Journal of Electronics and Information Science (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2024.090319
ISSN 2371-9524 Vol. 9 Num. 3

Research on Crop Planning Based on Data Mining and Genetic Algorithms

Shaosong Zhang[1], Yuhang Chen[1], Liping Du[1,*]

[1]College of Science, North China University of Science and Technology, Tangshan, China
[*]Corresponding author: liping.du@163.com

Keywords: Data mining, linear programming, orthogonal experiment, genetic algorithm

Abstract: Data mining techniques can be employed to extract information that is not immediately apparent from large amounts of data, and to construct predictive models based on this extracted information. These models can then be used as a basis for decision-making. In order to expand the scope of its application, this paper combines data mining with genetic algorithms and orthogonal experiments and applies it to the optimization of planting decisions. In particular, this study initially gathered and structured data on planting conditions, crop sales, per-mu yields, planting costs, and selling prices in a village through data mining techniques and subsequently analyzed the intrinsic relationships between these variables. On this basis, this paper constructs a planning function with the goal of maximizing profits and uses genetic algorithms to solve optimization problems. Overall, this study has successfully applied data mining techniques to practical planting decision-making problems, which not only has strong practicality, but also provides a reference for solving other complex planning problems. In the future, further exploration of the integration of additional optimization algorithms into the data-driven decision-making analysis framework may yield more comprehensive solutions.

1. Introduction

Data mining is an effective technique for decision-making analysis that can extract implicit information from large data sets and construct predictive models to inform decision-making processes. While the majority of extant research concentrates on the deployment of data mining in industrial and economic contexts, there is a paucity of scholarship examining its utilisation in the domain of agricultural production, particularly in the realm of crop planning decisions. Crop planning, as a fundamental aspect of agricultural production, necessitates a comprehensive assessment of multiple variables, including planting conditions, yield prediction, cost control, and market demand [1,2]. In light of the aforementioned, it is evident that identifying an optimal planting strategy within the context of a complex crop planning problem has become a pressing concern.

To extend the reach of data mining technology, this study fuses data mining with optimization techniques, including genetic algorithms and orthogonal experiments, to develop a comprehensive optimization model for crop planting decisions. Specifically, first, data mining techniques are employed to collect and analyze the planting conditions, crop yields, planting costs, sales, and other related factors in a village. This is done in order to reveal the intrinsic correlations and influencing factors. Subsequently, a crop planting planning function with the objective of maximizing profit is constructed, and the genetic algorithm is employed to optimize and resolve this function, thereby identifying the optimal planting combination and resource allocation plan. The genetic algorithm is capable of searching globally, which circumvents the limitation of traditional optimization methods that are prone to becoming trapped in local optima. This provides a more reliable foundation for decision-making [3,4,5].

Furthermore, this paper introduces specific constraints inherent to the crop planting process, including land area, climate change, planting cycle, and other factors, thereby enhancing the model's practical applicability [6,7]. Further optimization of parameter configuration and experimental schemes is achieved through the use of orthogonal experimental design and other methods, thereby ensuring the efficiency and robustness of the model. This study offers a scientific foundation for practical crop planting decisions and presents novel approaches for addressing other intricate planning issues.

This paper presents a flowchart of the research process, as shown in Figure 1, which provides a clear illustration of the overall framework and ideas underlying the research. The research is divided into four principal phases:

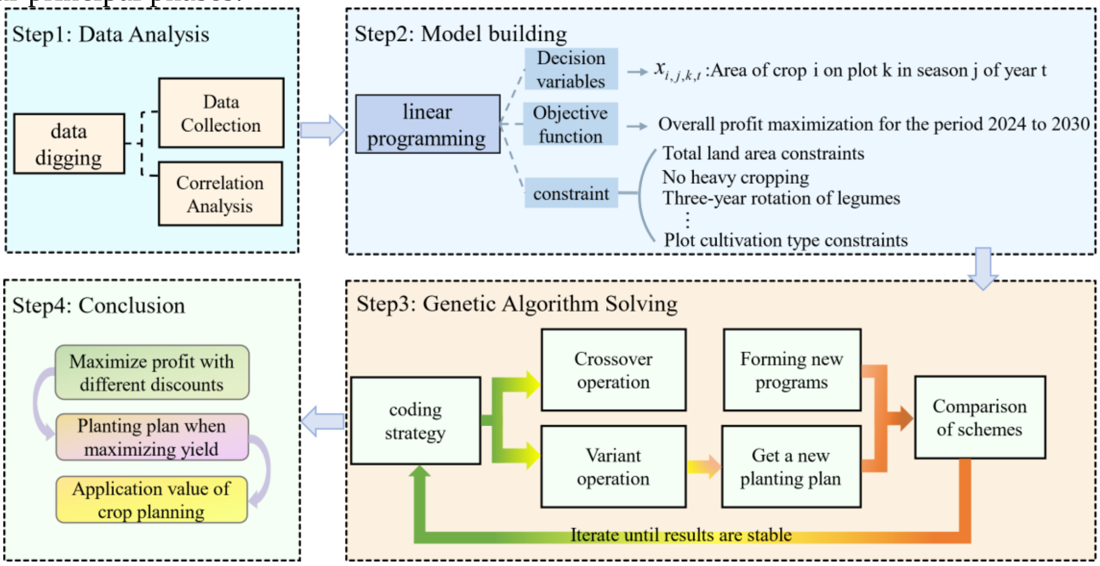

Figure 1: Study process diagram

2. Data mining

2.1 Data Collection and Mining

This study collected data on the yield, cultivated area, per-mu output, value, and total cost of various crops in China over the years. The data was plotted using rice, wheat, corn, and potatoes as examples (see Figure 2). Additionally, the study gathered data pertaining to planting practices from a particular village, encompassing the extent of cultivated land, the types of planting employed, the sales volume of various crops, the per-mu output, the costs associated with planting, and the prices at which crops were sold. Through field surveys and data inquiries, this paper elucidates the multiple constraints inherent in planting plans. In accordance with the patterns of crop growth, it is imperative that each crop not be cultivated continuously in the same field (including greenhouses) to prevent a reduction in yield. Furthermore, given that the root system of legumes can promote the growth of other crops, the planting strategy requires that legumes be planted at least once every three years on each plot, commencing in 2023. Furthermore, the planting plan must consider the convenience of planting and field management. For instance, the area planted to each crop in each season should not be excessively dispersed, and the area planted to each crop in a single plot should not be excessively limited.

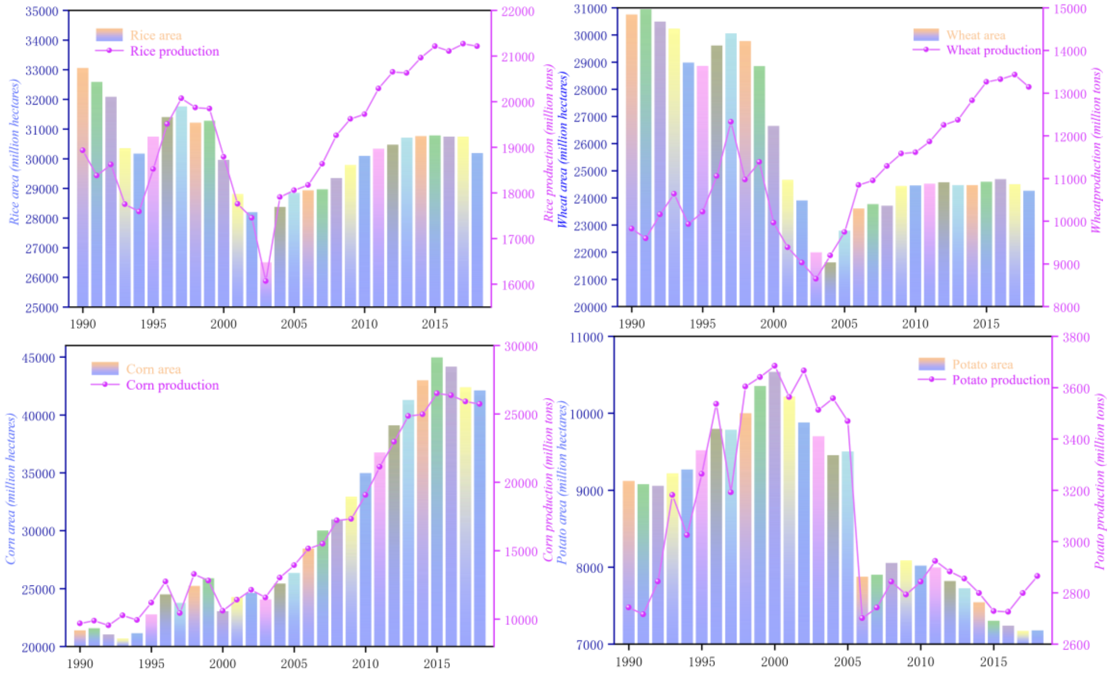

Figure 2: Trends in yields and cultivated areas for rice, wheat, maize and potatoes

The four charts illustrate the trends in the acreage and output of China's primary crops, including rice, wheat, maize, and potatoes, from 1990 to 2018. It can be observed that the acreage of rice exhibits a fluctuating trend, whereas the yield per mu has continued to increase. This suggests that despite fluctuations in acreage, the yield per mu of rice is gradually increasing, which may reflect advancements in agricultural technology. Since the late 1990s, wheat acreage has exhibited a gradual decline, whereas yields have continued to rise, with a particularly marked increase occurring after 2000. This is also associated with enhanced production efficiency and technological advancements. Both the acreage and yields of maize have increased significantly, reflecting China's growing demand for maize in agriculture. This may be related to the expansion of the feed and biofuel industries. The acreage of potato cultivation has demonstrated a downward trajectory, exhibiting a pronounced decline following the year 2000. This phenomenon may be associated with shifts in planting areas.

In summary, these data demonstrate that China's agricultural sector has become increasingly dependent on scientific and technological advancement, sophisticated management practices, and policy guidance over the past few decades. This has led to a significant increase in the output of major crops, while the acreage under cultivation has remained relatively stable or experienced a decline. This shift not only signifies an enhancement in food security but also reflects the transformation and adjustment of the agricultural structure.

2.2 Correlation analysis

In practical terms, there is a clear correlation between expected sales volume and both sales price and planting cost. A linear relationship exists between expected sales volume and per-mu yield; as such, this paper considers the relationship between per-mu yield and the aforementioned variables. In order to fully consider all factors, this paper employs a correlation analysis. Prior to conducting a correlation test, it is essential to perform a normal distribution test on the data to ascertain its normality.

Given the limited number of samples, this study employs the Shapiro-Wilk test to evaluate the normality of annual yield, sowing area, national average per-mu main product yield, total per-mu output value, and total per-mu cost. A P-value greater than 0.05 indicates that the data are normally distributed, whereas a P-value less than or equal to 0.05 indicates that the data do not conform to a normal distribution. A normality test was conducted using data from 2018, and the results are presented in Table 1.

Table 1: Normal distribution test table

	Kolmogorov–Smirnov test		Shapiro-Wilk test	
	Statistics	Significance	Statistics	Significance
Annual output	0.299	0.004	0.761	0.003
Sown area	0.303	0.003	0.744	0.002
Output of main products per mu	0.388	0.000	0.628	0.000
Total output value per mu	0.238	0.058	0.823	0.017
Total cost per mu	0.195	0.200	0.855	0.043

As evidenced by the aforementioned table, the p-values associated with annual output, sowing area, output of main products per mu, total output value per mu, and total cost per mu are all less than the significance level of 0.05. This indicates that these variables exhibit a significant departure from normal distribution, as determined by both the Kolmogorov-Smirnov and Shapiro-Wilk tests. Consequently, the assumption of normality cannot be upheld. This indicates that the data distribution is skewed or exhibits other characteristics that do not conform to a normal distribution. Consequently, statistical analysis methods based on a normal distribution cannot be applied directly. Moreover, it is possible that this non-normality may also be present in the data from other years, which will not be further demonstrated here.

Given that the Pearson correlation coefficient is contingent upon the data obeying a normal distribution, it is not a suitable analytical tool in this case. In this instance, the Spearman correlation coefficient is employed for correlation analysis, with the corresponding heat map presented in Figure 3.

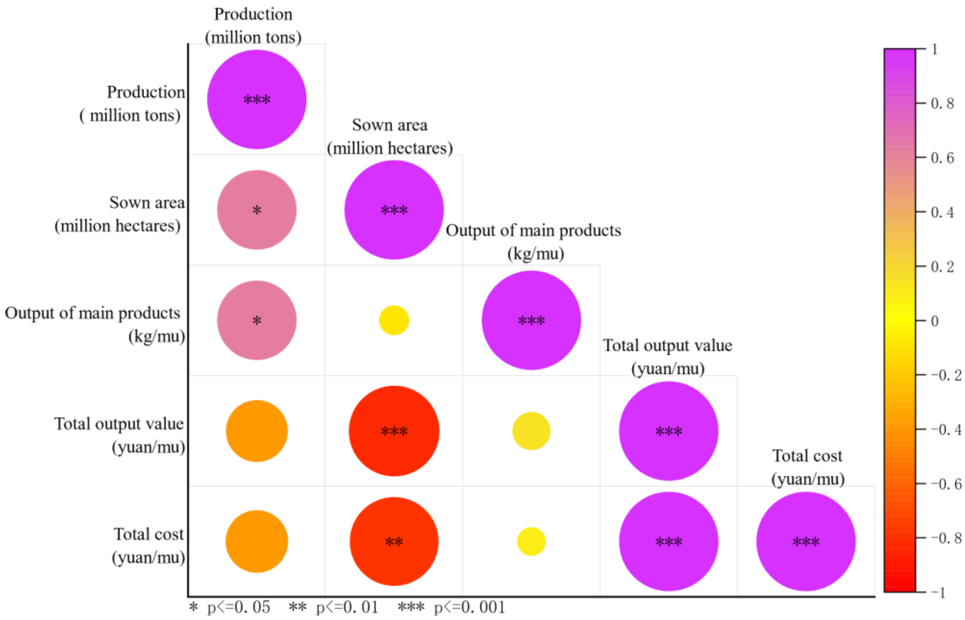

Figure 3: Spearman's correlation coefficient heat map

As illustrated in Figure 3, the heat map depicts the correlation between variables, represented by color and the size of the circles. The color bar ranges from dark purple to orange, indicating the strength of the correlation. The color purple indicates a positive correlation, whereas orange denotes a negative correlation. The size of the circle is indicative of the absolute value of the correlation coefficient, which in turn indicates the strength of the correlation between the two variables. In particular, a notable correlation is evident between yield, sown area, main product yield, total value of production, and total cost.

A highly positive correlation is evident between yield and sowing area, main product yield and total value of production. The correlation is extremely significant $(p \leq 0.001)$, with a dark purple color indicating a strong relationship. This suggests that an expansion in the area sown and an increase in the yield of the main product both exert a beneficial influence on yield and the value of production.

The total cost is negatively correlated with the other variables, particularly with the total value of production and yield. This indicates that as costs increase, yield and income may decrease. Consequently, it can be posited that higher inputs do not always lead to higher returns.

3. Establishment and resolution of planting patterns

The objective of agricultural production in rural regions is to achieve the greatest economic benefit while simultaneously minimizing the inherent risks of farming and facilitating the sustainable growth of the rural economy. Linear programming is an effective method for solving goal optimization problems. In particular, the objective is to direct the formulation of the most advantageous planting plan, whereby the total profit is optimized (i.e., the revenue derived from crop sales is minimized in conjunction with the costs associated with planting). In light of the aforementioned considerations, the model must also take into account potential income adjustments contingent on varying sales scenarios, including instances of slow-moving or discounted agricultural product sales. To this end, the model is designed to adapt the income items included in the objective function, thereby ensuring the identification of an optimal planting plan irrespective of the prevailing sales circumstances. In this regard, this paper proposes the introduction of a planning model to address the aforementioned challenges.

3.1 Decision variable

In the context of linear programming, decision variables are defined as the area of each crop per season in different fields or greenhouses. These variables are all continuous, non-negative real numbers, which satisfy the requirements of the linear programming model.

The decision variable, $x_{i,j,k,t}$, is employed to delineate the distribution and acreage of crops cultivated in disparate seasons across each plot. It represents the acreage of crop i on plot k in season j of year t. It is a valuable tool for optimizing planting decisions, ensuring compliance with plot suitability, seasonal restrictions, and crop rotation requirements, and achieving an efficient allocation of resources.

3.2 Objective function

The objective is to maximize total profit over a seven-year period from 2024 to 2030. This can be achieved by maximizing total crop yield minus cultivation costs.

Total cost of the i crop in year t, $C_{i,t}$:

$$C_{i,t} = C_i \sum_{j=1}^{2} \sum_{k=1}^{54} X_{i,j,k,t} \tag{1}$$

Among them, the unit planting cost of crop i (yuan/mu) is denoted by C_i, and the planting area of crop i in the jth quarter of the tth year (mu) is denoted by $x_{i,j,k,t}$. This formula calculates the total cost by multiplying the total area of crop i in year t on all plots and in all seasons by its unit planting cost.

Sales revenue from the i crop in year t, $R_{i,t}$:

$$R_{i,t} = \min(\sum_{j=1}^{2}\sum_{k=1}^{54} Y_{i,k} \times x_{i,j,k,t}, 0.8D_i) \times P_i \tag{2}$$

This equation represents the sales revenue of crop i in year t. Based on the relevant data for 2023, the sales price of each crop will fluctuate due to market supply and demand. Therefore, 80% of the 2023 production is taken as the expected sales volume and the average value of the sales price range is taken as the forecast sales price to formulate a reasonable planting strategy. First, the sum of the yields of all plots and seasons, $\sum_{j=1}^{2}\sum_{k=1}^{54} Y_{i,k} \times x_{i,j,k,t}$, is calculated, where $Y_{i,k}$ is the unit yield of crop i on plot k. This is then compared with 80% of market demand $(0.8D_i)$ and the smaller of the two is taken, which is the actual sales volume of the crop in the event that it cannot be sold. Finally, the yield is multiplied by the unit selling price P_i to obtain the actual sales revenue.

The objective is to maximise profits:

$$\max(\sum_{t=2024}^{2030}\sum_{i=1}^{41}(R_{i,t} - C_{i,t})) \tag{3}$$

The objective function represents the maximisation of total profit between 2024 and 2030, which is the sum of the difference between sales revenue and costs of each crop each year.

In summary, the model takes into account the possibility of crop obsolescence by calculating costs and sales revenue, ensuring that market demand is fully considered in the production process and that the profitability of growing crops is optimised.

3.3 Constraints

There are several constraints in the problem, which can be expressed as linear equations or inequalities and solved using a linear programming model.

(1) Total area constraint: In the first or second quarter of each year, the total area planted with all crops in each quarter must not exceed the suitable area that can be planted in that quarter.

$$\sum_{i=1}\sum_{k=1} x_{i,j,k} \le \sum_{k=1} S_k, \quad \forall i, j, k \tag{4}$$

$\sum_{k=1} S_k$ is the total area of all plots. When $j = 1$, the total area available for planting crops in the first season is expressed as $\sum_{i=1}\sum_{k=1} x_{i,1,k}$; when $j = 2$, it represents the total area available for planting crops in the second season, expressed as $\sum_{i=1}\sum_{k=1} x_{i,2,k}$.

(2) The same crop may not be planted continuously in the same plot or greenhouse. For crops that can be grown in the first season only, they shall not be replanted in the same plot in the following two consecutive years; for crops that can be grown in the first and second seasons, they shall not be replanted in the same plot in the following two consecutive seasons.

Where L_1 indicates a set of plots suitable for growing crops in the first season only, and L_2

indicates a set of plots suitable for growing crops in both the first and second seasons. $x_{i,j,k,t}$ indicates the area of plot k in the jth quarter of the year where crop i is planted.

$$x_{i,j,k,t} \times x_{i,j,k,t-1} = 0, \quad \forall k \in L_1, j,k,t \geq 2024 \tag{5}$$

When $k \in L_1$ and $j=1$, the planting area $x_{i,j,k,t}$ in the first and second quarter of the same plot in year t must satisfy the condition that at least one of them is 0.

$$x_{i,1,k,t} \times x_{i,2,k,t} = 0, \quad \forall k \in L_2, j,k,t \geq 2024 \tag{6}$$

(3) Three-year rotation restriction for pulse crops: Legume crops must be grown on the same parcel at least once every three years. The letter J indicates a group of legume crops.

$$\sum_{i \in J} \sum_{j=1}^{2} \sum_{s=0}^{2} x_{i,j,k,t-s} > 0, \quad \forall k,t \geq 2025 \tag{7}$$

(4) Single plot area limit: The planting area of each crop in a single plot must not be less than the set minimum area $S_{\min}$ (0.2 acres) and must not exceed the maximum area of the plot. When the planting area reaches the maximum value of the plot, it means that the plot is completely used for planting a single crop.

$$S_{\min} \leq x_{i,k} \leq S_k \tag{8}$$

Among them, $S_{\min}$ is the minimum area of crop i on plot k, and the area described by the area of plot k is S_k.

(5) Avoid over-fragmentation: To ensure that planting is concentrated, you can set the number of plots for each crop per season not to exceed the set value M. Introduce 0-1 variables:

$$y_{i,j,k,t} = \begin{cases} 1, & S_{min} \leq x_{i,j,k,t} \leq S_k \\ 0, & x_{i,j,k,t} = 0 \end{cases} \tag{9}$$

The area planted with crop i on plot k or in greenhouse k in year t is defined as $y_{i,j,k,t} = 1$ when the following constraints are satisfied; otherwise, $y_{i,j,k,t} = 0$:

$$\sum y_{i,j,k,t} \leq M, \quad \forall i,j,t \tag{10}$$

Where $\sum y_{i,j,k,t}$ is the quantity of all crop types grown on the kth plot or greenhouse in the tth year, up to a maximum of M.

(6) Seasonal restrictions: Vegetable growing is affected by season. Some vegetables can be grown in both the first and second seasons, while others are limited to one season. Except for Chinese cabbage, white radish, and carrots, other vegetables can be grown in one or two seasons; Chinese cabbage, white radish, carrots, and edible mushrooms can only be grown in the second season.

$$x_{i,1,k,t} = 0, \quad if \quad i = 35, 36, \cdots, 41 \tag{11}$$

(7) Restrictions on crop types for parcels: Different types of land parcels have specific requirements for growing crops. Flat land, dry land, terraced land, and hillside land are suitable for growing grain crops other than rice for only one season each year; irrigated land is suitable for growing rice for one season or vegetables for two seasons, with Chinese cabbage, radish, or carrot being planted in the second season. Ordinary greenhouses are suitable for growing two crops per year,

with the first season for growing vegetables other than Chinese cabbage, radish or carrot, and the second season for growing edible mushrooms only. Smart Greenhouses are suitable for growing two seasons of vegetables other than Chinese cabbage, radish or carrot.

$$x_{i,j,k,t}=0, \quad if \quad i \notin L_k \tag{12}$$

L_k represents the set of crops suitable for planting on the kth plot. If a crop is not included in L_k, the area planted with that crop on the kth plot, $x_{i,j,k,t}$, is recorded as 0.

3.4 Model overview

To summarize, the following linear programming model is established:

$$\max \sum_{t=2024}^{2030}\sum_{i=1}^{41}\left[\min(\sum_{j=1}^{2}\sum_{k=1}^{54}Y_{i,k}\times x_{i,j,k,t},0.8D_i)\times P_i - C_i\sum_{j=1}^{2}\sum_{k=1}^{54}x_{i,j,k,t}\right] \tag{13}$$

$$s.t.\begin{cases}\sum_{i=1}\sum_{k=1}x_{i,j,k}\le\sum_{k=1}S_k, & \forall i,j,k\\ x_{i,j,k,t}\times x_{i,j,k,t-1}=0, & \forall k\in L_1, j,k,t\ge 2024\\ x_{i,1,k,t}\times x_{i,2,k,t}=0, & \forall k\in L_2, j,k,t\ge 2024\\ x_{i,2,k,t}\times x_{i,1,k,t-1}=0, & \forall k\in L_2, j,k,t\ge 2024\\ \sum_{i\in J}\sum_{j=1}^{2}\sum_{s=0}^{2}x_{i,j,k,t-s}>0, & \forall k,t\ge 2025\\ S_{min}\le x_{i,k}\le S_k & \\ \sum y_{i,j,k,t}\le M, & \forall i,j,t\\ x_{\mathrm{i,2,k,t}}=0, & if\ i=35,\cdots,41\\ x_{i,j,k,t}=0, & if\ i\notin L_k\end{cases} \tag{14}$$

3.5 Model solving based on improved genetic algorithms

In the agricultural production process, factors such as weather changes, fluctuations in market demand, and uncertainty in crop yields can have a significant impact on yields. This kind of dynamic and uncertain nature is difficult to fully address with traditional linear programming models. In addition, the problem of rural crop planning involves multiple decision variables and complex constraints, such as multiple plots, multiple crops, multiple seasonal crop rotation constraints, legume rotation, and plot area allocation. The problem is multi-periodic and high-dimensional.

Genetic algorithms are intelligent optimization algorithms suitable for handling large-scale, complex, multi-objective optimization problems. They can effectively search for global optimal solutions in complex constraints and dynamic environments. Therefore, this paper uses genetic algorithms to solve the crop planting planning problem in order to improve the optimization results and deal with its complexity and uncertainty.

By simulating the processes of natural selection, crossing, and mutation, the genetic algorithm effectively optimizes the planting season and area of crops to maximize total profit. The algorithm uses a fitness function to select planting plans with excellent performance in the objective function. These high-quality plans have a higher probability of being passed on to the next generation, gradually approaching the optimal solution. After replacing several sets of parameters, the optimal parameters were selected based on a comprehensive consideration of factors such as running time, spatial cost, prevention of excessive mixing of feasible decoding, and accuracy of results. The optimal parameters are listed in Table 2.

Table 2: Genetic algorithm optimal parameter table

Parameter name	Population size	Crossover probability	Mutation probability	Genetic generation
parameter value	50	0.7	0.3	800

First, a set of initial planting plans is randomly generated and evaluated using a fitness function to screen out the better ones. Then, the planting strategies are further optimized by selecting the plans that passed the fitness evaluation again based on the objective function. On this basis, the selected plans are crossed with a crossover probability of 0.7 and mutated with a mutation probability of 0.3 to generate two new planting plans. Then, among the three newly generated plans, the 50 with the best performance are selected and the above process is repeated. The optimization process is iterated 800 times to finally obtain the optimal planting plan.

The selection operation selects the most profitable planting plan from the current population according to the fitness function, so that it has a greater chance of survival and reproduction in the next generation, thereby improving the profitability of the plan. The crossover operation generates new solutions by swapping the planting information in the selected solution, thereby improving the global search capability and expanding the scope of exploration in the solution space. The mutation operation introduces random mutations into the newly generated plan, increases the diversity of the population, helps avoid situations where the plan satisfies only the locally most profitable situation, and improves the local search ability.

Through the effective combination of crossover and mutation, the genetic algorithm combines global search with local search, can comprehensively explore the vast solution space, gradually optimize the planting plan, and finally find the optimal planting season and area arrangement to maximize planting benefits.

(1) Coding strategy: The decision variable $x_{i,j,k,t}$ is affected by multiple factors such as land parcel, year, season, and crop. This multi-dimensionality makes it difficult to directly use traditional coding methods to represent the decision variable. In this case, using multi-matrix chromosome coding can more effectively represent the decision variable. Matrix chromosome encoding makes it easier to perform operations such as crossover and mutation in genetic algorithms. By encoding different decision variables into their own matrices, we can operate genetic algorithms with more flexibility. This method decomposes the complex search space into multiple smaller subspaces, which helps the genetic algorithm search for the optimal solution more effectively. Each matrix can be dedicated to a specific aspect of the decision variable, making the search more precise and targeted.

(2) Crossover operation: After evaluating the solutions by fitting the function, the better solutions are selected for crossover operation. In the selected solution, a part of any year or plot is crossed and swapped with a part of another selected solution to generate a new solution.

(3) Mutation operation: In the mutation operation, the optimal solution obtained by evaluating the fitness function is changed according to a predefined mutation probability. When it is decided to mutate a solution, one year or plot in that solution is randomly selected for adjustment. For each plot, the following operations are performed with a certain probability:

If the random number is less than 0.5, a new value is randomly generated according to the location of the plot, and it is assigned to the corresponding position in the new solution with an upper limit of the area, achieving a random value addition operation. If the random number is greater than 0.8, the value of the corresponding position in the new solution is set to 0, which is a random value deleting operation.

However, since this operation may generate infeasible protocols or poor-quality protocols (with lower returns), it is necessary to comprehensively evaluate the newly generated protocols and the original protocols after crossover and mutation are completed, screen out the better protocols, and further iteratively optimize them to improve the quality of the final solution.

(4) Iteration: After extensive evaluation and screening, the outstanding individuals enter the next generation group. Iteration repeats the above steps until the stopping condition is reached, gradually optimizing the quality of the solution and the convergence speed.

3.6 Model solution results

The genetic algorithm simulates the mechanisms of natural selection and genetic variation to continuously iteratively optimize the planting plan so that crop yields gradually increase and approach the global optimal solution. Figure 4 shows the selling profit curve with the number of iterations under different selling discounts. The three curves in the figure correspond to the changing trends of crop yields with iterative optimization under different selling strategies. As can be seen from the figure, the profit curves under different sales discount strategies show a gradual convergence trend, indicating that after a certain number of iterations, the genetic algorithm can effectively find a planting plan that maximizes revenue.

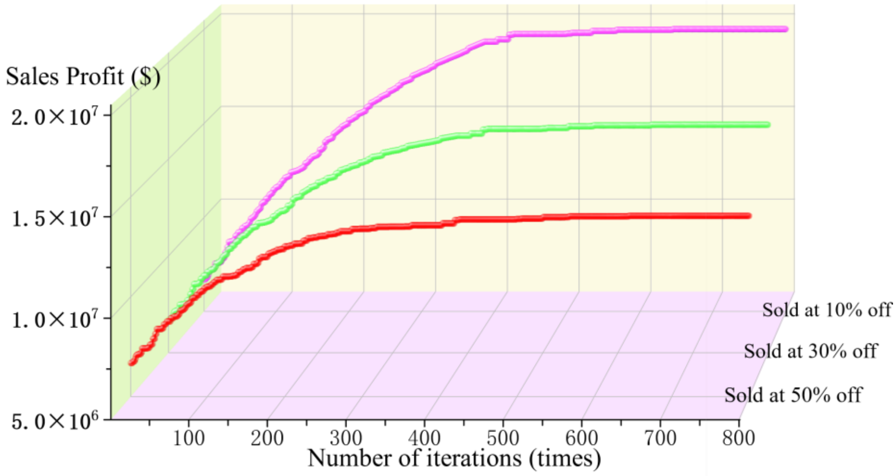

Figure 4: Profit curve as a function of the number of iterations.

As the number of iterations of the genetic algorithm increases, the sales profit gradually stabilizes. This shows that under given constraints, the genetic algorithm can find a local optimal solution, thus bringing the planting structure and marketing sales strategy to a state of equilibrium. In particular, under the 90% discount sales strategy and the 70% discount sales strategy, the profit reaches a plateau earlier, indicating that the sales strategy at this time is more reasonable and the optimization effect is more significant.

The three curves in Figure 3 represent different sales discount strategies. The figure shows that for the same number of iterations, the 10% discount strategy yields the highest profit, followed by the 30% discount strategy, while the half-price discount strategy yields the lowest profit. This shows that while higher discounts (such as 30% and half-price discounts) may increase sales, they do not have the same effect on total profit as a relatively stable pricing strategy (such as a 10% discount). Therefore, an appropriate discount strategy plays an important role in improving the overall profitability of crops.

In the first 100 iterations, all three profit curves show a rapid growth rate, indicating that the genetic algorithm is effective in improving the planting plan in the early stages. However, as the number of iterations increases, the growth rate of the curves gradually slows down and stabilizes around 800 iterations. This pattern of change reflects the process by which the genetic algorithm gradually approaches the optimal solution. According to the characteristics of the genetic algorithm, the rapid growth in the early stage is due to the high diversity of individuals in the population, which can search a wide range of solution spaces. In the later stage, the curve gradually converges, indicating that the individuals in the population tend to agree and the space for improvement decreases.

Although the genetic algorithm demonstrated strong optimization capabilities in this study, the convergence of the profit curve shows that the profit curve is more volatile when the discount strategy is low (such as a half-price sale). This may mean that under certain specific conditions, the local optimal solution of the algorithm will be disturbed by external factors (such as fluctuations in market demand, fluctuations in crop yields, etc.). Therefore, in practical applications, it may be necessary to include more external factors or use multi-objective optimization methods to further improve the stability and applicability of the algorithm.

4. Conclusions

The planting decision optimization model based on data mining, genetic algorithm and orthogonal experiments proposed in this study has achieved remarkable results in both theoretical innovation and application practice. In terms of theoretical innovation, the model integrates several analysis techniques such as data-driven analysis, heuristic optimisation and experimental design to form a novel decision support framework. On the one hand, data mining provides the basis for analysing the intrinsic relationship of key factors in planting decisions, which lays a solid foundation for subsequent modelling. On the other hand, genetic algorithms and orthogonal testing give the model the ability to optimise the solution and uncertainty analysis, so that it can be better adapted to the complex reality of crop production. The organic combination of several techniques expands the application scope of data mining in the field of decision optimisation. In terms of application practice, the model makes the optimal decision according to the specific planting situation of a village, which not only provides a reliable decision basis for actual production, but also provides a reference method for solving other agricultural planning problems. In the future, we can further explore how to integrate more optimisation algorithms into the data-driven decision analysis framework so that it can play a role in a wider range of application scenarios.

References

[1] Li Bingmei, Jiang Fengxia. Greenhouse vegetable cultivation technology and the development trend of smart agriculture[J]. Agricultural Engineering Technology, 2023, 43(5): 24-25.

[2] Liu Yan, Wang Wei, Lu Yang, et al. Climatic suitability of the kale-radish-cabbage three-harvests-a-year scale planting model in northern Xianyang[J]. Journal of Agronomy, 2023, 13(4): 66.

[3] Zhang Hua, Long Cheng, Hu Siyang, et al. Topology verification of distribution networks based on hierarchical clustering method and Pearson's correlation coefficient[J]. Power System Protection and Control, 2021, 49(21): 88-96.

[4] Zhu Hainan, Wang Juanjuan, Chen Bingbing, et al. multi-objective planning of integrated electricity-gas energy system considering economy and carbon emission[J]. Journal of Shanghai Jiao Tong University, 2023, 57(4): 422.

[5] Huang Shuzao, Tian Junwei, Qiao Lu, et al. UAV path planning based on improved genetic algorithm[J]. Computer Applications, 2021, 41(2): 390.

[6] Xu Li, Liu Yunhua, Wang Qifu. Application of adaptive genetic algorithm to robot path planning[J]. Journal of Computer Engineering & Applications, 2020, 56(18).

[7] Vadhwani D , Thakor D .Prediction of extent of damage in vehicle during crash using improved XGBoost model[J].International journal of crashworthiness, 2023, 28(3/4):299-305.DOI:10.1080/13588265.2022.2075101.