Frontiers in Economics and Management Volume 6 Issue 6, 2025
ISSN: 2692-7608 DOI: 10.6981/FEM.202506_6(6).0013

A Hybrid Forecasting Model Combining Autoregressive Integrated Moving Average and Gaussian Process Regression

Jiale He*

School of Economics, Yunnan University, Kunming 650500, China

*Correspondence should be addressed to Jiale He: 20241040074@stu.ynu.edu.cn

Abstract

The prediction of population mortality rates is fundamental to economic development, healthcare resource allocation, and the construction of pension systems. Based on the population mortality rate data of China from 1978 to 2022, this paper establishes an ARIMA-GPR hybrid model to forecast the population mortality rate. Initially, ARIMA models with different combinations of p and q are employed to capture the linear trends of the original time series. The five best-performing combinations are selected as input features for the Gaussian Process Regression (GPR) model, which then outputs the predicted values after further fitting. The results indicate that the ARIMA-GPR hybrid model significantly improves the forecasting accuracy compared to the standalone ARIMA model.

Keywords

ARIMA; Gaussian Process Regression; Hybrid Forecasting Model; Time Series Analysis; Machine Learning.

1. Introduction

As the aging population in China intensifies, changes in population mortality rates are receiving increasing attention. According to data released by the National Bureau of Statistics of China, by 2024, the population of elderly individuals aged 60 and above is expected to reach approximately 310 million, accounting for 22.0% of the total population. This proportion is anticipated to continue rising, which will have a profound impact on China's demographic structure and economic development. The Chinese government attaches great importance to this issue, having incorporated the response to population aging into the national strategy. The government has also adopted multi-dimensional measures such as optimizing fertility policies, improving the elderly care service system, and introducing delayed retirement policies, all aimed at mitigating the negative impact of increasing aging on China's development.However, as of now, population mortality rate, as a critical statistical indicator, not only reflects the health status of a country's population but also reveals the societal challenges posed by aging trends [1]. Changes in population mortality rates have far-reaching implications for economic development, healthcare resource allocation, and the construction of eldercare systems. Accurately forecasting changes in population mortality is not only beneficial for the government in formulating evidence-based population and public health policies but also helps relevant institutions proactively develop response strategies.In recent years, with rapid societal changes, the trend of population mortality has become increasingly complex. The multifaceted influences of economic development, urbanization, and other factors have caused population mortality rates to no longer follow a simple or uniform trend. This complexity presents challenges for traditional population mortality forecasting models such as the Lee-Carter model and ARIMA model, which are often inadequate in addressing complex, non-linear abrupt changes in data. These models often fail to account for the non-linear characteristics and

ISSN: 2692-7608 DOI: 10.6981/FEM.202506_6(6).0013

evolving trends within time series, leading to biased predictions.Therefore, establishing a scientifically sound and reliable forecasting model is of significant practical and theoretical importance. It plays a crucial role in enhancing our understanding of the underlying mechanisms governing population mortality trends and in improving the foresight and scientific rigor of policy-making, particularly in the context of a rapidly aging population.

In practical problems, we often face the challenge of modeling nonlinear and uncertain systems. Traditional linear regression models have significant limitations when handling nonlinear data. Linear models, such as Autoregressive Integrated Moving Average (ARIMA), are effective in addressing short-term dependencies in stationary time series data. However, they fall short in capturing non-stationary abrupt changes, representing nonlinear relationships in economic contexts, and quantifying the uncertainty in forecast results [2]. Meanwhile, Gaussian Process Regression (GPR), as a typical method of Bayesian nonparametric estimation, can handle nonlinear data effectively through the selection of kernel functions and parameter adjustments, leading to excellent prediction performance. Furthermore, GPR offers significant advantages in noise observation and uncertainty system modeling. It can derive the posterior distribution of the target prediction by applying a Gaussian prior distribution to the function space of nonparametric regression [3].

In current research, the ARIMA (Autoregressive Integrated Moving Average) model has been proven to provide good forecasting results across various fields of production and daily life. The GPR (Gaussian Process Regression) model, an important method in the field of machine learning regression in recent years, is based on rigorous statistical learning theory and has demonstrated significant superiority in handling certain nonlinear abrupt changes. Zhang et al. (2003) explored a hybrid model combining ARIMA with neural networks, specifically using Artificial Neural Networks (ANNs) to capture the nonlinear relationships hidden in the residuals of the ARIMA model's prediction fitting [4]. Ci Bicong et al. (2022) established an ARIMA-LSTM model, which produced better prediction results compared to single models [5]. Yao Jinhai et al. (2022) developed an SVM-ARIMA model to predict the residential consumption price index and achieved good results [6]. Zhang Fengming et al. (2025) optimized the Gaussian Process Regression model using a Gaussian filtering algorithm, significantly improving prediction accuracy [7]. From the extensive research conducted by various scholars, it is evident that a scientifically sound and reasonable model combination approach can significantly enhance model performance, achieving better results and improving prediction accuracy compared to standalone models.

This paper proposes the ARIMA-GPR combined model, incorporating the concept of model ensemble. The model first utilizes the ARIMA model and employs a grid search method to select different combinations of p and q within a given range for modeling. It outputs the predicted values for each combination, which are then used to form feature vectors for the Gaussian Process Regression (GPR) on the training set. The actual values in the training set are then set as the target vectors for training the GPR model. After determining the kernel function and relevant parameters, the ARIMA-GPR combined model is used to forecast the time series for a future period.

Theoretically, the ARIMA-GPR model combines the strengths of both Autoregressive Integrated Moving Average (ARIMA) and Gaussian Process Regression (GPR). On one hand, ARIMA is well-suited for handling linear, stationary time series data; on the other hand, GPR is effective in managing small sample sizes and complex, non-linear economic data, addressing the shortcomings of ARIMA in dealing with non-linear data, abrupt changes, and other related issues.

ISSN: 2692-7608 DOI: 10.6981/FEM.202506_6(6).0013

2. Materials and Methods

2.1. Data Source

The data for this study primarily comes from authoritative sources such as *the China Statistical Yearbook* (1999–2023) and *the China Population and Employment Yearbook* (2008–2017). Since the statistical data is not segmented by age groups, this paper analyzes and forecasts the overall population mortality rate. Based on previous research by scholars, it is recommended to select a data interval with the largest possible time span to ensure higher forecast robustness. Therefore, this study uses the population mortality rate data of China from 1978 to 2022 as the research subject, with a unit length of one year.

2.2. Data Preprocessing

In this study, the population mortality rate data from 1978 to 2022 was first subjected to outlier detection and correction to ensure the validity of the data. Additionally, since most machine learning and statistical models are sensitive to the scale of the data during prediction, data normalization, which maps the original data to a unified scale, can significantly improve the optimization process of model solving. Therefore, to ensure the reliability and efficiency of the data, this paper adopts the Z-score normalization method to process the data, transforming the original data into a standard normal distribution with a mean of 0 and a standard deviation of 1, thereby optimizing the model's convergence efficiency.

3. ARIMA and Gaussian Process Regression Models

3.1. ARIMA(p,d,q)

The ARIMA model is a widely used statistical method for time series analysis, consisting of three components: Autoregressive (AR), Integration (I), and Moving Average (MA) [8]. The core objective of the ARIMA model is to analyze the statistical characteristics of a time series, establish a model to capture linear trends, and forecast future developments.

In the ARIMA model, the autoregressive component reveals the dynamic relationship between the current observation and past observations. The integration component stabilizes the original time series by eliminating non-stationarity. The moving average component describes the relationship between the current observation and historical forecast errors, adjusting predictions based on lagged residuals.

By integrating these three components, ARIMA constructs a modeling framework parameterized by the autoregressive order p, the differencing order d, and the moving average order q.

$$Y_t = C + \sum_{i}^{p} \varphi_i Y_{(t-i)} + \sum_{j}^{q} \theta_j \varepsilon_{(t-j)} + \varepsilon_t \quad (1)$$

The model construction process consists of the following steps:a)Stationarity Testing:The stationarity of the time series is evaluated using the Augmented Dickey–Fuller (ADF) test. If the series is found to be non-stationary, differencing is applied to achieve stationarity.b)Model Identification:Based on the order of integration required to achieve stationarity, the value of d is determined. The autoregressive order p and the moving average order q are set to vary within the range of 0 to 3. A grid search algorithm implemented in Python is utilized to automatically fit the ARIMA model by evaluating combinations of (p,d,q).c)Model Selection:The optimal model is selected based on the minimum value of the Akaike Information Criterion (AIC), ensuring a balance between model fit and complexity.d)Model Diagnostics:The residuals of the

ISSN: 2692-7608 DOI: 10.6981/FEM.202506_6(6).0013

fitted model are subjected to the Ljung–Box test. If the residuals exhibit white noise characteristics, the model is deemed adequately fitted.

3.2. GPR

Gaussian Process Regression (GPR) is a flexible, non-parametric Bayesian approach to regression modeling [9]. At its core, GPR utilizes a Gaussian process to capture the distributional relationship between inputs and outputs, allowing for uncertainty quantification in predictions and adaptability to complex, nonlinear patterns [10].

The input set *X* and output set *f(x)* are assumed to follow a prior distribution of a Gaussian process:

$$f(x) \sim GP(\mu(x), K(X, X)) \tag{2}$$

Here, $\mu(x)$ represents the mean function, which is typically assumed to be 0, indicating no prior preference, i.e., $\mu(x) = 0$. *K(X,X)* represents the covariance matrix between the input vectors X, which can be computed using a covariance function. The specific form of the function is defined by the kernel function.

Since the observed values *y* typically contain noise, an independent Gaussian noise term ε is introduced. The expression for the regression model is:

$$\begin{cases} y = f(x) + \varepsilon \\ \varepsilon \sim N(0, \sigma^2 I) \end{cases} \tag{3}$$

In the formula, the observation error ε is Gaussian white noise with a mean of 0 and a variance of σ^2.

Thus, the prior distribution of the output vector *y* is:

$$y \sim N(0, K(X, X) + \sigma^2 I) \tag{4}$$

For a new input point x^*, the corresponding predicted value is denoted as y^*. The Gaussian process assumes that the output vector y of the training samples and the predicted value f_* of the test samples follow a joint Gaussian distribution, as detailed below:

$$\begin{bmatrix} y \\ f^* \end{bmatrix} \sim GP\left\{0, \begin{bmatrix} K(X, X) & K(x^*, X) \\ K(X, x^*) & K(x^*, x^*) \end{bmatrix}\right\} \tag{5}$$

Here, *K(X,X)* represents the covariance matrix between the training samples, $K(x^*,X)$ and $K(X,x^*)$ are the covariance matrices between the training samples and the test samples, and $K(x^*,x^*)$ is the covariance matrix between the test samples.

Based on the conditional distribution properties of the Gaussian distribution, the posterior distribution of the predicted value f^* for the test sample is given by:

$$f^* \mid y \sim GP(K(X, x^*)K^{-1}y, K(x^*, x^*) - K(K, x^*)K^{-1}K(x^*, K)) \tag{6}$$

ISSN: 2692-7608 DOI: 10.6981/FEM.202506_6(6).0013

$$E(f^*) = K(X, x^*)K^{-1}y \tag{7}$$

$$var(f^*) = K(x^*, x^*) - K(X, x^*)K^{-1}K(x^*, X) \tag{8}$$

Here, $E(f^*)$ and $var(f^*)$ represent the mean and variance at point x^*, respectively.

When using Gaussian Process Regression for prediction, it is necessary to train the hyperparameters involved in the model. This study employs a commonly used kernel function combination in machine learning, which is the constant kernel multiplied by the Gaussian kernel, plus the white noise kernel. The specific form is as follows:

$$K(X, X^*) = \sigma_f^2 exp[\frac{-(X - X^*)^2}{2l^2}] + \sigma_n \delta(X, X^*) \tag{9}$$

Here, σ_f^2 represents the signal variance, l is the length scale parameter, σ_n is the noise standard deviation, and $\delta(X, X^*)$ is the Kronecker function. The parameter set $\theta = \{\sigma_f^2, l, \sigma_n\}$ represents the hyperparameters. In this study, these hyperparameters are solved using the maximum likelihood estimation method, and the logarithmic form of the function is as follows:

$$logp(f^* \mid X, \theta) = -\frac{1}{2} y^T K^{-1} y - \frac{1}{2} log \mid K \mid - \frac{n}{2} log 2\pi \tag{10}$$

After obtaining the specific values of the hyperparameters, the mean $E(f^*)$ and variance $cov(f^*)$ corresponding to the prediction point x^* can be calculated using equations (6), (7), and (8) to obtain the final predicted value.

4. ARIMA-GPR Hybrid Model

4.1. Partitioning The Time Set

This study divides the population mortality rate data from 1978 to 2022 into three parts: the ARIMA training set, the GPR training set, and the test set. The ARIMA training set is used to train the parameters of the ARIMA model, the GPR training set is used to train the parameters of the GPR model, and the test set is used to evaluate the prediction performance of the ARIMA-GPR hybrid model. The prediction set consists of the population mortality rate for the next 10 years.

4.2. Building ARIMA Model

First, the data is standardized. The standardized data is then tested using the Augmented Dickey-Fuller Test (ADF), which reveals that the time series is non-stationary. Subsequently, the series is transformed to achieve stationarity by differencing until it becomes stationary. The results of the original series and the first-order differenced series are shown in Table 1 and Table 2, respectively:

Table 1. ADF Test for the Original Time Series

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-1.311649	0.623
Test critical Values:	1% Level	-3.593	
	5% Level	-2.932	
	10% Level	-2.604	

ISSN: 2692-7608 DOI: 10.6981/FEM.202506_6(6).0013

Table 2. ADF test for the first-order differenced time series

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-4.605097	0.000
Test critical Values:	1% Level	-3.593	
	5% Level	-2.932	
	10% Level	-2.604	

From the obtained results, it can be observed that the absolute value of the ADF statistic after first-order differencing is 4.605097, which is greater than the critical value of 2.932 at the 5% significance level. Therefore, we fail to reject the null hypothesis, and the very small p-value indicates that the original time series is stationary after first-order differencing. Hence, the differencing parameter ddd for the ARIMA(p, d, q) model is determined to be 1.

Next, a grid search method is employed to exhaustively explore the values of p and q. When both p and q are set to 0, the ARIMA model degenerates into a random walk model. Since ARIMA(0,1,0) is meaningless for the combined prediction model and negatively impacts the prediction accuracy, this case is excluded. During the search process, the values of p and q are set within the range of 0 to 4. Then, the error metrics for all p and q combinations are evaluated, and the top five best-performing combinations are selected as the feature sequences. The test results are shown in Table 3.

Table 3. Comparison of Different Models

Error Metric	MSE	RMSE	MAE
ARIMA(3,1,2)	0.055460	0.235499	0.161449
ARIMA(0,1,4)	0.054898	0.234303	0.164262
ARIMA(1,1,4)	0.055670	0.235945	0.162423
ARIMA(4,1,2)	0.055740	0.236093	0.163169
ARIMA(3,1,1)	0.055113	0.234762	0.169050

4.3. Building GPR Model

After determining the ARIMA(p,1,q) model for training, the input vectors for the GPR training set X_1, the test set X_2, and the prediction set X_3 are obtained through training using three different datasets: the ARIMA training set, the ARIMA training set + GPR training set, and the ARIMA training set + GPR training set + test set.

$$X_1 = [X_1^{1*}, X_1^{2*}, X_1^{3*}, X_1^{4*}, X_1^{5*}]^T \tag{11}$$

$$X_2 = [X_2^{1*}, X_2^{2*}, X_2^{3*}, X_2^{4*}, X_2^{5*}]^T \tag{12}$$

Here, X_n^m represents the output vector X_n obtained from the ARIMA model with the mmm-th combination of p and q.

The time series values on the GPR training set and test set are respectively used as the target vectors Y_1 and Y_2. Next, the input vector X_1 and the target vector Y_1 are fed into the GPR model to obtain the parameters. The input vector X_1 is then substituted into the GPR model, with the determined hyperparameters, to obtain the predicted values Y_2^* for model accuracy assessment. Finally, the input vector X_3 is input into the model to obtain the predicted population mortality rates for the prediction set.

ISSN: 2692-7608 DOI: 10.6981/FEM.202506_6(6).0013

4.4. Experimental Evaluation Metrics

This study selects the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) as the evaluation metrics for the prediction performance of each model. The calculation formulas for MSE, RMSE, and MAE are as follows:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \tag{13}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{14}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}| y_t - \hat{y}_t | \tag{15}$$

In the above formulas, y_t represents the actual population mortality rate data, $\hat{y}_t$ represents the predicted population mortality rate data, and n denotes the length of the time series.
The combination approach of the ARIMA Model and GPR Model is shown in Figure 1.

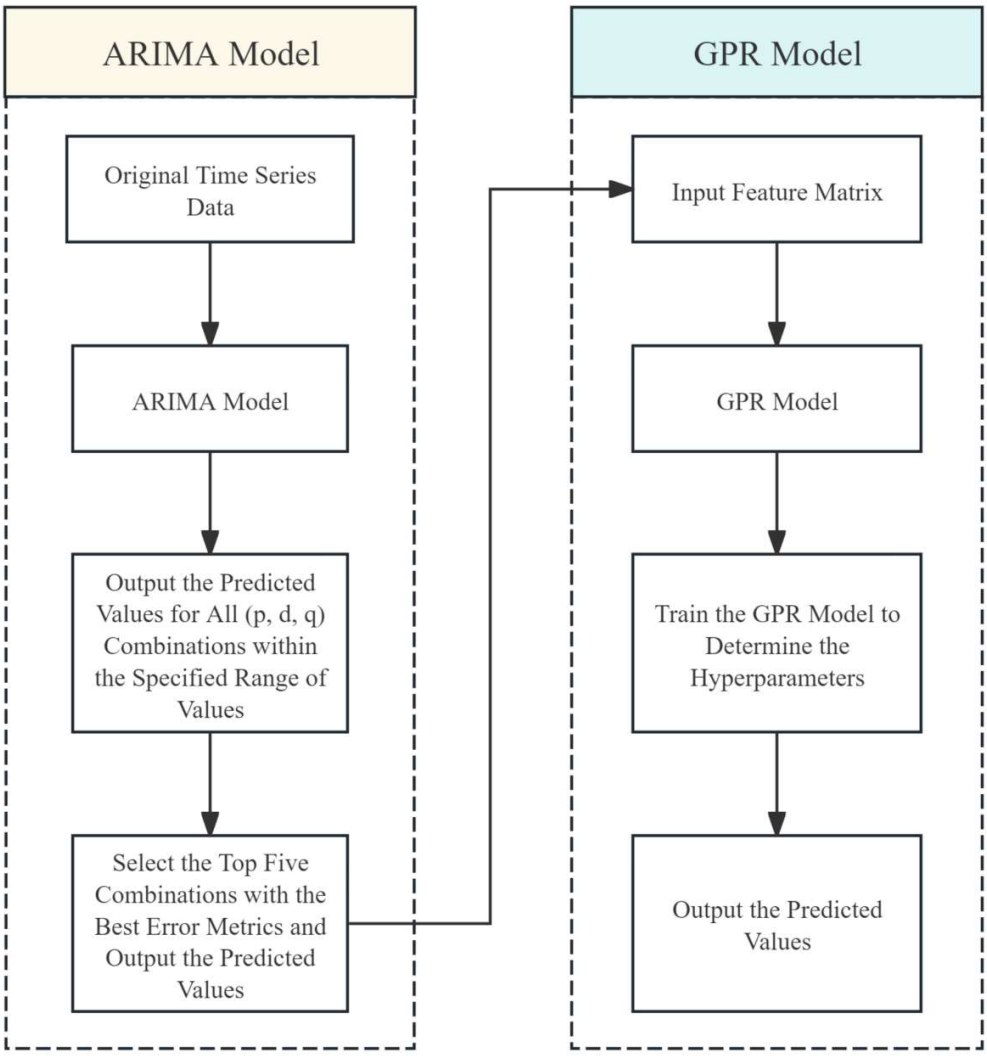


Figure 1. ARIMA-GPR Model Combination Diagram

ISSN: 2692-7608 DOI: 10.6981/FEM.202506_6(6).0013

5. Model Evaluation and Comparison

To validate the effectiveness of the model combination approach, this study compares the MSE, RMSE, and MAE of the ARIMA-GPR hybrid model with those of the ARIMA model. The comparison between the predicted values and actual values for both the ARIMA model and the ARIMA-GPR hybrid model is shown in Figure 1, while the comparison of the three evaluation metrics (MSE, RMSE, MAE) is presented in Table 4.

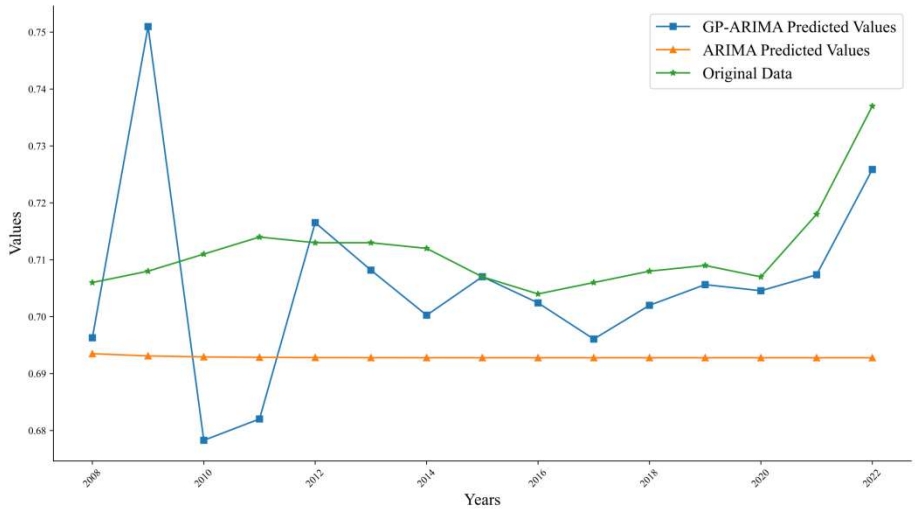


Figure 2. Comparison of the original data and the forecasted data of ARIMA and ARIMA-GPR

Table 4. Comparison of MSE, RMSE and MAE

Model	MSE	RMSE	MAE
ARIMA-GPR	0.0003065	0.0175090	0.0121700
ARIMA	0.0004082	0.0202054	0.0186579

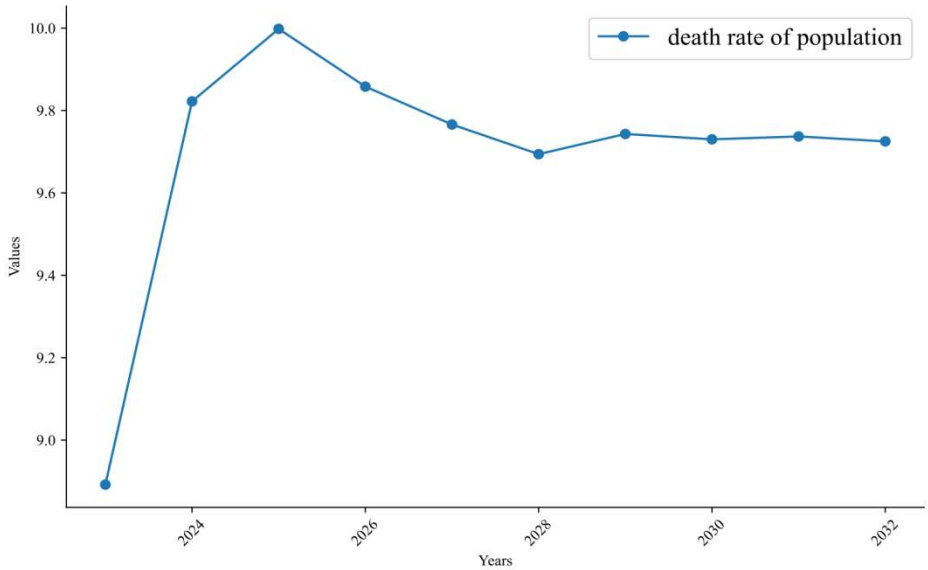


Figure 3. Predicted Mortality Rates for the Next Decade by ARIMA-GPR Model

ISSN: 2692-7608 DOI: 10.6981/FEM.202506_6(6).0013

As shown in Figure 2, the ARIMA-GPR hybrid model exhibits significantly higher fitting prediction accuracy than the ARIMA model, with the predicted curve closely following the trend of the actual data curve, thereby reducing the overall prediction error. As seen in Table 4, the proposed ARIMA-GPR model has smaller MSE, RMSE, and MAE values, which are 0.0003065, 0.0175090, and 0.0121700, respectively, indicating that the prediction error is relatively small. Therefore, the subsequent research selects the ARIMA-GPR model to forecast the population mortality rate from 2023 to 2033. The fitting prediction results are shown in Figure 3.

6. Conclusion

ARIMA and other machine learning models have been widely used in various fields of production and daily life. However, with the increasing complexity of data such as population mortality rates in recent years, relying solely on a single model is insufficient to accurately capture the underlying trends in the data. This paper innovatively proposes a hybrid model combining ARIMA and GPR for population mortality rate prediction. A comparison of the experimental results between ARIMA and ARIMA-GPR leads to the following conclusions:

1)Effectiveness of the ARIMA-GPR Hybrid Model

The ARIMA-GPR hybrid model proposed in this paper fully leverages the strengths of the ARIMA model in capturing linear trends and the GPR model's ability to capture nonlinear abrupt changes. This approach corrects the large errors often seen in single models when dealing with complex time series data, effectively reducing prediction errors and significantly improving forecasting accuracy. Experimental results show that the combination of ARIMA and GPR can effectively enhance the prediction accuracy of population mortality rates.

2)Optimizing Mortality Surveillance Systems through Granular Data Collection Strategies

To enhance the robustness and generalizability of mortality prediction models, future investigations should focus on establishing a high-resolution mortality database incorporating multidimensional demographic stratifications, including but not limited to age-specific, gender, socioeconomic status, and geographic distributions. The methodological framework should emphasize both temporal extension of observation periods to capture complete mortality transition patterns and spatial expansion to encompass regional heterogeneity. Such systematic data augmentation would facilitate: a) improved model calibration through enriched covariate matrices; b) enhanced external validation across diverse population strata; and c) robust performance in dynamic epidemiological contexts.

3) Transfer of the Model Combination Concept

In subsequent research, the serial combination approach of the two models, as proposed in this study, should be further extended to other model integration scenarios to validate both the effectiveness and scientific validity of this combined modeling methodology. Specifically, future investigations could explore: a) alternative sequential architectures with different model ordering configurations; and b) integration with additional model types beyond the current framework.

References

[1] Yan Zeng, Xi Chen, Yinglu Deng. Innovative Dynamic Population Mortality Prediction and Its Application [J]. Systems Engineering - Theory & Practice, 2016, 36(7): 1710-1718.

[2] Xiaofeng Wu, Yingmei Yang, Yaotong Chen. Combined ARIMA Forecasting Model Based on BP Neural Network Error Correction [J]. Statistics and Decision, 2019, 35(15): 65-68.

[3] Zhikun He, Guangbin Liu, Xijing Zhao, et al. A Survey of Gaussian Process Regression Methods [J]. Control and Decision, 2013, 28(8): 1121-1129+1137.

ISSN: 2692-7608 DOI: 10.6981/FEM.202506_6(6).0013

[4] Zhang G P. Time Series Forecasting Using a Hybrid ARIMA and Neu- ral Network Model [J].Neurocomputing,2003,(50).

[5] Bicong Ci, Pinyi Zhang. Financial time series forecasting based on ARIMA-LSTM model [J]. Statistics & Decision, 2022, 38(11): 145-149.

[6] Jinhai Yao, Jiajun Zou. Construction and numerical simulation of SVM-ARIMA model for CPI prediction [J]. Statistics & Decision, 2022, 38(21): 48-52.

[7] Fengming Zhang, Qian Su, Zhixing Deng, et al. Deformation prediction and application research of metro station foundation pit based on GF-GPR [J]. Journal of Hefei University of Technology (Natural Science Edition), 2025, 48(4): 563-569.

[8] Jinhai Yao. Stock index forecasting research based on ARIMA and information granulation SVR combination [J]. Operations Research and Management Science, 2022, 31(5): 214-220.

[9] Wenting Zong, Zhinong Wei, Guoqiang Sun, et al. Short-term load interval forecasting based on improved Gaussian process regression model [J]. Proceedings of the CSU-EPSA, 2017, 29(8): 22-28.

[10] YU J. State of health prediction of lithium-ion batteries:multiscale logic regression and Gaussian process regression ensemble[J]. Reliability engineering & system safety,2018,174:82-95.