Research on optimal crop planting strategy based on K-means cluster analysis

Xinyi Sun*

School of Finance, Capital University of Economics and Business, Beijing, China, 100070

*Corresponding author: 15383449422@163.com

Abstract. This study explores the issue of crop planting strategies in a rural area in North China, which is facing many challenges from traditional agricultural planting models. In response to this situation, this study aims to provide scientific and reasonable crop planting strategies for the village through mathematical modeling to promote rural revitalization and achieve sustainable agricultural development. In formulating the optimal planting strategy for the period from 2024 to 2030, we took into account the substitution and complementarity between crops, as well as the correlation between sales volume, sales price, and planting cost, and introduced a new random variable - the Spearman correlation coefficient. By calculating the Spilman correlation coefficient for each crop, we obtained a strong correlation between expected sales volume and planting cost. Then, we used K-means clustering analysis to classify crops, resulting in three different categories that achieve complementarity within each class and substitution between classes. This article included the correlation coefficient and adjustment parameters in the objective function and obtained the optimal crop planting scheme through Monte Carlo simulation.

Keywords: K-means cluster analysis, Monte Carlo simulation, Optimal planting strategy, Linear regression.

1. Introduction

With the deepening of the rural revitalization strategy, agricultural development is encountering new opportunities and challenges. How to rationally utilize limited land resources to develop organic agriculture, enhance production efficiency, and mitigate planting risks has emerged as a crucial topic in rural economic development. Current literature on rural organic farming in mountainous areas primarily focuses on land resource utilization, crop selection, and planting strategies.

However, existing research has some gaps and limitations:①Fragmentation in Analysis: Most literature conducts isolated analyses, failing to explore the interrelations and overall impact on rural economic development.②Lack of Integrated Strategies: There is a dearth of integrated strategies that combine crop selection, cropping technology, land management, and ecological conservation into a cohesive framework.

The main marginal contributions of this study are as follows:①Holistic Analysis: By examining the relationships among crops, land use, and various indicators, this study offers a comprehensive understanding of the interrelated factors influencing rural organic farming in mountainous areas.② Data-Driven Insights: This study analyzed historical data using statistical and predictive models to identify trends, patterns, and potential risks, informing more data-centric cropping strategies.③ Climate and Terrain Adaptation: Special attention is given to the unique climate and terrain conditions of northern China's mountainous areas to ensure the proposed strategies are suitable for these specific challenges.④Long-Term Sustainable Development: This study emphasizes not only economic profitability but also ecological and social sustainability in agricultural practices, ensuring the solutions possess sustainable development power.

In summary, this study aims to bridge the gap in existing literature by providing a comprehensive, data-driven, and sustainable framework for developing rural organic agriculture in the mountainous areas of northern China. By overcoming these limitations, it contributes to a deeper understanding and more effective practice strategies for promoting rural economic development and ensuring sustainable agricultural practices in these regions.

2. Materials and Methods

2.1. Data Acquisition and Preprocessing

The data used in this study were obtained from an open source website (https://www.mcm.edu.cn/), specifically including the basic situation of the existing arable land and crops in rural areas, as well as the 2023 crop planting and related statistical data in rural areas.

The data were preprocessed. Firstly, the study conducted outlier detection on the yield per mu, planting cost, and selling price, all of which were performed using the boxplot method. The specific content is shown in Figure 1. According to this analysis, the overall outliers are relatively few. Among them, the least outliers were found in planting cost, the most outliers were found in yield per mu, and the yield per mu and planting cost were relatively large (the red "+" marks the three columns of outlier data).

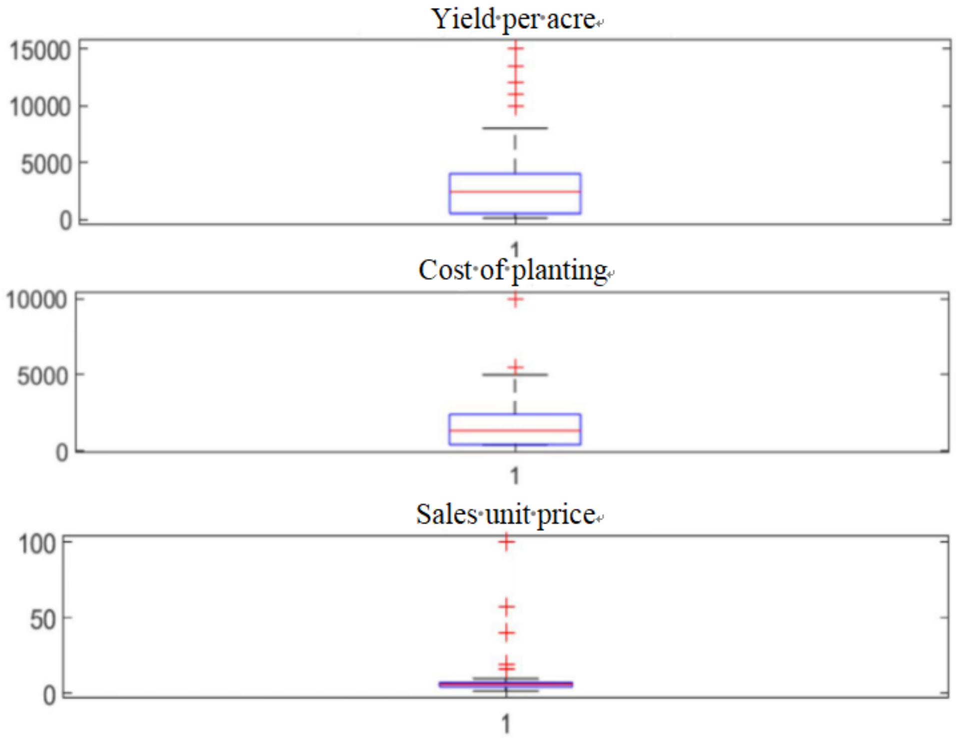

Figure 1 Boxplot of yield per mu, planting cost and sales unit price

Second, descriptive statistics were performed on the data. The specific results are shown in Table 1. It can be seen that the distribution of yield per mu, planting cost and selling price all show relatively obvious fluctuations and non-normal characteristics, especially the distribution of selling price is the most extreme. The reason may be caused by uncertainties in agriculture, such as climatic conditions, market demand, and production costs. The peak values and high skewness of selling prices indicate that there may be some very high selling prices in the market, which may be related to the market scarcity of some specific crops.

Table 1 Descriptive statistics

	Mean value	median	Standard deviation	kurtosis	skewness
Yield per acre	2990.84	2400.00	3047.93	6.44	1.78
Cost of planting	1686.07	1320.00	1636.14	13.47	2.64
Sales unit price	8.20	6.00	11.91	37.85	37.85

2.2. Method Introduction

2.2.1 K-means Clustering Algorithm

The K-means clustering algorithm is an iterative solution algorithm for clustering analysis0, which is widely used among all clustering algorithms due to its simplicity and efficiency. Its specific steps are as follows: First, the data is preliminarily divided into K groups, and then K random objects are selected as the initial clustering centers. Then, the distance between each object and each seed clustering center is calculated, and each object is assigned to the nearest clustering center. The clustering center and the objects assigned to it represent a cluster [8]. After assigning each sample, the clustering center of the cluster will be recalculated based on the existing objects in the cluster [9]. This process will be repeated until a termination condition is met. The termination condition can be no (or the minimum number) of objects being reassigned to different clusters, no (or the minimum number) of clustering centers changing, or the error square sum reaching a local minimum [10].

In order to explore the potential patterns among different crops, identify groups of crops with similar characteristics, find alternative and complementary relationships between crops, this study performs clustering analysis on the crops in different fields and seasons, using the K-means clustering method [11]. Since the type of field and planting season of the same crop are different, resulting in different yields, costs, and unit prices, this study classifies the same crop into several different categories of crops and processes them accordingly, generating 107 types of crops for data analysis. The results are shown in Tables 2 and 3.

Table 2. Cluster analysis

	The final cluster center			Analysis of variance	
	1	2	3	F	Significance
Expected sales volume	10967	1000	2271	91.023	<0.001
Cost of planting	4432	10000	1346	56.882	<0.001
Sales unit price	7.21	100	7.35	67.698	<0.001

Table 3. Specific categories of cluster analysis

Type 1	59	cucumber	70	eggplant	80	water spinach	95	eggplant	105	water spinach
	62	water spinach	77	cucumber	88	pleurotus	102	cucumber		
Type 2	89	morel								
Type 3	1	soya beans	11	buckwheat	21	wheat	31	soya beans	41	buckwheat
	2	black bean	12	pumpkin	22	corn	32	black bean	42	pumpkin
	3	ormosia bean	13	sweet potato	23	millet	33	ormosia bean	43	sweet potato
	4	mung bean	14	naked oats	24	sorghum	34	mung bean	44	naked oats
	5	crawling bean	15	barley	25	broomcorn millet	35	crawling bean	45	barley
	6	wheat	16	soya beans	26	buckwheat	36	wheat	46	rice
	7	corn	17	black bean	27	pumpkin	37	corn	47	cowpea

8	millet	18	ormosia bean	28	sweet potato	38	millet	48	canavalia
9	sorghum	19	mung bean	29	naked oats	39	sorghum	49	kidney bean
10	broomcorn millet	20	crawling bean	30	barley	40	broomcorn millet	50	Potato
51	tomato	63	yellow heart	74	cabbage	86	elm mushroom	99	cabbage
52	eggplant	64	celery	75	romaine lettuce	87	mushroom	100	romaine lettuce
53	spinach	65	cowpea	76	brassica chinensis	90	cowpea	101	brassica chinensis
54	green pepper	66	canavalia	78	lettuce	91	canavalia	103	lettuce
55	cauliflower	67	kidney bean	79	pepper	92	kidney bean	104	pepper
56	cabbage	68	potato	81	yellow heart	93	potato	106	yellow heart
57	romaine lettuce	69	tomato	82	celery	94	Tomato	107	celery
58	brassica chinensis	71	spinach	83	chinese cabbage	96	spinach		
60	lettuce	72	green pepper	84	white radish	97	green pepper		
61	pepper	73	cauliflower	85	carrot	98	cauliflower		

According to Table 2, the 107 crops can be divided into three categories according to the expected sales volume, planting cost and sales unit price, and the results of each category are referred to Table 3. Based on this analysis, it is known that crops within each category are complementary, and there is substitution between classes.

2.2.2 Correlation analysis

Crop substitution means that when the price of a crop falls, consumers turn to other alternative crops, resulting in a decrease in the demand for the original crop and an increase in the demand for the alternative crop. Crop complementarity means that when the price of the original crop falls, consumers purchase the complementary products of the original crop, resulting in an increase in the demand for the original crop.

In order to measure the subjection and complementarity among various crops and among the expected sales volume, sales price and planting cost, Pearson correlation coefficient or Spearman correlation coefficient are generally used for analysis. When the data satisfy linear relationship and the results are significantly normal distributed, Pearson correlation coefficient is used. Otherwise, Spearman correlation coefficient [6] is used. Now the correlation analysis of the expected sales volume, sales price and planting cost in 2023 is carried out.

Firstly, scatter plots of yield per mu, planting cost, and sales unit price are drawn, as shown in Figure 2. In this view, the yield per mu (that is, the expected sales volume) is non-linearly related to both the planting cost and the unit price of sales.

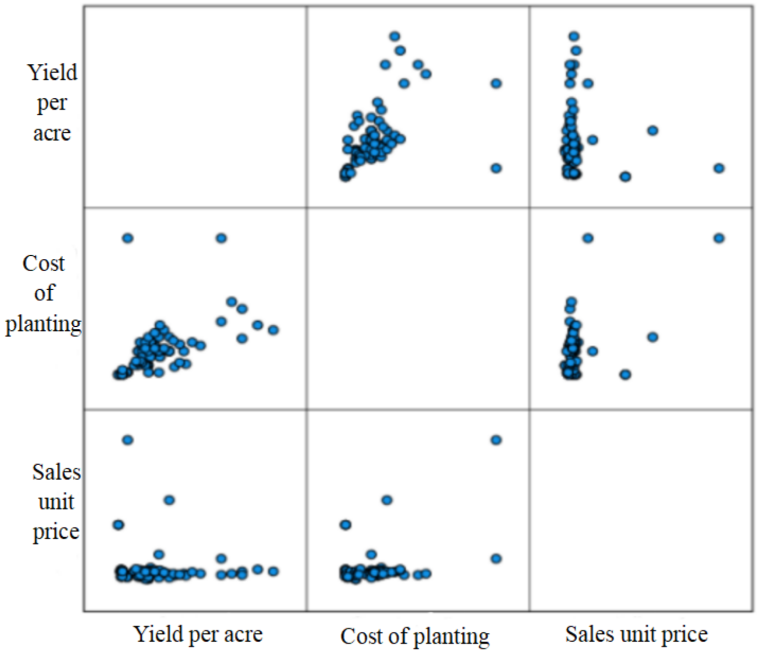

Figure 2 Scatter plot of yield per mu, planting cost and sales unit price

Table 4 Normality test of yield per mu, planting cost and sales unit price

Test of normality			
	Kolmogorov-Smirnov		
	Statistics	Degree of freedom	Significance of significance
Expected sales	0.171	107	<0.001
Planting cost	0.207	107	<0.001
Sales unit price	0.409	107	<0.001

The confidence level for this study is assumed to be 95%. According to Table 4, the significance level of yield per mu, planting cost and unit sales price is less than 0.001, indicating that the null hypothesis should be rejected, and yield per mu, planting cost and unit sales price do not conform to the normal distribution.

Then, the correlation coefficient is calculated. Let U and V be two sets of data, di is the rank difference between Ui and Yi, and the Spearman (rank) correlation coefficient is:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i}{n(n^2-1)} \tag{1}$$

Thus, Spearman correlation coefficients of yield per mu, planting cost and sales unit price were calculated according to SPSS, and the specific results are shown in Table 5.

Table 5 Correlation of yield per mu, planting cost and sales unit price

Correlation				
		Expected sales volume	Cost of planting	Sales unit price
Expected sales volume	Correlation coefficient	1.000	0.834	-0.160
	significance		<0.001	0, 100
Cost of planting	Correlation coefficient	0.834	1.000	0.011
	significance	<0.001		0.912
Sales unit price	Correlation coefficient	-0.160	0.011	1.000
	significance	0.100	0.912	

There is a significant correlation between expected sales volume (i.e., yield per mu) and planting cost, and a strong positive correlation, correlation coefficient; The significance level of expected sales volume and sales unit price is greater than 0.05, indicating that there is no significant correlation between them; The significance level of planting cost and sales unit price is as high as 0.912, indicating that there is no significant correlation between them. From this point of view, the expected sales volume and planting cost are complementary.

2.2.3 Monte Carlo simulation

Monte Carlo simulation is a numerical method to estimate the expected value of the solution to a problem by taking many random samples, which is widely used in uncertainty analysis and prediction of complex systems. This method estimates the distribution and expected value of the output of the system by simulating possible combinations of input variables and repeating the calculation several times. In agricultural planting planning, Monte Carlo simulation can be used to deal with the complexity and uncertainty of future yield, planting cost, selling price and expected sales volume [7]. The specific steps are as follows:

① Determine the preset probability distribution.

②Monte Carlo sampling: simulate the fluctuations of different crops under four scenarios of yield per mu, sales volume, sales price and planting cost every year to help decision makers evaluate the potential benefits and risks of different planting schemes. Each simulation generates different random variables according to a preset probability distribution.

③ Multi-scenario synthesis: through repeated calculation of a large amount of data, the profit distribution results of each planting plan in the future from 2024 to 2030 were obtained, and its expected value was taken as the optimal planting plan.

3. Establishment of linear regression model

3.1. Introduction of random variables

The expected sales volume, yield per mu, planting cost and selling price all change, so the correlation coefficient fluctuates up and down over time, generating new random variables. This paper assumes that the fluctuation range of the correlation coefficient is ±0.1.

$$\rho_{j_1,j_2,t} = \rho_{j_1,j_2,2023} \times (1+R) \tag{2}$$

3.2. Establishment of objective function

In 2023, there is a strong and significant relationship between expected sales volume (that is, yield per acre) and planting costs. Specifically, the impact of yield per mu on planting cost is obvious, and the change of yield per mu will directly affect the input required for planting, such as the cost of fertilizer, labor and equipment.

Therefore, when considering strategies to maximize profits, we need to focus on how to optimize profits by increasing the yield per mu.

In summary, the strong correlation between yield per mu and planting cost indicates that cost can be effectively affected by adjusting yield, thus playing an important role in profit optimization [14]. This indirect influence mechanism emphasizes the importance of managing yield to control costs to achieve maximum profit when developing planting strategies [15].

Unsalable causes waste:

When total production is less than or equal to the expected sales volume, all production is sold at the standard price.

$$Profit^{(1)} = \sum_{i=1}^{54} \sum_{j=1}^{41} \sum_{t=2024}^{2030} \left[\left(X_{i,j,t} Y_{j,t} - \alpha \cdot \rho_{j_1,j_2,t} \right) \times Price_{j,t} - X_{i,j,t} \times Cost_{j,t} \right] \tag{3}$$

When the total output is greater than the expected sales volume, the excess output will be wasted directly, which will reduce the total profit.

$$Profit^{(2)} = \sum_{i=1}^{54} \sum_{j=1}^{41} \sum_{t=2024}^{2030} \left[\left(Sale_{j,t} - \alpha \cdot \rho_{j_1,j_2,t} \right) \times Price_{j,t} - X_{i,j,t} \times Cost_{j,t} \right] \tag{4}$$

More than part of the reduced price sale:

When total production is less than or equal to the expected sales volume, all production is sold at the standard price.

$$Profit^{(3)} = \sum_{i=1}^{54} \sum_{j=1}^{41} [X_{ij} Y_j Price_j + (X_{ij} Y_j - Sale_j) \times Price_j \times 50\% - X_{ij} Cost_j] \tag{5}$$

When the total production is greater than the expected sales volume, the excess production will be sold at a 50% reduction of the 2023 sales price, which will increase the total profit.

$$\begin{aligned} Profit^{(4)} = \sum_{i=1}^{54} \sum_{j=1}^{41} \sum_{t=2024}^{2030} \Big[& \left(X_{i,j,t} Y_{j,t} - \alpha \cdot \rho_{j_1,j_2,t} \right) \times Price_{j,t} \\ & + \left((X_{i,j,t} Y_{j,t} - Sale_{j,t}) + \beta \cdot \rho_{j_1,j_2,t} \right) \times Price_{j,t} \times 50\% \\ & - X_{i,j,t} Cost_{j,t} \Big] \end{aligned} \tag{6}$$

Where α and β are parameters, which are used to control the influence of correlation coefficients on the objective function. Larger values of α, β indicate more emphasis on risk control, while smaller values indicate more emphasis on returns. Let ρj1,j2,t denote the correlation coefficient between crop j1 and j2 in year t [17] .

3.3. Determination of constraints

Correlation constraints of crops:

Specifically, this correlation constraint means that when there is a negative correlation between different crops, it may lead to an uneven allocation of planting, that is, resources are excessively concentrated on a single crop. This uneven distribution will reduce the amount of other crops planted, which will lead to a series of problems of lower profits. The specific mathematical expression is as follows.

$$Yield_{j_1,t} + Yield_{j_2,t} \leqslant Maximum\ total\ amount_t \tag{7}$$

4. Conclusion

This study utilizes K-means cluster analysis as a fundamental approach to classify different types of crops based on their growth attributes, environmental prerequisites, and historical profitability. Through partitioning the crops into distinct clusters, the study acquires valuable perceptions regarding which groups of crops are more prone to flourish under similar circumstances, facilitating more informed planting decisions.

Furthermore, the research incorporates the concept of the correlation coefficient of random variables to scrutinize the relationships among various factors influencing crop yields and profits. This statistical metric enables a deeper comprehension of how alterations in one variable, such as rainfall or temperature, could impact another, such as crop yield or market price. By analyzing these correlations, the study can identify patterns and trends that can be exploited to enhance agricultural productivity.

To predict crop yields and profits more precisely, the study formulates a linear regression model. This model takes into consideration various explanatory variables, such as soil fertility, water availability, and market demand, and employs statistical techniques to assess the relationship between these variables and the dependent variables of interest, namely crop yield and profit. By leveraging this model, the study can forecast potential outcomes under diverse scenarios, enabling farmers to make more data-driven decisions.

Additionally, the study employs Monte Carlo simulation to generate random numbers and scenarios. This potent computational technique allows for the exploration of a broad spectrum of potential outcomes by repeatedly sampling from the defined probability distributions of the input variables. By conducting numerous simulations, the study can estimate the distribution of possible crop profits and identify the planting scheme that maximizes the expected profit. This approach offers a more comprehensive and robust analysis compared to traditional deterministic methods.

Ultimately, the discoveries of this study provide a theoretical foundation for determining the optimal planting scheme that maximizes crop profits. By integrating K-means cluster analysis, correlation coefficient analysis, linear regression modeling, and Monte Carlo simulation, the study presents a holistic framework for agricultural decision-making. The practical implications of this research are substantial, as it can assist farmers in allocating their resources more effectively, increase their profitability, and contribute to the rural revitalization and sustainable development of agriculture.

4.1. The scenarios are highly diverse and comprehensive

The model takes into account a multitude of key factors in its design, such as sales volume, yield per acre, planting costs, and selling prices, among others. The model also incorporates uncertainty analysis to better mirror the complexity of actual agricultural production. Moreover, the model is founded on Monte Carlo simulation and the simulated annealing algorithm to analyze the trends of crop sales volume, planting costs, and selling prices in the coming years, thereby augmenting the model's adaptability.

4.2. Disadvantages -- High model complexity

The model integrates a variety of variables, and Monte Carlo simulation is used for random value, which leads to a large amount of data and high complexity, and the solution process may take a long time.

References

[1] L K S ,Zhehui L ,E J R , et al.Determining County-Level Counterfactuals for Evaluation of Population Health Interventions : A Novel Application of K-Means Cluster Analysis.[J].Public health reports (Washington, D.C. : 1974),2021,137(5):333549211030507-333549211030507.

[2] Ying J ,Zongwei L ,Chunmei Y , et al.Dugong chirp type classification based on fundamental contour extraction and hierarchical cluster analysis[J].Applied Acoustics,2024,217109812-.

[3] Xianrong J ,Fenghua L .Retraction notice to “K-means clustering analysis and evaluation for internet of acoustic environment characteristics” [Cogn. Syst. Res. 52 (2018) 603–609][J].Cognitive Systems Research,2023,82

[4] Yanyun Z ,Peng X ,Yifan Z , et al.Spectra measurement and clustering analysis of global horizontal irradiance for solar energy application[J].Renewable Energy,2024,222119813-.

[5] Tao Y ,Zhang Q ,Chen Q , et al.An intelligent approach using micro-seismic monitoring signal clustering and an optimized K-means model to guide the selection of support patterns in underground mines[J].Tunnelling and Underground Space Technology incorporating Trenchless Technology Research,2024,154106095-106095.

[6] [1]Li X ,Ma Y ,Zhou Q , et al.Sparse large-scale high-order fuzzy cognitive maps guided by spearman correlation coefficient[J].Applied Soft Computing,2024,167(PA):112253-112253.2.

[7] Fei X ,Lou Z ,Sheng M , et al.Source-oriented stochastic health risk assessment of toxic metals in soil via a hybrid model and Monte Carlo simulation[J].Ecotoxicology and Environmental Safety,2024,286117209-117209.

[8] Du D ,Wang J ,Li J .What drives intercity venture capital investment? A comparative analysis between multiple linear regression and random forest[J].Humanities and Social Sciences Communications,2024,11(1):1207-1207.

[9] Meng K ,Gai Y ,Wang X , et al.Transfer learning for high-dimensional linear regression via the elastic net[J].Knowledge-Based Systems,2024,304112525-112525.

[10] Pareto A .On a correlation coefficient based on the L1-norm[J].Statistical Papers,2024,(prepublish):1-21.