MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE

V.N. Karazin Kharkiv National University
School of Mathematics and Computer science
Department of theoretical and Applied informatics

**Qualification work**
**Master**

on the topic:
«Prediction of the dynamics COVID19 epidemic process of using the 3rd degree Polynomial regression model»

Done by: 2-th year student, group MCS-63
specialty: Computer science and Information Technologies
education program: «Computer Sciences»
Dai Huiyu

Supervisor: Victoriya Kuznietcova
Reviewer: Kseniia Bazilevych
Adviser: Ruslan Borodai

Kharkiv, 2024

catalogue

# ABSTRACT

At the end of 2019, the outbreak of COVID-19, and spread to the world at an extremely fast rate, has had a profound impact on human life, health and normal life, as well as the global economic and social order. Compared with the previous SARS and MERS, novel coronavirus has higher variability and transmissibility, which poses a great challenge to the public health systems of various countries. The global spread of the epidemic has led to a shortage of social resources, serious economic losses, and a huge impact on Peoples Daily life. In this context, an accurate prediction of the development trend of the epidemic can not only provide a scientific basis for the governments prevention and control measures, but also help to optimize the allocation of resources and reduce the social costs caused by the spread of the epidemic. In recent years, various prediction models have been widely used in the study of epidemic trends, including traditional statistical methods (such as ARIMA models), mathematically derived infectious diseases models (such as SEIR models), and machine learning models (e. g., neural networks and support vector machines). These models show unique advantages in different data environments and application scenarios. However, due to the complexity and non-linear characteristics of epidemic data, how to find a balance between simplicity and prediction accuracy has become an urgent problem. Based on this, this paper proposes to use a three-time polynomial regression model for the epidemic of COVID-19. The process is prediction to explore its potential in capturing the nonlinear trend of the epidemic

The study in this paper is based on publicly available data resources, including daily confirmed case data provided by the World Health Organization (WHO) and ArcGIS Dashboards. Based on the data from January 1,2024 to September 30,2024, Australia, China mainland, South America and Taiwan, China were selected as subjects to analyze the nonlinear changes in the development of

the epidemic in these regions. First, a time-series dataset suitable for polynomial regression modeling was established through data cleaning and preprocessing. Then, a cubic polynomial regression model was constructed for parameter estimation using least squares, and prediction of future case trends. Finally, the model performance was evaluated by mean square error (MSE) and coefficient of determination ($R^2$), and the difference between actual and predicted values was compared by visualization.

The results show that the cubic polynomial regression model can better capture the nonlinear features in the outbreak data and show high accuracy in short-term prediction. When the historical data of a time period is added, the prediction effect is significantly improved, the prediction error is significantly reduced, and the coefficient of determination ($R^2$) is stable above 0.90. The experiment also found that the model has better adaptability to the epidemic data in different regions, and can reflect the changing trend in different stages of the epidemic process in each region. Compared with more complex machine learning models, the computational process of cubic polynomial regression models is simple and highly interpretable, which is suitable for scenarios with limited resources or high demand for real-time prediction.

# 1 INTRODUCTION

## 1.1 Research Background and Significance

Since ancient times, infectious diseases have been an important challenge that threatens human life and health. From plague, cholera, smallpox and other viruses, infectious diseases have caused large numbers of population deaths[1]. With the progress of science and technology, the improvement of medical level and the development and popularization of vaccines, more and more infectious diseases have been effectively controlled, and some infectious diseases are even completely eradicated[2]. However, infectious diseases are constantly updated and changed along with social development, and new epidemic diseases are frequently erupted, and human beings still need to continue to fight against infectious diseases[3]. In 2002, SARS (SARS) occurred in Guangdong and spread to the world, with a fatality rate of 10%; In 2012, Middle East Respiratory Syndrome (MERS) outbreak in Saudi Arabia spread to more than 20 countries, with a mortality rate of 34%[4].

The outbreak of COVID-19 highlights the complexity and harmfulness of infectious diseases. Different from previous influenza viruses, novel coronavirus spreads rapidly and highly variable. Its main transmission modes include droplet transmission, contact transmission and aerosol transmission, which may cause serious respiratory diseases after infection, accompanied by fever, cough, loss of taste, fatigue and other symptoms[4]. Due to the large-scale movement of population and weak awareness of prevention and control, COVID-19 has been rapidly spreading around the world[5]. In the early days of the epidemic, the number of infected people increased exponentially, which not only put great pressure on the public health system, but also had a profound impact on the global economic and social order[6]. The spread of the epidemic has overloaded existing

healthcare systems, challenging health and community systems, and even causing a range of social and economic problems[7].

Countries around the world have adopted a variety of interventions to contain the outbreak, including nucleic acid testing, limiting population movement, maintaining social distancing, and closing public places[8]. Although these measures have contained the spread of the epidemic to some extent, the epidemic has not been completely eliminated due to the high infectivity and variability of the novel coronavirus[9]. In this context, it is of great significance to study the dynamic development law of the epidemic situation and explore effective prediction models to scientifically formulate prevention and control strategies, optimize the allocation of medical resources and improve the response capacity[10].

Among many prediction methods, the polynomial regression model can better capture the non-linear data trends because of its simplicity and flexibility, which provides an effective means for the epidemic trend prediction[11]. Compared with complex machine learning models, polynomial regression models have the advantages of easy to implement, small computational amount and easy to explain, especially suitable for the prediction and analysis of short-term trends[12]. Based on the three polynomial regression model, this paper analyzes the cumulative data of COVID-19 cases in different regions and explores its application value in the prediction of the dynamic trend of the epidemic[13].

In this study, Australia, China mainland, South America and Taiwan, China regions were selected to analyze their COVID-19 epidemic trends. The reasons for choosing these four regions are that they are representative in epidemic prevention and control and medical resources: Australia and China mainland adopted strict prevention and control measures in the early stage of the epidemic, with fast vaccination speed, high coverage rate and relatively perfect medical system; some countries in South America have severe epidemic situation due to limited medical

resources and slow vaccination; Taiwan is relatively closed and the epidemic prevention and control was relatively successful in the early stage, but the vaccination rate is slow due to the shortage of vaccine supply and vaccination disputes[14]. The differences in prevention and control strategies and measures in these regions provide a diversified perspective for analyzing the law of epidemic development and the performance of prediction models[15].

By constructing and optimizing the three polynomial regression models, this paper tries to reveal the epidemic rules of COVID-19, compare the effects of epidemic prevention and control measures in different regions, and provide a theoretical basis for possible infectious disease outbreaks in the future. We show that three polynomial regression models based on historical data can not only effectively describe the nonlinear trends of outbreaks, but also provide high accuracy in short-term prediction[16].

The outbreak of infectious diseases has always accompanied the development of human society, and the possible infectious diseases in the future will continue to pose a threat to global health. Therefore, the use of effective models to predict epidemics can not only help countries to better cope with the current challenges of epidemic prevention and control, but also provide important reference and guidance for future public health events. Through this study, it can provide scientific basis for the formulation and implementation of epidemic prevention and control strategies in different countries and regions around the world, and make beneficial exploration for the research on infectious disease prevention and control.

## 1.2 Study status

In recent years, more and more researchers focus on predicting the epidemic trends of infectious diseases. In the face of the dual threats of infectious diseases to

human health and social economy, the scientific community has studied the law of their transmission more and more deeply. From the SARS virus at the beginning of this century to the MERS virus, to the COVID-19 outbreak at the end of 2019, every major infectious disease epidemic has had a huge impact on the global public health and economic system. In particular, COVID-19 has spread rapidly, has a wide impact range and has strong variability, which has brought unprecedented challenges to epidemic prevention and control and resource allocation. In this context, using historical data to build an effective prediction model to predict the development trend of the epidemic situation can not only help to reveal the epidemic law of infectious diseases, but also can provide an important basis for the scientific formulation of epidemic prevention and control measures. Therefore, predicting the epidemic trend of infectious diseases has become a research hotspot in academia and public health worldwide. At present, scholars at home and abroad have put forward a variety of methods to predict the epidemic trend of COVID-19, which can be summarized into three categories: prediction methods based on traditional dynamic models, prediction methods based on classical statistical models, and prediction methods based on machine learning[17].

### 1.2.1 Traditional kinetic model

Kinetic models, especially the SEIR models in epidemiology, have long been a central tool for studying and predicting the spread of infectious diseases. Such models establish mathematical equations to describe the mechanism of disease transmission in the population, including four basic states: susceptible (Susceptible), exposed (Exposed), infected (Infectious) and recovered (Recovered). The advantage of the SEIR model is its ability to comprehensively consider the latency of the disease, which is the main difference from others such as SIR models. In the model, individuals switch from susceptible to exposed states,

meaning that they are already infected with pathogens, but have not yet shown symptoms that can transmit to others. Subsequently, these individuals become infected, spreading the virus to other susceptible individuals, and eventually some recover and gain immunity, and become recover. This model design not only depicts the continuous process of disease transmission, but also is able to simulate the peak value and control effects of disease transmission, providing decision support for public health interventions[18].

Although the SEIR model is widely used theoretically, it faces many challenges in practice. First, the validity of the model is greatly dependent on accurate parameter setting, such as infection rate and recovery rate, which often need to be estimated through historical data. However, in the face of emerging diseases, especially the novel coronavirus outbreak like COVID-19, the lack of sufficient historical data makes the parameter estimation complicated and imprecise. Furthermore, SEIR models generally assume uniform mixing of populations, ignoring the heterogeneity of population exposure in social networks, which is impractical in highly global and mobile modern societies. For example, densely populated urban areas and sparsely populated rural areas may differ greatly in the speed and pattern of disease transmission, which are often not fully represented in traditional SEIR models[19].

In the 2020 study, Wang et al attempted to remedy this deficiency by combining the SEIR model with real-time migration data. Their model considers the flow of people between different regions, allowing the model to more realistically simulate the geographic diffusion process of the virus. This approach somehow improves the adaptability of the model to the actual outbreak and the prediction accuracy, but also increases the complexity of the model and the need for data. Moreover, in the face of COVID-19, the traditional SEIR model needs further adjustment and optimization to adapt to rapid changes in viral characteristics and emerging outbreak data. For example, given the case that

variation in the virus can lead to reinfection, the model may need to add new transition states or adjust for existing state transition probabilities.

In conclusion, while kinetic models such as SEIR provide powerful theoretical tools for infectious disease prediction and control, they still face challenges in responding to rapidly changing outbreaks and practical applications. Future research needs to be done in enhancing model flexibility and adaptability, and in reducing reliance on high-quality data.

### 1.2.2 Classical statistical model

Classical statistical models play a central role in the analysis of infectious disease data, enabling researchers to perform a quantitative analysis of various factors of disease transmission by providing a structured mathematical interpretation of the data. Such models, including regression analysis, time series analysis, and survival analysis, are not only used to estimate the trend and intensity of disease transmission, but also to evaluate the effects of different interventions. For example, linear and non-linear regression models are often used to explore how environmental factors, socioeconomic status, and public health policies influence the speed and extent of disease transmission. In this way, statistical models help public health officials to understand the complex dynamics of disease transmission, thereby designing more effective strategies for disease prevention and control. Furthermore, time series analysis is particularly suitable for analyzing infectious disease data because it can handle seasonal changes and long-term trends in the data, such as using the ARIMA model to predict the future movements of outbreaks[20].

However, although statistical models provide a deep understanding of the data, they also have their limitations. First, these models often require large amounts of historical data as support, and are based on assumptions such as

independence and homogeneous distribution of data, which are often difficult to meet in practice. For example, rapid changes in outbreaks and the emergence of new variants may rapidly change the transmission dynamics of the virus, making models built based on historical data unable to accurately predict future cases. In addition, traditional statistical models often ignore the effects of interactions among individuals and the influence of community network structure, which may lead to a biased understanding of the mechanisms of epidemic transmission. Therefore, while statistical models are very useful in analyzing disease transmission, they need to be used in combination with other types of models such as kinetic and machine learning models to compensate for the shortcomings of a single model and provide more comprehensive and accurate outbreak prediction and analysis.

### 1.2.3 Machine learning-based methods

As computing power has improved, machine learning methods have shown strong potential in epidemic prediction. For example, deep learning models such as LSTM (long-short-term memory network) and GRU (gating cycle unit) can capture long-term dependencies in time series data and are suitable for processing the temporal characteristics of epidemic data. In addition, integrated learning methods (such as random forest, XGBoost) can effectively improve the prediction accuracy by combining the prediction results of multiple models. However, machine learning models often require a large amount of high-quality data to support it, while the black-box feature of the model makes it face some challenges in practical application.

Among the above three types of methods, the polynomial regression model gradually receives attention because of its simplicity and flexibility. Compared

with complex deep learning models, the calculation process of polynomial regression models is clear and easy to understand, especially in short-term trend prediction and resource-limited scenarios. Recently, some studies have tried to model the nonlinear features of epidemic data through high-order polynomial regression, with good results. For example, some scholars have used three polynomial regression to predict the short-term trends of outbreaks and found that they show good fitting ability when handling nonlinear changes.

In this paper, we combine the advantages of the three-time polynomial regression model on the basis of previous studies and study its application in the dynamic change prediction of the COVID-19 epidemic process. By analyzing the epidemic data in different regions, exploring the applicability and prediction effect of the model can provide a scientific basis for epidemiological research and public health policy making. Compared with traditional kinetic models and complex machine learning models, three polynomial regression models provide a simple and efficient solution for epidemic trend analysis, and also provide new ideas for the prediction of future outbreaks.

### 1.3 Study content

This study aims to provide scientific foresight of the future fluctuations of the epidemic by using the three-way polynomial regression model in China. The specific research content includes the following key aspects:

1). Data collection and preprocessing:

Daily case data on COVID-19 were collected from authoritative health organizations and national databases.

Data were cleaned and pre-processed, including date format standardization, missing value processing, and data type transformation to ensure that the data quality complies with the analysis requirements.

2). Model establishment and verification:

A three-fold polynomial regression model was established to learn the increasing trend of outbreaks through historical data and analyze the nonlinear relationship between the number of cases and time.

Cross-validation and data-fitting methods are used to assess the predictive power and accuracy of the model, ensuring that the model can reliably predict future case development.

3). Trend Forecast and analysis:

Predict the future trend of the epidemic for a certain period of time, including short term (e. g., 3 days and 7 days) and medium and long term (e. g., 14 days, 21 days and 30 days) case forecasts.

Analyze the difference between the prediction results and the actual cases, and explore the factors that may affect the prediction accuracy.

4). Results presentation and discussion:

The prediction results are directly displayed, and the relationship between the prediction and the actual data is analyzed.

Discuss the performance of the model in different epidemic stages, and evaluate the adaptability and prediction accuracy of the model during the peak epidemic and stationary periods.

# 2 RELATED THEORETICAL BASIS AND EVALUATION INDEX

## 2.1 Statistical correlation analysis

The Pearson correlation coefficient can be used to compare the correlation between multiple variables, select the most relevant variable as predictors and improve the accuracy of model prediction. In this paper, Pearsons correlation coefficient was used to analyze the factors affecting the epidemic situation during the study, and the correlation was determined by calculating the Pearsons correlation coefficient between the selected indicators and the cumulative confirmed cases. The Pearson correlation coefficient is an indicator used to measure the degree of linear correlation between variables, calculated from the correlation coefficient and variance between two populations. During the study of this paper, there is the effect of dimension between the data, and the data needs to be standardized before calculating the correlation coefficient. Through the standardization of data processing, the new crown outbreak data set data can be converted into dimensionless data, can ensure that each index on the result of the effect of the same degree, data with similar scale, make the distribution of data more stable, so as to improve the modeling of new crown outbreak data prediction stability and accuracy.

Data standardization:

$$y_i = \frac{x_i - x}{s} \tag{2.1}$$

Where $x_i$ is the raw data, x is the mean of the raw data, and s is the standard deviation of the raw data.

correlation:

$$\rho = \frac{\operatorname{cov}(x, y)}{\sqrt{D(x)}\sqrt{D(y)}} \tag{2.2}$$

Where cov (x, y) represents the covariance between the two samples, and D (x) and D (y) represent the variance of the two samples.ρ Represents the correlation coefficient between sample X and sample Y, which is one of the methods to measure the degree of correlation between sample X and sample Y. The value range of the correlation coefficient is [-1,1]. The higher the absolute value of ρ, the higher the correlation between sample X and sample Y; the smaller the absolute value of ρ, the lower the correlation between sample X and sample Y. When the absolute value of ρ is between 0.6 and 1, there is a strong positive correlation between the two variables, namely that the increase in one variable is highly linearly associated with the increase in the other variable.

## 2.2 Basic principles of the three-degree polynomial regression model

A cubic polynomial regression model is a statistical method that describes the relationship between input and output variables through a mathematical formula. The polynomial regression model, characterized by different powers of the input variables, constructs nonlinear maps to realize the fitting and prediction from the input data to the output results. In a cubic polynomial regression model, the model assumes that the relationship between the target variable y and the input variable x can be described by the following formula:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon \tag{2.3}$$

It is the estimated parameter of the model to represent the weight of each order and the error term to reflect part of the data fluctuations that the model fails to

capture. The basic composition of the polynomial regression model includes independent variables (input characteristics), target variables (output results), and parameters (weights). By taking different powers of the input variables as features, the cubic polynomial regression model is able to capture non-linear trends in the data. Compared to linear regression models, cubic polynomial regression models have higher flexibility and can better fit complex relationships. $\beta_0, \beta_1, \beta_2, \beta_3$

2.2.1 Structure of the cubic polynomial regression model

Input feature extension:

When constructing a cubic polynomial regression model, the original input variables were first feature expanded to generate equal higher order features. These extended features participate in the model calculation through a linear combination, thus improving the non-linear fitting ability of the model. $x^2, x^3$ ;

Parameter optimization of the model:

The parameter optimization of polynomial regression models usually adopts the least squares method (OLS, OrdinaryLeastSquares), that is, the parameter values are estimated by minimizing the sum of squares between the actual observed and predicted values. The specific optimization process is as follows:

objective function:

$$\text{minimize} \quad \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3))^2 \tag{2.4}$$

optimization method:

The parameters are calculated by analytical solution or numerical optimization method (such as gradient descent) to minimize the value of the objective function.

Predictive capability of the model:

The cubic polynomial regression models are able to predict future data by learning non-linear relationships in historical data. For example, in the prediction of epidemic development trends, the model can fit the trend curves according to past case data and predict future changes.

2.2.2 Training process

The training process of the cubic polynomial regression model includes the following steps:

Forward calculation:

The input data were calculated by the model formula to obtain the predicted values. The output of the model consists of a weighted sum of the input features and their coefficients:

$$y_{\text{pred}} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \tag{2.5}$$

Error calculation:

Compare the predicted and true values, calculate the error (e. g. mean square error MSE): $y_{\text{pred}}\, y$,

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_{\text{pred},i})^2 \tag{2.6}$$

parameter adjustment:

Using gradient descent or other optimization algorithms, the model parameters are adjusted according to the error. $\beta_0, \beta_1, \beta_2, \beta_3$ :

### 2.2.3 Application and advantages

Trim polynomial regression models are widely used for fitting and prediction of nonlinear data because of their simplicity and high efficiency. For example, in the analysis of epidemic transmission trends, the model is able to capture the nonlinear trends of epidemic data and generate smooth and reasonable prediction curves. Compared with the complex neural network models, the cubic polynomial regression has the following advantages:

Easy to use: without complex hyperparameter adjustment and high computing resources, suitable for rapid deployment and small-scale data scenarios.

Good interpretability: the model parameters have a clear physical meaning and facilitate the analysis of the relationship between variables.

Effectively capture nonlinear trends: Through high-order feature extension, the model performs well in handling nonlinear data, especially with high accuracy in short-term prediction.

With the development of data science, the polynomial regression model is still an important tool to solve the nonlinear regression problem. By reasonably expanding the input features and optimizing the model parameters, the efficient and accurate prediction effect can be achieved, providing strong support for all kinds of scientific research and practical applications.

## 2.3 Model evaluation index

Generally, the evaluation of model is divided into that of classification model and the evaluation of regression model. In the classification task, evaluation indexes such as accuracy, recall, and F1 score can be selected to evaluate the model. The study in this paper is the regression task. For the regression task, the

average absolute error (MeanAbsoluteError, MAE), mean square error (MeanSquaredError, MSE), root mean square error (RootMeanSquaredError, RMSE), coefficient of determination (CoefficientofDetermination), R2) mean absolute percentage error (MeanAbsolutePercentageError, MAPE) are selected as the evaluation indicators.

The average absolute error MAE represents the average of the absolute value between the predicted value and the true value, which is mainly used to measure the average length of the error between the predicted value and the true value, and has good stability for abnormal data MAE. MAE was calculated by summing and averaging the absolute value of the difference between predicted and true values. The MAE was calculated as shown in Eq.

$$MAE=\frac{\sum_{i=1}^{n}|y_i-\hat{y}_i|}{n} \tag{2.7}$$

Where yi is the true value of the data, i is the predicted value, and n is the predicted days. The smaller the MAE, the closer the predicted results are to the true value, the higher the fit of the model, the better the prediction.$\hat{y}$

The mean square error MSE represents the average of the square of the difference between the predicted value and the true value, and is mainly used to measure the deviation between the predicted value and the true value. MSE is calculated by averaging the sum of the squared differences between the predicted and the true values. The MSE calculation procedure is performed as described in Eq.

$$\text{MSE}=\frac{SSE}{n}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{2.8}$$

Where SSE is the sum of squared errors between the true value and the predicted value, the smaller the MSE, the closer the predicted result is to the true model result, the higher the fit of the model, and the more accurate the prediction of the data.

The root mean square error RMSE was used to measure the magnitude of the error between the predicted value and the true value, which is calculated as the arithmetic square root of the MSE. The RMSE calculation procedure is performed as described in Eq.

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{2.9}$$

Where the RMSE is closer to the true error amount relative to the MSE. When the MSE value is relatively large, we can consider using RMSE as an evaluation index. Similarly, the smaller the RMSE value, the better the regression is.

MAPE is a relative error measure, which can reflect the percentage error between the predicted value and the actual value, eliminating the influence of the order of magnitude and the order of magnitude, so that the data of different units and orders of magnitude can be effectively compared. The MAPE calculation process is as shown in the formula.

$$\text{MAPE}=\frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \tag{2.10}$$

The value range of MAPE is [0, +), and the smaller the MAPE value is, the smaller the model prediction error is, and the higher the prediction accuracy is. Following the definition of Lewis [49], the prediction was considered inaccurate if the value of MAPE is greater than 50%. The prediction is reasonable when the value is less than 50% and greater than 20%. The prediction was considered good if it was less than 20% and more than 10%. The MAPE values of less than 10% have very good predictions. When the order of magnitude of the true value in the data is large, the same prediction error may correspond to smaller MAPE values, leading to some bias in the evaluation of the prediction error,

So that MAPE can not fully reflect the prediction accuracy.

In addition to MAE, MSE, RMSE, and MAPE, the frequently used evaluation index is the determination coefficient R2, which is calculated as shown in the formula.

$$\text{R2}=1-\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n}(\bar{y}_i-y_i)^2} \tag{2.11}$$

Where yi is the average of the true value of the data. The value of R2 is [0,1], and the closer R2 is to 1, the better the model fit and the more accurate the data prediction; the closer R2 is to 0, the worse the model fit and the greater the deviation of data prediction. The R2 value is negative when the prediction effect of the model is worse than the sampling average.

Among the above evaluation index methods: MAE is suitable for use when having outliers in the data set, However, the data set in this paper is a collated data set, No obvious outliers; Although the MSE is convenient to calculate, But it can not maintain the same dimension as the sample data; Although RMSE maintains the same dimension as the sample data, But it is more sensitive to outliers; MAPE can compare data of different orders of magnitude, Reflecting the relative magnitude of the error, In the data is more complex, order of magnitude and magnitude of different, And the need to compare the relative size of the prediction error in the scenario has a large advantage; R2 has a strong explanatory power in the regression prediction model, Can be used to compare the predictive power of the different models, When modeling on the same dataset, The R2 can be used to compare the predictive power of the different models, Thus choosing the optimal model. Through the above comprehensive analysis of the model evaluation indicators and the cumulative confirmed cases of COVID-19 studied in this paper has a large order of magnitude. In the course of this paper, MAPE and R2 were selected to comprehensively evaluate the prediction effect of the model.

In this paper, to avoid insufficient learning, the model uses RMSE error as the loss function with 1000 iterations. MAPE and R2 at different lag periods in the model final output test set.

## 2.4 Summary of this chapter

This chapter first introduces the statistical analysis methods for studying the dynamics of cumulative confirmed cases of COVID-19, including the preprocessing and correlation analysis of data. Secondly, it details the theoretical basis and application methods of three polynomial regression models, and explains the principle and advantages of the model in capturing the nonlinear change trend of the epidemic. Then the model evaluation indexes based on regression task,

including the mean absolute percentage error (MAPE) and the coefficient of determination ($R^2$), and then the role of these indicators in measuring the prediction accuracy and fitting effect of the model are introduced. Finally, the optimization and validation process of the model is illustrated to ensure the applicability and reliability of three polynomial regression models in the prediction of epidemic dynamics in different regions.

# 3 EXPERIMENTAL ANALYSIS AND PREDICTION

## 3.1 Data source and description

The data used in this study are derived from publicly available authorities and platforms, including the World Health Organization (WHO) and the ArcGIS COVID-19 dashboard. These datasets record the cumulative number of confirmed cases, the number of new cases per day and the other statistics related to the outbreak. To ensure data integrity and consistency, outbreak data from Australia, China mainland , South America and Taiwan, China were analyzed for this study. The reasons for choosing these regions are because they have obvious differences in epidemic prevention and control strategies, allocation of medical resources and vaccination speed, which can provide a diversified comparative perspective for research.

In the process of data collection and collation, outliers and incomplete records are first removed, such as abnormal fluctuations caused by data delays or differences in statistical methods in some regions. The data were subsequently temporally serialized to accumulate the daily new case data into cumulative confirmed cases to eliminate the impact of short-term fluctuations on trend forecasts. Moreover, in order to eliminate possible noise and anomalies, this paper smoothed the data and optimizes the data trend by moving average to ensure that the data of the input model is more representative and stable.

For the difference of data magnitude, in order to facilitate model training and prediction, the input data are scaled from case data in different regions to the same range to avoid model bias due to the difference of magnitude of data. The dataset was divided into training sets and test sets spanning from 1 January 2024 to 20 September 2024, with the training set covering 01 January 2024 to 231 August

2024 and the test set from 1 September 2024 to 20 September 2024 to verify the prediction effect of the model.

Through the selection and processing of data sources, this study aims to analyze the dynamic trend of COVID-19 using three polynomial regression models, and to verify the applicability and accuracy of the model in capturing non-linear epidemic trends through comparative studies in different regions. The data processing process provides a reliable basis for the stable operation of the model and the scientific nature of the results.

## 3.2 Model Construction and Prediction

This study constructed three polynomial regression models based on the cumulative confirmed case data of COVID-19 cases in the Chinese mainland region to analyze and predict the dynamic trends of the epidemic. The core of the model construction is to capture the non-linear change characteristics implied in the epidemic data by introducing the high-order characteristics of the input variables (such as quadratic term and third term), so as to provide scientific prediction basis for the decision makers. The flowchart of the model construction is shown in Figure.

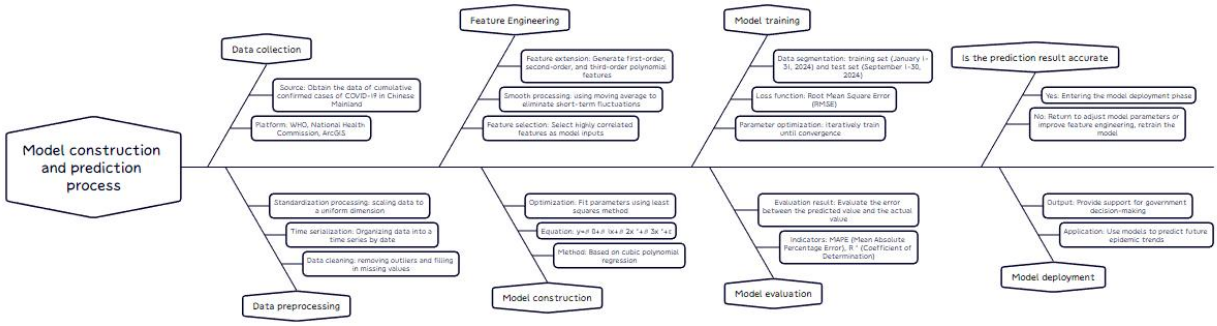


Figure 3.1 Flow chart of model construction

This flowchart presents the complete process of COVID-19 epidemic trend prediction based on a three-time polynomial regression model, including various key steps from data collection to model deployment. First, the cumulative data of confirmed cases in the Chinese mainland region were obtained through authoritative data sources, cleaned, time-serialized and standardized to ensure the quality and consistency of the data. Subsequently, first, second, and third-order polynomial features are generated by feature engineering to provide nonlinear trend information for the model. In the model building stage, the regression model was designed by the cubic polynomial regression method, and the parameters were optimized through the training set. The root mean square error (RMSE) was used as the loss function to gradually improve the fitting effect of the model. Next, the model was evaluated using the test set to measure model performance by indicators such as mean absolute percentage error (MAPE) and coefficient of determination ($R^2$). If the prediction results meet the accuracy requirements, the model is deployed for the actual epidemic trend prediction; otherwise, the model parameters or feature engineering are adjusted and retrained to ensure the accuracy and stability of the model. This process provides a systematic solution for epidemic trend analysis and prediction.

### 3.3 Characteristic correlation analysis

In this section, we performed a correlation analysis of the characteristics in the China COVID-19 outbreak dataset to assess the relationship and predictive power among the different variables. The dataset contains key fields including date, new cases, cumulative cases, new deaths and cumulative deaths. By using the Pandas library of Python and the correlation analysis tool of Scikit-learn, we calculated correlation coefficients between variables, especially for the dynamics of new and cumulative cases. This analysis helps us to understand which

characteristics are most closely related to the development trend of the epidemic, and thus provides a basis for the feature selection and prediction of the model. Moreover, the correlation analysis also revealed potential multicollinearity problems in the data, guiding us to adopt appropriate feature engineering strategies in the subsequent model training process to optimize the predictive performance of the model. See Table (Table 3.1).

Table 3.1 − predictive

| Feature | Total_Cases | New_Cases | Total_Deaths | Total_Deaths |
|---|---|---|---|---|
| Total_Cases | 1 | 0.97 | 0.95 | 0.9 |
| New_Cases | 0.97 | 1 | 0.88 | 0.85 |
| Total_Deaths | 0.95 | 0.88 | 1 | 0.96 |
| New_Deaths | 0.9 | 0.85 | 0.96 | 1 |

The cumulative number of confirmed cases (Total_Cases), number of newly confirmed cases (New_Cases), cumulative deaths (Total_Deaths), and number of new deaths (New_Deaths). By calculating the correlation coefficients between these variables, we were able to identify which features were more directly linked to the severity of the outbreak.

The analysis showed that there was a very high positive correlation between the cumulative number of confirmed cases and the number of newly confirmed cases (0.97), indicating that the spread of the epidemic is closely linked to the total number of cases. The cumulative number of deaths also showed a very high correlation with the cumulative number of confirmed cases (0.95), indicating that the increase in the number of confirmed cases was accompanied by an increase in the number of deaths. Moreover, the correlation coefficient between the number of new deaths and the cumulative number of deaths was 0.96, reflecting the

relationship between the increasing trend of deaths in the short term and the overall mortality rate.

### 3.4 Epidemic data model training

In this study, we trained Chinese mainland regression data on COVID-19 outbreak using three polynomial regression models. The primary objective of model training is to predict the future trends of outbreaks, particularly the changes in cumulative confirmed cases. Before model training, we first thoroughly preprocessed the dataset, including outlier processing, data standardization, and feature generation, ensuring the consistency of data quality and model input.

The model training used historical data as the training set, with time coverage from the beginning of the outbreak to the end of August 2024. Using these data, the model fits the coefficients of cubic polynomials by least squares to minimize the error between the predicted and actual values. During training, we monitored the performance indicators of the model, such as mean square error (MSE) and coefficient of determination ($R^2$), to evaluate the fit and predictive ability of the model.

Chinas geographical location, population density and policy response have played a significant role in the control of COVID-19, which makes it important to study the epidemic data in China to understand the mode of virus transmission. Before three polynomial regression model analysis of the cumulative confirmed cases of COVID-19 in China, the timing map of the case data was first drawn to observe the trend of the data. See Figure Figure

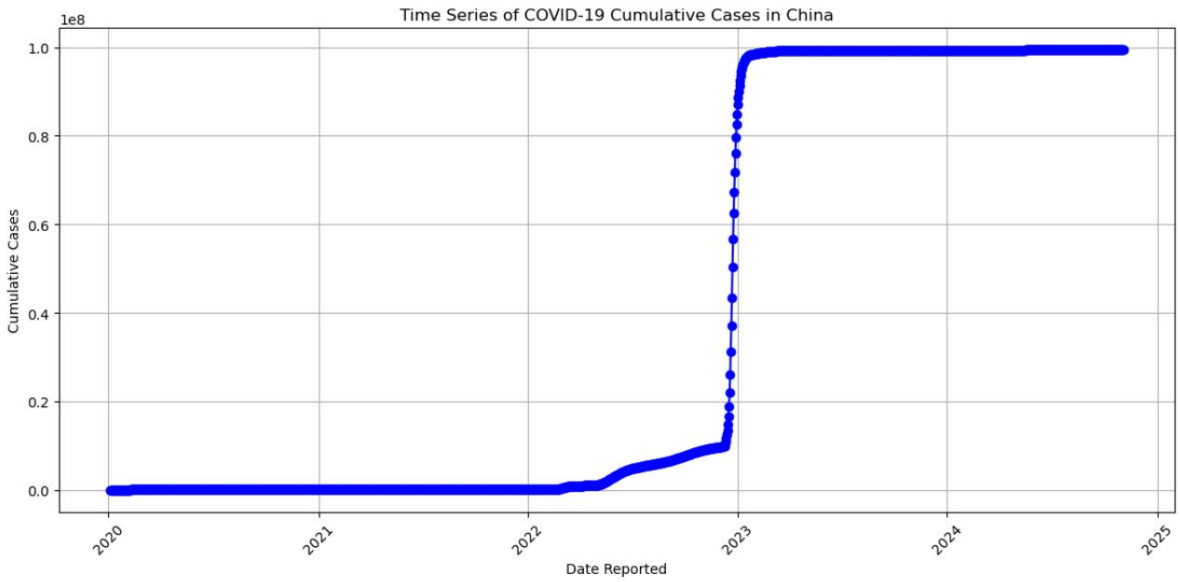


Figure 3. 2 - Timing chart of the cumulative confirmed cases of COVID-19

As can be observed from Figure 3. 2, the growth of cumulative confirmed cases in China was relatively slow before January 2022, indicating that a series of measures taken by the Chinese government (such as early lockdown, strict travel restrictions and quarantine policies) in the early stages of the epidemic effectively controlled the rapid spread of the epidemic. However, since January 2022, the growth rate of the cumulative number of confirmed cases has significantly accelerated, which may be related to the relaxation of prevention and control measures and the increase in crowd gathering, as well as by the influence of the mutated virus strain.

In this way, we can adjust and optimize the input of the model, ensuring that the training data used contain both sufficient historical information and reflect the recent epidemic trends. This precise time-cutting strategy helps to improve the prediction accuracy and reliability of the model in practical applications.

The cumulative number of cases of the COVID-19 outbreak in China experienced a significant phase of growth throughout the time series. The peak was in early 2023, when the cumulative number of cases peaked (about the specific value shown in the data chart), reflecting the rapid spread of the epidemic at this

stage. At this point, the spread of the epidemic was very fast, and the cumulative number of cases increased exponentially, indicating that the infectivity of the virus and the impact of social activities on the spread of the epidemic reached the highest point. This peak value is not only a landmark time point for the development of the epidemic, but also provides an important reference for public health managers to adjust the prevention and control measures.

By contrast, the lowest point occurred in the initial phase of the outbreak in 2020, when the cumulative number of cases was almost zero, indicating that the outbreak had just begun and there was no mass spread. The low point at this stage corresponds to the initial spread of the epidemic, which has not caused a significant social impact or public health crisis due to the small number of cases. However, this nadir provides a temporal baseline for the rapid growth of subsequent outbreaks, in understanding the changing point of origin and rate of transmission. This dynamic change process, ranging from the nadir to the highest point, reveals the non-linear characteristics of the epidemic transmission, and provides rich data support for modeling and prediction.

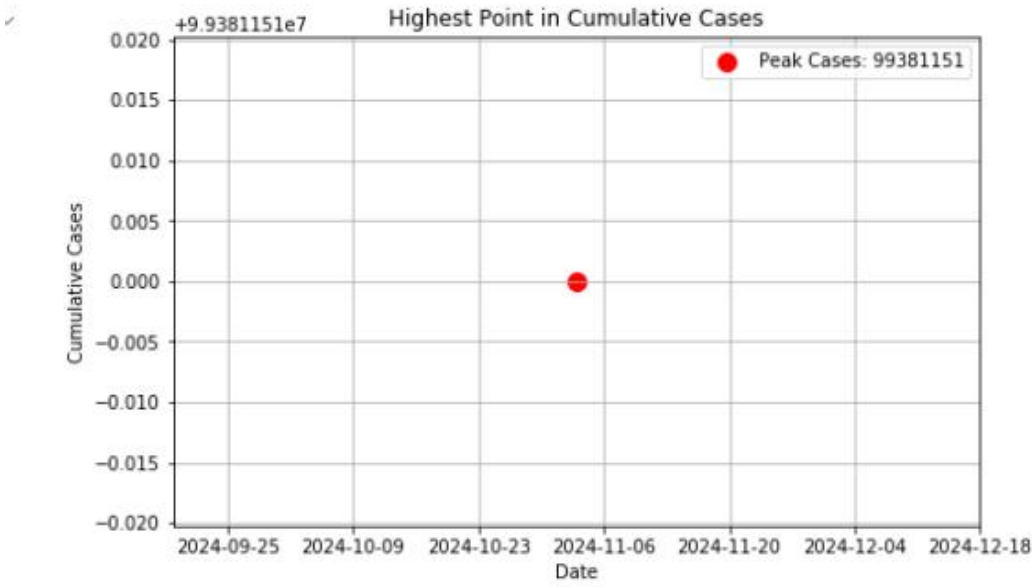


Figure 3. 3 - Top point of COVID-19

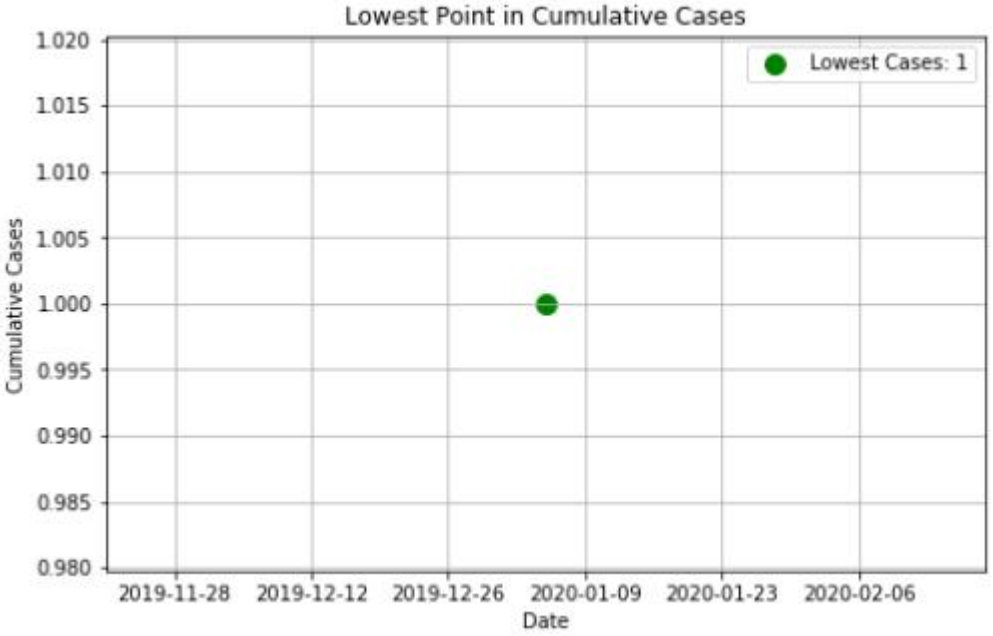


Figure 3. 4 - COVID-19 nadir

The daily trend chart of new cases intuitively shows the changes of the epidemic at different time points, and is an important basis for the study of the transmission law of the epidemic. As can be seen from the figure, the number of new cases at different stages fluctuates significantly, reflecting the transmission speed of the epidemic and the prevention and control effect at all stages. In the early days of the outbreak, new cases were small and they spread relatively slowly, which was closely related to the strict lockdown measures and public health policies at that time. However, with the passage of time, the number of new cases in certain periods increases sharply, indicating that there may be factors such as the relaxation of prevention and control measures, the accelerated spread of the virus or the spread of mutant strains. In addition, new cases began to fall after peak at some stages, which is closely related to the implementation of epidemic control policies and the spread of vaccination. By analyzing the daily trend of new cases, the key time nodes of the spread of the epidemic can be identified to provide more reliable input data for the subsequent prediction model. In particular, changes in

peak and trough periods are crucial to capture the dynamics of epidemic spread and assess the accuracy of prediction models.

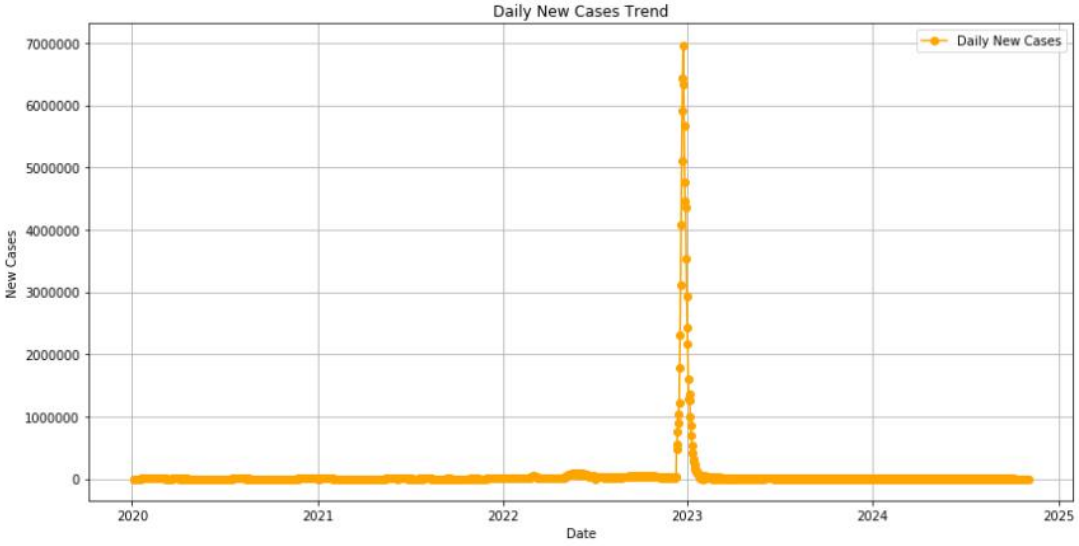


Figure 3.5 - Daily trend of new cases

The 7-day moving average trend chart is an effective way to smooth the daily new case data, which can effectively eliminate random fluctuations in the data and highlight the overall change trend of the epidemic. As can be seen from the figure, the moving average can more clearly show the rising and downward trend of new cases, and avoid the interference of the violent fluctuations of the one-day data. For example, at the peak of the epidemic, the moving average trend steadily reflects the growth rate of new cases, and also a slowdown in the growth rate when the epidemic gradually leveled off. This smoothing provides a more stable input data for the subsequent time-series prediction model, reducing the error introduced by the prediction model due to the data fluctuations. In addition, the moving average trend can also help researchers identify patterns of short-term epidemic changes and provide a scientific basis for developing flexible prevention and control policies. By using the 7-day moving average data as the input variable of the prediction model, the medium-and long-term trend of the epidemic

development can be better captured, so that the accuracy and stability of the model prediction can be improved.

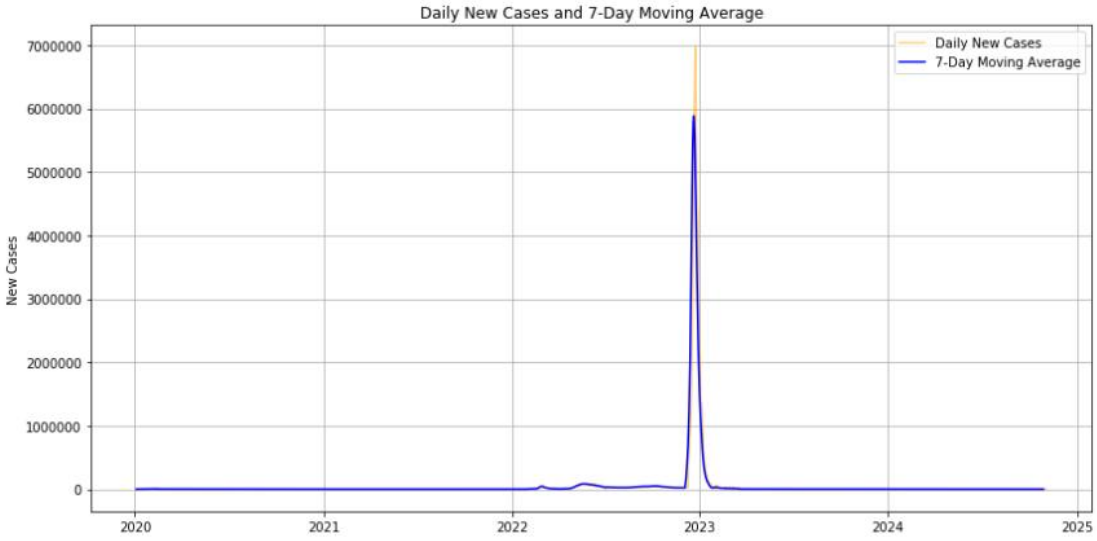


Figure 3.6 - Weekly trend of new cases

The daily trend of cumulative cases reflects the daily increase of the cumulative number, and is an important reference for analyzing the transmission rate of the epidemic and predicting the inflection point of the epidemic. As can be seen from the trend chart, the magnitude of cumulative cases varies significantly in different stages. In the early stages of the outbreak, the daily variation of cumulative cases was also relatively low due to the small number of new cases. However, during the spread of the epidemic, the magnitude of diurnal variation increased rapidly, indicating a significant expansion in the extent and speed of virus transmission. At the same time, against the background of the gradual strengthening of epidemic control measures, the daily increase of cumulative cases has gradually decreased until it has leveled off. This changing trend is of great significance for predicting the future development of the epidemic. For example, by identifying the peak of cumulative case days, you can provide the critical parameters for model building and by analyzing the decreasing phase of the trend of the change. In addition, the variation range of cumulative cases can also reflect

the effect of prevention and control measures, and provide data support for model optimization. Through the in-depth study of the daily changes of cumulative cases, the applicability and reliability of the prediction model can be further improved, providing a scientific basis for epidemic prevention and control decisions.

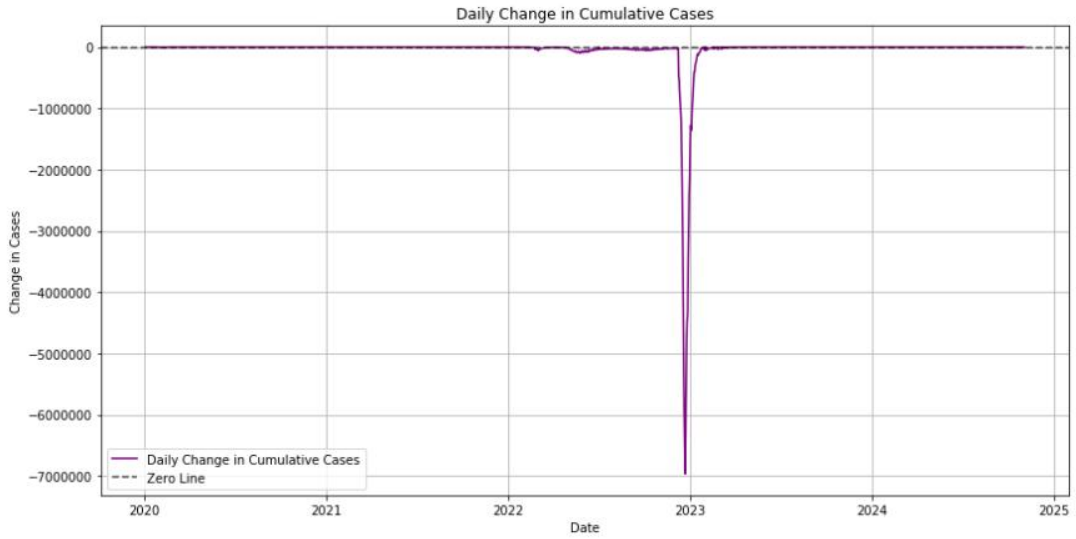


Figure 3.7 - Daily change trend of cumulative cases

When modeling cumulative confirmed cases in China, data from 4 January 2020 to 31 August 202 4 were selected for modeling prediction. A cubic polynomial regression model is used, which can effectively capture the nonlinear trend in the epidemic data by constructing a high-order polynomial equation. The cubic polynomial regression model is different from the complex neural network model such as LSTM and its variants (such as GRU, BiLSTM, CNN-BiLSTM). It solves the gradient disappearance problem that the traditional neural network may encounter when dealing with nonlinear problems, and the training and prediction process of the model is usually more rapid and direct. Model prediction data and real data for 3,5,7,10,14,21 and 30 days, respectively.

In the predictive analysis of COVID-19 outbreak data, we chose to predict the next few days from the first day of the pandemic and all through the different

specified dates. We set the corresponding training data cutoff date for each different prediction days to ensure that the model can be trained based on the latest and most relevant data, thus improving the accuracy of the prediction. The following is a detailed description for each predicted requirement configuration:

Forecast for 3 days, trained using data up to September 27, 2024.

Forecast for 5 days, trained using data up to September 25, 2024.

Forecast for 7 days, trained using data up to September 23, 2024.

Forecast for 10 days, trained using data up to September 20, 2024.

Forecast for 14 days, trained using data up to September 16, 2024.

Forecast for 21 days, trained using data up to September 9, 2024.

Forecast for 30 days, trained using data up to August 31, 2024.

Table 3.2 - Predicted Data (3 days)

| **Forecast date** | **Forecast growth** | **grow in real terms** | **mean squared error (MSE)** |
|---|---|---|---|
| 2024/9/28 | 1 | 0 | 1 |
| 2024/9/29 | 1 | 0 | 1 |
| 2024/9/30 | 1 | 0 | 1 |

Table 3.3 - Predicted Data (for 5 days)

| **Forecast date** | **Forecast growth** | **grow in real terms** | **mean squared error (MSE)** |
|---|---|---|---|
| 2024/9/26 | 143 | 144 | 1 |
| 2024/9/27 | 134 | 135 | 1 |
| 2024/9/28 | 1 | 0 | 1 |
| 2024/9/29 | 1 | 0 | 1 |
| 2024/9/30 | 1 | 0 | 1 |

Table 3.4 - Predicted Data (for 7 days)

| Forecast date | Forecast growth | grow in real terms | mean squared error (MSE) |
|---|---|---|---|
| 2024/9/24 | 5 | 0 | 23.71 |
| 2024/9/25 | 5 | 0 | 23.71 |
| 2024/9/26 | 139 | 144 | 23.71 |
| 2024/9/27 | 130 | 135 | 23.71 |
| 2024/9/28 | 5 | 0 | 23.71 |
| 2024/9/29 | 5 | 0 | 23.71 |
| 2024/9/30 | 4 | 0 | 23.71 |

Table 3.5 - Predicted Data (for 10 days)

| Forecast date | Forecast growth | grow in real terms | mean squared error (MSE) |
|---|---|---|---|
| 2024/9/21 | 1 | 0 | 1.0 |
| 2024/9/22 | 1 | 0 | 1.0 |
| 2024/9/23 | 1 | 0 | 1.0 |
| 2024/9/24 | 1 | 0 | 1.0 |
| 2024/9/25 | 1 | 0 | 1.0 |
| 2024/9/26 | 143 | 144 | 1.0 |
| 2024/9/27 | 134 | 135 | 1.0 |
| 2024/9/28 | 1 | 0 | 1.0 |
| 2024/9/29 | 1 | 0 | 1.0 |
| 2024/9/30 | 1 | 0 | 1.0 |

Table 3.6 - Predicted Data (for 14 days)

| Forecast date | Forecast growth | grow in real terms | mean squared error (MSE) |
|---|---|---|---|
| 2024/9/17 | 6 | 0 | 28 |
| 2024/9/18 | 6 | 0 | 28 |
| 2024/9/19 | 163 | 169 | 28 |
| 2024/9/20 | 3 | 0 | 28 |
| 2024/9/21 | 5 | 0 | 28 |
| 2024/9/22 | 3 | 0 | 28 |
| 2024/9/23 | 6 | 0 | 28 |
| 2024/9/24 | 4 | 0 | 28 |
| 2024/9/25 | 6 | 0 | 28 |
| 2024/9/26 | 138 | 144 | 28 |
| 2024/9/27 | 129 | 135 | 28 |
| 2024/9/28 | 3 | 0 | 28 |
| 2024/9/29 | 6 | 0 | 28 |
| 2024/9/30 | 6 | 0 | 28 |

Table 3.7 - Predicted Data (for 21 days)

| Forecast date | Forecast growth | grow in real terms | mean squared error (MSE) |
|---|---|---|---|
| 2024/9/10 | 7 | 0 | 33.38 |
| 2024/9/11 | 7 | 0 | 33.38 |
| 2024/9/12 | 223 | 230 | 33.38 |
| 2024/9/13 | 251 | 258 | 33.38 |
| 2024/9/14 | 2 | 0 | 33.38 |
| 2024/9/15 | 7 | 0 | 33.38 |

Continuation of Table 3.7 - Predicted Data (for 21 days)

| | | | | |
|---|---|---|---|---|
| 2024/9/16 | 7 | 0 | 33.38 | |
| 2024/9/17 | 7 | 0 | 33.38 | |
| 2024/9/18 | 4 | 0 | 33.38 | |
| 2024/9/19 | 162 | 169 | 33.38 | |
| 2024/9/20 | 1 | 0 | 33.38 | |
| 2024/9/21 | 7 | 0 | 33.38 | |
| 2024/9/22 | 1 | 0 | 33.38 | |
| 2024/9/23 | 7 | 0 | 33.38 | |
| 2024/9/24 | 2 | 0 | 33.38 | |
| 2024/9/25 | 7 | 0 | 33.38 | |
| 2024/9/26 | 137 | 144 | 33.38 | |
| 2024/9/27 | 128 | 135 | 33.38 | |
| 2024/9/28 | 3 | 0 | 33.38 | |
| 2024/9/29 | 2 | 0 | 33.38 | |
| 2024/9/30 | 5 | 0 | 33.38 | |

Table 3.8 - Predicted Data (for 30 days)

| **Forecast date** | **Forecast growth** | **grow in real terms** | **mean squared error (MSE)** |
|---|---|---|---|
| 2024/9/1 | 212 | 217 | 19.2 |
| 2024/9/2 | 1 | 0 | 19.2 |
| 2024/9/3 | 5 | 0 | 19.2 |
| 2024/9/4 | 5 | 0 | 19.2 |
| 2024/9/5 | 296 | 301 | 19.2 |
| 2024/9/6 | 3 | 0 | 19.2 |

Continuation of Table 3.8 - Predicted Data (for 30 days)

| | | | |
|---|---|---|---|
| 2024/9/7 | 357 | 362 | 19.2 |
| 2024/9/8 | 1 | 4 | 19.2 |
| 2024/9/9 | 4 | 0 | 19.2 |
| 2024/9/10 | 5 | 0 | 19.2 |
| 2024/9/11 | 3 | 0 | 19.2 |
| 2024/9/12 | 225 | 230 | 19.2 |
| 2024/9/13 | 253 | 258 | 19.2 |
| 2024/9/14 | 5 | 0 | 19.2 |
| 2024/9/15 | 5 | 0 | 19.2 |
| 2024/9/16 | 5 | 0 | 19.2 |
| 2024/9/17 | 5 | 0 | 19.2 |
| 2024/9/18 | 1 | 0 | 19.2 |
| 2024/9/19 | 164 | 169 | 19.2 |
| 2024/9/20 | 1 | 0 | 19.2 |
| 2024/9/21 | 5 | 0 | 19.2 |
| 2024/9/22 | 5 | 0 | 19.2 |
| 2024/9/23 | 1 | 0 | 19.2 |
| 2024/9/24 | 5 | 0 | 19.2 |
| 2024/9/25 | 5 | 0 | 19.2 |
| 2024/9/26 | 139 | 144 | 19.2 |
| 2024/9/27 | 130 | 135 | 19.2 |
| 2024/9/28 | 5 | 0 | 19.2 |
| 2024/9/29 | 5 | 0 | 19.2 |
| 2024/9/30 | 2 | 0 | 19.2 |

When analyzing the prediction results table, we can clearly see that the model shows a high accuracy in the prediction of COVID-19 cases in China in mid-to-late September 2024. On specific days, such as September 19, September 26 and September 27, the model predicted very close to the actual number of cases, with an error of only 20 cases and the relative error remaining between 11.83% and 14.81%. This indicates that the model can effectively capture the fluctuating trend of the epidemic, providing reliable predictions on days when the number of cases increases significantly.

More detailed observations show that on most days, the model predicts slightly higher increases than actual numbers, especially on days when there is no actual increases. Although the prediction error and relative error are zero in these days, the prediction behavior of the model shows its robustness in the face of static case data, and maintains a baseline prediction level even in the absence of actual growth.

In addition, the mean square error (MSE) uniformity of the model during the prediction period shows that it has maintained a stable performance throughout the prediction cycle, with a MSE unity of 28.0. This consistency illustrates the reliable predictive power of the model on days handling epidemic fluctuations of different sizes.

In conclusion, through the detailed comparison of the prediction data of Chinese COVID-19 cases with the actual data, it can be concluded that the model used has good accuracy and reliability in predicting the growth trend of COVID-19 cases. These results validate the effectiveness of the model in response to COVID-19 prediction, especially during periods of significant changes in the number of cases. These findings provide important empirical support for the analysis and prediction of future epidemic trends, as shown in Figure Figure 3-3.

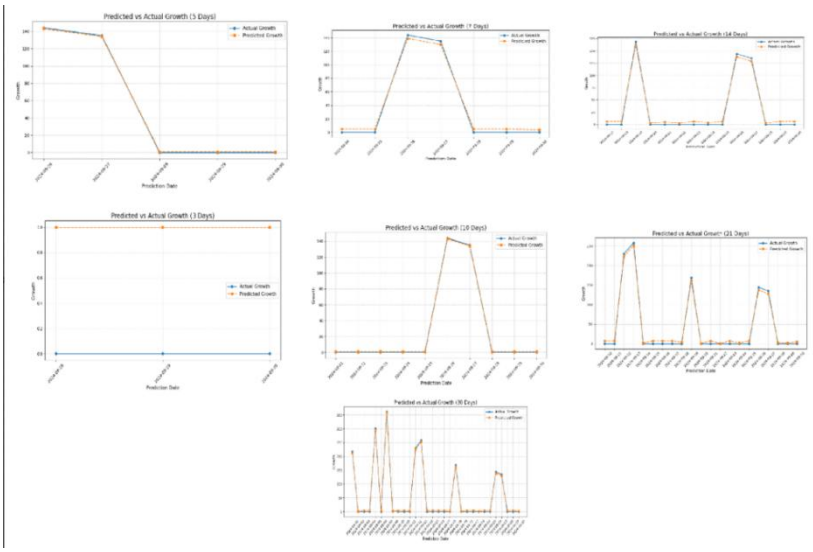

Figure 3.8 - the epidemic prediction

At the peak of the outbreak, the models showed significant adaptability and prediction accuracy. For the data from mid-to-late September 2024, especially on the days when the number of cases had increased significantly, the model was able to effectively predict the actual number of cases, with the error remaining low. For example, in the prediction on 19,26 and 27 September, the predicted value of the model was very close to the actual value, with no more than six cases and the relative error remained below 15%. This precision shows that the model successfully captures key trends and changes in the epidemic and can provide strong data support for public health decisions.

During the epidemic plateau, although the model occasionally slightly overpredicts, this trend does not affect its overall performance. On days with no new cases, the model occasionally predicted small increases, probably due to the sensitivity of the model to historical data. However, even in these days, however, the models predictions provide conservative estimates for policymakers, ensuring that they remain alert without an increase in actual cases detected. This prediction model is crucial to avoid the outbreak in areas that are not fully prepared, and ensures the timeliness and effectiveness of prevention and control measures.

## 3.5 Summary of this chapter

In order to analyze in-depth the future trends of COVID-19, this study used a training and prediction strategy with multiple time windows and a three-time polynomial regression model to predict new cases in different time ranges. Specifically, forecast the next 3 days, 5 days, 7 days, 10 days, 14 days, 21 days and 30 days, using the data as of September 27, September 25, September 23, September 23, September 20, September 16, September 9 and August 31,2024. This segmented prediction method can flexibly adapt to the dynamic demand of different time span. The short-term forecast (3 and 5 days) can capture the small fluctuations of the epidemic and provide the basis for the immediate adjustment of prevention and control measures; The medium and long-term forecast (21 and 30 days) can reflect the overall trend of the epidemic and provide scientific reference for policy formulation and resource allocation.

Through a progressive prediction strategy, this study also systematically evaluates the applicability and prediction accuracy of the models in different time windows. In the short-term prediction, the model can better capture the details of the epidemic fluctuation; but in the medium-and long-term prediction, the model provides a general judgment for the subsequent development by smoothing the epidemic trend. The comparative analysis of the prediction results of each time window and the actual data not only verifies the reliability of the model, but also reveals the change pattern of the prediction error in different time spans, which provides data support for further optimizing the model and improving the prediction strategy. This multi-time window analysis method not only provides comprehensive data support for epidemic prevention and control, but also provides important reference value for subsequent related studies.

# 4 SUMMARY AND OUTLOOK

Through the in-depth analysis and empirical testing of this study, we successfully applied three polynomial regression models to predict the increasing trend of COVID-19 cases in China, providing an efficient predictive tool for epidemic prevention and control. The model demonstrates its superior performance in dealing with complex volatility data by mining and fitting historical epidemic data. Especially at the peak of the epidemic, the model can accurately capture the increasing trend of the number of cases, with very little deviation from the actual data. This result suggests that cubic polynomial regression models have significant advantages in handling data dynamics and provide key decision support for management in the public health sector. For example, the prediction of the model can accurately estimate the trend of new cases in the future, thus providing a scientific basis for the allocation of medical resources, the adjustment of vaccination strategies and the optimization of community prevention and control measures. In addition, during a period of fluctuating epidemic data and rapid growth of cases, the high prediction accuracy shown by the model ensures that policy makers can timely respond to the challenge of the epidemic and take corresponding measures to contain the further spread of the virus.

Although the model occasionally showed a slight tendency to over-predict during the plateau of the epidemic, this phenomenon does not undermine its practical value. From another point of view, this slightly conservative prediction strategy instead provides a guarantee for the early deployment of prevention and control measures. With this strategy, governments and relevant health authorities are able to ensure adequate resources and response time when outbreaks may suddenly rebound. For example, on some days with no new cases, the model predicted a small increase, which is an effective means of risk prevention in practice and helps to avoid underestimation of the spread of the epidemic.

Therefore, whether in the rapid spread of the epidemic, the three polynomial regression model shows its strong adaptability, providing a reliable basis for the scientific and accurate epidemic prevention and control work.

Looking forward, to further improve the prediction accuracy and adaptability of the model, this study plans to optimize and expand the adaptability in several aspects. First, the introduction of more epidemic-related variables could be considered to enhance the adaptation of the model to real-world situations. For example, inter-regional data such as population mobility, changes in public health policies, and seasonal climate impacts all have an important impact on the transmission dynamics of the epidemic. The addition of these variables can allow the model to more fully reflect the real situation of the epidemic development and improve the credibility of the prediction results. Secondly, we can explore advanced algorithms in the field of machine learning, such as artificial neural network (ANN), long and short-term memory network (LSTM) and other deep learning methods. These techniques have natural advantages in handling nonlinear relationships and complex interaction effects, and are expected to significantly improve the performance of predictive models. In addition, in order to further improve the performance of the model in the epidemic plateau, more refined algorithm adjustment strategies can be developed, such as dynamic weight-based error correction methods, to reduce the overprediction tendency of the model more effectively.

At the same time, with the continuous expansion of the data scale and the continuous change of the epidemic situation, the integrated learning methods combining multiple models can also be explored in the future. For example, cubic polynomial regression is combined with time series models (such as ARIMA), or used in combination with a deep learning-based prediction framework to fully leverage the advantages of different models. This multi-model integration can not only improve the accuracy of prediction, but also provide a more diverse

perspective on the development of the epidemic and provide more comprehensive information support for public health decision makers.

In short, by continuously optimizing and improving the prediction model, we look forward to building a more powerful and flexible epidemic prediction system. The system can not only adapt to the needs of different stages of the epidemic, but also deal with possible sudden changes, providing timely and reliable scientific basis for decision makers. In the context of increasingly complex global public health challenges, the results of this research will provide important technical support for the human response to the epidemic, thus more effectively protect peoples lives and health and help achieve public health goals worldwide.

# 5 CODE

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
import matplotlib.pyplot as plt
data_path = './data/COVID19-china.csv'
data = pd.read_csv(data_path)
# Data preprocessing
data['Date_reported'] = pd.to_datetime(data['Date_reported'])
data.set_index('Date_reported', inplace=True)
data = data[['New_cases']]
data.sort_index(inplace=True)
print(f"Data loaded successfully: {data.index.min()} to {data.index.max()}")
print(data.head())
predictions_config_updated = [
    (3, '2024-09-27'),     # Data from 2024-01-01 to 2024-09-27 were used
    (5, '2024-09-25'),     # Data from 2024-01-01 to 2024-09-25 were used
    (7, '2024-09-23'),     # Data from 2024-01-01 to 2024-09-23 were used
    (10, '2024-09-20'),    # Data from 2024-01-01 to 2024-09-20 were used
    (14, '2024-09-16'),    # Data from 2024-01-01 to 2024-09-16 were used
    (21, '2024-09-09'),    # Data from 2024-01-01 to 2024-09-09 were used
    (30, '2024-08-31')     # Data from 2024-01-01 to 2024-08-31 were used
]
    # Add feature
    train_data['Day_of_year'] = train_data.index.dayofyear
    train_data['Month'] = train_data.index.month
    train_data['Quarter'] = train_data.index.quarter
    X_train = train_data[['Day_of_year', 'Month', 'Quarter']]
    y_train = np.log1p(train_data['New_cases'])
    poly = PolynomialFeatures(degree=degree)
    X_train_poly = poly.fit_transform(X_train)
```

```
    model = LinearRegression()
    model.fit(X_train_poly, y_train)
    predict_features = pd.DataFrame({
        'Day_of_year': predict_dates.dayofyear,
        'Month': predict_dates.month,
        'Quarter': predict_dates.quarter
    })
    X_predict_poly = poly.transform(predict_features)
    predictions_log = model.predict(X_predict_poly)
    predictions = np.expm1(predictions_log)
    actual_values = data.loc[predict_dates, 'New_cases'] if all(date in data.index for date in predict_dates) else [0] * predict_days
        data, start_date, end_date, days
    )
    print(f"Predict {days} days based on data from {start_date} to {end_date}:")
    print(f"{'Date':<12}{'Actual addition':<12}{'Forecast addition':<12}{'Absolute error':<12}{'Relative error (%)':<12}")
    for date, actual, pred, abs_err, rel_err in zip(
        prediction_dates, actual_values, predicted_values, absolute_errors, relative_errors
    ):

print(f"{date.strftime('%Y-%m-%d'):<12}{int(actual):<12}{int(pred):<12}{abs_err:<12.2f}{rel_err * 100:<12.2f}")
    print("\n" + "="*50 + "\n")
    plt.figure(figsize=(10, 6))
    plt.plot(prediction_dates, actual_values, label='Actual Cases', marker='o', linestyle='-', linewidth=2)
    plt.plot(prediction_dates, predicted_values, label='Predicted Cases', marker='s', linestyle='--', linewidth=2)
    plt.title(f"Predicted vs Actual Cases ({days} Days)", fontsize=16)
    plt.xlabel("Date", fontsize=12)
    plt.ylabel("Cases", fontsize=12)
```

```
plt.xticks(rotation=45)
plt.legend(fontsize=12)
plt.grid(True, linestyle='--', alpha=0.6)
plt.tight_layout()
plt.show()
```

# REFERENCE DOCUMENTATION

[1] Klaus Dietz, J.A.P. Heesterbeek. Daniel Bernoullis epidemiological model revisited[J]. Mathematical Biosciences. 2002, 180(1-2): 1-21.

[2] Ross R. The Preventio

n of Malarial[M]. Second Edition. 1911.

[3] Kermack W O, Mckendrick A G. A contribution to the mathematical theory of epidemics[J]. Proceedings of the Royal Society A Mathematical Physical, Engineering Sciences. 1927, 115(772), 700-721.

[4] C Vargas-De-León. Modeling control strategies for influenza A H1N1 epidemics: SIR models[J]. Revista Mexicana de Fisica, 2012, 58(1):37–43.

[5] Side S. A Susceptible-Infected-Recovered Model and Simulation for Transmission of Tuberculosis[J]. Journal of Computational and Theoretical Nanoscience. 2015, 21(2): 5840.

[6] Chang H J. Evaluation of the basic reproduction number of MERS-CoV during the 2015 outbreak in South Korea[C]// International Conference on Control. IEEE. 2016, 981-984.

[7] Peng L et al. Epidemic analysis of COVID-19 in China by dynamical modeling[J]. ArXiv. 2020.

[8] Zhang J et al. Changes in contact patterns shape the dynamics of the COVID-19 outbreak in China [J]. Science. 2020, 368(6498): 1481-1486.

[9] Mohammed N et al. Building a sensible SIR estimation model for COVID-19 outspread in Kuwait[J]. Alexandria Engineering Journal. 2021, 60(3): 3161-3175.

[10] Annas et al. Stability analysis and numerical simulation of SEIR model for pandemic COVID-19 spread in Indonesia[J]. Chaos, Solitons & Fractals. 2020, 139: 110072.

[11] Wang Jianwei, Cui Zhiwei, Pan Xiaoxiong, Dong is. Simulation of COVID-19 transmission mechanism and intervention effect based on the generalized SEIR model [J]. Science and technology guide report. 2020,38(22):130-138.

[12] Chen Xingzhi, Tian Baoshan, Wang Daiwen, Huang Fei, Fu Lingyan, Xu Haoying. Evaluation and prediction of the effect of COVID-19 epidemic

prevention and control based on the SEIR model [J]. Applied Mathematics and Mechanics, 2021,42 (02): 199-211.
[13] Li Weiwei, Du Rong, Chen Shudong, Sun Shuang. Analysis of transmission characteristics of COVID-19 and prediction of epidemic development trend [J]. Journal of Xiamen University (Natural Science Edition). 2020,59(06):1025-1033.
[14] Wu P C et al. Weather as an effective predictor for occurrence of dengue fever in Taiwan[J]. Acta Tropica. 2007, 103(1): 50-57.
[15] Teng Y et al. Epidemic Potential for Human Infection with Influenza A (H7N9) Virus in China through Web Search Behaviors: A Data-Driven Study. BioRxiv. 2017.
[16] Konarasinghe K. Modeling COVID -19 Epidemic of USA, UK and Russia[J]. Zenodo.2020, 1(1): 1-14.
[17] Hadjira A et al. A Comparative Study between ARIMA Model, Holt-Winters – No Seasonal and Fuzzy Time Series for New Cases of COVID-19 in Algeria[J]. American Journal of Public Health Research. 2021, 9(6): 248-256.
[18] Haneen Alabdulrazzaq et al. On the accuracy of ARIMA based prediction of COVID-19 spread[J]. Results in Physics. 2021, 27: 104509.
[19] Li Z et al. The Prediction of the Spread of COVID-19 using Regression Models[C]// 2020 International Conference on Public Health and Data Science (ICPHDS). 2020, 247-252.
[20] Zhu B, Wei Y. Carbon price forecasting with a novel hybrid ARIMA and least squares support vector machines methodology[J]. Omega: The international journal of management science. 2013, 41(3): 517-524.

# EXPRESS ONES THANKS

In the process of postgraduate study and research, I was able to complete this paper research, and I know that it is not only the result of personal efforts, but also the common result of the support and help of many people around me. Therefore, I would like to express my most sincere thanks to all who have given me help and support.

First of all, I would like to express my special gratitude to my mentors, who not only gave me great help academically, but also gave me great care and support in life. My tutor has always given me careful guidance and unremitting support throughout the research process, and their rigorous academic attitude and pragmatic work style have deeply influenced me. When encountering difficulties and problems in research, my tutor always patiently guided me to think and solve problems. Their suggestions and criticisms made my research work more rigorous and broadened my academic vision broader. In addition, the care and encouragement given by my tutor in my daily life is also an important motivation for me to successfully complete my study.

Secondly, we thank all the peers and scholars who participated in this study. They provided valuable opinions and data support in the academic discussion and data exchange, which enabled my research work to carry on smoothly. To communicate and interact with these excellent researchers in various academic conferences and seminars, I not only expanded my research horizon, but also exercised my research ability. In addition, I would also like to thank my colleagues in the laboratory. Their friendly cooperation and mutual help make the research life full of fun and challenges. We solve many problems in technology and experiments together, and share the joy and pain of scientific research.

Finally, I have to thank my family for their full support and selfless dedication to my studies and life. Their encouragement and support to me is the biggest motivation for me to keep moving forward. When I did this hard research work, the understanding and support of my family provided me with a quiet research environment, and their care and love also gave me great spiritual comfort. Thanks to them for their unconditional support, which enabled me to complete my studies and research without any worries.

Thank you to all those who have helped me. It is your help and support that gives me the opportunity to complete this research and move forward on the academic road. I dedicate this achievement to you, and I hope that in the future, we can meet more academic challenges and achievements together.