Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 12, No. 1, 2025

Spatial Prediction of Mortality Based on Double-layer Gaussian Process Ensemble Regression Model

Shengxian Wang *

Department of economics, Beijing Technology and Business University, Beijing, China
*** Corresponding author Email:** 1131452764@qq.com

Abstract: Mortality rate prediction is one of the core aspects of risk pricing and product design in the insurance industry. Its accuracy directly determines the rationality of premiums, the adequacy of reserves, and the solvency assessment for life insurance, annuity products and other personal insurance products. Currently, there is temporal and spatial heterogeneity in China's population mortality rate, and implementing a traditional unified rate system is difficult to match the actual risk distribution, which may lead to systematic pricing deviations and other problems. This paper proposes an ensemble learning model based on Gaussian process regression, integrating regional factors as feature variables. On the one hand, it uses ensemble learning methods to further optimize the prediction results of the Gaussian process regression model; on the other hand, it uses Gaussian process regression to quantitatively estimate the uncertainty of future predictions. In the actual data analysis, this paper first verified that there are differences in population mortality rates among 31 provinces in China; secondly, it conducted a Monte Carlo experiment on the proposed method and found that this method can fit nonlinear functions; then, based on the historical mortality rates of 31 provinces in China, it conducted real data analysis and prediction to obtain mortality rate prediction intervals for uncertainty quantification estimation. From the prediction results, this method can well support spatial prediction of population mortality rates, thereby providing more effective decision-making basis for insurance companies to set premiums for personal insurance products.

Keywords: Gaussian Process Regression Model (GPR); Regional Mortality Prediction; Uncertainty Quantification; Stacking Ensemble Learning.

1. Introduction

The mortality rate prediction is the core cornerstone of the insurance industry, directly determining the pricing of personal insurance products, capital reserves, and the long-term operational stability. As the underlying logic of actuarial science, mortality rate data provide key parameters for the rate determination of personal insurance products such as life insurance, annuity, and health insurance. Its accuracy directly affects the risk exposure and profit structure of insurance companies - overestimating the mortality rate will lead to overly high premium pricing, weakening market competitiveness; underestimating it may trigger systemic interest rate losses or even solvency crises.

There have been numerous studies on mortality prediction models both at home and abroad, mainly divided into static mortality models and dynamic mortality models. The static mortality model encompasses the De Moivre model (1729), the Gompertz model (1825), the Makeham model (1860), and other classic models. However, the static mortality rate model has certain limitations. It fails to take into account the changing factors of future mortality rates. As a result, such models are usually only applicable for fitting existing data and are difficult to play an effective role in predicting future mortality rates. The dynamic mortality models include the Lee-Carter model (1992), the CBD model (2006), the Renshaw-Haberman model (2003) and other models. Among them, the Lee-Carter model and the CBD model respectively conduct modeling work by applying the logarithmic transformation and Logistic transformation to mortality rates. These two models have significant advantages and are widely used in related research. Milidonis et al. (2011) proposed the Markov mechanism transition stochastic mortality model, which expanded the Lee-Carter model through mechanism transition and conducted empirical research analysis using US population data.[10] Wang Tingting (2014) used the Lee-Carter model as the foundation and took the male mortality rates of China and Japan as the data samples for empirical research. By using the cointegration theory, she constructed the China-Japan mortality time factor cointegration model. This model achieved the prediction of future male mortality rates in China by correcting the errors of the predicted values of Japanese male mortality rates, providing a new idea for the research on mortality rate prediction in China.[2] Wang Zhigang et al. (2016) provided a comprehensive and complete theoretical elaboration of the Lee-Carter model, gave the expressions of the distribution and prediction of the Lee-Carter model, and proved that compared with the traditional interval prediction expressions, the interval prediction expressions obtained based on the complete theoretical model were more ideal.[3] Li Yaojie (2016) used the Lee-Carter model to calculate mortality rates and conducted unbiased prediction of the results through random simulation methods, effectively solving the problem of underestimated bias in mortality rate prediction caused by simple extrapolation models.[6] After a comparative analysis of eight commonly used mortality models, including the Lee-Carter model, the CBD model, the RH model, etc. Wang Xiaojun and Lu Qian (2020) found that the CBD model performed excellently in fitting and predicting the mortality rate of the elderly in the China's mainland. Its prediction interval was reasonable, and the survival curve conformed to the actual situation. In response to the problems of limited data volume and large fluctuations in the mortality rate of people above the retirement age in the China's mainland. [5]Wang Xiaojun and Zhao Xiaoyue (2021) adopted the CBD model to construct a

Logistic multi-population model suitable for the mortality rate of the elderly.[4] With the advent of the era of big data, machine learning models have gained favor among researchers. Qiao Chunjuan et al. (2020) analyzed the influencing factors of the health status of the elderly through the XGBoost algorithm, established a two-year transition probability matrix for the elderly, and established a BP neural network model to obtain the one-step transition probability matrix for the elderly.[8] Subsequently, Qiao Chunjuan et al. (2023) adopted a deep neural network model to price long-term care insurance products. [9] The achievements of scholars at home and abroad on mortality prediction models are quite abundant. However, there are still relatively few studies on the nonlinear dynamic trend of mortality rates and the integration of geographical factors into the models. Based on this, this paper proposes to use a two-layer Gaussian process ensemble regression model to incorporate geographical factors. On one hand, it accurately predicts the mortality rates of 31 provinces, enabling insurance companies to more accurately estimate future liability for compensation and avoid redundant or insufficient reserves. On the other hand, by leveraging the Bayesian characteristics of the Gaussian process regression model (GPR), it outputs the confidence intervals of the population mortality rates in each province, conducting uncertainty quantification and providing probabilistic support for risk management. Moreover, through empirical research, it has been proved that the performance of this model is excellent and the results are effective.

Based on this, this paper proposes to use a two-layer Gaussian process ensemble regression model to incorporate geographical factors. On one hand, it accurately predicts the mortality rates of 31 provinces, enabling insurance companies to more accurately estimate future liability for compensation and avoid redundant or insufficient reserves. On the other hand, by leveraging the Bayesian characteristics of the Gaussian process regression model (GPR), it outputs the confidence intervals of the population mortality rates in 31 provinces, conducting uncertainty quantification estimation and providing probabilistic support for risk management. Moreover, through empirical research, it has been proved that the prediction results and performance of this model are more outstanding.

2. Theory and Methodology

2.1. Gaussian Process Regression

Gaussian Process Regression (GPR) is a Bayesian method for nonlinear regression, which can be used to achieve nonlinear compensation. Gaussian process is modeled as $f(x)$ with mean function $m(x)$ and covariance kernel $k(x, x')$:

$$f(x) \sim GP\big(m(x), k(x, x')\big) \tag{1}$$

Given n data points: $X = \{x_1, x_2, \cdots, x_n\}$ and the corresponding observed value $Y = \{y_1, y_2, \cdots, y_n\}$, it can form a multivariate matrix:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \sim N\left(\begin{bmatrix} m(x_1) \\ m(x_2) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1,x_1)k(x_1,x_2)\cdots k(x_1,x_n) \\ k(x_2,x_1)k(x_2,x_2)\cdots k(x_2,x_n) \\ \vdots \\ k(x_n,x_1)k(x_n,x_2)\cdots k(x_n,x_n) \end{bmatrix} + \sigma^2 I \right) \tag{2}$$

Where σ^2 is variance, I is identity matrix.

Based on the observational data, the posterior distribution can be obtained. Given new data points x_*, the posterior distribution of y_* corresponding to this output is:

$$y_* | X, y, x_* \sim N(\mu_*, \sigma_*^2) \tag{3}$$

$$\mu_* = m(x_*) + k_*^T (K + \sigma^2 I)^{-1} (y - m) \tag{4}$$

$$\sigma_*^2 = k(x_*, x_*) - k_*^T (K + \sigma^2 I)^{-1} k_* \tag{5}$$

$$k_* = [k(x_*, x_1), k(x_*, x_2), \cdots, k(x_*, x_n)] \tag{6}$$

Where K is covariance matrix, $K_{ij} = k(x_i, x_j)$; m is mean vector, $m = [m(x_1), m(x_2), \cdots, m(x_n)]$.

The Gaussian process regression model not only provides point prediction values, but also gives the confidence interval of mortality rate with a certain confidence level, quantifying the uncertainty of the prediction:

$$y_{n+1} \in \left(\mu_* - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma_*}{\sqrt{n}} + \mu_* + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma_*}{\sqrt{n}} \right) \tag{7}$$

2.2. Double-layer Gaussian Process Ensemble Regression

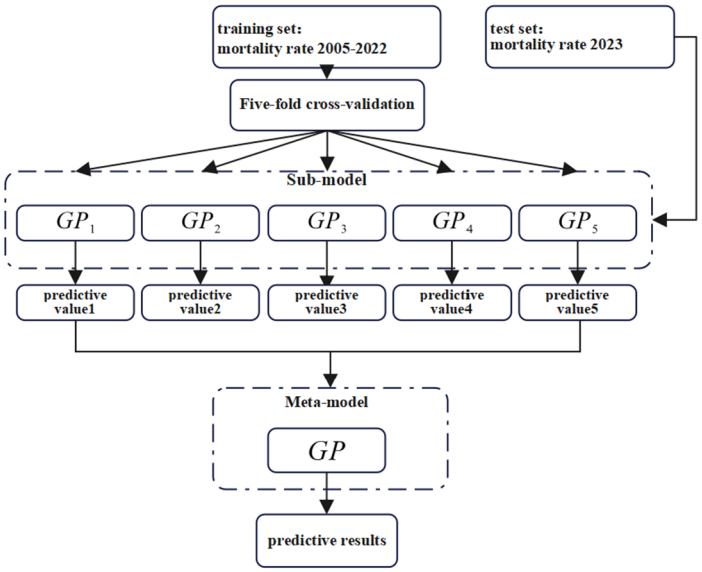

Figure 1. schematic diagram of DGPR

Stacking ensemble learning model is a hierarchical integration architecture that integrates different algorithm models. Here, multiple single Gaussian process regression models are first grouped to form sub-models for the first layer of parallel learning data. Then, the results obtained from each sub-model are input into the second layer of meta-model for training, thereby obtaining a complete two-layer Gaussian process ensemble learning model. The model uses the prediction data of the sub-models as the training data of the meta-model, reducing the generalization error and overfitting problem of a single model and improving the prediction accuracy.

3. Empirical Analysis

3.1. Simulation Data Analysis

The simulated data consists of 200 samples. The independent variable X is a random sample drawn from the standard normal distribution, and the dependent variable Y is generated by the following formula:

$$Y_i = \sin\left(\frac{3}{5}\pi X_i\right) + \epsilon_i \quad (8)$$

Where ϵ_i is Random noise, following a normal distribution $N(0, 0.1^2)$.

In the simulated data set, 80% of the data is randomly selected as the training set, while 20% of the data is set aside as the test set. In order to evaluate the predictive performance of the proposed model, three commonly used evaluation metrics were adopted, including: Mean Squared Error(MSE), Mean Absolute Error (MAE), Mean Relative Error (MRE).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (9)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (10)$$

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \quad (11)$$

Where n is sample size; y_i is the true value of sample i; $\hat{y}_i$ is the predicted value of sample i.

Based on the simulated data, a Monte Carlo experiment was conducted: The experiment found that the predicted values of the model were basically consistent with the true values, and the simulation effect was good; as well as that the double-layer Gaussian process regression model can be used to fit nonlinear functions.

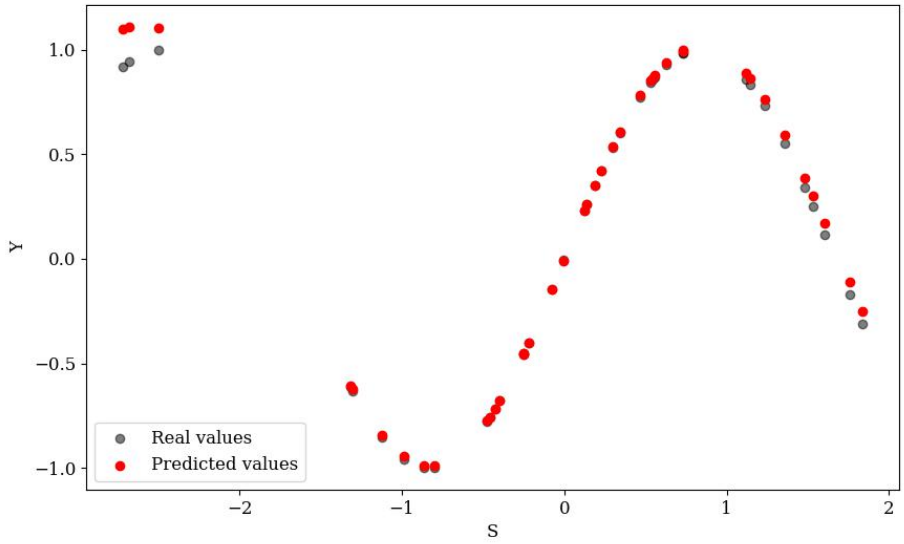

Figure 2. The actual values and predicted values of the simulated data

The DGPR model was compared with the Random Forest, GPR, XGBoost, and GBDT models in terms of model performance. It was found that the mean square error, mean absolute error and mean relative error of the DGPR model were all smaller than those of other models, and it had excellent performance and could predict the mortality rate more accurately.

3.2. Real Data Analysis

3.2.1. Data Sources and Pre-analysis

The data selected for this article are from the "National Statistical Yearbook" regarding the mortality rates of the population in 31 provinces of China from 2005 to 2023. During the pre-analysis of this data, it was discovered that:

The dispersion degree of mortality rates among the 31 provinces of China fluctuated greatly from 2005 to 2023. The overall dispersion degree showed an upward trend, indicating that the mortality rate differences among regions were increasing. This might be due to the expansion of population migration scale in China, the uneven spatial distribution of medical resources, and regional differences such as environmental differences, which led to a significant enhancement of spatial heterogeneity in mortality rates among provinces. As a result, there was a systematic deviation between the unified premium rate system and the actual risk structure. When designing personal insurance products, insurance companies should not only set premiums uniformly but also take geographical factors into account and formulate premiums separately for different regions.

By fitting the population mortality rates of 31 provinces in China, it is found that the trends of population mortality rates vary among different provinces. However, they have all shown an upward trend in recent years. For each province, the prediction model should be independent rather than uniform. But based solely on the historical mortality rate data of each province over the past 19 years, with insufficient data volume and insufficient training for the model, it is impossible to accurately predict future results. According to Figure 5, it can be observed that the curves of some provinces are very similar. Therefore, we can draw on their data to expand the sample

size and obtain more accurate population mortality rate data.

Table 1. Performance comparison between DGPR, GPR, RF, XGBoost and GBDT

	MSE-mean	MSE-variance	MAE-mean	MAE-variance	MRE-mean	MRE-variance
DGPR	0.0131	0.0075	0.0860	0.0166	8.1144	7.3171
RF	0.0281	0.0198	0.1204	0.0193	8.1208	7.4164
GPR	0.4679	0.0578	0.6077	0.0527	1.0000	0.0000
XGBoost	0.0572	0.0162	0.1882	0.0255	6.3024	5.6146
GBDT	0.1977	0.0250	0.3821	0.0335	3.7801	3.1188

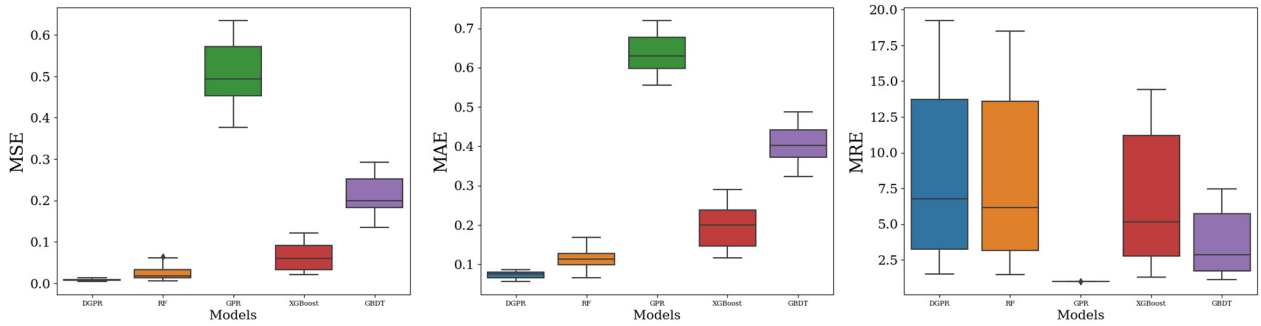

Figure 3. Box plot for performance comparison

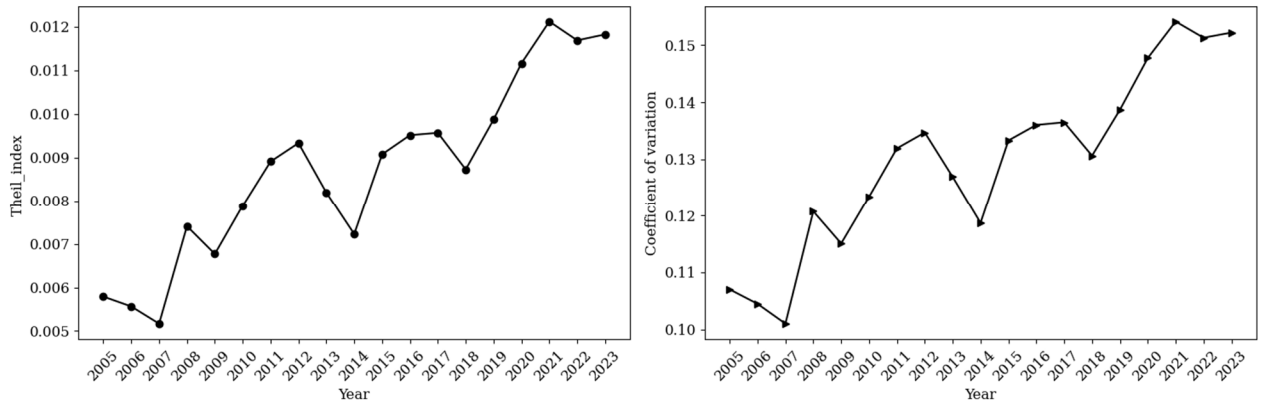

Figure 4. The Theil coefficient and coefficient of variation of mortality rates in 31 provinces of China

Figure 5. The mortality trend in 31 provinces of China

3.2.2. Analysis of Experimental Results

The real data experiment uses the provincial-level population mortality rates from 2005 to 2002 as the training set, and the provincial-level population mortality rates in 2023 as the test set to predict the provincial-level population mortality rates in 2023:

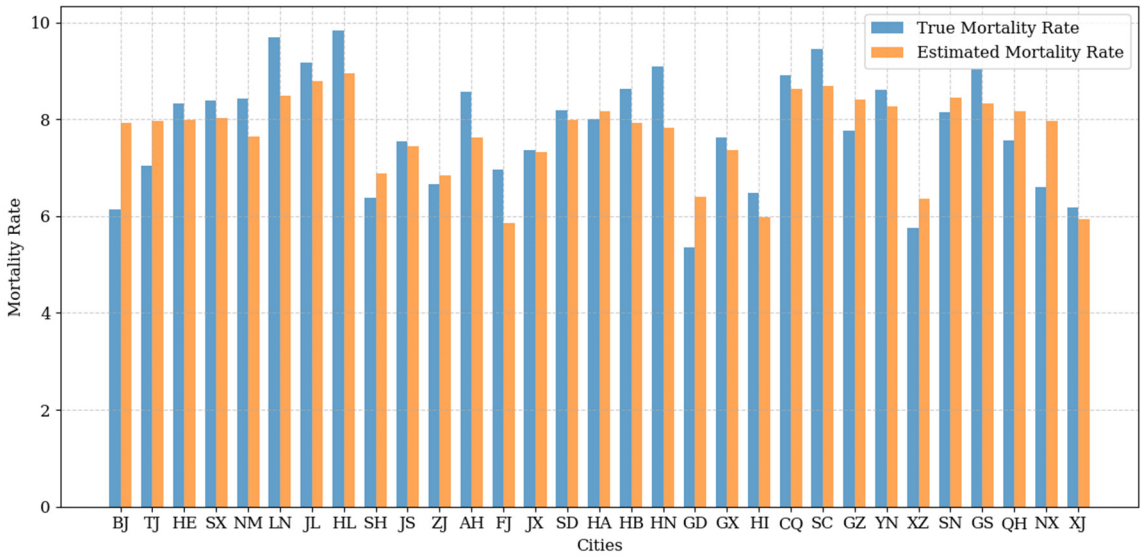

Figure 6. The actual and predicted values of mortality rate in 2023

The experimental results not only yielded the predicted values for 2023, but also quantitatively estimated the uncertainty of the mortality rates in each province:

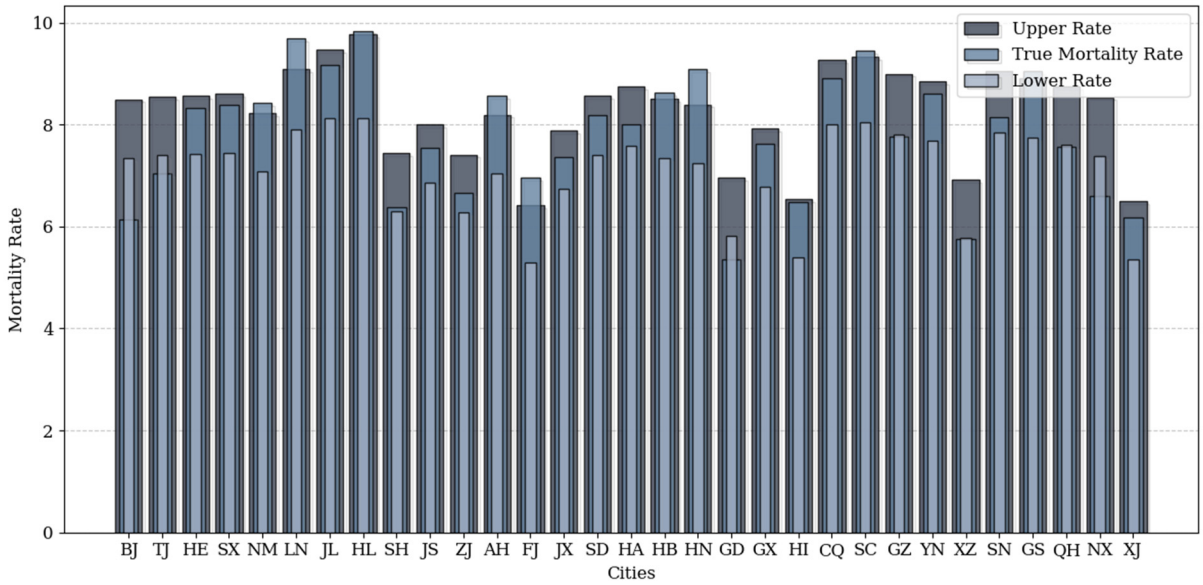

Figure 7. The predicted range of mortality rate for 2023

As shown in Figure 7, it can be observed that the actual mortality rates of most provinces fall within the confidence intervals, indicating that the predicted confidence intervals are reasonable and can encompass the actual data. Insurance companies set life insurance premiums based on mortality rates. If they can predict the possible range of future mortality rates rather than a single value, they can better assess risks. If the predicted mortality rate range for a province is wide, it indicates higher uncertainty, and the insurance company may need to set higher premiums to cover potential risks. Conversely, if the range is narrow, it indicates more reliable prediction and can set premiums more precisely. Moreover, confidence intervals can help insurance companies buffer risks. If the actual mortality rate exceeds the predicted interval, the insurance company may need to adjust capital reserves or reinsurance strategies. For provinces like Liaoning, Anhui, and Henan where the actual data do not fall within the confidence intervals significantly, further analysis may be needed to identify reasons, such as whether there are special population structures, distribution of medical resources, or other external factors, so as to adjust the model or formulate targeted risk management measures when calculating premiums.

In order to evaluate the predictive performance of the model under real data, three commonly used evaluation metrics were also adopted: Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Relative Error (MRE):

Table 2. Performance comparison between DGPR, RF, ARIMA, XGBoost and GBDT

	MSE	MAE	MRE
DGPR	**0.5676**	**0.6285**	**0.1593**
RF	**0.9902**	**0.8172**	**0.1674**
ARIMA	**4.3426**	**1.8061**	**0.2197**
XGBoost	**4.1543**	**1.8859**	**0.2365**
GBDT	**1.1849**	**0.9097**	**0.1614**

By comparing the mean square error, absolute error and relative error of each model, it is found that the DGPR model has smaller errors than those of Random Forest, ARIMA, XGBoost and GBDT, and the prediction results are more accurate.

4. Conclusion and Outlook

4.1. Conclusion and Recommendations

Due to the significant spatial-temporal heterogeneity of mortality rates in 31 provinces of China caused by reasons such as population mobility, uneven distribution of medical resources and environmental differences, the traditional unified rate model is difficult to match the actual risk distribution. In this paper, a two-layer Gaussian process ensemble regression model (DGPR) is proposed, which takes geographical factors into account and accurately predicts the mortality rates of 31 provinces. Insurance companies can more accurately estimate future liability for compensation and avoid redundant or insufficient reserves. Moreover, by leveraging the Bayesian characteristics of the Gaussian process regression model (GPR), the confidence intervals of population mortality rates in each province are output to quantify uncertainty and provide probabilistic support for risk management.

For insurance companies, the regional differences in mortality rates will directly affect the pricing of products and the assessment of reserves of their branches, thereby influencing the company's solvency. Therefore, when setting the rates, the mortality rate differences among different provinces should be taken into account, and different prices should be implemented for different provinces. Based on the current trend of population mortality rates, insurance companies should consider the overall increase in mortality rates and fully grasp the population mortality patterns in each province of China. By using the confidence intervals of mortality rates in different provinces to quantify uncertainty, precise pricing and differentiated rates should be implemented. Identify provinces with high uncertainty and set higher premiums to cover potential risks, while appropriately reducing premiums in regions with low uncertainty and low mortality rates to enhance market competitiveness. Conduct capital reserves and solvency management, reserve more reserves for provinces with high uncertainty to ensure solvency adequacy. For extreme risks outside the prediction range, transfer part of the risks through reinsurance.

4.2. Outlook

It conducts research and prediction on the regional differences of population mortality rates, but there are still many aspects that need improvement. First, there is no precise research on how to determine the number of base models when multiple Gaussian process regression models are adopted. In the future, more complex combined kernel functions should be selected to conduct more complex mortality rate predictions. Second, the data used only includes mortality rates by province, without considering the mortality rates of different genders and age groups within each province. In the future, it is necessary to further analyze regional mortality rates by gender and age group to provide more specific data for insurance companies. Third, due to the limitations of research data, this paper only analyzed the predicted mortality rates and did not deeply explore the specific reasons for the differences in population mortality rates among different regions. In future research, it is hoped to obtain more data and be able to analyze the dynamic evolution of population mortality rates in different provinces more comprehensively and specifically.

References

[1] Cheng Gongpin, Shen Shijie, Xu Dongni. Pricing Strategies for Long-Term Care Insurance Products in China Based on Combined Machine Learning Models [J]. Insurance Research, 2024, (12): 57-71.

[2] Wang Tingting. Extension Research on Mortality Models and Prediction of China's Population Mortality Rate [D]. Zhejiang University, 2014.

[3] Wang Zhigang, Wang Xiaojun, Zhang Xuebin. Theoretical Distribution and Interval Prediction of the Lee-Carter Model [J]. Mathematical and Statistical Methods in Management, 2016, 35 (03): 484-493.

[4] Wang Xiaojun, Zhao Xiaoyue, Chen Huimin. Joint Modeling and Consistent Forecasting of Mortality Rates for Elderly Populations with Multiple Demographic Characteristics [J]. Population and Economy, 2021, (02): 45-56.

[5] Wang Xiaojun, Lu Qian. Research Progress on Dynamic Mortality Models [J]. Journal of Applied Probability and Statistics, 2020, 36(04): 415-440.

[6] Wu Xiaokun, Li Yaojie. Exponential Extrapolation Prediction of Mortality Rate and Deviation Correction Based on Lee-Carter Model [J]. Statistics and Decision, 2016, (20): 19-21.

[7] Li Yangzheng. Estimation and Dynamic Research on Population Mortality Rates in Chinese Provinces [D]. Southwest University of Finance and Economics, 2023.

[8] Qiu Chunjuan, Guan Huilin, Qian Linyi, Wang Wei. Pricing Research on Long-Term Care Insurance: Based on XGboost Algorithm and BP Combined Neural Network Model [J]. Insurance Research, 2020, (12): 38-53.

[9] Qiu Chunjuan, Liu Shoushan, Zhang Nan. Research on End-to-End Long-Term Care Insurance Pricing Model Based on Deep Neural Network [J]. Insurance Research, 2023, (12): 71-81.

[10] A. Milidonis,Y. Lin, S. H. Cox. Mortality Regimes and Pricing[J]. North American Actuarial Journal ,2011, 15(2) :266 – 289.