Problem Chosen	2024	Team Control Number
Project two	MCM/ICM Summary Sheet	SZO

Analysis and Research on population of North Korea

Summary

This paper presents a comprehensive analysis and forecasting of population changes in North Korea. Our study begins by examining the current population dynamics, particularly analyzing the male-to-female ratio. This foundational data provides the basis for our subsequent predictive analyses.

To project future demographic trends, we employed four distinct modeling approaches: polynomial fitting, interpolation fitting, logistic model fitting, and Grey prediction modeling. Each model offers unique advantages and perspectives on population forecasting, catering to different aspects of demographic changes.

Polynomial Fitting: This model helps in understanding non-linear trends in historical data. We used the divided difference table to find the best fit that captures the complexities in the population changes over time. **Interpolation Fitting**: By applying interpolation methods, This model is particularly useful for estimating missing data points and refining the granularity of demographic insights. **Logistic Model Fitting:** Given its efficacy in handling population growth, the logistic model was used to forecast the population size considering the carrying capacity of the environment, which is crucial for predicting population stabilization or decline. **Grey Prediction Model:** This model is effective in dealing with small data sets and poor information, characteristic of the data scarcity in North Korean demographics. The Grey model provides reliable predictions under these constraints by using a minimal amount of data to forecast future trends.

Finally, the paper concludes with a comparative analysis of these four models. We evaluated each model's effectiveness based on its mean squared error with the given data and official prediction. Our findings indicate that while each model has its strengths, the choice of model depends heavily on the specific requirements of the demographic indicators being forecasted and the quality of available data.

This comprehensive study not only enhances our understanding of North Korea's demographic trends but also contributes to the methodology of population forecasting under constraints of data availability and quality. The insights gained from this research could assist policymakers and researchers in making informed decisions regarding demographic planning and policy-making in contexts similar to North Korea.

Keywords: North Korea; Population model; Population forecast

Contents

1 Analysis of the Problem 2

2 Model Pre-assumption 4

3 Modeling 4

3.1 method 1: Polynomial fitting 4

3.1.1 Expression of fitting function 4

3.1.2 Fitting figure 5

3.1.3 Prediction comparison 5

3.2 method 2: Logistic Model 6

3.2.1 Analysis 7

3.3 method 3:Interpolation method 7

3.3.1 Brief Introduction 7

3.3.2 Procedure 8

3.3.3 Analysis of the Model 8

3.4 method 4:Grey prediction model(GM(1,1)) 8

3.5 Model Evaluation 11

3.5.1 Assumption 11

3.5.2 Procedure 11

4 Reference 12

1 Analysis of the Problem

North Korea is located in the northern part of the Korean Peninsula in East Asia. The population of North Korea is approximately between 25 million and 26 million. North Korea's economy is mainly planned, but its economic development has been relatively lagging due to sanctions and its own policy restrictions for a long time. The male and female population structure is relatively balanced, but it is also influenced by traditional cultural concepts and family policies, such as encouraging childbirth. The North Korean government attaches great importance to education and healthcare, providing free basic education and healthcare services. However, due to resource constraints and outdated technology, the level of education and healthcare is relatively low. North Korea has long faced political and economic sanctions from the international community, and its relations with neighboring countries are complex, especially with countries such as South Korea, China, and the United States, which have received much attention. North Korea has a long history and unique cultural traditions, while also being influenced by the ideology and propaganda of the Workers' Party of Korea.

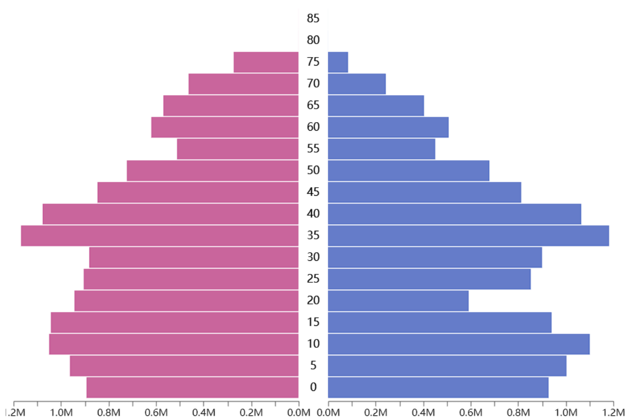

Figure 1: Male and Female

The above chart shows the current demographic structure of North Korea's male and female age groups.

Age Distribution: The population age distribution in North Korea appears relatively stable, showing a typical pyramid shape with a larger proportion of younger population gradually decreasing with age.

Gender Disparity: Overall, there is a roughly equal distribution of male and female population across different age groups. However, slight variations can be observed in certain age brackets, such as between the ages of 20 and 29, where the male population seems slightly higher than the female population. This may be influenced by socio-economic and cultural factors.

Population Aging: As age increases, the population size decreases, particularly evident in the age group of 70 and above, where there is a sharp decline in population numbers. This suggests that North Korea may be facing the challenge of population aging,

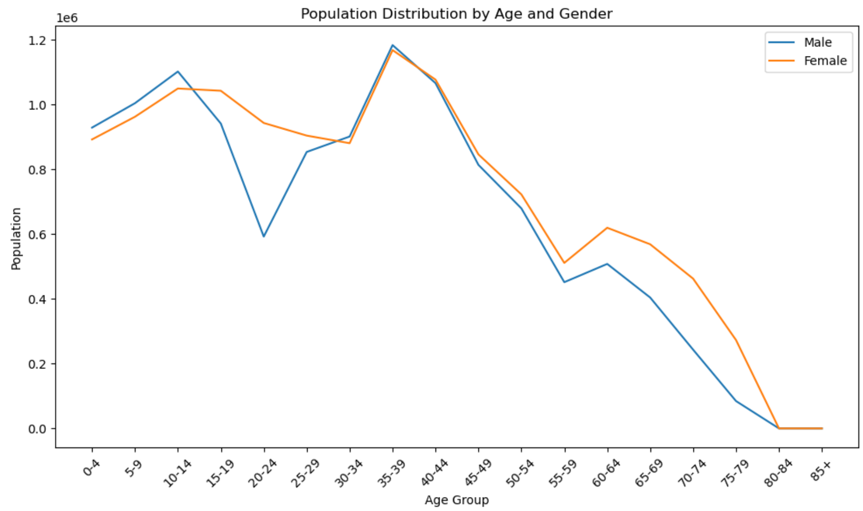

Figure 2: Age distribution-time

necessitating appropriate policies and measures to address this issue.

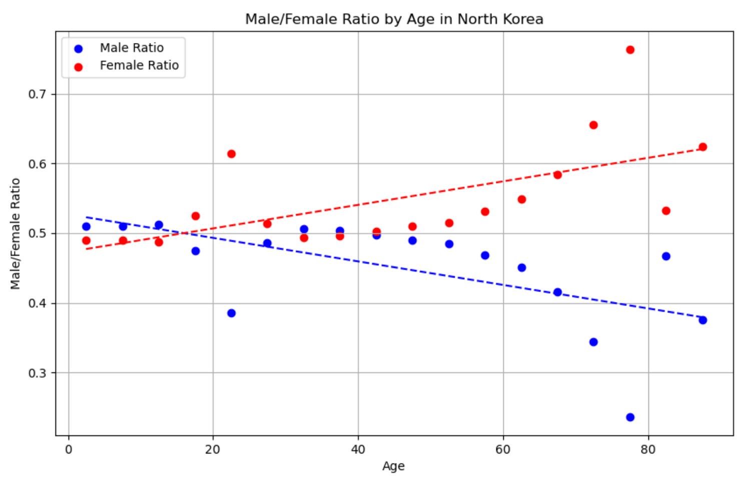

Figure 3: male, female ratio

The male and female ratios are roughly balanced at younger ages (0-20 years old), indicating a balanced sex ratio at birth. From around 20-60 years old, the prime working ages, the female ratio is slightly higher than males, possibly due to population losses from war, disasters, or other causes impacting male numbers more. After 60 years old, the female population significantly outnumbers males, and this gender gap widens with increasing age, likely reflecting lower life expectancy for males. Overall, the population pyramid shape appears relatively normal, suggesting a stable natural population growth pattern.

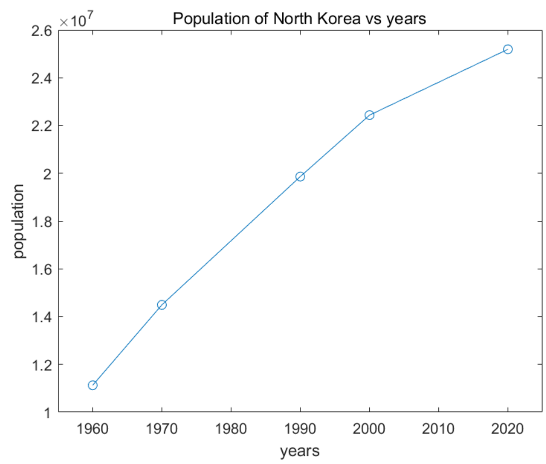

Figure 4: Population of North Korea vs years

2 Model Pre-assumption

1. We assume that population forms a one-dimensional function versus t, $i.e.$ we ignore other factors like terrible environment, survival supplies, and so on.
2. We assume that there is no inward or outward movement of population in the country
3. We assume the data that we searched is reliable.

3 Modeling

3.1 method 1: Polynomial fitting

3.1.1 Expression of fitting function

We use the polynomial fitting to analysis this question. Using difference table first we have

Years	Population	Δ	Δ^2	Δ^3	Δ^4
1960	11,127,017	3.3662×10^5	-2.2711×10^3	47.2761	-1.9860
1970	14,493,242	2.6849×10^5	-380.0967	-71.8850	
1990	19,863,008	2.5709×10^5	-3.9743×10^3		
2000	22,433,862	1.3785×10^5			
2020	25,190,961				

Table 1: Difference Table of Population

We can see that the difference table in the third column alternate positive and negative, approaching 0. We assume the relationship of population and years is a 2-order polynomial. Using the function $polyfit$ in Matlab, we have the parameters of the 2-order polynomial, so we write down our first fitting equation:

$$f(x) = -2099.72151515258 \cdot x^2 + 8592667.48485270 \cdot x - 8764242081.09509 \quad (1)$$

3.1.2 Fitting figure

Using the equation (1), we can draw the graph that compare the official data points and our predicting point in the same figure. Figure 5 shows the relationship of years and population with our prediction of polynomial method.

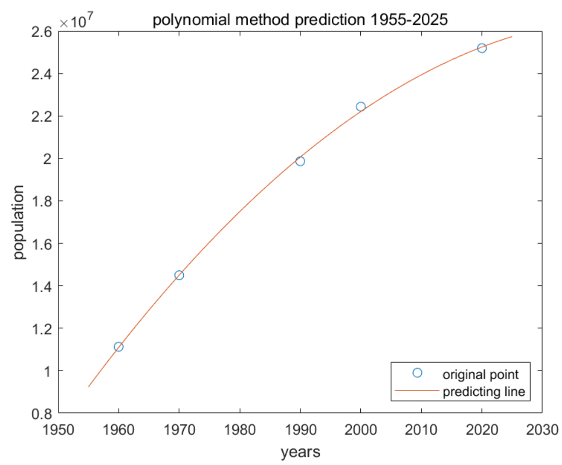

Figure 5: polynomial method prediction 1955-2025

3.1.3 Prediction comparison

Next, we want to observe the difference between our prediction and the official prediction.

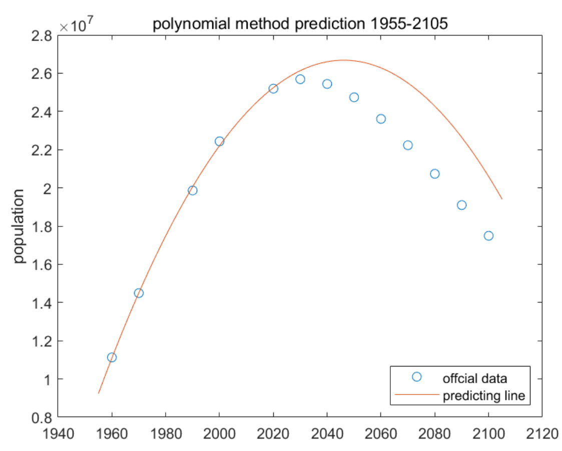

Figure 6: polynomial method prediction 1955-2105

Figure 6 shows the relationship between the fitting curve obtained by equation (1) and the official predicting data given by the government of North Korea. Our model is obviously a little bit higher than the official one. But both of the two models show the same trend of the population in North Korea that it will reach a peak in the next decade and drop down.

3.2 method 2: Logistic Model

Our second idea is to use the logistic model to fit because the background of our problem is to analyze the population and logistic model is a famous way in such area.

We first deal with the data and create a expression as below

$$logP = a \times year + b \tag{2}$$

Then we utilize linear regression and **"fitlm"**function to fit logistic model. Finally we got the value of the two parameters a and b displayed in the table below

a	0.0136
b	16.3171

Table 2: fitting value of logistic model

And then by transforming the data by **exp**, we got the fitting line as displayed in figure 7.

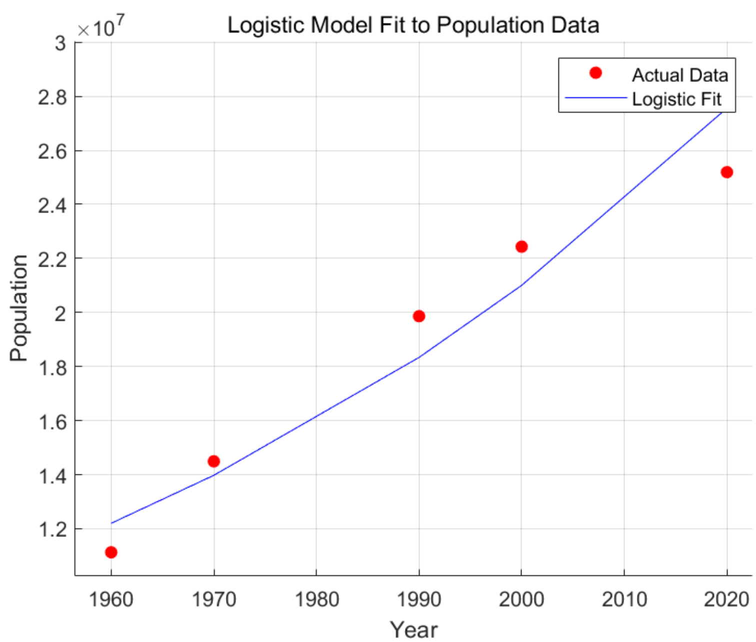

Figure 7: Logistic model

For our purpose, we then use the model to predict the following years by the expression:

$$P = e^{a\times(y-y_1)+b} = e^{0.0136\times(y-y_1)+16.3171} \tag{3}$$

Then we followed the equation and substituted $y = 2020, \cdots 2100$ to the equation and then got the prediction result. Finally, we added the official prediction to our figure 8 to compare the results.

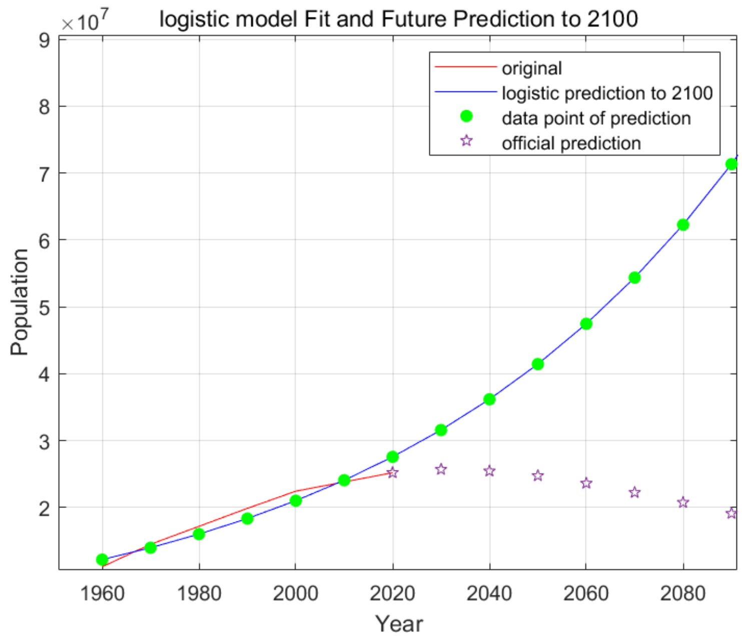

Figure 8: Logistic model prediction

From figure 8, we can find that the logistic model showed an increasing tendency of the population.

3.2.1 Analysis

As our assumption ignores the factor of survival material and other ingredients. Thus the logistic model will show a buoyant trend. Comparing with the official prediction, the two lines show different trend which implys that our logistic model is not perfectly appropriate for our requirement of predicting the population of South Korea.

3.3 method 3:Interpolation method

3.3.1 Brief Introduction

Linear Interpolation focuses on a one-dimensional data set and performs numerical estimation according to the value of the two nearest data points of the point that we need to predict.

3.3.2 Procedure

To transact the interpolation, our team utilized "Pchip" interpolating method and used function **interp1** in Matlab to help us. In the first step, we finished the interpolation according to the given data and then stored the interpolate function in matlab. Secondly, we generate the predicting value of population in the after years from 2020 to 2100 by substituting the value of years to the function. Eventually, to visualize the outcome, we plotted figure 9 below

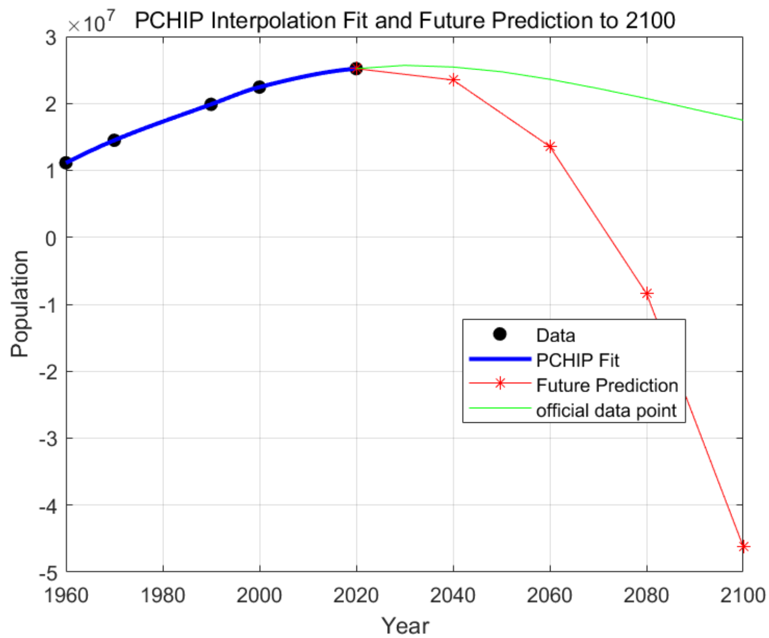

Figure 9: Visualization of Interpolation

3.3.3 Analysis of the Model

By using the interpolation method, we are able to conclude from the figure that the fitting is reasonable while the point we want is staying inside the range of the given data set. However, the prediction becomes aberrant when the time is close to 2100 because the population cannot be minus as common sense. Additionally, in this case, our prediction is far from the official prediction and even ridiculous.

Thus, when the evaluation is required, the result indicates that interpolation is useful when analyzing the point inside the range of the given set, but it is not precise to predict the point outside the range of the given set. Therefore, we should be scrupulous in using such a method.

3.4 method 4:Grey prediction model(GM(1,1))

The model theory The brief principle of the GM (1,1) prediction model is to first use the accumulation technique to make the data have an exponential law, then establish a first-order differential equation and solve it, and then reduce the obtained result to obtain the grey prediction value, thereby predicting the future. Step 1: Before establishing the grey prediction model, it is necessary to ensure the feasibility of the modeling method, that is, to perform a level comparison test on the known raw data. Set the initial non negative data sequence as

$$X^{(0)} = \left\{x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n)\right\}$$

Only when all$\sigma(k)$ fall within the calculation range can the model be established. The calculation and judgment formulas for grade ratio are:

$$\sigma(k) = \frac{x^{(0)}(k-1)}{x^{(0)}(k)}, \sigma(k) \in \left(e^{-\frac{2}{n+1}}, e^{\frac{2}{n+1}}\right)$$

The first-order accumulation sequence of $x^{(0)}$ obtained through accumulation operation can weaken the disturbance of $x^{(0)}$:

$$x_k^{(1)} = \sum_{i=1}^{k} x_i^{(0)}, k = 1, 2, \cdots, n$$

The sequence generated by the nearest neighbor mean of $X^{(1)}$ is $Z^{(1)}$

$$Z^{(1)} = \left\{z^{(1)}(2), z^{(1)}(3), \cdots, z^{(1)}(n)\right\}$$
$$z^{(1)}(k) = \frac{1}{2}\left(x^{(1)}(k) + x^{(1)}(k-1)\right)$$

Therefore, the corresponding differential equation of the GM (1,1) model can be obtained as follows:

$$x^{(0)}(k) + az^{(1)}(k) = b$$

where the$z^{(1)}$is the background value of GM(1,1).

Step 2: Construct data matrix B and data vector Y, which are respectively

$$\boldsymbol{B} = \begin{bmatrix} -z(2) & 1 \\ -z(3) & 1 \\ \vdots & \vdots \\ -z(n) & 1 \end{bmatrix} \quad \boldsymbol{Y} = \begin{pmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{pmatrix}$$

The least squares estimation parameter column of the grey differential equation satisfies

$$\boldsymbol{u} = [a \quad b]^{\mathrm{T}} = \left(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\right)^{-1} \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}$$

Among them, a mainly controls the development trend of the system, known as the development coefficient; The size of b reflects the relationship between data changes and is called the grey action quantity.

Step 3: Establish a model and solve for the generated and restored values. By solving according to the formula, a predictive model can be obtained

$$\hat{x}^{(1)}(k) = \left[x^{(0)}(1) - \frac{b}{a}\right] \mathrm{e}^{-a(k-1)} + \frac{b}{a}$$
$$k = 1, 2, \cdots, n$$

After subtraction, the restored predicted value is obtained.

So the role of this model:Grey prediction is a method of predicting systems containing uncertain factors. Grey prediction identifies the degree of differences in development trends among system factors through correlation analysis, and generates and processes the original data to find patterns of system changes. It generates data sequences with strong regularity, and then establishes corresponding differential equation models to predict the future development trends of things. We could use it to predict the population of North Korea.

The analysis results are as follows:

索引项	原始值	级比值	平移转换后序列值	平移转换后级比值
1900	4734292	-	29925253	-
1910	4785216	0.989	29976177	0.998
1920	5765649	0.83	30956610	0.968
1930	6902698	0.835	32093659	0.965
1940	8268915	0.835	33459876	0.959
1950	9903460	0.835	35094421	0.953
1960	11127017	0.89	36317978	0.966
1970	14493242	0.768	39684203	0.915
1990	19863088	0.73	45054049	0.881
2000	22433862	0.885	47624823	0.946
2020	25190961	0.891	50381922	0.945

Figure 10: Grade comparison test result table

索引项	原始值	预测值	残差	相对误差（%）
1900	4734292	4734292	0	0
1910	4785216	2945099.724	1840116.276	38.454
1920	5765649	4780005.099	985643.901	17.095
1930	6902698	6734574.622	168123.378	2.436
1940	8268915	8816612.242	-547697.242	6.624
1950	9903460	11034430.846	-1130970.846	11.42
1960	11127017	13396885.45	-2269868.45	20.4
1970	14493242	15913408.553	-1420166.553	9.799
1990	19863088	18594047.8	1269040.2	6.389
2000	22433862	21449506.095	984355.905	4.388
2020	25190961	24491184.338	699776.662	2.778

Figure 11: Model fitting results table

The average relative error of the model is 10.889%, indicating good fitting performance. The result shows that the population of North Korea in the next stage, that is 2030(because the step is 10 years) is

预测阶数	预测值
1	27731226.944

Figure 12: Result in 2030

it is 27731226.

We can loop this operation, predict one stage at a time, and add the predicted value to the data for recursive analysis.

3.5 Model Evaluation

Here we choose to calculate **Mean Squared Error** to compare the three different models. For the grey prediction model, we will adjust the order of historical data (with a 20 year interval) to predict the population in 2020 and compare it with real data to determine the effectiveness of the model.

3.5.1 Assumption

- We assume that the data from the official prediction is precise.

3.5.2 Procedure

As the line created by the interpolation method will pass through every given data point, so we took the prediction point into consideration and assumed the official data point is precise. So we utilize the equation

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(real_n - predict_n)^2 \tag{4}$$

Then after calculation, we finally got the MSE values of the three model. Thus, by com-

method	Value of MSE
Polynomial	4.43×10^{12}
Interpolation	5.62×10^{14}
Logistic	8.22×10^{14}

Table 3: Comparison of MSE

paring the MSE value, we can reach a preliminary conclusion that the **Polynomial model** is the most appropriate model for the prediction of the population of South Korea. And here the reason why the interpolation model has a lower MSE value than the value of the Logistic model is that the fitting line passes through all the given points and averts the error of the given point.

For the GM(1,1), the result shows:

It is 28115330. Comparison with real data, that is 25867467, the percentage error is

$$\frac{|28115330 - 25867467|}{25867467} \times 100\% = 8.7\%$$

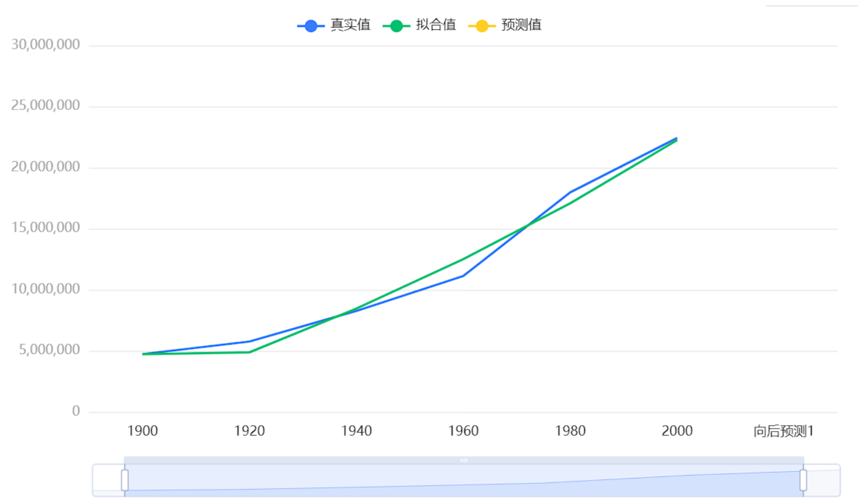

Figure 13: Fitting result

预测阶数	预测值
1	28115330.634

Figure 14: Result in 2020

4 Reference

[1] https://zh.city-facts.com/north-korea/population. (n.d.).

```
    % given data
years = [1960, 1970, 1990, 2000, 2020];
population = [11127017, 14493242, 19863008, 22433862, 25190961];

% log transformation for year and population
years_diff = years - years(1);
log_population = log(population);

% linear regression to fit logistic model
% create a linear model: log_population = a * years_diff + b
X = years_diff';
y = log_population';

% fit with linear model
mdl = fitlm(X, y, 'y ~ x1');

% get coefficients
coefficients = mdl.Coefficients.Estimate;

% get parameters of logistic model
a = coefficients(2);
b = coefficients(1);

% utilize logistic function to predict population
population_pred = exp(a * years_diff + b);

% depict original data and fiting line
figure;
scatter(years, population, 'r', 'filled');
hold on;
plot(years, population_pred, 'b-');
xlabel('Year');
ylabel('Population');
title('Logistic Model Fit to Population Data');
legend('Actual Data', 'Logistic Fit');
grid on;
hold off


years_2100=[2020:10:2100];
population_2100=[25190961 25683112 25436579 24736617 23606927
    22229696 20734133 19098762 17492412];
```

```
future=1960:10:2100;
future_dealed=future-1960;
P_predict=exp(a.*future_dealed+b);
figure;
plot(years, population, 'r');
hold on;
plot(future, P_predict, 'b-');
scatter(future, P_predict, 'g','filled');
plot(years_2100,population_2100,'p')
xlabel('Year');
ylabel('Population');
title('logistic model Fit and Future Prediction to 2100');
legend('original','logistic prediction to 2100','data point of
    prediction','official prediction')
grid on;
hold on;
```

method 1: polynomial fitting code

```
years=[1960 1970 1990 2000 2020];
population=[11127017 14493242 19863008 22433862 25190961];

plot(years, population,'o-');
xlabel('years')
ylabel('population')
title('Population of North Korea vs years')
xlim([1955 2025])

%polynomial method
dt=difference_table(years,population)
dt(:,3:5)
disp(['we can see that the difference table in the third column
    alternate positive and negative, approaching 0']);
disp(['we assume the population is a polynomial and we use 2-order to
    fit it.'])

poly_fit=polyfit(years,population,2)

%ploting predict point
plot(years, population,'o');
hold on
plot(1955:2025,polyval(poly_fit,1955:2025));
xlabel('years')
ylabel('population')
title('polynomial method prediction 1955-2025')
legend('original point','predicting line','Location','southeast')
hold off
```

```

%predict till 2100
years_2100=[years 2030:10:2100];
population_2100=[population 25683112 25436579 24736617 23606927
   22229696 20734133 19098762 17492412];

plot(years_2100, population_2100,'o');
hold on
plot(1955:2105,polyval(poly_fit,1955:2105));
xlabel('years')
ylabel('population')
title('polynomial method prediction 1955-2105')
legend('offcial data','predicting line','Location','southeast')
hold off
```

function used in method 1

```
function Y = difference_table(x_val,y_val)%compute the difference
   table of two row vector
% x_val and v_val are two row vector
if length(x_val) == length(y_val)
    Y=zeros(length(x_val));
    Y(:,1) = y_val';
    for i = 2:length(x_val)
        z=0;
        for j = 1:(length(x_val)-i+1)
            Y(j,i) = (Y(j+1,i-1) - Y(j,i-1)) / (x_val(j+i-1)-x_val(j));
            if Y(j,i) ~= 0
                z = 1;
            end
        end
        if z == 0
            disp(['This data set equals to zeros at the ', num2str(i),
               'th column. So we could use a ', num2str(i-2), ' order
               polynomial to fit it.']);
            break
        end
    end
else
    error('Length of two entries is different.')
end

end
```

```
% given data
years = [1960, 1970, 1990, 2000, 2020];
population = [11127017, 14493242, 19863008, 22433862, 25190961];

```

```
% log transformation for year and population
years_diff = years - years(1);
log_population = log(population);

% linear regression to fit logistic model
% create a linear model: log_population = a * years_diff + b
X = years_diff';
y = log_population';

% fit with linear model
mdl = fitlm(X, y, 'y ~ x1');

% get coefficients
coefficients = mdl.Coefficients.Estimate;

% get parameters of logistic model
a = coefficients(2);
b = coefficients(1);

% utilize logistic function to predict population
population_pred = exp(a * years_diff + b);

% depict original data and fiting line
figure;
scatter(years, population, 'r', 'filled');
hold on;
plot(years, population_pred, 'b-');
xlabel('Year');
ylabel('Population');
title('Logistic Model Fit to Population Data');
legend('Actual Data', 'Logistic Fit');
grid on;
hold off


years_2100=[2020:10:2100];
population_2100=[25190961 25683112 25436579 24736617 23606927 22229696
    20734133 19098762 17492412];


future=1960:10:2100;
future_dealed=future-1960;
P_predict=exp(a.*future_dealed+b);
figure;
plot(years, population, 'r');
hold on;
plot(future, P_predict, 'b-');
```

```
scatter(future, P_predict, 'g','filled');
plot(years_2100,population_2100,'p')
xlabel('Year');
ylabel('Population');
title('logistic model Fit and Future Prediction to 2100');
legend('original','logistic prediction to 2100','data point of prediction','official prediction')
grid on;
hold on;
```

```
%Given
x = [1960 1970 1990 2000 2020];
y = [11127017 14493242 19863008 22433862 25190961];

%                        'pchip'
xq = linspace(min(x), max(x), 400); %
interpFit = interp1(x, y, xq, 'pchip');

%                        2020210020
future_years = 2020:20:2100;
future_predictions = interp1(x, y, future_years, 'pchip', 'extrap');

%
figure;
plot(x, y, 'ko', 'MarkerFaceColor', 'k'); %
hold on;
plot(xq, interpFit, 'b-', 'LineWidth', 2); %
plot(future_years, future_predictions, 'r*-'); %
xlabel('Year');
ylabel('Population');
title('PCHIP Interpolation Fit and Future Prediction to 2100');
grid on;
hold on;


%
disp('Future Predictions from 2020 to 2100:');
for i = 1:length(future_years)
    fprintf('Year %d: %d\n', future_years(i), round(future_predictions(i)));
end


%
%
x = [1960 1970 1990 2000 2020];
y = [11127017 14493242 19863008 22433862 25190961];

```

```
%         'pchip'
interp_values = interp1(x, y, x, 'pchip');

%
ssd = sum((y - interp_values).^2);

%
disp(['Sum of Squared Deviations (SSD) for the interpolation model: ', num2str(ssd)]);

%%
years_2100=2020:10:2100;
population_2100=[25190961 25683112 25436579 24736617 23606927 22229696 20734133 19098762 17492412];
plot(years_2100,population_2100,'g')
legend('Data', 'PCHIP Fit', 'Future Prediction', 'official data point','Location', 'Best');
```

```
% calculate MSE

years_total=[1960 1970 1990 2000 2020 2030 2040 2050 2060 2070 2080 2090 2100];
population_total_offical=[11127017 14493242 19863008 22433862 25190961 25683112 25436579 24736617 23606927 22229696 20734133 19098762 17492412];

years_total_dealed=years_total-1960;

p_logistic=exp(0.0136.*years_total_dealed+16.3171); %get value of logistic

MSE_logistic=sum((p_logistic-population_total_offical).^2)./length(years_total_deal
    %MSE of logistic


x = [1960 1970 1990 2000 2020];
y = [11127017 14493242 19863008 22433862 25190961];

%                           'pchip'
xq = linspace(min(x), max(x), 400); %
interpFit = interp1(x, y, xq, 'pchip');

%                          2020210020
future_years = 2030:10:2100;
future_predictions = interp1(x, y, future_years, 'pchip', 'extrap');

MSE_interpolation=sum((future_predictions-population_total_offical(1,6:13)).^2)/len

```

```
p_polynomial= -2099.721515.*years_total.^2 + 8592667.485*years_total-
   8764242081.09509;

MSE_polynomial=sum((p_polynomial-population_total_offical).^2)/length(years_total);

disp(MSE_polynomial)
disp(MSE_interpolation)
disp(MSE_logistic)
```