [x_j, x_{j+1}] \quad (3)$$

The spline function $S(x)$ for each interval is found by performing an indefinite integral on each segment.

$$S(x) = g_j \frac{(x_{j+1}-x)^3}{6h_j} + g_{j+1} \frac{(x-x_j)^3}{6h_j} + (y_j - \frac{g_j h_j^2}{6}) \frac{x_{j+1}-x}{h_j} + (y_{j+1} - \frac{g_{j+1} h_j^2}{6}) \frac{x-x_j}{h_j} \quad (4)$$

4.1.3. IQR Method for Detecting Outliers

Determining the Positions of Q1, Q2, Q3, and the Quartiles

$$\begin{cases} Q1 = \frac{(n+1)}{4} \\ Q2 = \frac{2(n+1)}{4} \\ Q3 = \frac{3(n+1)}{4} \end{cases} \quad (5)$$

To calculate the standard IQR:

$$IQR_{standard} = (Q3 - Q1) \times 0.7413 \quad (6)$$

4.1.4. Growth Rate Calculation

Month-on-Month Growth Rate:

$$G_i = \frac{Y_i - Y_{i-1}}{Y_{i-1}} = \frac{Y_i}{Y_{i-1}} - 1, i = 1, 2, \dots, n \quad (7)$$

Base Year Growth Rate:

$$G_i = \frac{Y_i - Y_0}{Y_0} = \frac{Y_i}{Y_0} - 1, i = 1, 2, \dots, n \quad (8)$$

Where Y_0 is the observed value for the fixed base period used for comparison.

Average Growth Rate:

$$\bar{G} = \sqrt[n]{\left(\frac{Y_1}{Y_0}\right)\left(\frac{Y_2}{Y_1}\right)\dots\left(\frac{Y_n}{Y_{n-1}}\right)} - 1 \quad (9)$$

n represents the number of period values used for the month-on-month growth rate.

Using the above formulas, the month-on-month growth rates for various regions can be calculated, and then the annual average growth rate for healthcare institutions in each region can be determined.

4.1.5. Result Analysis

The regions with the fastest growth in the number of healthcare institutions are Hebei, Henan, Guizhou, Shandong, and Hunan. The regions with the slowest growth in the number of healthcare institutions are Tibet, Ningxia, Tianjin, Shanghai, Qinghai, and Hainan.

Table 2: Average Annual Growth Rate

Region	Average Annual Growth Rate (%)	Region	Average Annual Growth Rate (%)
Beijing	1.95%	Henan	4.87%
Tianjin	2.71%	Hubei	2.75%
Hebei	4.63%	Hunan	4.35%
Shanxi	4.73%	Guangdong	4.14%
Inner Mongolia	3.59%	Guangxi	3.81%
Liaoning	0.08%	Hainan	1.50%
Jilin	1.04%	Chongqing	3.16%
Heilongjiang	1.84%	Sichuan	3.40%
Shanghai	0.17%	Guizhou	3.15%
Jiangsu	2.33%	Yunnan	3.09%
Zhejiang	3.57%	Tibet	2.11%
Anhui	3.07%	Shaanxi	4.03%
Fujian	3.85%	Gansu	3.85%
Jiangxi	4.18%	Qinghai	3.69%
Shandong	4.42%	Ningxia	3.03%
Xinjiang	3.02%		

The regions with the highest average annual growth rates include Henan, Hunan, Guangdong, Shaanxi, Shanxi, Jiangxi, and Shandong. The northeastern regions and Hainan have the lowest average annual growth rates. From Figure 4, we can clearly observe that the central regions of China have a higher growth rate for healthcare institutions. The high growth rate of healthcare institutions in certain regions is tentatively attributed to factors such as high GDP, large population base, and the growing demand for healthcare resources.

4.2. Model Establishment and Solution for Issue 2

Issue 2 requires a discussion on the disparities in the allocation of healthcare resources across different regions, as well as the factors influencing healthcare resources. Here, the indicators for the allocation of healthcare resources in each region are defined as: number of healthcare institutions per 10,000 people, number of healthcare institution beds per 1,000 people, medical technology expenditure, and the number of healthcare technology professionals per 1,000 people. The influencing factors for healthcare resource allocation are defined as: government financial expenditure, resident population size, proportion of aging population, gross domestic product (GDP), and GDP per capita. Given that there are multiple variables for healthcare resource allocation, principal component analysis (PCA) is used for dimensionality reduction. This is followed by an analysis of the correlation between healthcare resources and other influencing factors to simplify the problem.

Correlation analysis examines the degree of correlation between pairs of variables. The Pearson correlation coefficient is suitable for quantitative data that follows a normal distribution, while the Spearman correlation coefficient is used for ordered variables or when the data does not meet the normality assumption. Therefore, a normality test is performed on the data. If the results indicate a

normal distribution, the Pearson correlation coefficient is used; otherwise, the Spearman correlation coefficient is applied[2].

Additionally, Issue 2 requires studying the trends in the number of healthcare institutions and regional differences. This can be done by comparing the annual growth rates of healthcare institutions in different regions (as derived from Issue 1). Moran's I index can also be used to analyze the spatial disparities in the number of healthcare institutions.

4.2.1. Dimensionality Reduction Based on Principal Component Analysis

To simplify the problem, this study first reduces the dimensionality of the four indicators of healthcare resource allocation by using Principal Component Analysis (PCA) with SPSSPRO software. The first step is to conduct the KMO and Bartlett tests to determine whether PCA is applicable. For the KMO value, a value above 0.8 is considered highly suitable for performing PCA. For Bartlett's test, if the p-value is less than 0.05 and the null hypothesis is rejected, this indicates that PCA can be performed[2]. Otherwise, PCA is not suitable.

Next, by analyzing the variance explained table and the scree plot, the number of principal components is determined. The variance explained table is used to calculate the contribution rate of each principal component to the variables and decide whether adjustments to the principal component data are necessary. By analyzing the component loadings and the heatmap, the importance of each latent variable within the principal components can be evaluated. The component matrix is then analyzed to derive the principal component formulas and their corresponding weights.

(1) Principal Component Analysis Model Theory

First, the four indicators (number of healthcare institutions per 10,000 people, number of healthcare institution beds per 1,000 people, medical technology expenditure, and the number of healthcare technology professionals per 1,000 people) for the 31 regions are standardized. A sample matrix a is constructed with 31 samples (regions) and 4 indicators, creating a matrix of size 31×4 .

$$a = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{14} \\ a_{21} & a_{22} & \cdots & a_{24} \\ \vdots & \vdots & \ddots & \vdots \\ a_{311} & a_{312} & \cdots & a_{314} \end{bmatrix} \quad (10)$$

The matrix is then standardized by calculating the mean and standard deviation for each column to obtain the standardized data.

$$A_{ij} = \frac{a_{ij} - \frac{\sum_{i=1}^n a_{ij}}{n}}{\sqrt{\frac{\sum_{i=1}^n (a_{ij} - \bar{a}_j)^2}{n-1}}} \quad (11)$$

The contribution rate is calculated as follows:

$$r_i = \frac{\lambda_p}{\sum_{k=1}^p \lambda_k}, p = 1, 2, \dots, 4 \quad (12)$$

Where r_i represents the eigenvectors of the covariance matrix of the standardized samples, and λ represents the eigenvalues.

The cumulative contribution rate is calculated as:

$$\sum A = \frac{\sum_{k=1}^i \lambda_k}{\sum_{k=1}^p \lambda_k}, (i = 1, 2, \dots, p) \quad (13)$$

(2) Detailed Steps of Principal Component Analysis

Step 1: Conduct KMO and Bartlett Tests

The results of the KMO and Bartlett tests are as follows:

Table 3: KMO and Bartlett's Test

KMO Value		0.878
Bartlett's Test (Approx. Chi-Square)	approximate chi-square	6.854
	df	6
	P	0.019

The KMO value is 0.878, indicating that the data is suitable for principal component analysis. Additionally, Bartlett's Test of Sphericity shows a significant p-value of 0.019, which rejects the null hypothesis, confirming that PCA is valid.

Step 2: Determine the Number of Principal Components

The number of principal components is determined using the scree plot and the variance explanation rate.

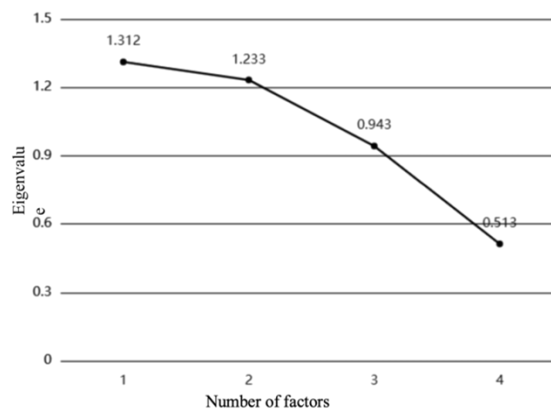


Figure 1: Scree Plot

From the scree plot, it can be observed that the eigenvalues for components 1 and 2 are greater than 1, with eigenvalues of 1.312 and 1.233, respectively. However, the eigenvalues for components 3 and 4 are less than 1, at 0.943 and 0.513, respectively.

Table 4: Variance Explanation Table

Component	Eigenvalue	Variance Explanation Rate (%)	Cumulative Variance Explanation Rate (%)
1	1.312	32.791	32.791
2	1.233	30.817	63.608
3	0.943	23.573	87.181
4	0.513	12.819	100

From the above table, it can be seen that the variance explanation rate for components 3 and 4 exceeds 80%. Each variable is well explained by the components. Therefore, based on this, the number of principal components is determined to be 4.

Step 3: Final Model

By analyzing the importance of latent variables in each principal component and using factor score coefficients, the factor formula is derived to obtain the final model's equation.

$$F_1 = 0.328X_1 + 0.308X_2 + 0.236X_3 + 0.128X_4 \quad (14)$$

4.2.2. Spearman Correlation Coefficient Analysis

As previously mentioned, the dimensionality of the healthcare resource allocation data was reduced using Principal Component Analysis (PCA). The next step involves conducting a correlation analysis. The normality of the data will first be tested, and if the data follows a normal distribution, the Pearson correlation coefficient will be used. If not, the Spearman correlation coefficient will be applied.

(1) S-W Normality Test

Since the sample size is less than 5000, the Shapiro-Wilk (S-W) test is used. The results are as follows:

Table 5: Significance Test Results

Variable Name	S-W Test
Comprehensive Score of Healthcare Resource Allocation	0.956 (0.230)
Government Financial Expenditure	0.941 (0.086*)
Resident Population Size	0.945 (0.111)
Aging Population Ratio (%)	0.974 (0.635)
Gross Domestic Product (GDP)	0.840 (0.000***)
GDP Per Capita	0.810 (0.000***)

From the data in the table, it can be seen that the p-values for GDP per capita and Gross Domestic Product (GDP) are 0.000***, indicating statistical significance and leading to the rejection of the null hypothesis. Therefore, the data does not follow a normal distribution.

(2) Calculating the Correlation Coefficient

$$r_s = \frac{Cov(R_i, Q_i)}{\sqrt{Var(R_i)Var(Q_i)}} \quad (15)$$

Where R_i represents the rank of the first variable, Q_i represents the rank of the second variable; for a sample of size n , the Spearman rank correlation coefficient r_s can be simplified as:

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum (R_i - Q_i)^2 \quad (16)$$

The test statistic is:

$$t_{obs} = r_s \sqrt{\frac{n-2}{1-r_s^2}} \quad (17)$$

The t-statistic approximately follows a t-distribution with $n-2$ degrees of freedom. Therefore, when $|t_{obs}| > |t_{\frac{\alpha}{2}}(n-2)|$, the null hypothesis is rejected, indicating that the Spearman rank correlation coefficient is significant.

From the above graph, it can be seen that GDP per capita and the other five indicators are positively correlated with healthcare resource allocation. Furthermore, GDP per capita, Gross Domestic Product (GDP), the aging population ratio, and government financial expenditure significantly impact healthcare resource allocation, showing a linear relationship. However, the resident population size does not significantly affect healthcare resource allocation, and there is no linear correlation.

4.2.3. Trend Analysis and Moran's I Index Analysis

(1) Exploratory Data Analysis (EDA)

First, exploratory data analysis (EDA) is conducted to preliminarily understand the differences in healthcare resource allocation across regions. The following steps can be taken:

Calculate descriptive statistics to observe the overall trends and distribution patterns across different regions. According to the calculations, regions such as Liaoning, Shanghai, Hebei, and Henan have a large number of healthcare institutions over the past three decades. This is likely due to high population density, leading to greater demand for healthcare institutions, or due to rapid economic development, substantial government healthcare expenditure, and favorable policies. On the other hand, regions like Jiangsu, Yunnan, Zhejiang, and Ningxia have fewer healthcare institutions, possibly because of lower population density, smaller healthcare needs, or smaller regional areas with limited space for establishing healthcare institutions.

(2) Using Moran's I Index to Assess the Relationship Between Healthcare Institution Numbers and Regional Differences

The global Moran's I index is used to analyze whether there is a spatial autocorrelation in the overall distribution of healthcare institutions. If the index shows statistical significance, further analysis using the local Moran's I index can be performed to understand the detailed spatial relationships. By calculating Moran's I index, it is possible to determine the spatial distribution of healthcare resources and identify whether some regions have an over-concentration or dispersion of resources[4]. Moran's I Index is calculated as:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (b_i - \bar{b})(b_j - \bar{b})}{\sum_i (b_i - \bar{b})^2} \quad (18)$$

Where b_i and b_j represent the number of healthcare institutions in regions i and j respectively, $\bar{b}$ is the mean number of healthcare institutions, w_{ij} is the spatial weight between region i and j , typically defined based on spatial distance.

Using MATLAB, the computed Moran's I index is found to be -0.0115, indicating negative spatial autocorrelation. This suggests a diffusion effect in the distribution of healthcare institutions, meaning that the number of healthcare institutions in neighboring regions shows disparity.

4.2.4. Exploring the Policy Implications Behind Healthcare Resource Allocation

The continuous improvement of healthcare resources in China is largely attributed to the implementation of various people-oriented policies. These policies aim to expand the capacity and improve the distribution of high-quality healthcare resources across regions, while promoting the balanced development of healthcare services. Public hospital reforms, which are guided by a public welfare-oriented approach, have also played a key role. Additionally, the cultivation of talent in emerging industries, the enhancement of GDP, and the advancement of technological levels contribute to the development of key industrial chains, such as medical equipment and advanced materials, through effective systems and mechanisms, facilitating technological breakthroughs and the application of research outcomes.

The deepening of healthcare and pharmaceutical system reforms has led to the implementation of a health-first development strategy and the improvement of the public health system. This includes promoting the coordinated development and governance of healthcare, health insurance, and pharmaceuticals. Efforts are being made to advance provincial-level coordination of basic medical insurance, deepen the reform of medical insurance payment methods, and improve systems for major illness insurance and medical assistance, along with strengthening the supervision of health insurance funds and reinforcing grassroots healthcare services.

In addition, China is working to improve its social security system by developing a unified national pension system, enhancing the nationwide public social security service platform, and ensuring the security and value growth of social security funds. Policies are also in place to rationally adjust the

financing and benefit mechanisms of basic pension and healthcare insurance, gradually increasing the basic pension for rural and urban residents. The country is advancing the provincial-level coordination of basic medical insurance, deepening payment method reforms, and strengthening the regulatory framework for health insurance funds. These measures are part of the broader efforts to actively address the challenges posed by population aging.

4.3. Model Establishment and Solution for Problem Three

Problem Requirements: Based on 33 years of data, this problem requires forecasting the number of healthcare institutions in each region for the next five years and evaluating the model's prediction accuracy, the reliability of the forecast results, and their implications for future healthcare resource allocation. The changes in the number of healthcare institutions are represented by dynamic sequences, and future predictions require the use of the ARIMA model for analysis.

First, a stationarity test is performed on the raw dataset. If the sequence is non-stationary, differencing can be applied to make the series stationary. The ARIMA model can be broken down into an AR(p) model, a differencing model, and an MA(q) model. The orders of the three parameters need to be determined. The appropriate values of p and q are obtained by calculating the autocorrelation and partial autocorrelation coefficients, while the differencing order d is determined using the ADF test. The optimal AR model is found by minimizing the AIC value, using SPSSPRO.

4.3.1. Handling Series Stationarity

Stationarity requires that the fitted curve obtained from the sample time series continues to follow the same pattern in the future. In other words, the sequence's expectation and the correlation coefficient (dependency) should remain unchanged, meaning that the value at time t depends on past information. Stationarity is crucial for time series analysis. If a time series is non-stationary, it is typically transformed into a stationary series through differencing.

Using SPSSPRO for analysis, based on the variable, it was found that when the differencing order is 1, the significance p-value is 0.000***, which is significant at the 0.01 level, thus rejecting the null hypothesis. This confirms that the series is stationary.

4.3.2. White Noise Test

The ARIMA model requires that the model exhibit pure randomness, meaning that the residuals of the model should be white noise. To check this, the model diagnostic table is examined, and the Q-statistic p-value ($P > 0.05$) is used to test whether the residuals are white noise[5]. The optimal parameters are automatically selected based on the AIC information criterion. The model results show an ARIMA (0,1,1) model diagnostic table. Based on the variable, the Q-statistic results are analyzed. If the Q6 value is not significant at the given level, the hypothesis that the residuals are white noise cannot be rejected. Conversely, if the result is significant, the hypothesis is rejected. Additionally, the closer the model's R^2 value is to 1, the better the fit, indicating that the model meets the required standards.

Through SPSSPRO analysis, it was found that the Q6 value for each variable was not significant, so the hypothesis that the residuals are white noise cannot be rejected. At the same time, the goodness of fit (R^2) is close to 1, indicating that the model meets the required standards.

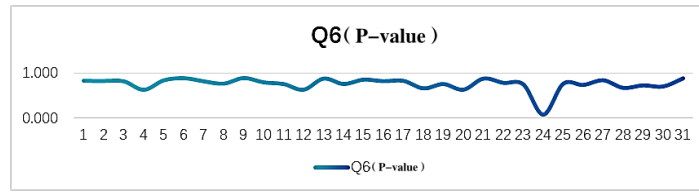


Figure 2: Significance of Q6 Values

4.3.3. Establishing the ARIMA Model

Time series models are suitable for analyzing trends, cycles, and random fluctuations in individual time series data. For the number of healthcare institutions in each region, the ARIMA model is used for modeling.

The mathematical expression of the ARIMA(p,d,q) model is as follows:

$$(1 - \sum_{k=1}^p \phi_k B^k)(1 - B)^d Y_t^{(i)} = (1 + \sum_{k=1}^q \theta_k B^k) \varepsilon_t^{(i)} \quad (19)$$

Where $Y_t^{(i)}$ represents the number of healthcare institutions in region i over the 33 years ($i = 1, 2, \dots, 31, t = 1, 2, \dots, 33$),

$\hat{Y}_t^{(i)}$ represents the forecast value for the number of healthcare institutions in region i for the next five years, ϕ_k are the autoregressive coefficients, θ_k are the moving average coefficients, d is the differencing order, $\varepsilon_t^{(i)}$ is the white noise sequence.

4.3.4. Forecast Results

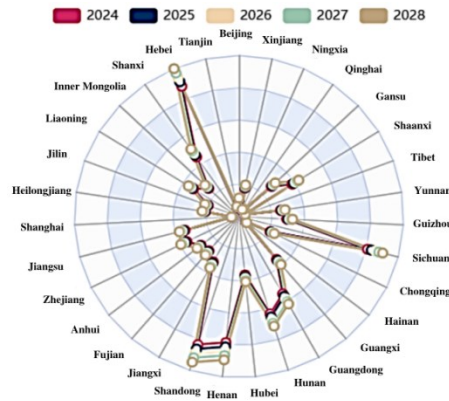


Figure 3: Radar Chart of Prediction Results

The results show that in the next five years, regions such as Hebei, Sichuan, Shandong, Henan, Hunan, and Guangdong will continue to maintain a leading position in terms of the number of healthcare institutions.

Uncertainty Factors:

1) Population Mobility

There are differences in the level of economic development across regions in China. Cities or regions with higher levels of economic development have more advanced facilities, resources, and can provide more employment opportunities, thus attracting more talent in the healthcare sector to contribute to the local healthcare system. This may influence the future distribution of healthcare resources.

2) Policy Changes

The direction of local policy changes can have a significant impact. Policies that improve the living standards of healthcare workers and promote the development of healthcare will lead to rapid growth in healthcare resources in those regions. Therefore, the improvement of local policies greatly affects the allocation of healthcare resources across regions.

4.3.5. Model Evaluation

(1) MSE Evaluation

MSE (Mean Squared Error) is highly sensitive to the magnitude of prediction errors, meaning even small errors significantly impact the MSE value. Therefore, it effectively reflects the model's predictive ability and is simple and easy to understand. The MSE value is the average of the squared differences between the predicted and actual values. The smaller the MSE value, the higher the accuracy of the model.

$$MSE = \frac{1}{n} \sum_{i=1}^p (Y_t^{(i)} - \hat{Y}_t^{(i)})^2 \quad (20)$$

(2) RMSE Evaluation

RMSE (Root Mean Squared Error) is the square root of MSE. The smaller the RMSE value, the higher the accuracy of the model.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^p (Y_t^{(i)} - \hat{Y}_t^{(i)})^2} \quad (21)$$

(3) MAE Evaluation

MAE (Mean Absolute Error) is the average of the absolute errors, which reflects the actual prediction errors. The smaller the MAE value, the higher the accuracy of the model.

$$MAE = \frac{1}{n} \sum_{i=1}^p |Y_t^{(i)} - \hat{Y}_t^{(i)}| \quad (22)$$

Model Evaluation Results:

Table 6: Model Evaluation Metrics

	MSE	RMSE	MAE	MAPE	R ²
Training Set	4.157	2.039	1.331	0.067	0.906

The Mean Squared Error (MSE) for the training set is 4.157, the Root Mean Squared Error (RMSE) is 2.039, the Mean Absolute Error (MAE) is 1.331, and the Mean Absolute Percentage Error (MAPE) is as low as 0.067. Additionally, the R² (Coefficient of Determination) is 0.906. These indicators collectively demonstrate that the model exhibits an extremely high level of goodness-of-fit and predictive accuracy on the training data.

4.4. Model Establishment and Solution for Problem 4

There are multiple factors that influence the number of healthcare institutions. A normality test is performed on the dataset, and if the result indicates normal distribution, the Pearson correlation coefficient is used for the correlation analysis. Otherwise, the Spearman correlation coefficient is applied. Subsequently, a Grey Prediction Model is used to predict future data. Prior to establishing the GM(1,1) Grey Prediction Model, a Level Ratio Test is conducted on the time series data[6]. If the test is passed, it indicates that the sequence is suitable for constructing a Grey model; if it fails, the sequence is adjusted via a "shifting transformation" to meet the requirements of the level ratio test. The model is mainly evaluated through the posterior difference ratio C-value.

4.4.1. Normality Test

The Shapiro-Wilk (S-W) test is applied for small sample sizes (typically $n < 5000$) to assess normality. If the result shows no significance ($P > 0.05$), the data is considered to follow a normal distribution; otherwise, it does not.

Normality Test Results:

Table 7: S-W Test Significance Results

Variable Name	Sample Size	S-W Test
Number of Healthcare Institutions (China)	14	0.967(0.838)
Proportion of Elderly Population (%)	14	0.931(0.315)
Total Population (persons)	14	0.891(0.083*)
GDP of China (Trillion Yuan)	14	0.96(0.718)
Total Healthcare Expenditure as % of GDP	14	0.923(0.244)
Per Capita Healthcare Expenditure (Yuan)	14	0.926(0.272)

The p-values for all variables show no significance ($P > 0.05$), indicating that the dataset follows a normal distribution. Therefore, Pearson's correlation coefficient is used for the correlation analysis.

4.4.2. Pearson Correlation Coefficient

Pearson correlation analysis was performed using SPSSPRO. The results are as follows:

The proportion of elderly population is highly positively correlated (0.9) with the number of healthcare institutions, GDP, the proportion of total healthcare expenditure to GDP, and per capita healthcare expenditure. This indicates that the acceleration of the process, the expansion of the population scale, and the overall increase in total quantity significantly increase the demand for healthcare services, thereby promoting the growth of healthcare institutions and driving up expenditures.

4.4.3. Using the Grey Prediction Model

(1) Level Ratio Test

The average time series is established as follows:

$$x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(14)) \quad (23)$$

The level ratio $\lambda(k)$ is calculated as:

$$\lambda(k) = \frac{x^{(0)}(k-1)}{x^{(0)}(k)}, \lambda = (\lambda(2), \lambda(3), \dots, \lambda(14)) \quad (24)$$

The level ratios are calculated as follows:

Table 8: Level Ratio Calculation Results

Year Index	Level Ratio				Shifted Level Ratio	
	Healthcare Institutions	Total Healthcare Expenditure as % of GDP	Elderly Population Ratio	Total Population	GDP	Per Capita Healthcare Expenditure
2010	0.982	0.972	0.978	0.994	0.959	0.991
2011	1.004	0.956	0.968	0.993	0.975	0.965

Table 8: (continued).

2012	0.975	0.978	0.969	0.994	0.974	0.972
2013	0.993	0.973	0.96	0.993	0.965	0.975
2014	0.998	0.923	0.962	0.995	0.974	0.95
2015	1	0.958	0.972	0.993	0.973	0.966
2016	0.997	0.983	0.947	0.994	0.961	0.962
2017	0.989	0.983	0.958	0.996	0.967	0.963
2018	0.99	0.964	0.944	0.997	0.96	0.942
2019	0.985	0.937	0.933	0.999	0.989	0.949
2020	0.992	1.06	0.951	1	0.948	0.965
2021	0.998	0.981	0.953	1.005	0.974	0.969
2022	0.965	0.996	0.92	1.004	0.965	0.971

From the above table, it can be observed that all the level ratios fall within the range $(e^{-2/(n+1)}, e^{2/(n+1)})$, i.e., (0.875, 1.143), indicating that this sequence is suitable for constructing a Grey prediction model.

(2) GM(1,1) Model Construction

The original data $x^{(0)}$ is accumulated to obtain $x^{(1)}$. Then, we construct the data matrix B and the data vector Y :

$$B = \begin{bmatrix} -\frac{1}{2}(x^{(1)}(1) + x^{(1)}(2)) & 1 \\ -\frac{1}{2}(x^{(1)}(2) + x^{(1)}(3)) & 1 \\ \vdots & \vdots \\ -\frac{1}{2}(x^{(1)}(13) + x^{(1)}(14)) & 1 \end{bmatrix}, Y = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(14) \end{bmatrix} \quad (25)$$

The development coefficient $\hat{a}$ and the grey action quantity $\hat{b}$ are calculated as follows:

$$\hat{u} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (B^T B)^{-1} B^T Y \quad (26)$$

Table 9: Development Coefficient and Grey Action Quantity

	Development Coefficient	Grey Action Quantity
Number of Healthcare Institutions	-0.008	936546.274
Total Healthcare Expenditure as % of GDP	-0.027	5.001
Elderly Population Ratio	-0.049	8.029
Total Population	-0.004	135396.52
GDP	-0.032	167.074
Per Capita Healthcare Expenditure	-0.04	5290.208

The model is constructed:

$$\frac{dx^{(1)}(t)}{dt} + \hat{a}x^{(1)}(t) = \hat{b} \quad (27)$$

Solve to get:

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{\hat{b}}{\hat{a}}\right)e^{-\hat{a}k} + \frac{\hat{b}}{\hat{a}} \quad (28)$$

4.4.4. Forecasting Results

Year	Number of Healthcare Institutions in China (units)
2024	980198
2025	980189
2026	980181
2027	980173
2028	980164
2029	980156
2030	980148
2031	980134
2032	980131
2033	980123

4.4.5. Reflections on Healthcare Policy, Resource Allocation, and System Construction

(1) Insights for Healthcare Policy Formulation, Resource Allocation, and System Construction:

1. Establish Good Social Support: Social support is moderately positively correlated with self-management. Establishing good social support can enhance treatment outcomes and improve the quality of life for patients.

2. Necessary Changes to Traditional Healthcare and Disease Prevention Systems: In the coming years, the number of healthcare institutions in China will continue to increase. However, regional disparities in healthcare resources and imbalances in medical development will become increasingly prominent. Rather than simply focusing on increasing the number of healthcare institutions, it is more important to promote local healthcare levels and provide sufficient healthcare resources.

3. Building a Healthcare Security System with Chinese Characteristics

(2) Challenges in the Growth of Healthcare Institutions in China:

1. Aging Population and the Decline of Medical Talent: As the aging population intensifies, there is a corresponding decrease in the availability of healthcare professionals. Based on the forecasted values, it is evident that the number of healthcare institutions in China will continue to rise in the coming years. However, issues such as regional disparities, imbalance, and lack of coordination will become more pronounced. Instead of focusing solely on stimulating the development of the healthcare sector, it is more important to promote self-efficacy among residents and improve their physical health.

2. The Application of Artificial Intelligence (AI) in Healthcare: While the application of AI in healthcare improves medical services, it also introduces new challenges to the traditional doctor-patient relationship. These challenges include issues such as safety, responsibility attribution, transparency, workload, and the humanization of AI systems.

3. Uncertainties Related to Emerging and Sudden Infectious Diseases: Reliable preventive systems are crucial in dealing with emerging and sudden infectious diseases. It is important to diversify information sources and improve early warning technologies. Healthcare institutions must prepare for sudden public health emergencies by ensuring proper prevention, control, and treatment measures are in place, while also ensuring that disease control organizations are adequately supported. In response to these challenges, healthcare institutions need to adopt comprehensive measures, including strengthening internal management, optimizing resource allocation, integrating advanced technologies, improving staff training, and reinforcing regulatory compliance, to manage risks and improve service quality.

(3) Countermeasures

1. Further promote infrastructure development, advance healthcare system reform, and enhance residents' sense of well-being and security.
2. China's experience shows that limited medical resources, particularly doctors, are insufficient to meet patient demand for consultations, especially in emergency situations. A feasible solution is to integrate support systems within healthcare services to enable timely, efficient, and high-quality diagnosis and treatment anywhere and anytime.
3. There are significant disparities in consumption potential across the four major regions in China, resulting in unbalanced development. Therefore, regional integration should be emphasized, with stronger economic zone development to improve healthcare levels across regions.
4. The goal is to serve national strategies and people's livelihoods. The focus should be on "uniting efforts, emphasizing practical results, and making key breakthroughs" to contribute significantly to the healthy development of China's aging society. This includes advancing high-quality theoretical and empirical research on healthy aging, exploring chronic disease prevention and control strategies with Chinese characteristics, and contributing to the overall goal of "Healthy Aging" to support a healthy China and the rejuvenation of the nation.

References

- [1] Zhou, P., Chen, J. H., Lin, Q. F., et al. (2018). Trajectory interpolation method based on AIS. *Journal of Jimei University (Natural Science Edition)*, 23(6), 443–447. <https://doi.org/10.19715/j.jmuzr.2018.06.07>
- [2] Xue, C. W., Liu, T., Deng, L. B., et al. (2024). Improvement of the DBSCAN classification differential evolution algorithm based on principal component analysis. *Modern Electronic Technology*, 47(16), 171–179. <https://doi.org/10.16652/j.issn.1004-373x.2024.16.026>
- [3] Sun, X. Y., Wang, Y., Feng, X. M., et al. (2024). Competitive status and spatial correlation analysis of the healthcare market in Beijing. *Chinese Hospital Management*, 44(1), 19–22.
- [4] Zhang, Y. K. (2014). Hospital spatial agglomeration evaluation and prediction based on attractiveness and Moran index. *Journal of Southeast University (Philosophy and Social Science Edition)*, 16(S2), 85–87. <https://doi.org/10.13916/j.cnki.issn1671-511x.2014.s2.027>
- [5] Sun, C. Y., Chen, X. Y., Long, Q. P., et al. (2024). Epidemic trend characteristics of newly discovered HIV/AIDS cases in Longgang District, Shenzhen from 2015 to 2022 and quantitative prediction using a grey model. *Journal of Youjiang Medical University for Nationalities*, 1–7. Retrieved August 19, 2024, from <http://kns.cnki.net/kcms/detail/45.1085.R.20240805.1006.002.html>
- [6] Wu, J. N. (2024). Research on the development of the express delivery logistics industry in Dalian based on the grey prediction model. *Journal of Heihe University*, 15(7), 58–60 + 79.