Integrated Model and Algorithm Research on Cargo Volume Forecasting and Personnel Scheduling Based on SARIMA and Genetic Algorithm (GA)

Jinghong Wang *, Lihao Gao

School of Mathematics and Information Science, Guangxi University, Nanning, China, 530004

* Corresponding Author Email: wangjinghong0407@163.com

Abstract. In the context of the swift growth of the logistics industry, the accurate prediction of cargo volumes and the rational arrangement of personnel scheduling have become essential for logistics companies to enhance operational efficiency and curtail costs. This study introduces a comprehensive model to address these challenges effectively. For cargo volume forecasting, the study developed an ARIMA-based time series model and further proposed a SARIMA model to account for seasonal variations. The SARIMA model, which incorporates seasonal differencing and seasonal autoregressive moving average terms, demonstrated significant improvements in prediction accuracy, especially for data with distinct seasonal patterns. In terms of personnel scheduling, the study formulated an optimization model utilizing the genetic algorithm (GA). This model aims to minimize the total number of man-days and balance the actual hourly man-efficiency per day by simulating natural selection and genetic mechanisms. The results from the study indicated that the GA-based model could substantially reduce scheduling costs while meeting service requirements. The integration of the SARIMA model for forecasting and the GA model for scheduling provides a robust solution for cargo volume prediction and staff scheduling in logistics sorting centers, offering a significant contribution to operational optimization.

Keywords: Staff Scheduling Optimization, SARIMA Model, Genetic Algorithm, Machine Learning Algorithm.

1. Introduction

In the rapid development of the courier industry today, supply chain management (SCM) as a concept has been developed by leaps and bounds [1], the sorting volume prediction of the sorting center and the rational arrangement of the sorters has become an important research topic. Accurate prediction of sorting volume in a sorting center is the basis for subsequent management and decision-making. If managers can predict the sorting volume of the sorting center in the future period of time in advance, they can arrange resources in advance, formulate a reasonable arrangement of sorters, and optimize the operational efficiency of the sorting center and cost control, thus improving the sorting efficiency and reducing the sorting cost.

Prior to this, some scholars have done related research. For the prediction of sorting volume, the random forest model can be predicted under the premise of strong noise immunity, each new case is evaluated by each tree in the algorithm, and in the case of classification, the prediction result is the average of all the predictions of the majority class (the class that receives the most "votes" from a single tree) or regression case [2]. AdaBoost as a fusion algorithm of Boosting, which fuses a series of weakly learned models together into a strong classifier by cascading them to achieve prediction and successfully improve the accuracy of weakly learned algorithms [3], is very popular due to its ability to learn patterns from data and infer solutions from unknown data. However, there are fewer studies on artificial neural networks for supply chain goods sorting prediction [4]. As for the problems related to sorting staff scheduling allocation [5], planning-type models are usually used for allocation, and the most common linear programming models can be used when the conditions are less complex and the data do not change drastically and abruptly [6]. In contrast, greedy algorithms can make optimal allocations at each stage [7].

However, in the context of logistics sorting, the random forest model and the Adaboost model are not so sensitive in predicting the changes in the volume of goods due to time changes and due to special time nodes, and artificial neural network can be very good at predicting based on learning, but considering that the volume of goods in logistics is strongly affected by the seasonality, the pure artificial neural network can not satisfy the requirements of accurate prediction. So we further propose a SARIMA model based on the ARIMA model to better handle the seasonal changes of cargo volume data. By introducing seasonal differences and seasonal autoregressive moving average terms, the SARIMA model achieves a significant improvement in prediction accuracy, especially when dealing with data with obvious seasonal fluctuations. Meanwhile, considering the need to integrate the fluctuation prediction of cargo volume to reasonably allocate the personnel scheduling, simple linear programming cannot meet the requirements of the complex situation, and the greedy algorithm lacks the consideration of the future prediction situation, so we proposed an optimization model based on the Genetic Algorithm (GA) [8] for the personnel scheduling. With the objective of minimizing the number of man-days and balancing the actual hourly man-efficiency per day, the model searches for the optimal staffing plan by simulating natural selection and genetic mechanisms. The model has been proven to significantly reduce scheduling costs while meeting service requirements.

2. The basic fundamentals of SARIMA and Genetic Algorithm (GA) models

2.1. The structure of SARIMA

SARIMA [9] (Seasonal Autoregressive Integrated Moving Average) is a time series forecasting method that is an extension of the ARIMA model. It includes a seasonal component and is suitable for data with seasonal patterns. The SARIMA model consists of four main components:

(1) Seasonal component: In the same way that the ARIMA model handles non-seasonal trends and cyclical fluctuations through ordinary differencing, the SARIMA model handles distinct seasonal patterns through seasonal differencing: this means subtracting observations from the current observation up to the m-period for D times. For example, for monthly data, if there is annual seasonality, there might be 12.

(2) Autoregressive component: this is an autoregressive component that correlates with previous observations in the time series

$$AR(p) = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t \tag{1}$$

(3) Differential part: this is a differential operation to deal with non smoothness

$$I(d) = (1 - B)^d X_t \tag{2}$$

(4) Moving average component: The moving average component describes the autocorrelation of the model error terms. Here, it represents the past model error term. Coefficients are model parameters that describe the contribution of each historical error term to the current forecast error. These parameters help the model to capture the stochastic fluctuation characteristics in the series, and the magnitude of q indicates the time span of the error correlation under consideration

$$MA(q) = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{3}$$

2.2. The structure of Genetic Algorithm (GA) models

Genetic Algorithm (GA) is a global optimization algorithm based on a natural evolutionary process, which is very suitable for solving complex and high-dimensional optimization problems. In the following, the text will elaborate on the steps of the GA algorithm in solving the logistics personnel scheduling problem.

1) Chromosome coding

According to the decision variables defined in the previous section, each individual (solution) is represented as a 360-dimensional real vector dvx, which is the chromosome encoding in the GA algorithm. Each decision variable dvxi corresponds to a gene on the chromosome, and its value range is a non-negative integer, indicating the number of regular or temporary workers in the corresponding time period.

2) Population initialization

N initial solutions (individuals) are randomly generated to form the initial population. The gene value of each individual is a random number obeying a certain probability distribution. The population size N is an important parameter, which needs to be selected appropriately according to the problem size and computational resources, generally 100-500.

3) Fitness function

The fitness function serves as an index to evaluate the strengths and weaknesses of individuals, which directly affects the convergence of the genetic algorithm and the quality of the solution. The fitness function can be directly used as the objective function of the optimization model.

4) Selection operation

The selection operation selects the "best" individuals from the current population to serve as the parents of the next generation. Commonly used methods include roulette selection and tournament selection. In this problem, we can use tournament selection. Randomly select k individuals and choose the one with the highest fitness as the parent. k is an adjustable parameter, generally 2-8.

5) Crossover operation

The crossover operation generates new children based on the parent individuals to further explore the solution space. For the real number encoding of this problem, linear combinatorial crossover can be used. That is, a crossover probability between [0,1] is randomly generated, and then the offspring obtains a part of the parent and a part of the parent, and the proportion of the parent's part to the total gene length is the crossover probability. The crossover probability is also a parameter that needs to be debugged and is usually taken as 0.5-0.9.

6) Variation operation

The variation operation searches locally in the solution space to ensure that the algorithm does not fall into local optimality. For this problem, we can use random perturbation mutation. That is, we add a random perturbation amount to each gene that obeys a normal distribution. [10] The probability of variation also needs to be debugged, usually taking 0.01-0.1.

7) Stopping condition

Appropriate iteration termination conditions are set, such as the maximum number of iterations, population convergence, and so on. When the stop condition is satisfied, the current optimal solution is output as the final staff scheduling program.

In general, the optimization process based on a genetic algorithm includes the steps of coding, initialization, fitness evaluation, selection, crossover, mutation, and so on. Through careful design and parameter debugging of these steps, the staff scheduling plan that meets the demand and minimizes the cost can be effectively solved.

3. Results

3.1. Solution of SARIMA-based model for future cargo volume forecasting

We preprocessed the data and performed an ADF data smoothness assessment, the results are shown in table 1.

(1) Data preprocessing

Daily and hourly data (i.e., data from SC48 sorting centers in Appendix I and Appendix II) of SC48 sorting centers were preprocessed, and the daily data were sorted by date order, and the hourly data were sorted by date and time in descending order, and there were no missing values in any of the data.

(2) ADF data smoothness assessment

The ADF test p-value for the daily data is 0.9917, much larger than 0.01, indicating that the daily data are very non-smooth; the ADF test p-value for the hourly data is 0.0222, larger than 0.01 but smaller than 0.05, indicating that there is a 95% confidence level that the daily data are smooth, but they are not smooth at the 99% confidence level. For smoother data, both sets of data were differenced.

(3) Differential Order Determination

The series that is smooth at a 99% confidence level is obtained by constant looping to get the least number of differencing orders and the p-value of the ADF test at the corresponding order as shown in the table below :

Table 1. Differential orders required for stabilization and p-values under the corresponding differentials

	difference in order	p-value
Daily data	1	1.929×10^{-5}
Hourly data	1	9.645×10^{-19}

It can be seen that after the corresponding differencing, the ADF test p-values are all much less than 0.01, which indicates that after differencing both sets of data have reached smoothness at 99% confidence level.

(4) Parameter estimation and model fitting

The process of fitting the SARIMA model is similar to that of ARIMA but requires additional consideration of the estimation of seasonal parameters. This is usually done through maximum likelihood estimation or other suitable optimization algorithms and may use an information criterion (e.g., AIC or BIC) to select the best model configuration.

Because of the large number of parameters in SARIMA, in this paper, we use auto_arima in Python to automate parameter selection, setting seasonal=True and specifying the m-value, to automatically find the optimal p, d, q, P, D, and Q

(5) Performance evaluation

The SARIMA model can also be evaluated using statistical metrics such as NMSE, in addition to focusing on whether the model adequately captures seasonal variations in the data. Performing residual analysis to ensure that the residuals are close to white noise is also an important part of assessing how well the model fits.

3.2. Solving the SARIMA model

Randomly selected sorting site SC481 and tested the model at site SC48.

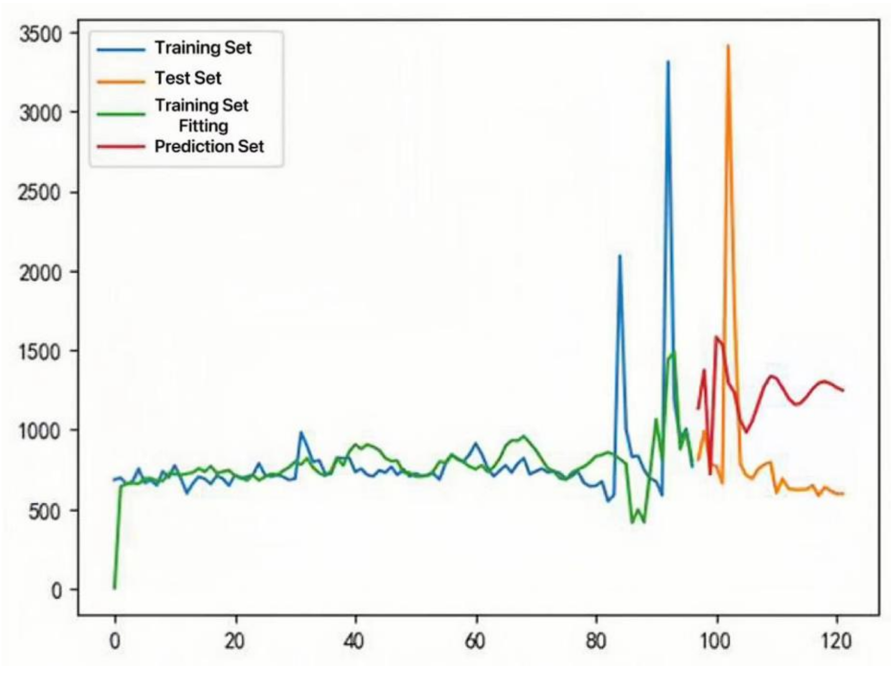

Figure 1. Predictive effects on daily data

As can be seen, the predictive effect on the daily data is specifically shown in Figure1 by the absence of many unidentified peaks and the prediction of true value peaks around 100. the NMSE value of 0.4654 shows that there is still room for improvement.

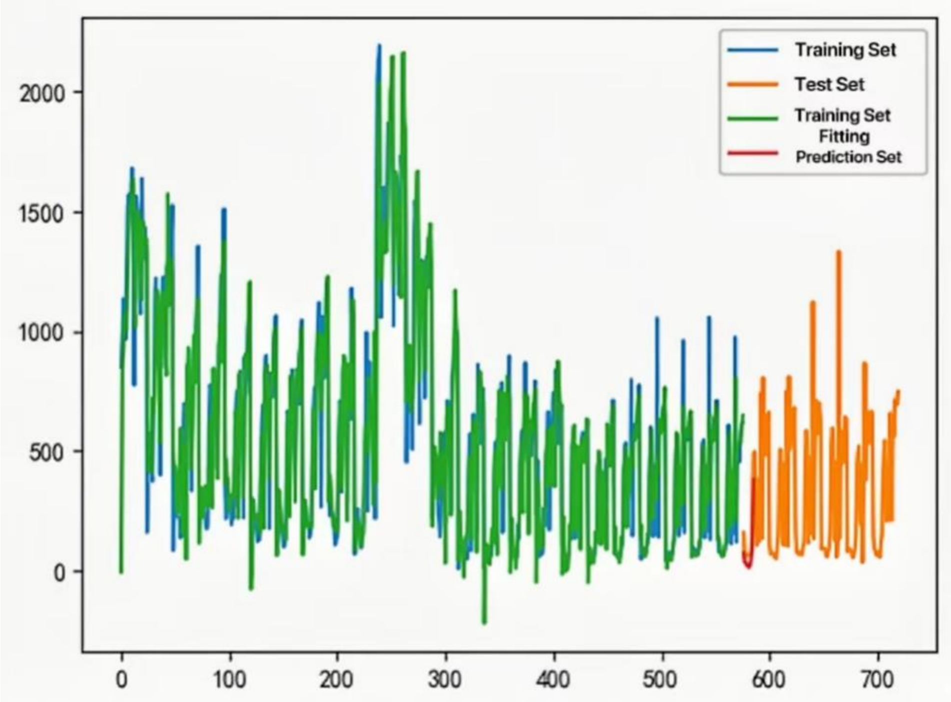

Figure 2. Predictive effects on hourly data

As can be seen, the prediction on hourly data shown in Figure2 is fantastic, almost perfectly predicting the data on the test set. The evaluation metric NMSE is 0.01693, indicating that the results have been almost perfectly predicted.

Generalize the model to all sites and make predictions

The study checked the NMSE of each sorting center for the daily data. The NMSE ranges from 0 to 1, with lower values indicating better performance. It is observed that the NMSE values are in the range of 0 to 0.2, indicating that the predictions are still very good.

3.3. Establishment of personnel allocation model for sorting center based on GA algorithm

(1) Decision variables

In the design of decision variables for optimization problems, a reasonable way of expressing the variables must be determined according to the characteristics of the actual problem. In the current logistics scheduling problem, the allocation of regular workers and temporary workers for each working day and each working period within 30 days must be determined. Therefore, a reasonable design for the decision variables is presented as follows.

The decision variable is defined the decision variable as a 30-day × 6 shifts × 2 types of labor = 360-dimensional vector X, where:

$$dvX = \{dvx_1, dvx_2, \dots, dvx_{360}\} \tag{4}$$

where: dvx1~dvx6 represents the number of regular workers in 6 shifts on day 1.

dvx_1~dvx_6represents the number of regular workers in the 6 shifts on day 1,

dvx_7~dvx_{12} represents the number of temporary workers in 6 shifts on day 1, and

dvx_{13}~dvx_{18} represents the number of regular workers in 6 shifts on day 2, and

dvx_{19}~dvx_{24} represents the number of temporary workers in 6 shifts on day 2, ...

dvx_{355}~dvx_{360} represents the number of temporary workers for 6 shifts on day 30.

With such a definition of decision variables, we can calculate the number of regular and temporary workers in any time period according to the formula. The formula is as follows.

$$pdvx_{\mathrm{j,k}} = \begin{cases} dvx_{6\times2\times(\mathrm{j}-1)+1}, 0 \le k \le 4 \\ \sum_{i=1}^{2} dvx_{6\times2\times(\mathrm{j}-1)+i}, 5 \le k \le 7 \\ \sum_{i=2}^{3} dvx_{6\times2\times(\mathrm{j}-1)+i}, 8 \le k \le 11 \\ \sum_{i=2}^{4} dvx_{6\times2\times(\mathrm{j}-1)+i}, k = 12 \\ \sum_{i=3}^{4} dvx_{6\times2\times(\mathrm{j}-1)+i}, k = 13 \\ \sum_{i=3}^{5} dvx_{6\times2\times(\mathrm{j}-1)+i}, 14 \le k \le 15 \\ \sum_{i=4}^{6} dvx_{6\times2\times(\mathrm{j}-1)+i}, 16 \le k \le 19 \\ \sum_{i=5}^{6} dvx_{6\times2\times(\mathrm{j}-1)+i}, 20 \le k \le 21 \\ dvx_{6\times2\times(\mathrm{j}-1)+6}, 22 \le k \le 23 \end{cases} \tag{5}$$

$$tdvx_{\mathrm{j,k}} = \begin{cases} dvx_{6\times2\times(\mathrm{j}-1)+7}, 0 \le k \le 4 \\ \sum_{i=1}^{2} dvx_{6\times2\times(\mathrm{j}-1)+i+6}, 5 \le k \le 7 \\ \sum_{i=2}^{3} dvx_{6\times2\times(\mathrm{j}-1)+i+6}, 8 \le k \le 11 \\ \sum_{i=2}^{4} dvx_{6\times2\times(\mathrm{j}-1)+i+6}, k = 12 \\ \sum_{i=3}^{4} dvx_{6\times2\times(\mathrm{j}-1)+i+6}, k = 13 \\ \sum_{i=3}^{5} dvx_{6\times2\times(\mathrm{j}-1)+i+6}, 14 \le k \le 15 \\ \sum_{i=4}^{6} dvx_{6\times2\times(\mathrm{j}-1)+i+6}, 16 \le k \le 19 \\ \sum_{i=5}^{6} dvx_{6\times2\times(\mathrm{j}-1)+i+6}, 20 \le k \le 21 \\ dvx_{6\times2\times(\mathrm{j}-1)+12}, 22 \le k \le 23 \end{cases} \tag{6}$$

where pdvxj.k denotes the number of regular workers in the kth hour of the jth day, tdvxj,k denotes the number of temporary workers in the kth hour of the jth day, and dvxi denotes the value of the ith element of the decision variable.

(2) Objective function

The objectives of the problem are discussed, with the first being the requirement to minimize the number of scheduled man-days, and the second being the requirement to equalize the actual hourly manpower effect per day. For the first objective, the total number of person-days is obtained by adding up all the elements of the decision variable; the second objective is evaluated using the actual hourly person-effect variance per day. Both components are smaller the better the result.

Given that this is a multi-objective optimization problem, the study proposes to use the weighting method for model construction. The so-called weighting method is to determine a weight for the function value of each objective. The weight is generally determined based on practical experience, as well as the magnitude of each objective function, and then obtained when solving.

In summary, the objective function is given by the following equation:

$$\min_{dvx}\ weight_1 \sum_{i=1}^{360} dvx_i + weight_2 \sum_{j=1}^{30} Var(\{\frac{G_{j,k}}{pdvx_{j,k} + tdvx_{j,k}} \mid 0 \le k \le 23\}) \tag{7}$$

where weight1 denotes the weight of the first component, weight2 denotes the weight of the second component and Gj,k denotes the number of goods in the kth hour of the jth day.

(3) Constraints

The constraints are as follows: the maximum hourly picking capacity should be greater than the hourly shipment volume; the maximum hourly manpower efficiency of temporary workers is 20 parcels per hour, and the maximum hourly manpower efficiency of regular workers is 25 parcels per hour; the number of regular workers in each picking center is 60; and each decision variable should be a natural number.

For the hourly person-efficiency constraints of the two types of workers, since we are obtaining the optimal combination, we directly use the maximum value in calculating the hourly picking capacity of the picking centers in order to integrate the volume constraint and the hourly person-efficiency constraint. The above constraints are formulated to obtain the following expression:

$$s.t.\begin{cases} 25pdvx_{j,k} + 20tdvx_{j,k} \geq G_{j,k}, 1 \leq j \leq 30 \wedge 0 \leq k \leq 23 (C1) \\ x_i \leq 60, 1 \leq i\%12 \leq 6 (C2) \\ x_i \in N, 1 \leq i \leq 360 (C3) \end{cases} \tag{8}$$

3.4. Results of GA Algorithm-Based Staff Assignment Model for Sorting Centers

The application of the Genetic Algorithm (GA) in this study has yielded effective solutions for the personnel scheduling problem within sorting centers. Here's a detailed outline of the process and its outcomes:

(1) Chromosome Encoding

Each solution is encoded as a chromosome, represented by a 360-dimensional vector, where each dimension corresponds to a decision variable. These variables denote the number of permanent and temporary workers assigned to each of the 30 days and 6 shifts. This encoding scheme allows for a comprehensive representation of the workforce allocation over the scheduling period.

(2) Population Initialization

The initial population of chromosomes is randomly generated, with each gene value following a specific probability distribution. The population size, typically ranging from 100 to 500, is determined based on the complexity of the problem and the available computational resources.

(3) Fitness Function

The fitness function evaluates the quality of each chromosome, directly influencing the convergence of the GA and the solution quality. It incorporates two main objectives: minimizing the total man-days and ensuring the balance of actual hourly efficiency across all shifts and days.

(4) Selection Operation

Tournament selection is employed to select the fittest individuals from the current population. This method involves randomly selecting a subset of individuals and choosing the one with the highest fitness score as a parent for the next generation.

(5) Crossover Operation

Crossover combines the genetic information of two parent chromosomes to produce offspring. A crossover probability, usually between 0.5 and 0.9, determines the extent of genetic exchange between the parents.

(6) Mutation Operation

Mutation introduces random changes to the offspring's genes to maintain diversity in the population and prevent premature convergence. The mutation probability, typically ranging from 0.01 to 0.1, controls the frequency of these changes.

(7) Stopping Conditions

The GA process is terminated when a set number of iterations have been reached or when the population has converged sufficiently. The optimal solution at this point is then proposed as the final staff scheduling plan.

The implementation of the GA in Matlab has been fine-tuned through continuous testing and adjustment. The weights for the objective function, weight1 and weight2, were determined to be 5x10^-8 and 0.5, respectively, after extensive calibration. The resulting staff allocation has been validated against the service requirements and cost constraints, demonstrating its feasibility and effectiveness.

This section now provides a more detailed account of the GA application in the context of the study, highlighting the specific parameters and outcomes relevant to the staff scheduling problem in sorting centers.

4. Conclusions and outlooks

Based on the SARIMA model, we solve the logistics cargo volume prediction problem, which performs well in handling data with significant seasonal fluctuations. Based on a Genetic Algorithm (GA) optimization model, we solve the staff scheduling problem. To minimize the number of man-days and balance the actual hourly man-efficiency per day, the model searches for the optimal staffing solution by simulating natural selection and genetic mechanisms. The model is proven to be able to significantly reduce the scheduling cost while meeting the service requirements. In summary, the model and algorithm proposed in this study provide an effective solution for cargo volume prediction and staff scheduling in logistics sorting centers. Future research can further explore and improve the model in terms of its generalization ability, computational efficiency, and integration with other decision factors. Through continuous optimization and innovation, we expect to contribute more scientific wisdom to the development of the logistics industry.

Meanwhile, in the future, it is believed that the above model will have a more matching application in cross-industry, and the results of this research can be extended to other logistics centers or similar industries that need staff scheduling optimization, such as catering and medical care.

References

[1] Bilgili M, Pinar E. Gross electricity consumption forecasting using LSTM and SARIMA approaches: A case study of Türkiye [J]. Energy, 2023, 284: 128575.

[2] Cheng J, Tiwari S, Khaled D, et al. Forecasting Bitcoin prices using artificial intelligence: Combination of ML, SARIMA, and Facebook Prophet models [J]. Technological Forecasting and Social Change, 2024, 198: 122938.

[3] Ariyanti V P, Yusnitasari T. Comparison of Arima and Sarima for forecasting crude oil prices [J]. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 2023, 7 (2): 405-413.

[4] Vince A. A framework for the greedy algorithm [J]. Discrete Applied Mathematics, 2002, 121 (1-3): 247-260.

[5] Lambora A, Gupta K, Chopra K. Genetic algorithm-A literature review [C]//2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon). IEEE, 2019: 380-384.

[6] Cui T, Shi Y, Lv B, et al. Federated learning with SARIMA-based clustering for carbon emission prediction [J]. Journal of Cleaner Production, 2023, 426: 139069.

[7] Agyemang E F, Mensah J A, Ocran E, et al. Time series based road traffic accidents forecasting via SARIMA and Facebook Prophet model with potential changepoints [J]. Heliyon, 2023, 9 (12).

[8] Alhijawi B, Awajan A. Genetic algorithms: Theory, genetic operators, solutions, and applications [J]. Evolutionary Intelligence, 2024, 17 (3): 1245-1256.

[9] Sohail A. Genetic algorithms in the fields of artificial intelligence and data sciences [J]. Annals of Data Science, 2023, 10 (4): 1007-1018.

[10] Acampora G, Chiatto A, Vitiello A. Genetic algorithms as classical optimizer for the quantum approximate optimization algorithm [J]. Applied Soft Computing, 2023, 142: 110296.