Research on Crop Planting Strategies Based on K-means Clustering and Optimisation Models

Tianyu Yuan *

College of Arts and Sciences, China University of Petroleum-Beijing at Karamay, Karamay, China

* Corresponding Author Email: xiaoyz000@qq.com

Abstract. This paper investigates the strategy of optimisation model for rural crop cultivation with a view to develop the best cultivation plan for the next seven years. An optimisation model with the objective of profit maximisation has been developed by taking into account the expected sales volume, acreage, cost of cultivation and selling price of various crops and their uncertainty risks. The model incorporates the complex constraints of the actual production process and takes into account multiple uncertainties such as changes in market dynamics and fluctuations in climatic conditions by analysing four scenarios, namely optimal, suboptimal, intermediate and stochastic, which makes the model much more practical. In addition, this paper applies the K-means clustering method to explore the substitution and complementary relationship between different crops, which makes the model more in line with the actual market law. The results show that a reasonable planting strategy can not only enhance the economic benefits of crops, but also promote the efficient use of resources and sustainable development of the environment. The optimisation model proposed in this paper provides scientific guidance for agricultural decision makers and provides a reference for the production strategies of other products. Future research can be further extended to other regions and crop types to improve the overall strategic layout of agricultural production and promote the high-quality development of agriculture.

Keywords: Nonlinear planning models, Gaussian noise, K-means clustering, Spearman correlation coefficient.

1. Introduction

The sustainable development of agriculture cannot be separated from the planting strategy of crops, so the optimization of the planting strategy of crops is of strategic significance for the development of agriculture. In this paper, we study the strategy of a rural planting optimization model to develop the optimal planting plan for the next 7 years by considering the expected sales volume, mu yield, planting cost and selling price of various crops and their uncertain risks. Firstly, this paper establishes an optimization model with profit maximization for the next 7 years as the optimization objective by incorporating the complex constraints of the real production process. Then, through the discussion of the four scenarios of optimal, inferior, intermediate, and stochastic situations, multiple uncertainties such as changes in market dynamics and fluctuations in climatic conditions are considered in the planning model, which greatly enhances the practicality of the model. Finally, with the help of K-means clustering method, this paper discusses the substitution and complementary relationship between different kinds of crops, which makes the model more in line with the actual market law. The model not only provides decision makers with scientific guidance on crop cultivation, but also provides a reference for the production strategy of other products.

2. Data collection and pre-processing

2.1. Data collection

To explore the application of game theory hybrid particle model in planting strategy, this paper selects a farm in Huabei area as the research object. Through in-depth analysis of the problem, it is concluded that there is the most important relationship between mu yield, planting cost, market sales, sales price and other factors and planting profit. Therefore, this paper first collects and collects the

corresponding data of yield per mu, planting cost, average sales unit price, and expected sales volume of each crop in North China [1].

2.2. Data preprocessing

(1) Missing value handling

In the original data, although theoretically there were 18 vegetables that could be grown in two seasons, the actual records showed that only eight of them (such as auberge, green pepper, baby bok choy, cucumber and lettuce, etc.) had records of cultivation in the second season of the smart greenhouses, while the cultivation status of the remaining 10 species in the second season was missing, making it impossible to directly project the sales volume of these species in the second season.

Therefore, it was confirmed that these 10 vegetables had no planting records in the second season, and therefore they were regarded as vegetables planted in one season only in the actual calculation.

(2) Data addition and logical division

The principle of 'grow in season, sell in season' was followed to ensure that the sales volume of vegetables in the two seasons were counted separately and not combined. For the eight vegetables recorded in the second season, for more detailed analysis, they are split into two separate categories for both seasons (e.g. auberge 1 for the first season and auberge 2 for the second season) and assumed to be grown in equal quantities in each season, so as to simulate the complete two-season growing data.

Considering that both watered plots and greenhouses have the ability to grow in two seasons, for simplicity of calculation, we assume that these plots are divided equally into two plots, each of which grows one season's crop independently.

(3) Data quantification and renumbering

To uniform management and analysis, all the vegetable types involved were renumbered to ensure that each vegetable (including its first and second season of cultivation) had a unique identifier. Vegetables such as auberges, peppers, baby bok choy, cucumbers, lettuce, etc. were further subdivided into two categories based on the growing season (e.g., auberge 2, peppers 2, etc.) and given new numbers (42-47) to facilitate differentiation and management.

(4) Data integration and preparation of analyses

The acreage data on the cultivation of crops in 2023 are integrated with the sales, costs, and other relevant data from the same year's statistics to ensure that all the information needed for the analysis is in the same data set.

During the integration process, any possible data inconsistencies or errors are checked and corrected to ensure the accuracy and reliability of the data.

The integrated data set will be used as the basis for subsequent calculations of expected sales volumes, growing costs and other key indicators for further analysis.

3. Establishment and solution of planning model

This section gives the optimal cropping scenario for the crop in 2024-2030 based on a 50% reduction in the sales price in 2023, which requires a new decision variable to be established to compare the size of production with the expected sales volume, so the optimization process is as follows:

(1) Establishment of decision variables

Let the total production for each of the next seven years be $Q_k = \sum_{i=1}^{82}\sum_{j=1}^{47} n_{ijk}x_{ijk}$.

(2) Establishing optimization objectives

$$maxZ_2 = \sum_{j=1}^{47}\sum_{k=2024}^{2030}\left(a_j min\left\{m_j, \sum_{i=1}^{82} n_{ijk}x_{ijk}\right\} - \sum_{i=1}^{82} c_{ij}x_{ijk}\right) - 0.5\left(Q_k - m_j\right) \quad (1)$$

(3) Establishment of constraints and model solution

In actual agricultural production, if there is a crop production in a certain year is much larger than its sales, the profit of planting may be a significant decline in the situation, so here to add a new

constraint, its total agricultural production in each of the next seven years should be less than or equal to 1.2 times the expected sales volume, then there is:

$$Q_k \leq 1.2m_j, j = 1, \cdots, 47 \tag{2}$$

The rest of the constraints are consistent with the first case, so in summary the following planning model is obtained:

$$maxZ_2 = \sum_{j=1}^{47}\sum_{k=2024}^{2030}\left(a_j min\left\{m_j, \sum_{i=1}^{82} n_{ijk}x_{ijk}\right\} - \sum_{i=1}^{82} c_{ij}x_{ijk}\right) - 0.5\left(Q_k - m_j\right) \tag{3}$$

$$s.t. = \begin{cases} Q_K \leq 1.2m_j, j = 1, \cdots, 47. \\ m_j = \sum_{i=1}^{82}\sum_{j=1}^{41} n_{ijk}x_{ijk}, k = 2023. \\ a_j = 0.5\left(max_{a_j} + min_{a_j}\right). \\ \sum_{i=1}^{82} y_{ijk} \leq 3, j = 1, \cdots, 47, k = 2024, \cdots, 2030. \\ \sum_{k}^{k+3} x_{ijk} > 0, i = 1, \cdots, 42, j \epsilon beans \\ \\ \sum_{i=1}^{82} x_{ijk} < s_i, j = 1, \cdots, 47, k = 2024, \cdots, 2030. \end{cases} \tag{4}$$

By solving the above, the optimal value of profit from the reduced-price sales program for the next 7 years is found to be 42,469,000 yuan [2-4].

4. Establishment and solution of dynamic programming model

4.1. Model establishment

The cultivation and marketing of various types of crops face significant challenges due to multiple uncertainties such as changing market dynamics and fluctuating climatic conditions. Specifically, these uncertainties are not only reflected in the expected sales volume, mu yield, planting costs and sales prices of crops, but also accompanied by planting risks that cannot be ignored. In the case of wheat and maize, for example, while these two crops have shown year-on-year growth in sales, the average annual growth rates are difficult to predict with any precision and can only be estimated within a wide range of 5% to 10%. Meanwhile, expected sales of other crops also face potential fluctuations of about 5 per cent up and down compared to 2023.

To cope with these complex and variable uncertainties and potential risks, the optimal planting scenarios for the period 2024 to 2030 were developed by considering the expected sales volume, acreage yield, planting costs, and selling prices.

To accurately quantify and analyze the range of fluctuations in the four core economic indicators of expected sales volume, acreage yield, planting cost and selling price, a quartile method was used in the academic study of planning for the next seven years, in which the fluctuation ranges of the indicators were subdivided into four well-defined scenarios: the optimal situation, the inferior situation, the intermediate situation, and the random situation.

(1) Optimization Scenario:

This scenario represents the theoretically optimal boundaries for each economic indicator, i.e. maximizing sales volume, peaking acre yield, minimizing planting costs, and optimizing sales price. The scenario is constructed to assess the potential benefits and efficiency of the planting activity under ideal market and optimal production conditions.

(2) Inferior Situation:

As the antithesis of the optimal scenario, the inferior scenario depicts the performance of the indicators under extremely unfavorable conditions, including a sharp drop in sales volume, a decline in acreage, a spike in planting costs, and a low sales price. The analysis of this scenario is crucial for identifying the risk points of the planting activities and developing corresponding risk management strategies.

(3) Intermediate Scenario:

This scenario is located between the optimal and the inferior and reflects a more general and realistic growing environment based on the degree of fluctuation of planting indicator data. The intermediate scenario provides a standard reference point for planting program development and helps to ensure the viability of the program.

(4) Stochastic Context:

To match the dynamics of the real world more closely, we introduced two stochastic scenarios that simulate random fluctuations in market and production conditions between optimal and intermediate, and between intermediate and inferior, respectively. This stochasticity not only increases the complexity of the model, but also improves its ability to predict the uncertainty of future planting activities. The analysis of stochastic scenarios allows us to assess the stability and relevance of cropping programs under different levels of market volatility.

4.2. Model solution

In summary, the optimal cultivation program for the next seven years can be planned by considering these four scenarios together. This scenario will effectively guide growers to make scientific judgements in the complex and changing market environment, in order to achieve long-term sustainable development of planting activities.

The stochastic scenario is effectively simulated by adding normally distributed random numbers to determine the optimal percentage of the program, and the percentage distribution of fluctuations in the five scenarios for each indicator type is shown in Table 1:

Table 1. Percentage of fluctuations in five scenarios for each indicator type.

	optimal situation	random situation	intermediate situation	random situation	shitty situation
Sales volume (wheat and maize)	10%	10% 75%	7.5%	7.5% 5%	5%
Sales volume (other)	5%	5% 0%	0%	0% -5%	-5%
yield per acre	10%	10% 0%	0%	0% -10%	-10%
Cost of cultivation	4%	4% 5%	5%	5% 6%	6%
Food prices	1%	1% 0%	0%	0% -1%	-1%
Vegetable price	6%	6% 5%	5%	5% 4%	4%
Edible Mushroom Prices	-1%	-1% -3%	-3%	-3% -5%	-5%
Price of morel mushrooms	-4%	-4% -5%	-5%	-5% -6%	-6%

The optimal values obtained for the five scenarios for the next 7 years are shown in Table 2:

Table 2. Table of optimal profit values for five scenarios over the next seven years.

	optimal situation	random situation	intermediate situation	random situation	shitty situation
profitable optimum	55455000	47914000	42317000	35772000	29107000

4.3. Analysis of model results

Among the different scenarios, the optimal value of profit in the optimal scenario differs from the inferior scenario and the random scenario 2, but the difference between the optimal scenario and the optimal value of profit in the random scenario 1 is relatively small.

According to the analysis, the optimal value of profit in the intermediate scenario is 42317000 yuan, which is significantly higher than the optimal value of profit in the first scenario in the first question, 39874000 yuan, and the difference between the optimal value of profit and the random scenario is relatively small.

In different scenarios, the difference between the optimal scenario and the inferior scenario is large, but the difference between the optimal scenario and the intermediate scenario is relatively small.

Despite the gap between the intermediate and optimal contexts, the profit optimal value of the intermediate context is still higher than the other cases. This suggests that the intermediate scenario provides a valuable reference point for developing a planting program, and that setting the optimal

planting program in and around the intermediate scenario is a more prudent choice due to the variations in market, climate, costs, prices, and other factors, as well as planting risks and uncertainties among crops.

To ensure the reliability of the planting program, it is advisable to set the optimal planting program in and around the intermediate scenario. Such a scenario would help to address risks such as market and climate and increase the reliability and stability of the cropping program.

5. Establishment and solution of alternative and complementary model

Then this paper in the previous dynamic planning model based on the need to consider a variety of crops there may be a certain degree of substitutability and complementarity between the expected sales volume and sales price, planting costs there is a certain degree of correlation between the two points. The following is a refinement of the substitutability, complementarity, and correlation.

If there are two crops, A and B, when the price of crop A increases and is significantly greater than the price of crop B, consumers will prefer to buy crop B to reduce the cost of consumption, resulting in a significant increase in the sales of crop B. Therefore, if there are two crops, A and B, consumers will prefer to buy crop B to reduce the cost of consumption. Therefore, there is a clear substitutability between crops A and B. This substitutability relationship is usually expressed as price elasticity in supply and demand models in the market.

If two crops, A and B, are cultivated in the same area at the same time, due to the mutually beneficial relationship between the two crops in terms of absorbing or metabolizing substances, the acreage yield of the two crops, A and B, will eventually increase significantly, which is called the existence of complementarities between the two crops.

Correlation describes the trend in the degree of change in two variables. In this case, the correlation between the expected sales volume and the sales price and cost of planting means that these variables will change together. Correlations are categorized as positive or negative, with a positive correlation meaning that when one variable is increasing, the other variable is also increasing, and a negative correlation meaning that when one variable is increasing, the other variable is decreasing.

5.1. Model establishment

(1) Cluster analysis

Based on the above refinement analyses of substitutability and complementarity respectively, all crops are here classified in a refined way, so the following K-means clustering method is used to cluster the data of the three indicators, namely, expected sales volume, sales price, and planting cost, of all crops three times.

K-means clustering is a widely used unsupervised learning algorithm for dividing a dataset into several non-overlapping clusters, so that data points within the same cluster have a high degree of similarity, while data points between different clusters have a large degree of difference.

K-means clustering is the method of minimizing the intra-cluster sum of squares:

$$J = \sum_{j'=1}^{K'} \sum_{x' \in C_j'} \|x' - \mu_j\|^2 \tag{5}$$

Where C_j' is the jth cluster, μ_j' is the centre of the jth cluster, x' is the data point in the cluster.

From the above K-means clustering method the crops were classified into the following three categories.

Categories 1: black beans, red beans, climbing beans, millet, buckwheat, peas, barley, rice, kidney beans, spinach, peppers, cauliflower, cabbage, oleander, cucumbers, lettuce, peppers, cabbage, yellow cabbage, celery, elm mushrooms, shiitake mushrooms, white mushrooms, morel mushrooms.

Categories 2: soya beans, mung beans, cereals, sorghum, pumpkin, sweet potatoes, cowpeas, beans, potatoes, tomatoes, auberges, baby bok choy, carrots.

Categories 3: wheat, maize, cabbage, white radish.

(2) Description of Spearman correlation analysis



Figure 1. Correlation heat map of grain vegetable edible mushroom crops.

Spearman correlation analysis is a statistical method used to assess the monotonic relationship between two variables and does not require the relationship to be linear, but rather assesses whether there is a consistent and monotonically varying relationship in the ranking of the variables. This makes Spearman correlation analysis suitable for non-normally distributed data or where the relationship between the variables is unclear and is therefore a non-parametric test [5, 6].

The Spearman correlation coefficient (usually denoted by) is calculated based on the ranking of the variables rather than the original values. Each value in the data is first replaced with its ranking and then the correlation coefficient between the rankings is calculated.

By solving for the Spearman correlation coefficients for the three metrics, the following three correlation heat maps were created shown in Figure 1.

The correlation heat map for food crops shows that there is a negative correlation between expected sales volume and sales price, and a somewhat positive correlation between expected sales volume and planting costs. The correlation heat map for vegetable crops shows that there is a slight negative correlation between expected sales volume and sales price, and a slight negative correlation between expected sales volume and planting cost, both of which are weak. From the heat map of the correlation of edible mushroom crops, there is a negative trend between the expected sales volume and sales price, and a slightly positive trend between the expected sales volume and planting cost.

Through the K-means clustering method to draw the clustering results analysis, here are three categories of various crops within the expected sales volume, unit sales price, cost of averaging process, so that the three categories of each type of each crop's expected sales volume, unit sales price, cost of the three indicators are consistent, so that the different types of crops can be ignored the differences in these indicators, and therefore the indicators of the same crops can be replaced by each other to meet the substitutability. Substitutability.

5.2. Model solution

From the above correlation analysis, the three heatmaps as a whole show that sales volume is positively correlated with cost and negatively correlated with unit price. Here, a perturbation coefficient is set as $\eta, \eta = 0.05$. Let the sales volume and cost increase by 0.05, and the unit price decrease by 0.05 to solve the problem. The objective function and constraints are shown below:

$$maxZ_1 = \sum_{j=1}^{47}\sum_{k=2024}^{2030}\left(a_j min\left\{m_j, \sum_{i=1}^{82} n_{ijk}x_{ijk}\right\} - 1.05\sum_{i=1}^{82} c_{ij}x_{ijk}\right) \tag{6}$$

$$s.t. = \begin{cases} m_j' = \sum_{i=1}^{82}\sum_{j=1}^{41} n_{ijk}x_{ijk}, k = 2023 \\ a_j' = 0.5\left(max_{a_j} + min_{a_j}\right) \\ \sum_{i=1}^{82} y_{ijk} \le 3, j = 1,\cdots,47, k = 2024,\cdots,2030 \\ \sum_{k}^{k+3} x_{ijk} > 0, i = 1,\cdots,82, j\epsilon bean^{\circ} \\ \sum_{i=1}^{82} x_{ijk} < s_i, j = 1,\cdots,47, k = 2024,\cdots,2030 \end{cases} \tag{7}$$

Where$a_j' = 0.95a_j, m_j' = 0.95m_j$.

Taking all these factors into account, the optimal value of the profit for the next 7 years is 47,325,000 yuan.

5.3. Sensitivity analysis

For the sensitivity analyses conducted for the model models described above, we delved into the specific impacts of key parameters or variables in the models on the objective function (e.g., total returns), focusing on the volume of sales of various crops as the central object of analysis.

First, the objective of the analysis was defined: to quantify the extent to which changes in sales volume affect total returns. To achieve this objective, the choice was made to add Gaussian noise to the sales volume data for each crop as a way of modelling external factors such as market fluctuations, changes in consumer preferences or supply chain uncertainty. Here, a standard deviation of 100 was preset to assess the robustness of the model.

Subsequently, five separate additions of Gaussian noise were made to the sales volume for each crop, and after each addition the model was run, and the results of total returns were collected. This step ensured that the model response was captured under different simulation scenarios and the sensitivity analysis plotted is shown in Figure 2:

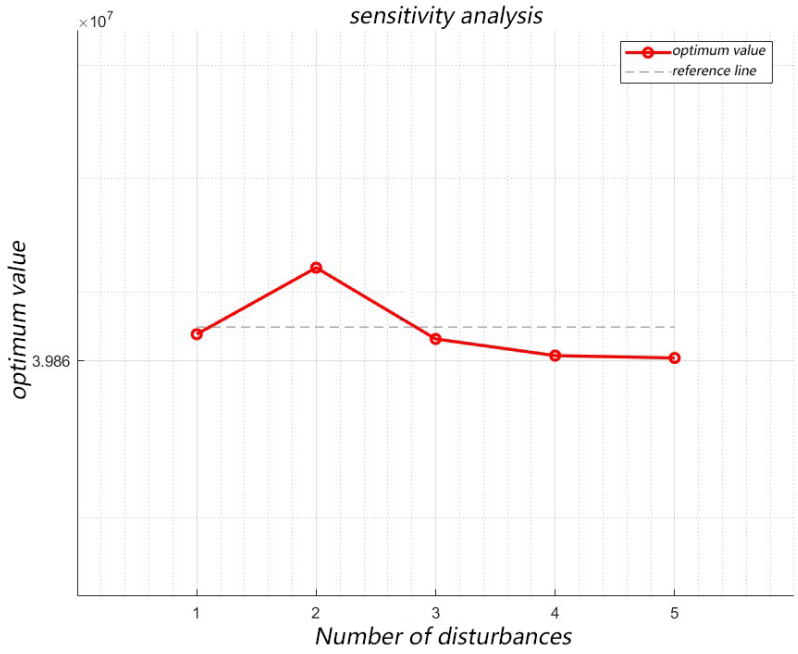

Figure 2. Curve of changes in the correlation between expected sales volumes and sales prices of food crops, vegetables, and edible mushroom crops.

As shown in Figure 2, although the sales volume changes after each addition of noise, the fluctuation range of the total return curve is relatively small and mainly concentrated around the optimal value. This result shows that the model is robust to changes in sales volume, i.e., the model can keep the total return relatively stable in the face of market fluctuations and is not easily affected by the dramatic fluctuations of a single factor.

6. Conclusion

This study provides a practical strategy for the sustainable development of agriculture in a rural village by establishing a crop planting optimization model that covers a wide range of uncertainty factors. The core of the model lies in the objective of maximizing profits in the next seven years, and it comprehensively considers key factors such as expected sales volume, mu yield, planting cost and sales price, providing a scientific basis for crop planting decisions. Through in-depth analyses of optimal, inferior, intermediate, and stochastic scenarios, the study effectively copes with the uncertainties brought by the market and climate change and enhances the practicality and adaptability of the model.

In addition, the application of K-means clustering method reveals the substitution and complementary relationship between different crops, which provides more accurate market feedback for the optimization of planting strategies. The study shows that a reasonable planting strategy can not only improve the economic benefits of crops, but also promote the efficient use of resources and sustainable development of the environment. In summary, the optimization model proposed in this paper provides an important reference for agricultural decision makers, and future research can be further extended to other regions and crop types to improve the overall strategic layout of agricultural production and promote the high-quality development of agriculture.

References

[1] Jianzhao Feng, Yongqi Pan, Yuesong Xiong, et al. Rice critical fertility identification based on mRMR-XGBoost[J]. Journal of Agricultural Engineering, 2024, 40(15):111-118.

[2] J. Wu. Research on optimization of agricultural plantation structure based on linear programming model [J]. New Farmer, 2024, (03):28-30.

[3] Ye Zhennan. Research on the optimization of agricultural industrial structure in Campsis River Basin based on multi-objective control [D]. Zhejiang University, 2018.

[4] Chen Lei. Research on optimal planting decision based on swarm intelligence algorithm and crop model [D]. Nanchang University, 2024.

[5] He Jing, Li Weiping, Zhang Daiyao, et al. Correlation analysis between grounding resistance and soil volumetric water content based on Spearman's correlation coefficient [J]. Journal of Mountain Meteorology, 2024, 48(03):86-90.

[6] Zhang Yumeng, Su Yi, Wu Yongqi, et al. Research on the optimization of raw material ordering and transshipment based on multi-objective nonlinear programming [J]. Journal of Natural Science of Harbin Normal University, 2022, 38(03):51-58.