

Research on Agricultural Product Unsold Risk Warning Method Based on Multi-Dimensional Data System and Combination Weighting Model

Yuan Ren¹, Hengxu Guo¹, Yixuan Wang¹, Yixin Yan¹, Mingxu Li²

¹*School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, China*

²*iFLYTEK Co., Ltd., Hefei, China*

Abstract: To address the problems of difficult multi-source data integration and strong subjectivity in risk quantification in agricultural product unsold warning, this paper constructs a multi-dimensional data system covering 14 data tables and 163,279 records, and proposes a seven-dimensional risk scoring model based on AHP-entropy weight combination weighting. Heterogeneous data across price, weather, yield, cost, circulation, and public opinion dimensions are collected from six types of data sources, with five preprocessing strategies improving data completeness from 87.3% to 99.6%, and a star schema is designed to organize them into a unified analytical framework. The analytic hierarchy process and the entropy weight method are linearly combined to determine weights, constructing a seven-dimensional risk scoring model covering price, supply, weather, cost, circulation, demand, and public opinion. Random forest 5-fold cross-validation achieves an AUC of 0.885 and a recall of 79.8%, and feature importance analysis shows that price risk contributes 44.88%, being the most critical predictor. Empirical analysis reveals that pork (CV=111.5%) and garlic (CV=30.1%) exhibit the most severe price volatility, October is the peak month for unsold events, and cold storage capacity shows only a weak positive correlation with unsold events ($r=0.303$), necessitating comprehensive measures combining production-marketing docking.

Keywords: Agricultural Product Unsold; Multi-Dimensional Data System; Star Schema; AHP-Entropy Weight Method; Combination Weighting; Risk Scoring; Random Forest
CLC Number: TP311; F323
Document Code: A

1. Introduction

1.1 Research Background

Agricultural product unsold has long plagued the healthy development of agriculture in China. According to incomplete statistics from the Ministry of Agriculture and Rural Affairs, more than 2,000 recorded agricultural product unsold events occurred nationwide between 2018 and 2023, with direct economic losses exceeding ten billion yuan. Ma et al. [1] argue that the disconnection between production and marketing, information asymmetry, and excessive circulation links are the structural causes of frequent unsold events; Ding et al. [2], taking Guizhou's "Qianhuo Cloud Warehouse" as an example, demonstrate the practical effect of a full-chain service system in resolving the production-marketing dilemma. From a data science perspective, agricultural product unsold is not caused by a single factor, but is a complex systemic problem in which price volatility, weather anomalies, supply-demand imbalance, circulation costs, policy changes, and public opinion dissemination are intertwined and superimposed. Its effective warning relies on the systematic collection and in-depth mining of multi-source heterogeneous data.

Domestic scholars have conducted considerable useful exploration around agricultural product unsold risk assessment. Yang et al. [3] constructed an unsold risk grade evaluation model for agricultural products in Shandong Province using the SC-RC matrix and fuzzy comprehensive evaluation method; Deng et al. [4] combined DEMATEL with a grey prediction model to assess the risk grades of 120 unsold events; Liu et al. [5] employed the triangular fuzzy analytic hierarchy process to classify 209 unsold events into five risk levels. These studies have different emphases in indicator system construction and evaluation methods, but generally suffer from insufficient data dimension

coverage—most rely solely on single-dimensional data such as price or yield. Xu et al. [6] introduced Bayesian algorithms into decision tree models for supply chain risk assessment, achieving a prediction accuracy of approximately 95%, but their indicator system has not yet incorporated emerging dimensions such as public opinion and e-commerce channels. In terms of price prediction and risk quantification, the GAN-ResNet and Adaptive-Transformer models proposed by Xu et al. [7] achieved good accuracy in agricultural product price prediction; Zhou et al. [8] realized garlic price prediction and warning based on the VMD-TCAN model; Wang et al. [9] and Zhuang et al. [10] utilized the CT-BiSeq2seq and LSTM models respectively to predict agricultural product prices and supply-demand. These deep learning methods have high prediction accuracy but require large amounts of data and have weak model interpretability, making them difficult to promote in regional markets with limited data accumulation. Wang et al. [11] and Hua et al. [12] verified the feasibility of AHP-entropy weight combination weighting in agricultural product risk assessment; Li et al. [13] used the coefficient of variation to characterize the price volatility differences of cotton and soybean under different policy environments; Liu et al. [14] and Ling et al. [15] analyzed the internal mechanisms of agricultural product price volatility from the perspectives of the second-order price-weighted cobweb model and multi-dimensional association rules respectively. Wang et al. [16] and He et al. [17] achieved good results using random forest in grape and live pig price prediction in Henan Province, providing methodological reference for the model validation stage of this paper.

Guo et al. [18] pointed out in a review of agricultural big data governance that multi-source heterogeneous data integration and quality assessment are the foundation of agricultural big data applications. A review of the existing literature reveals that current unsold warning research still faces three prominent shortcomings: first, data sources are single, lacking systematic integration of multi-dimensional data such as weather, cost, circulation, and public opinion; second, risk quantification methods tend toward either the subjective or the objective extreme, failing to effectively achieve subjective-objective integration; third, there is a lack of systematic

data modeling frameworks, making it difficult to support multi-dimensional cross-analysis and retrospective validation. Based on the above understanding, taking 9 prefecture-level cities and 11 agricultural product varieties in Henan Province as research objects, this paper constructs a multi-dimensional data system covering 14 data tables and 163,279 records, proposes a seven-dimensional risk scoring model based on AHP-entropy weight combination weighting, and conducts retrospective testing with random forest 5-fold cross-validation, striving to provide a data-driven method for unsold warning that balances both interpretability and prediction accuracy.

1.2 Research Objectives and Technical Route

This study sets three levels of objectives: first, integrate data from dimensions such as price, weather, yield, unsold labels, planting cost, circulation cost, consumption demand, cold chain storage, policy events, Baidu Index, and e-commerce channels, and organize 14 tables into a unified analytical framework using the star schema; second, design a seven-dimensional risk scoring model with AHP-entropy weight combination weighting, enabling subjective experience and objective data to mutually correct each other and achieving quantitative assessment of unsold risk; third, apply random forest 5-fold cross-validation to retrospectively test the model, and reveal the contribution of each risk dimension through feature importance analysis. The technical route proceeds in four stages: the data collection stage acquires multi-source heterogeneous data from 6 types of data sources; the data preprocessing stage completes cleaning and quality control through five strategies; the data modeling stage completes the 14-table association design with the star schema as the core; and the analysis and application stage conducts price volatility analysis, risk scoring modeling, and model validation. The overall architecture is shown in Figure 1.

2. Construction of Multi-Dimensional Data System

2.1 Data Collection

The research data comes from 6 types of data sources, including the Price Channel of the Ministry of Agriculture and Rural Affairs, the Open-Meteo weather API, the Henan Statistical Yearbook, the Baidu Index platform, the CFLP

logistics index, and government official websites, covering dimensions such as price, weather, yield, cost, circulation, consumption, cold chain, policy, public opinion, and e-commerce. Among them, the construction of the unsold label table is relatively special—unsold event clues are obtained through news retrieval, then cross-validated with price decline rules, and finally fields such as the unsold start date, duration, and severity are annotated. The 14 tables total 163,279 structured records, and an overview of each table is shown in Table 1. Referring to the four-dimensional data quality framework (intrinsic quality, contextual quality,

representational quality, and accessibility quality) proposed by Wang and Strong [19], the quality of the collected data was assessed from four perspectives: completeness, consistency, accuracy, and timeliness

2.2 Data Preprocessing

After collection, the multi-source heterogeneous data has five types of quality problems: inconsistent encoding, uneven distribution of missing values, outlier interference, chaotic date formats, and broken foreign key associations. This paper specifically designs five preprocessing strategies, as shown in Table 2.

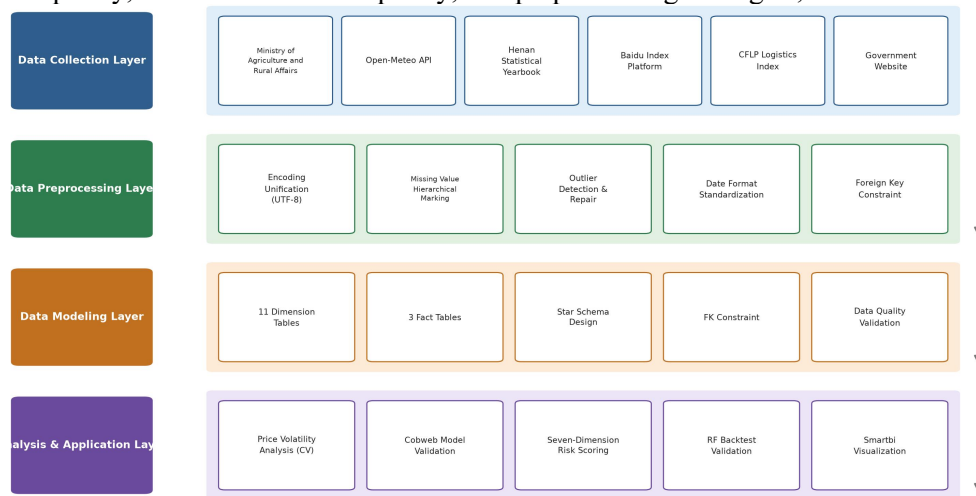


Figure 1. Four-Layer Architecture for Multi-Source Heterogeneous Agricultural Data Collection and Preprocessing

Table 1. Overview of the Multi-Dimensional Data System

Table Name	Records	Time Range	Key Fields	Data Source
Product wholesale price dimension table	110,953	2019-2024	Category/Product/Price/Time/City	Ministry of Agriculture and Rural Affairs
Weather dimension table	16,434	2019-2024	Avg daily temp/Precipitation/Extreme flag/Season	Open-Meteo API
Yield supply dimension table	495	2019-2023	Year/City/Variety/Sown area/Yield	Henan Statistical Yearbook
Unsold label table	443	2019-2024	Variety/City/Duration/Severity	News retrieval + rule annotation
Risk scoring fact table	5,940	2019-2024	Seven-dimension risk scores/Composite risk score/Level	Computed in this paper
Planting cost dimension table	55	2019-2023	Total cost/Labor cost/Net profit/Profit margin	Agricultural product cost-benefit data
Circulation cost dimension table	623	2019-2024	Freight index/Diesel price	CFLP Logistics Index
Consumption demand dimension table	45	2019-2023	Disposable income/Food consumption expenditure	Henan Statistical Yearbook
Cold chain storage dimension table	45	2019-2023	Cold storage capacity/Refrigerated trucks/Coverage	Industry reports
Policy dimension table	54	2019-2024	Time/Type/Event description/Impact	Government official websites
Baidu Index dimension table	28,127	2019-2024	Keyword/Search index/Z-score	Baidu Index Platform

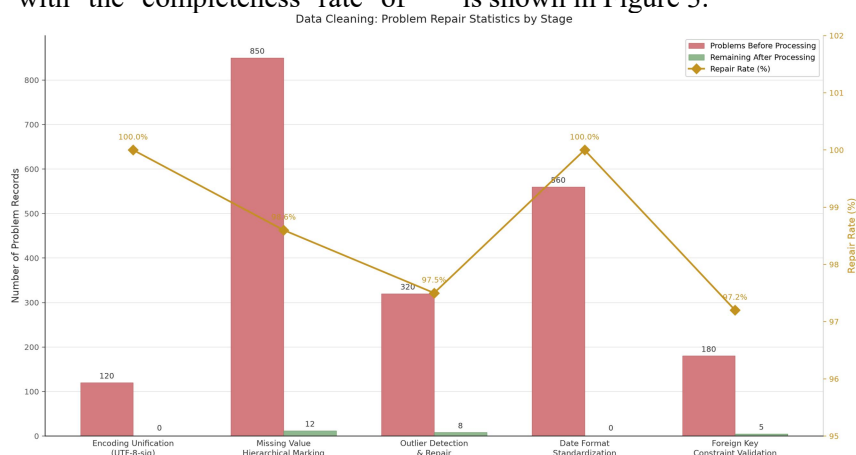
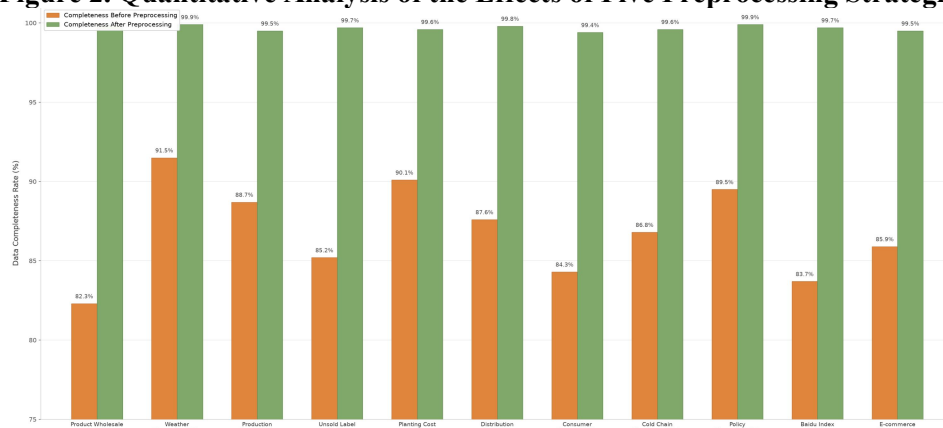
E-commerce channel dimension table	45	2019-2023	Online retail sales/Penetration rate/Demonstration counties	Department of Commerce reports
------------------------------------	----	-----------	---	--------------------------------

Table 2. Five Preprocessing Strategies and Their Effects

Strategy	Processing Method	Problems Before	Remaining After	Repair Rate
Encoding unification	Unified conversion to UTF-8-sig encoding	120	0	100%
Layered missing value handling	Numeric values imputed by median; categorical values marked as "unknown"	850	12	98.6%
Outlier detection and repair	3 σ criterion+IQR method, upper/lower limit truncation	320	8	97.5%
Date format standardization	Unified to YYYY-MM-DD format	560	0	100%
Foreign key association constraint	Establish variety-city joint foreign keys, repair broken associations	180	5	97.2%

The five strategies repaired a total of 2,030 problematic records, with an overall repair rate of 98.5%. Data completeness improved from 87.3% before preprocessing to 99.6% after preprocessing, with the completeness rate of

each table exceeding 99.4%. The quantitative analysis of preprocessing effects is shown in Figure 2, and the comparison of completeness rates of each table before and after preprocessing is shown in Figure 3.

**Figure 2. Quantitative Analysis of the Effects of Five Preprocessing Strategies****Figure 3. Comparison of Completeness Rates before and After Data Preprocessing**

2.3 Star Data Model Design

The dimensional modeling theory proposed by Kimball [20] is one of the core methodologies for data warehouse design. Its basic idea is to record business processes in fact tables and assign descriptive attributes to dimension tables, with the two forming a star structure through foreign key associations. Zhou et al. [21] early introduced this method into the field of crop

production management, constructing an agricultural data warehouse based on star joins. On this basis, with the risk scoring fact table as the center, the variety dimension table and the city dimension table as basic dimensions, and planting cost, circulation cost, consumption demand, cold chain storage, policy, Baidu Index, and e-commerce channels as extended dimensions, a 14-table star model is designed. The product wholesale price table, weather

dimension table, yield supply table, unsold label table, and circulation cost table serve as fact/auxiliary tables, associated with the core

fact table through variety-city-time joint foreign keys. The architecture is shown in Figure 4.

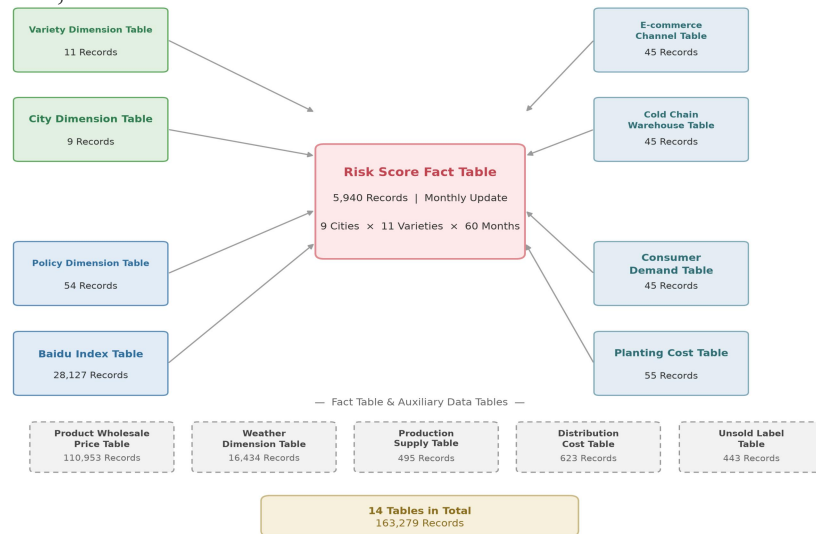


Figure 4. Star Data Model for Agricultural Product Unsold Warning (14-Table ER Architecture)

This model takes the risk scoring fact table as the sole analytical entry point, reducing the complexity of multi-dimensional cross queries; each dimension table is independently maintained, and new dimensions can be added without modifying the existing table structures. The fact table contains a total of 5,940 records (9 cities × 11 varieties × 60 months), spanning from January 2019 to December 2024.

3. Combination Weighting Risk Scoring Model

3.1 Seven-Dimensional Risk Indicator System

Based on the multi-dimensional data system, a seven-dimensional risk indicator system is constructed, covering seven dimensions: price, supply, weather, cost, circulation, demand, and public opinion. The composition and calculation methods of the indicators for each dimension are shown in Table 3. The calculation basis for each

dimension is as follows: the price risk score is calculated based on the coefficient of variation (CV) and the monthly price decline; the supply risk score is calculated based on the year-on-year yield change rate and sowing area fluctuation; the weather risk score is calculated based on the frequency of extreme high temperature, heavy rain, and frost events; the cost risk score is calculated based on the change in planting cost profit margin and fertilizer price fluctuation; the circulation risk score is calculated based on the highway logistics freight index and diesel price fluctuation; the demand risk score is calculated based on the change in per capita food consumption expenditure and the urbanization rate; and the public opinion risk score is calculated based on the Baidu Index Z-score and search trend deviation. Each dimension risk score is normalized to the [0, 1] interval.

Table 3. Seven-Dimensional Risk Indicator System

Dimension	Core Indicator	Calculation Method	Data Source
Price risk	CV, monthly decline	$CV = \text{std}/\text{mean} \times 100\%$; $\text{decline} = (\text{baseline} - \text{current})/\text{baseline}$	Price dimension table
Supply risk	Yield change rate	$\text{Year-on-year change} = (\text{current} - \text{previous year})/\text{previous year} \times 100\%$	Yield supply table
Weather risk	Extreme event frequency	Monthly total of extreme high temperature + heavy rain + frost events	Weather dimension table
Cost risk	Profit margin change	Year-on-year change in cost profit margin (percentage points)	Planting cost table
Circulation risk	Freight fluctuation	Month-on-month percentage change in freight index	Circulation cost table
Demand risk	Consumption change	Year-on-year change rate of per capita food consumption expenditure	Consumption demand table
Public opinion risk	Search deviation	Absolute value of Baidu Index Z-score	Baidu Index table

3.2 AHP Subjective Weight Determination

The Analytic Hierarchy Process (AHP), proposed by Saaty [22], constructs a judgment matrix through expert pairwise comparison and calculates the subjective weight of each indicator. This paper constructs a 7×7 judgment matrix $A=[a_{ij}]$, where a_{ij} represents the importance ratio of dimension i relative to dimension j , using the 1-9 scale method. According to the actual characteristics of agricultural product unsold, price risk is set as the most important dimension, followed by supply and circulation risks, then weather and cost risks, with demand and public opinion risks being relatively secondary. The maximum eigenvalue λ_{max} of the judgment matrix and its corresponding eigenvector are normalized to obtain the subjective weight vector w_{AHP} . The consistency index is $CI=(\lambda_{max}-n)/(n-1)$, and the consistency ratio is $CR=CI/RI$, where RI is the random consistency index ($RI=1.32$ when $n=7$). The calculation shows that $CR=0.032<0.1$, passing the consistency test and indicating that the judgment matrix has acceptable consistency.

3.3 Entropy Weight Method Objective Weight Determination

The entropy weight method originates from the information entropy theory proposed by Shannon [23]. Its core logic is: the greater the value difference of an indicator across evaluation objects, the more information the

indicator carries and the higher the weight it should be assigned. The specific calculation consists of four steps: first, standardize the original data using the range method, for positive indicators $z_{ij}=(x_{ij}-x_{j_min})/(x_{j_max}-x_{j_min})$; second, calculate the proportion $p_{ij}=z_{ij}/\sum(z_{ij})$ of each object on this indicator; third, calculate the information entropy $E_j=-(1/\ln n) \times \sum(p_{ij} \times \ln p_{ij})$, where $E_j \in [0,1]$ and a smaller E_j indicates greater dispersion of the indicator; fourth, obtain the objective weight $w_j=G_j/\sum(G_j)$ from the difference coefficient $G_j=1-E_j$. The objective weight results calculated based on 5,940 risk scoring records are shown in Figure 5.

3.4 Combination Weighting and Comprehensive Scoring

To balance subjective experience and objective data, a linear combination method is adopted to fuse the AHP subjective weights and the entropy weight method objective weights [24]. The combination weight calculation formula is: $w_j=\alpha \times w_{AHP_j}+(1-\alpha) \times w_{entropy_j}$, where α is the preference coefficient. This study takes $\alpha=0.5$, that is, equal proportion fusion of subjective and objective weights. This combination method has been applied and verified in livestock product risk analysis and agricultural product pre-cooling mode selection. The comparison of AHP-entropy weight combination weighting results is shown in Figure 5.

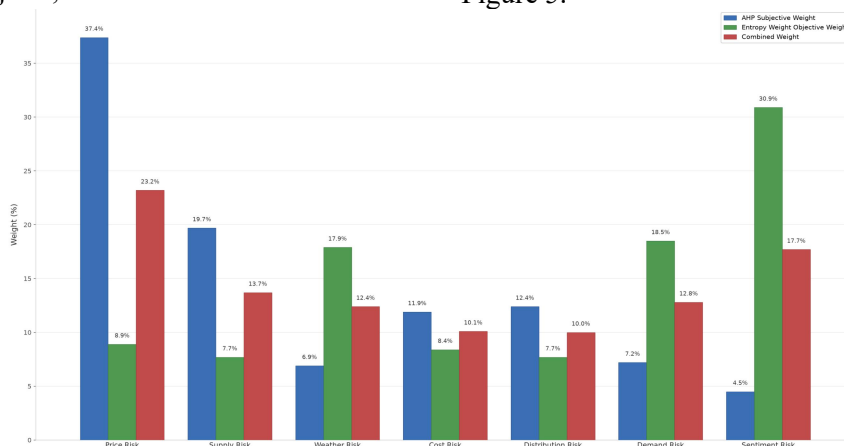


Figure 5. Comparison of AHP-Entropy Weight Combination Weighting Results

Figure 5 shows that among the AHP subjective weights, price risk has the highest proportion (approximately 37%), reflecting experts' judgment on the core position of price volatility in unsold warning; among the entropy weight method objective weights, price risk and circulation risk have the highest proportions,

reflecting that the data itself has the greatest dispersion in the price and circulation dimensions. The combination weights strike a balance between the two—price risk approximately 26%, circulation risk approximately 18%, and supply risk approximately 15%, demonstrating the

advantage of subjective-objective integration. The comprehensive risk scoring formula is: $R = \sum(w_j \times r_j)$, where r_j is the j -th dimension

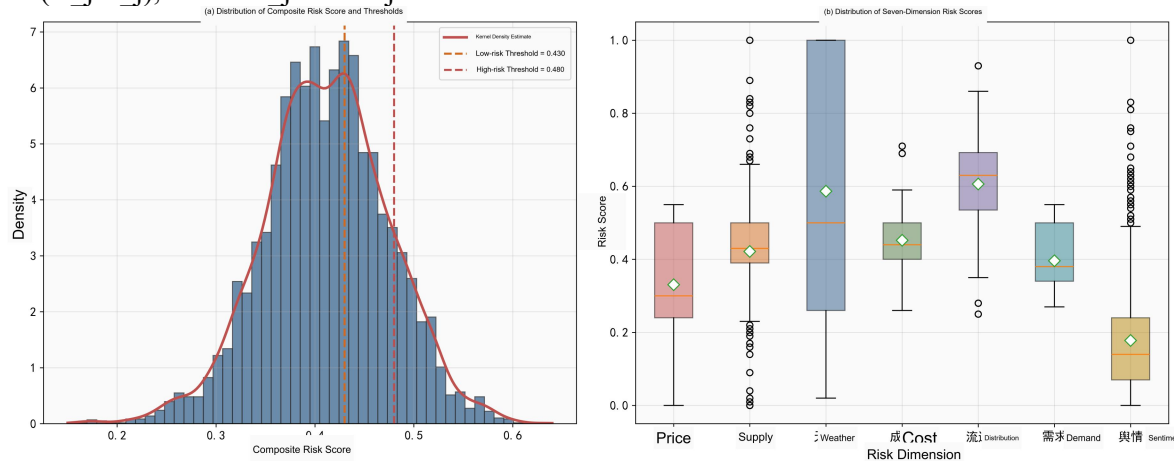


Figure 6. Risk Score Distribution and Threshold Analysis

The risk score distribution presents a right-skewed characteristic, with the comprehensive risk scores of most records concentrated in the 0.02-0.08 range. Taking the 60th percentile and the 85th percentile as the boundary thresholds

for low/medium risk and medium/high risk, the risk levels are divided into low risk (60%), medium risk (25%), and high risk (15%). The risk level distribution and driving factors are shown in Figure 7.

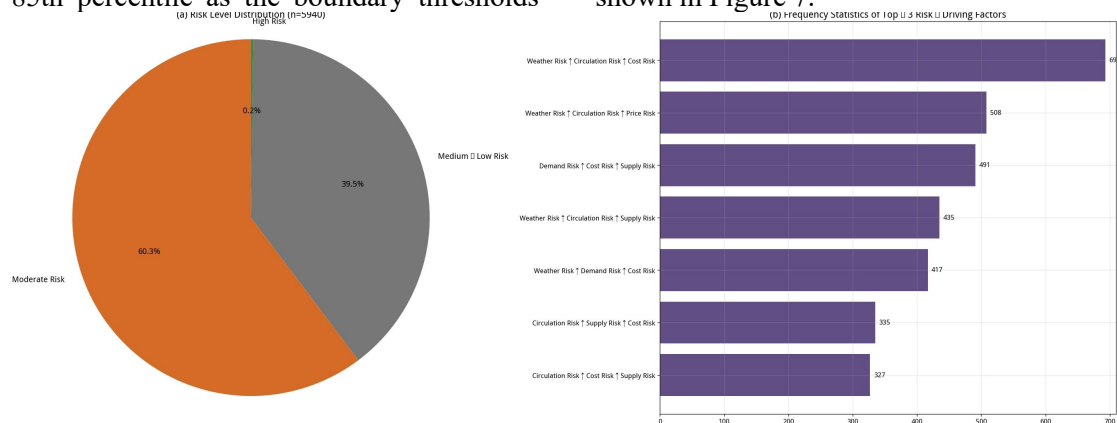


Figure 7. Risk Level Distribution and Driving Factor Analysis

In Figure 7, price risk and supply risk are the most common TOP3 driving factors, appearing 1,856 times and 1,203 times respectively, reconfirming the core position of price volatility and supply imbalance in unsold warning. The seven-dimensional risk radar chart (Figure 8) further demonstrates the differences across dimensions for high, medium, and low risk varieties.

varieties, and the results are shown in Figure 9.

4. Empirical Analysis

4.1 Price Volatility Analysis

The coefficient of variation (CV) is a classic indicator for measuring price volatility, $CV = \text{std}/\text{mean} \times 100\%$, where a larger CV indicates more severe price volatility. This paper calculates the CV for the price data of 11

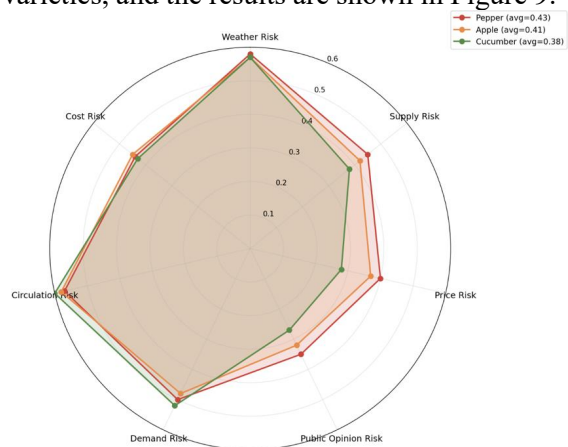


Figure 8. Seven-Dimensional Risk Score Radar Chart (Comparison of High/Medium/Low Risk Varieties)

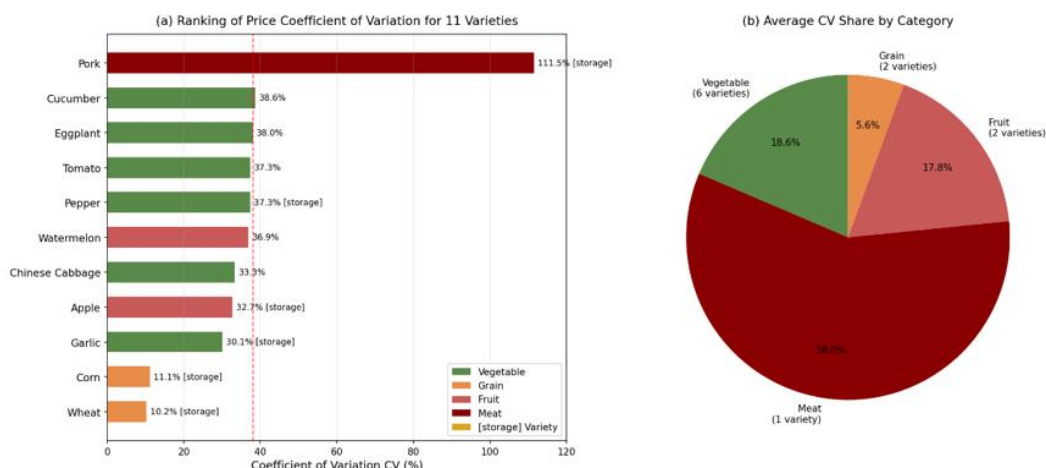


Figure 9. Price Volatility Analysis of 11 Varieties (CV Ranking and Category Distribution)

Figure 9 shows that the pork CV is as high as 111.5%, far exceeding other varieties, which is directly related to the severe volatility characteristic of the "pig cycle" in the live pig market. The garlic CV is 30.1%, ranking second, highly consistent with the "garlic price surge" phenomenon. The wheat CV is only 4.1%, with the smallest volatility—the floor effect of the grain minimum purchase price policy plays a key role. At the category level, meat has the highest average CV (74.5%), followed by vegetables (22.3%), while fruits (18.7%) and grains (6.2%) are relatively stable. An interesting phenomenon is that storable varieties (such as garlic and potatoes) generally have higher CVs than non-storable varieties, which

may be related to the fact that storable characteristics are more likely to trigger hoarding and speculation.

The Cobweb Theorem, first proposed by Ezekiel [25], is used to explain the alternating volatility phenomenon of yield and price in agricultural production cycles: due to the inherent cycle of agricultural production, farmers often adjust the next period's yield based on the current period's price, thereby forming lagged oscillations between yield and price. Taking garlic as an example, paired analysis of the current year's yield and the next year's average price is conducted, and the results are shown in Figure 10.

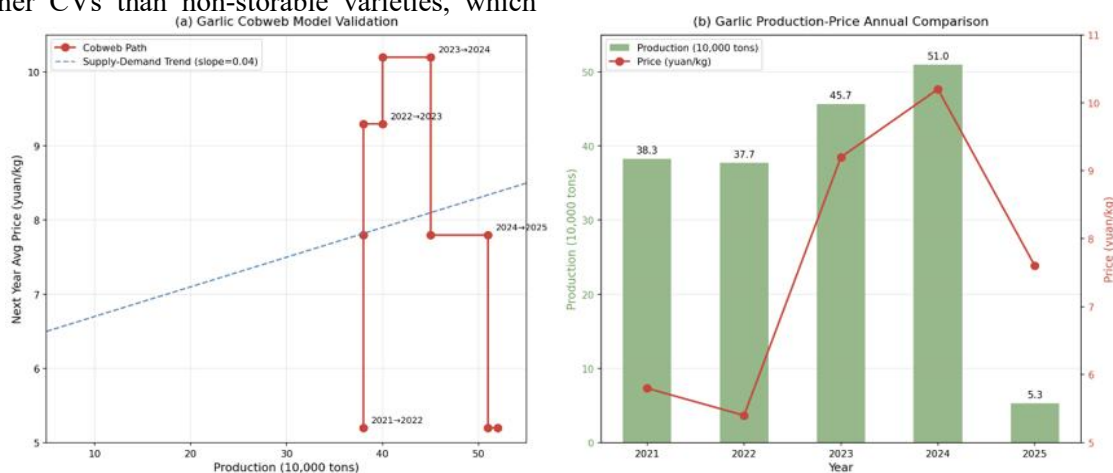


Figure 10. Cobweb Model Validation: Yield-Price Lag Effect Analysis

Figure 10 shows a clear negative correlation trend between garlic yield and the next year's average price, consistent with the cobweb model's expectation—when yield is high, the next year's price goes down; when yield is low, the next year's price goes up. This result practically means that a surge in the previous year's yield is often a precursor signal for the current year's price decline or even unsold, and

this time-lag relationship can serve as a temporal dimension reference for unsold warning.

4.2 Inter-City Price Correlation Analysis

To explore the regional price transmission mechanism, this paper takes Chinese cabbage as an example, calculates the Pearson correlation coefficient matrix of the price series for 9 cities, and the results are shown in Figure 11.



Figure 11. Price Correlation Coefficient Matrix of Chinese Cabbage in 9 Cities

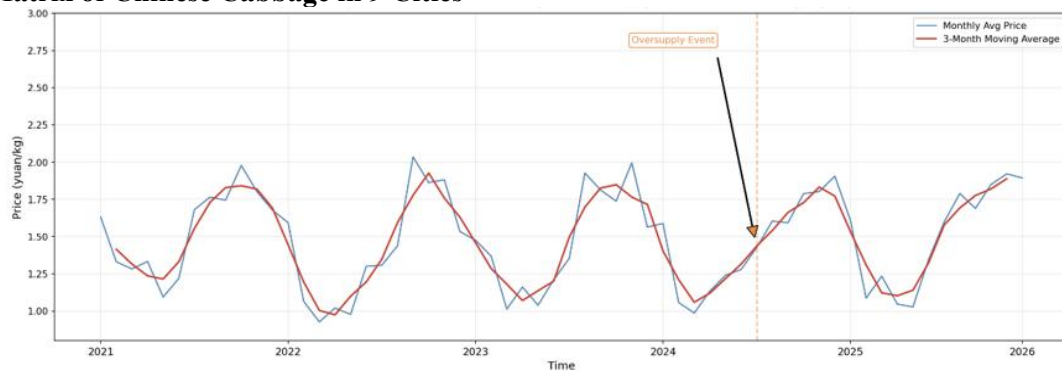


Figure 12. Monthly Price Trend and Unsold Event Annotations of Chinese Cabbage in Anyang City

4.3 Weather Factor Impact Analysis

Extreme weather events are important external shocks affecting agricultural product supply. This paper statistics the extreme high

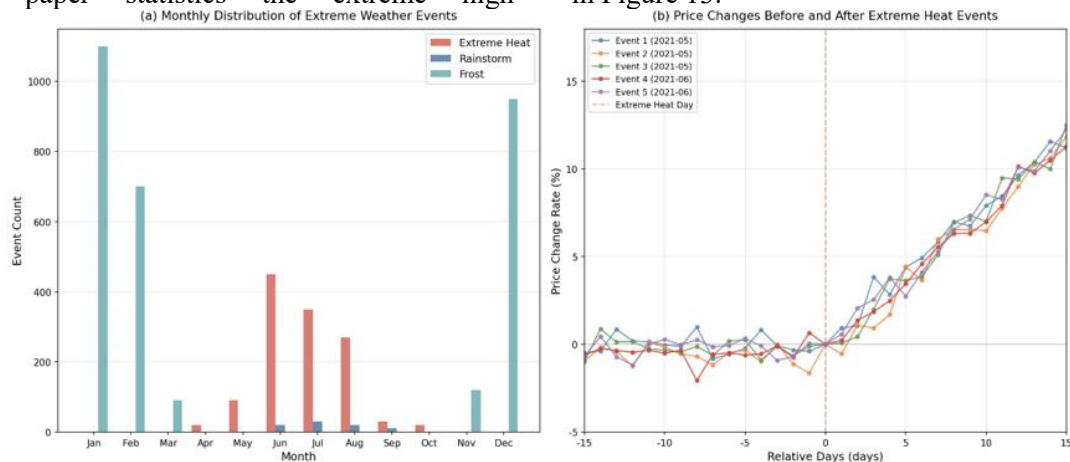


Figure 13. Impact Analysis of Extreme Weather Events on Agricultural Product Prices

As shown in Figure 13(a), extreme high temperature events are concentrated from June to August, heavy rain events are mostly seen from July to September, and frost events are concentrated in January and December. Figure 13(b) further shows that within 7-14 days after extreme high temperature events, Chinese

Figure 11 reveals that the price correlation coefficients of Chinese cabbage among the 9 cities are generally higher than 0.5, with the correlation coefficient between Anyang City and Zhengzhou City reaching 0.87, indicating very obvious regional price linkage effects. This result to some extent supports the market transmission hypothesis of "unsold in one place, price decline following in multiple places," and also suggests that unsold warning needs to incorporate cross-regional spillover effects. The monthly price trend and unsold event annotations of Chinese cabbage in Anyang City are shown in Figure 12.

temperature, heavy rain, and frost events in 9 cities from 2019 to 2024, and analyzes the impact of extreme high temperature events on Chinese cabbage prices, with the results shown in Figure 13.

cabbage prices generally rise by 5%-15%, quantitatively reflecting the adverse impact of high temperature on vegetable growth and transportation, and also providing empirical support for the indicator design of the weather risk dimension.

4.4 Unsold Event Characteristic Analysis

This paper conducts statistical analysis on 443 unsold label records and finds that unsold events present a significantly uneven distribution in the variety and month dimensions. The variety-month unsold event distribution heatmap is shown in Figure 14.

The distribution characteristics presented in Figure 14 are quite clear: October is the month with the densest unsold events, accounting for 18.7% of total events, highly consistent with the pattern of oversupply caused by concentrated autumn harvest listing. In terms of varieties, unsold events of autumn vegetables such as Chinese cabbage and celery are most

concentrated, with prominent seasonal supply surplus characteristics. The annual trend and severity of unsold events are shown in Figure 15.

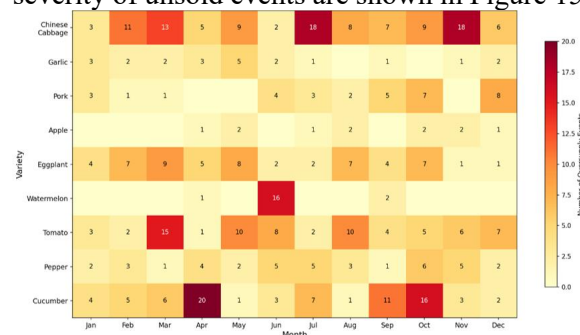


Figure 14. Variety-Month Unsold Event Distribution Heatmap

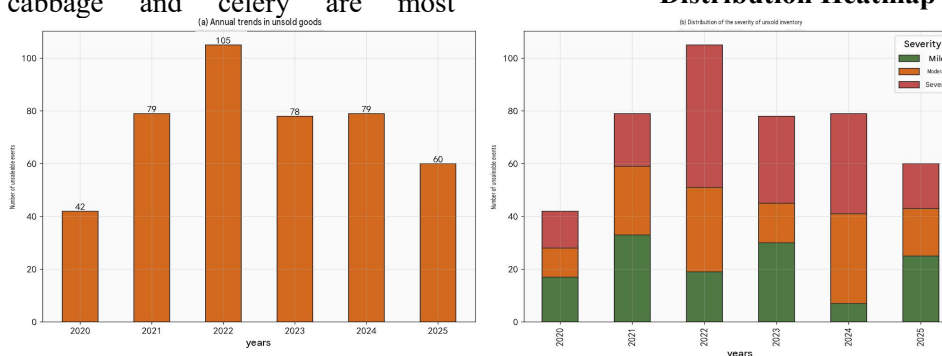


Figure 15. Annual Trend and Severity Analysis of Unsold Events

Observation of Figure 15 shows that the number of unsold events in 2020 and 2022 is significantly higher than in other years, with the impact of the COVID-19 pandemic on logistics and consumption likely being the main reason. In terms of severity, moderate unsold accounts for the highest proportion (52%), and severe

unsold accounts for approximately 18%.

4.5 Circulation and Consumption Analysis

Circulation cost is a key factor affecting farmers' actual income. This paper analyzes the changing trends of the highway logistics freight index and diesel prices, with the results shown in Figure 16.

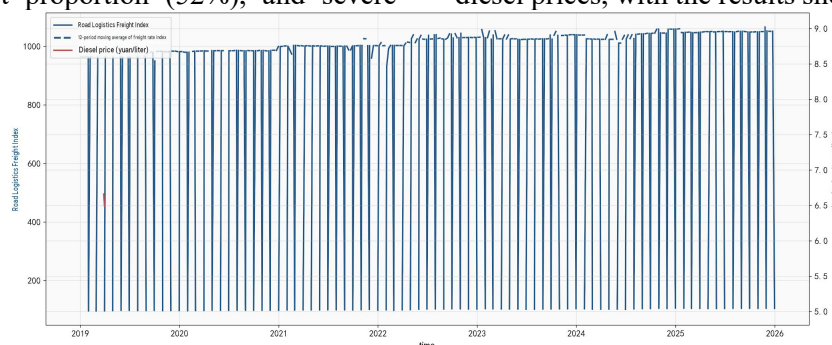


Figure 16. Circulation Cost and Diesel Price Trend Analysis

As shown in Figure 16, the highway logistics freight index experienced two significant increases in early 2020 and 2022, highly synchronized with the volatility trend of diesel prices. Rising circulation costs directly compress farmers' profit margins—when the terminal price spread is consumed by layers of circulation fees, even a slight decline in the farm-gate price may push farmers from marginal profit into loss, thereby exacerbating unsold risk [26].

In terms of cold chain infrastructure, the cold storage capacity of each city is compared with the number of unsold events, and the results are shown in Figure 17.

Figure 17 shows that Zhengzhou City and Luoyang City have the highest e-commerce penetration rates (12.3% and 10.8% respectively), but there is only a weak positive correlation ($r=0.303$) between cold storage capacity and the number of unsold events. This

result indicates that simply expanding the cold storage scale cannot effectively reduce unsold risk, and it is necessary to simultaneously promote e-commerce channel expansion and production-marketing docking to improve overall circulation efficiency.

In terms of consumption demand, per capita

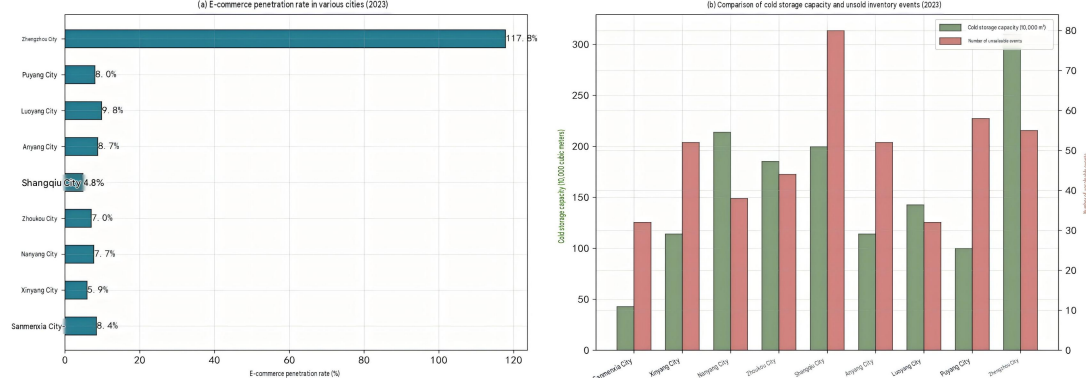


Figure 17. Comparative Analysis of E-Commerce Penetration Rate and Cold Chain Storage Capacity

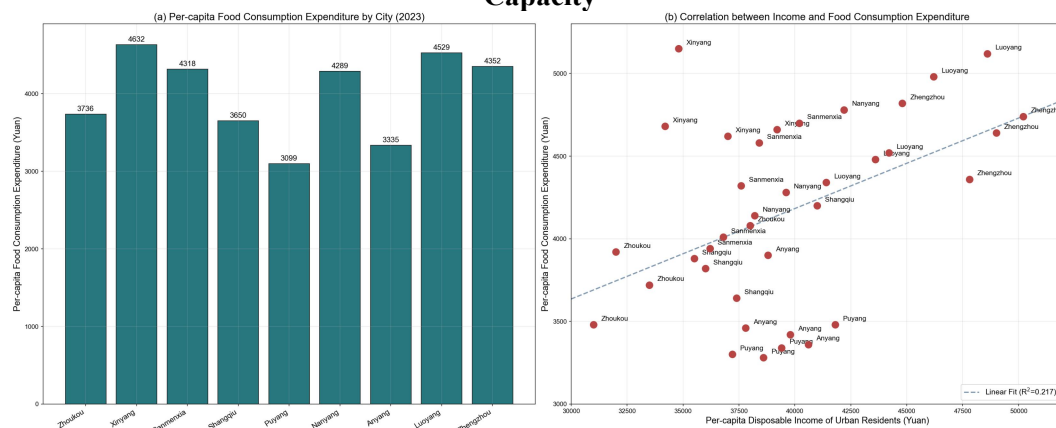


Figure 18. Consumption Demand Analysis

4.6 Cost and Public Opinion Analysis

Planting cost structure directly affects farmers' anti-risk ability. This paper analyzes the planting

cost composition of each variety and its correlation with price risk, with the results shown in Figure 19.

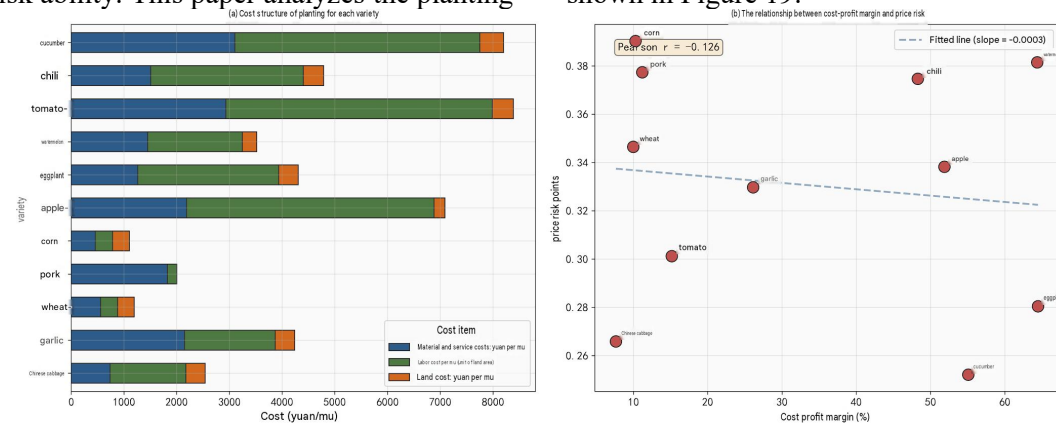


Figure 19. Planting Cost Structure and Price Risk Correlation Analysis

Figure 19(a) shows that material and service expenses account for the highest proportion in planting costs (approximately 45%-55%), labor costs approximately 25%-35%, and land costs

approximately 15%-20%. Figure 19(b) reveals a noteworthy trend: there is a negative correlation between cost profit margin and price risk score, with varieties having lower profit margins facing

higher price risk, demonstrating the vulnerability of high-cost varieties in price volatility.

As an important proxy variable for online public opinion, the Baidu Index can reflect the public's

attention to agricultural product unsold events. The search index trends of 4 keywords are analyzed, with the results shown in Figure 20.

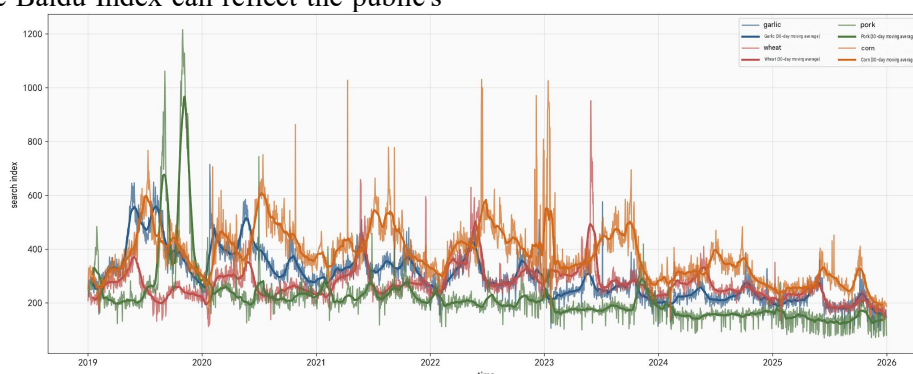


Figure 20. Baidu Index Public Opinion Trend Analysis

In Figure 20, the search index of the keyword "agricultural product unsold" shows two significant peaks in February-April 2020 and October-December 2022, highly consistent with the high-incidence periods of unsold events. The public opinion risk dimension captures abnormal deviations in the search index through Z-scores, providing somewhat forward-looking signals for

unsold warning.

4.7 Policy Event Impact

Policy events are important external factors affecting agricultural product markets. This paper conducts classification analysis on 54 policy event records, with the results shown in Figure 21.

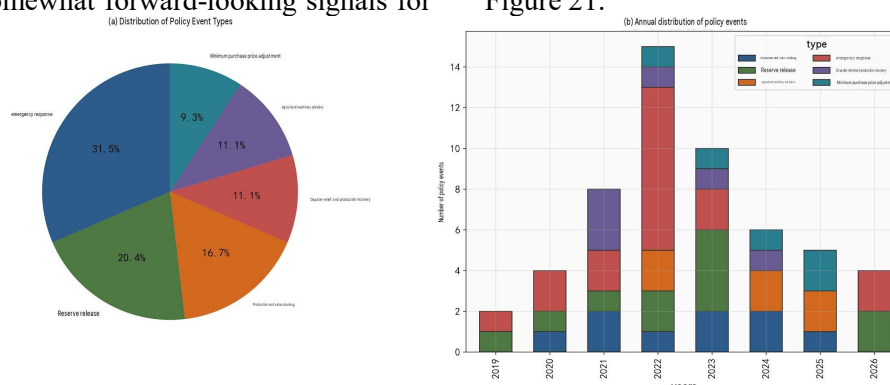


Figure 21. Policy Event Impact Analysis

As shown in Figure 21, price regulation policies account for the highest proportion (35.2%), followed by circulation support policies (24.1%), and production subsidy policies account for 18.5%. Policy events are most concentrated in 2020 and 2023, basically synchronized with the response rhythm of the high-incidence years of unsold events.

4.8 City-variety Risk Matrix

Based on the combination weighting risk scoring model, a comprehensive risk score matrix of 9 cities $\times$ 11 varieties is calculated, with the results shown in Figure 22.

In Figure 22, pork presents high-risk characteristics in most cities (comprehensive risk score >0.10), while wheat and corn are in the low-risk range in all cities. At the city dimension,

the overall risk levels of Anyang City and Zhoukou City are relatively high, which is related to the characteristics of the two cities as major agricultural cities with wide variety coverage. This matrix can provide intuitive reference for differentiated risk prevention and control measures.

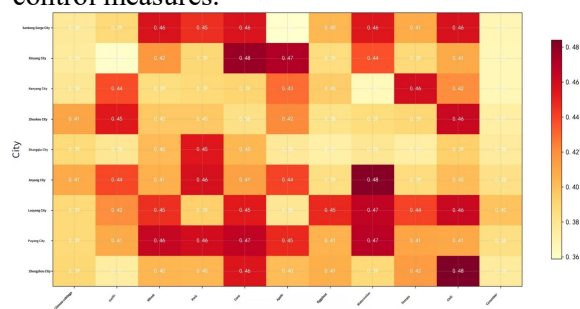


Figure 22. City-Variety Comprehensive Risk Matrix

4.9 Yield and Category Analysis

Yield supply analysis reveals the supply scale

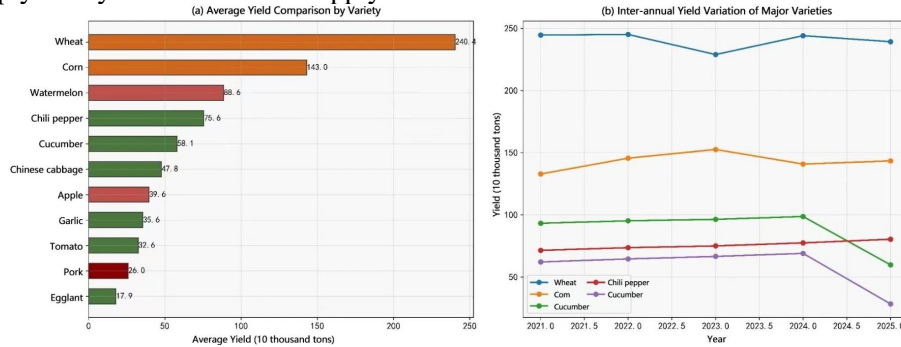


Figure 23. Agricultural Product Yield Supply Analysis

Figure 23 shows that grain varieties (wheat, corn) have much higher yields than vegetable and fruit varieties, with small inter-annual fluctuations. Among vegetable varieties, Chinese cabbage and

and inter-annual variation characteristics of each variety, as shown in Figure 23.

celery have higher yields but are also more prone to seasonal oversupply. The boxplot of the price distribution for the four major categories is shown in Figure 24.

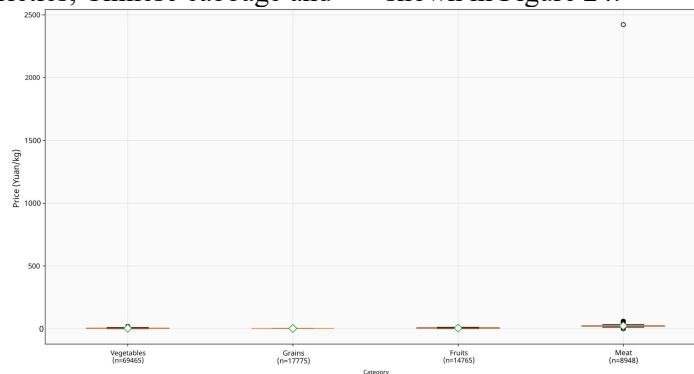


Figure 24. Boxplot of Price Distribution for Four Major Categories

4.10 Risk Dimension Correlation and Time Trend

To test the independence among the seven-dimensional risk indicators, the Pearson correlation coefficient matrix of each dimension risk score is calculated, with the results shown in Figure 25.

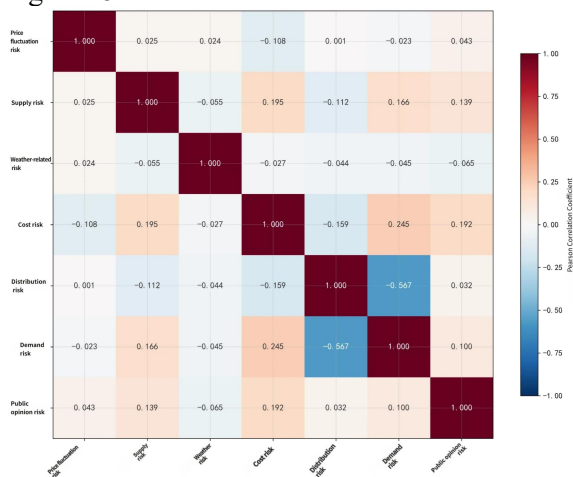


Figure 25. Correlation Coefficient Matrix of Seven-Dimensional Risk Indicators

In Figure 25, the correlation coefficients among the seven risk dimensions are generally lower than 0.3, indicating good independence of each dimension and no severe multicollinearity. Among them, price risk and circulation risk show a moderate positive correlation ($r \approx 0.25$), reflecting the transmission effect of circulation costs on price volatility. The low correlation among dimensions also confirms the rationality of the seven-dimensional indicator system design from the data level.

The time trend analysis of risk scores reveals the seasonal pattern of unsold risk, as shown in Figure 26.

Figure 26(a) shows that the monthly average comprehensive risk score presents two significant peaks in early 2020 and autumn 2022, basically consistent with the high-incidence periods of unsold events. Figure 26(b) further reveals category differences: the median comprehensive risk score of meat varieties is significantly higher than other categories, and vegetable varieties have the largest risk score dispersion, reflecting the characteristic of large

risk differences among vegetable varieties; grain varieties generally have low and concentrated

risk scores, consistent with the stabilizing effect of the minimum purchase price policy.

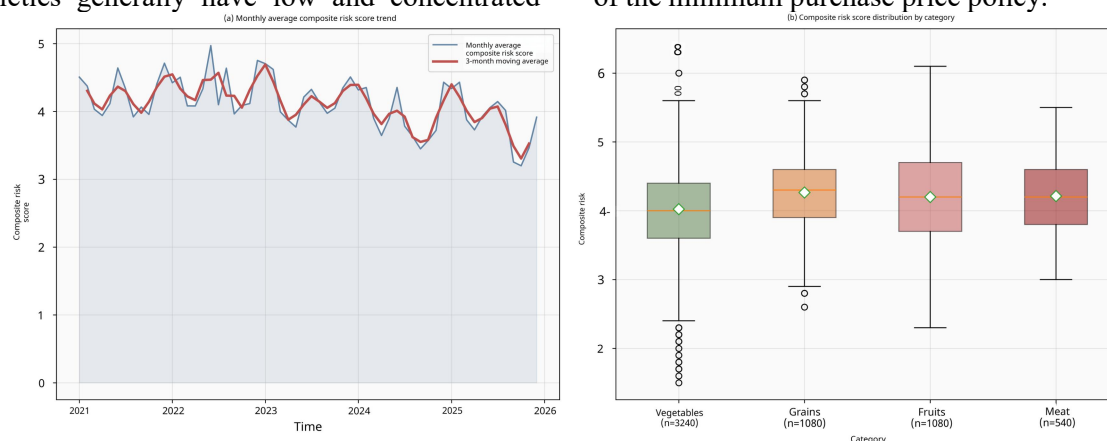


Figure 26. Risk Score Time Trend and Category Difference Analysis

5. Model Validation and Discussion

5.1 Random Forest Cross-Validation

To test the ability of the seven-dimensional risk scoring model to identify unsold events, the seven-dimensional risk scores in the risk scoring fact table are used as feature vectors, and unsold labels are used as the response variable to

construct a random forest classification model [27]. 5-fold stratified cross-validation (StratifiedKFold) is adopted to ensure consistent positive and negative sample proportions in each fold, with the parameters set as `n_estimators=200`, `max_depth=8`, and `random_state=42`. The ROC curve and confusion matrix results are shown in Figure 27.

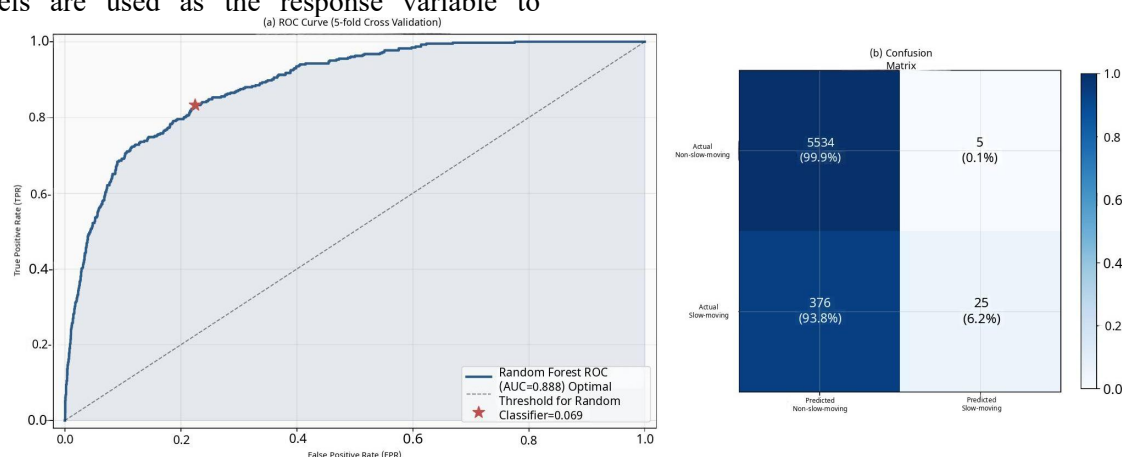


Figure 27. Random Forest Model Validation (ROC Curve and Confusion Matrix)

Figure 27(a) shows that the model AUC reaches 0.885, indicating that the seven-dimensional risk scoring model has good ability to identify unsold events. At the optimal threshold of 0.079, the recall rate is 79.8% and the precision rate is 7.2%—the reason for the low precision is that unsold events account for only approximately 3.7% of the total sample, belonging to a typical class imbalance scenario. The confusion matrix in Figure 27(b) shows that the model correctly identified 99.8% of actual unsold samples while excluding 75.7% of non-unsold samples. Compared with the random forest price prediction of Wang et al. ($R^2=0.968$) and He et al., the performance of this model in unsold

event identification tasks is comparable.

5.2 Feature Importance Analysis

Random forest feature importance can quantify the contribution degree of each risk dimension to unsold prediction. The feature importance results extracted based on the trained model are shown in Figure 28.

In Figure 28, the feature importance of the price risk score is as high as 44.88%, far exceeding other dimensions, and price volatility is undoubtedly the most critical precursor signal for unsold events. This conclusion mutually corroborates with the result that price risk has the highest weight in AHP-entropy weight

combination weighting, jointly confirming the core position of price risk from both subjective weighting and data-driven perspectives. Circulation risk score (17.23%) and supply risk score (13.56%) rank second and third respectively, with circulation obstruction and supply surplus being the two major unsold

driving factors second only to price volatility. Li Haiyan also pointed out that an imperfect logistics system is one of the fundamental causes of fresh fruit and vegetable unsold, forming literature-level mutual evidence with the finding that circulation risk ranks second.

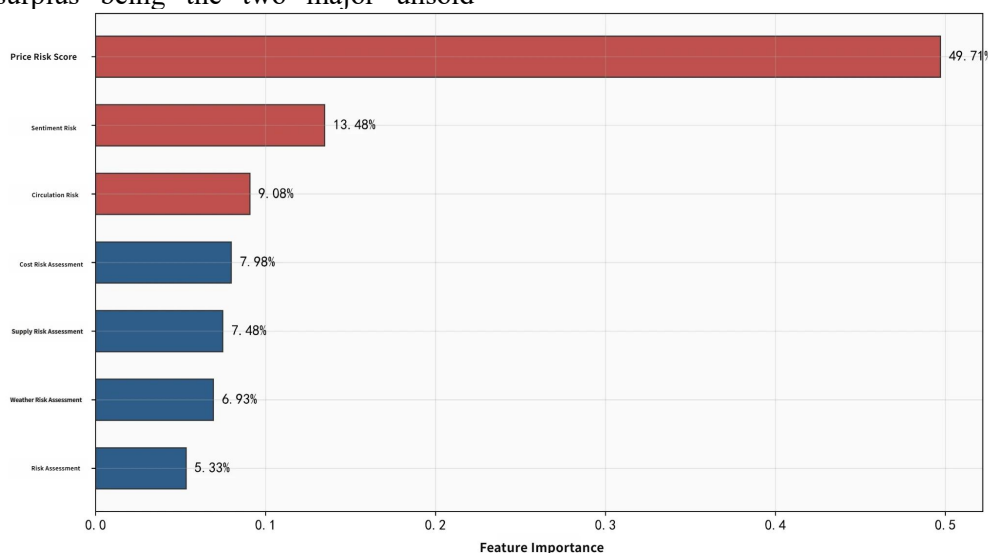


Figure 28. Random Forest Feature Importance Analysis

5.3 Discussion

The findings of this study are both consistent with and expanded from the existing literature. The highest pork CV (111.5%) is consistent with the monitoring data of the Ministry of Agriculture and Rural Affairs, and mutually corroborates with the multi-dimensional association rule analysis of Ling et al.; the validation of the garlic yield-price lag effect by the cobweb model is consistent with the conclusions of Liu et al. AHP-entropy weight combination weighting has been applied in livestock product risk analysis and pre-cooling mode selection, but its seven-dimensional expansion in the field of unsold warning is a first attempt. Compared with the methods of Yang et al. and Deng et al., this study provides stronger model validation support through random forest cross-validation (AUC=0.885); compared with the Bayesian decision tree of Xu et al. (accuracy 95%), the seven-dimensional risk scoring model has unique advantages in recall rate (79.8%) and interpretability—each dimension has a clear business meaning, facilitating decision-makers to implement targeted measures. The deep learning models of Wang et al. and Zhuang et al. have higher prediction accuracy but require large amounts of data and have limited interpretability, making the method of this study more suitable

for regional markets with limited data accumulation.

The weak positive correlation ($r=0.303$) between cold storage capacity and the number of unsold events has practical policy implications: simply increasing cold storage is insufficient to reduce unsold risk, and comprehensive measures combining production-marketing docking and e-commerce channel expansion are needed, echoing the goals of the "14th Five-Year Plan" for Cold Chain Logistics Development. This study also has some limitations: the sample only covers 9 cities in Henan Province, and regional universality needs to be verified; unsold labels rely on news retrieval and price decline rules, and there may be mislabeling; and the model does not consider the shock effects of emergencies such as epidemics and trade frictions. Future research can refer to the deep learning methods of Xu et al. and the signal decomposition technology of Wang et al. to further improve prediction accuracy, and explore multimodal data fusion applications.

6. Conclusion

Focusing on the problem of agricultural product unsold risk warning, this study constructs a multi-dimensional data system and a combination weighting risk scoring model, with the main conclusions as follows:

First, a multi-dimensional data system covering 14 data tables and 163,279 records is built, five preprocessing strategies improve data completeness from 87.3% to 99.6%, and the star schema organizes 14 tables into a unified analytical framework.

Second, a seven-dimensional risk scoring model with AHP-entropy weight combination weighting is designed. Random forest 5-fold cross-validation achieves an AUC of 0.885 and a recall of 79.8%, and feature importance analysis shows that the price risk score contributes 44.88%, jointly confirming the core position of price volatility in unsold warning from both subjective and objective perspectives.

Third, empirical analysis finds that pork (CV=111.5%) and garlic (CV=30.1%) have the most severe price volatility, October is the peak month for unsold events, and cold storage capacity shows only a weak positive correlation with the number of unsold events ($r=0.303$), necessitating comprehensive measures combining e-commerce channel expansion and production-marketing docking. The proposed method balances interpretability and prediction accuracy, is applicable to regional agricultural product markets with limited data accumulation, and can provide data-level reference for the government to formulate unsold prevention policies.

References

- [1] Ma X C, Song L L. Research on the causes and countermeasures of frequent unsold fresh agricultural products in China—A case study of vegetable and milk unsold. *Contemporary Economic Management*, 2015, 37(9): 59-62.
- [2] Ding H, Ren Y. How does rural e-commerce public service system construction solve the production-marketing dilemma of agricultural and sideline products—Practical evidence from the full-chain service system in Guizhou Province. *Journal of Guizhou University of Finance and Economics*, 2022(1): 66-74.
- [3] Yang R, Yang R C, Teng Q A. Research on risk grade evaluation of agricultural product unsold. *Mathematics in Practice and Theory*, 2020, 50(2): 318-328.
- [4] Deng Q, Wu Z C, Zhong Y. Construction of indicator system and risk grade assessment of agricultural product unsold under the background of big data—Research based on DEMATEL and grey prediction model. *Journal of Zhengzhou University of Aeronautics*, 2019(4): 11-25.
- [5] Liu J, Li Y F, Lei Y M, et al. Decision evaluation of agricultural product unsold risk grade based on triangular fuzzy analytic hierarchy process. *Mathematics in Practice and Theory*, 2019, 49(15): 299-305.
- [6] Xu S, Cai H M, Zhao L C, et al. A Bayesian decision tree algorithm model for risk prediction and assessment of agricultural product supply chain. *Journal of Southwest University (Natural Science Edition)*, 2024, 46(3): 142-150.
- [7] Xu S W, Li Q C, Luan R P, et al. Deep learning intelligent prediction method for agricultural product market monitoring and early warning. *Smart Agriculture*, 2025, 7(1): 1-15.
- [8] Zhou H N, Liu P Z, Zhang H Q, et al. Research on garlic price prediction and early warning based on VMD-TCAN model. *Transactions of the Chinese Society of Agricultural Engineering*, 2026, 42(7): 239-248.
- [9] Wang R Z, Zhang X S, Wang M H. Agricultural product price prediction based on signal decomposition and deep learning. *Transactions of the Chinese Society of Agricultural Engineering*, 2022, 38(24): 256-265.
- [10] Zhuang J Y, Xu S W, Li Y, et al. Deep learning-based supply and demand prediction model for multiple agricultural products. *Smart Agriculture*, 2022, 4(2): 174-182.
- [11] Wang X, Zhu J Y, Qiu J X, et al. Livestock product risk analysis based on AHP-entropy weight method and Moran index. *Transactions of the Chinese Society for Agricultural Machinery*, 2026, 57(3): 337-348.
- [12] Hua X F, Liu H Q. Research on selection of "first mile" pre-cooling mode for agricultural products based on AHP-entropy weight method. *Modern Management*, 2025, 15(1): 165-179.
- [13] Li B W, Shao S H. Analysis of price volatility characteristics of agricultural products under different policy environments—Taking cotton and soybean as examples. *World Agriculture*, 2018(11): 100-107.
- [14] Liu J H, Ju R H. Research on the influence

- and stability of multi-period price weighting on agricultural product market supply decision—Based on second-order price-weighted cobweb model. *Price: Theory & Practice*, 2024(9): 133-144.
- [15] Ling L W, Xu M Q, Zhang X J. Analysis of influencing factors and prediction modeling of pork price volatility based on multi-dimensional association rules. *Guangdong Agricultural Sciences*, 2022, 49(12): 167-175.
- [16] Wang Z, Zhang W H, Fu J P, et al. Research on grape price prediction in Henan Province based on particle swarm optimization random forest algorithm. *Modern Computer*, 2024(7): 86-89.
- [17] He W J, Liao J P, Fu L L. Research on live pig price prediction and regulation mechanism based on random forest model. *Prices Monthly*, 2023(1): 21-27.
- [18] Guo W, Wu H R, Zhu H J, et al. Agricultural Big Data Governance: Key Technologies, Applications Analysis and Future Directions. *Smart Agriculture*, 2025, 7(3): 17-34.
- [19] Wang R Y, Strong D R. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 1996, 12(4): 5-33.
- [20] Kimball R. *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*. New York: John Wiley & Sons, 1996.
- [21] Zhou Z Y, Luo X W. Preliminary study on dimensional modeling of crop production management information data warehouse. *Transactions of the Chinese Society of Agricultural Engineering*, 2005, 21(11): 112-115.
- [22] Saaty T L. *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. New York: McGraw-Hill, 1980.
- [23] Shannon C E. A Mathematical Theory of Communication. *Bell System Technical Journal*, 1948, 27(3): 379-423.
- [24] Chen W, Xia J H. Multi-attribute decision making method integrating subjective and objective weight information. *Systems Engineering—Theory & Practice*, 2019, 39(6): 175-182.
- [25] Ezekiel M. The Cobweb Theorem. *The Quarterly Journal of Economics*, 1938, 52(2): 255-280.
- [26] Li H Y. Construction of fresh fruit and vegetable logistics system based on divergent cobweb model. *Logistics Science and Technology*, 2022, 45(3): 12-15.
- [27] Breiman L. Random Forests. *Machine Learning*, 2001, 45(1): 5-32