International Journal of New Developments in Engineering and Society
ISSN 2522-3488 Vol. 8, Issue 2: 37-43, DOI: 10.25236/IJNDES.2024.080206

# Data Analysis and Pricing Model Exploration in Supermarket Vegetable Sales

**Zixin Zeng[1,#], Yuhao Wan[2,#], Yang Zhang[1,#]**

[1]*School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, 214122, China*
[2]*School of Internet of Things Engineering, Jiangnan University, Wuxi, 214122, China*
[#]*These authors contributed equally.*

***Abstract:*** *Due to the short shelf life of vegetables, this article aims to provide supermarkets with automatic pricing and replenishment decision support for vegetable products, aiming to optimize category structure, increase profitability, reduce loss rates, and enhance service quality. To this end, we employed Spearman's correlation model, hierarchical clustering model, integer programming model, and LSTM time series model to conduct in-depth research on the purchase quantity and pricing strategies of vegetable products. The research results indicate that: 1) through correlation analysis, we can understand the relationship between the sales volume of various vegetables; 2) through linear regression fitting and the establishment of LSTM time series models, we can better predict daily sales volume and daily replenishment volume, thereby providing a basis for pricing decisions; 3) in the process of vegetable pricing, combining genetic algorithm optimization with integer programming models can make pricing decisions more reliable. Additionally, adopting different pricing strategies for different vegetables based on consumers' purchasing psychology can help increase vegetable sales.*

***Keywords:*** *Long Short-Term Memory (LSTM) model, Hierarchical Clustering, Integer Programming, Genetic Algorithm*

## 1. Introduction

Due to the wide variety of vegetables and the special time of purchase, merchants need to make quick replenishment decisions based on limited information. Therefore, it is necessary to formulate replenishment and pricing strategies based on actual circumstances. Through automatic replenishment, the inventory turnover rate can be increased and waste can be reduced. A reasonable pricing strategy can help maximize sales and profit margins, and enhance the profitability of the supermarket. In addition, research on vegetable products can provide references for inventory management and pricing of other fresh products, promoting the overall development of fresh supermarkets. This article adopts effective optimization algorithms to establish mathematical models to provide support for supermarket decision-making.

In this study, the Data source is www.mcm.edu.cn/index_cn.html. We explore the correlation of vegetable sales, cluster analysis, cost-plus pricing[1], daily replenishment quantity prediction, and decision-making to maximize supermarket profits. Firstly, by calculating the Spearman correlation coefficient among six types of vegetables, we found a high correlation between chili and edible mushroom sales (correlation coefficient of 0.6). Secondly, using hierarchical cluster analysis to group over 250 types of vegetables, we found that the sales of broccoli, lotus root, and Wuhu green peppers are mutually influenced, while Chinese cabbage is less affected by other vegetables. Then, we plotted the curve of sales volume over time for each category and found that at the end of 2022, the sales of all vegetables increased significantly. Based on the cost-plus pricing method, we established a linear regression model and combined it with the LSTM time series model to predict daily replenishment quantities. The accuracy of the prediction results exceeded 90%. Additionally, we used integer programming models [2]to investigate daily replenishment totals and pricing strategies for vegetables. By using genetic algorithms, we obtained the optimal solution that maximizes supermarket profits. Finally, based on experimental results and multiple factors, we provided comprehensive suggestions for supermarkets to better formulate replenishment and pricing decisions for vegetable products.

ISSN 2522-3488 Vol. 8, Issue 2: 37-43, DOI: 10.25236/IJNDES.2024.080206

## 2. Exploration of Vegetable Relationships

### *2.1 Descriptive Statistical Analysis*

Firstly, descriptive statistical analysis was conducted on the data. The order of variance values of total sales of vegetables is cruciferous vegetables>aquatic root and stem vegetables>chili>solanaceous vegetables>leafy vegetables>edible fungi. Therefore, it is preliminarily judged that the sales distribution of individual products of cruciferous vegetables is relatively scattered and their correlation is relatively weak, while the sales distribution of individual products of edible fungi is relatively concentrated and their correlation is relatively strong. To visually show the sales proportion of each vegetable category, the following Figure 1 shows the pie chart:

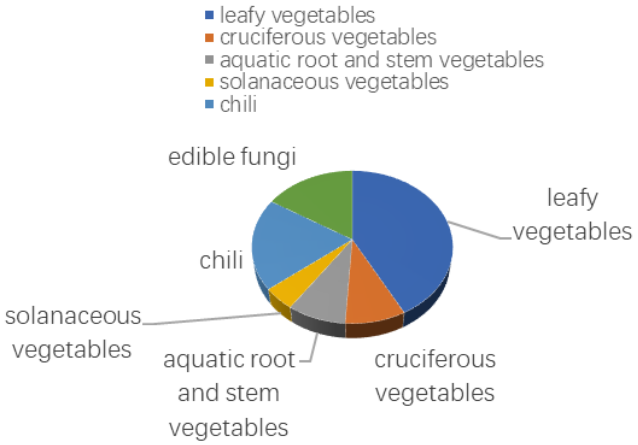


*Figure 1: Fan chart of the proportion of sales by vegetable category*

### *2.2 Spearman correlation analysis*

Analysis of the interrelationships between six categories: leafy vegetables, aquatic root and stem vegetables, solanaceous vegetables, chili, edible fungi, and cruciferous vegetables. There are three correlation coefficients in statistics that measure the relationship between data: Pearson[3], Spearman, and Kendall. The Spearman correlation coefficient is suitable for analyzing the relationships between variables that are categorical or have a distribution that is non-normal or unknown. Since vegetable categories are non-continuous variables, the Spearman correlation coefficient is used to analyze the relationships between categories.

The Spearman correlation coefficient ranges from -1 to +1. A value of -1 indicates complete inverse correlation, +1 indicates complete positive correlation, and 0 indicates no correlation between the data. The calculated correlation coefficients between the six vegetable categories are shown in Table 1 below:

*Table 1: Spearman correlation coefficient table for the six vegetable categories*

| | philodendron | cauliflower | Aquatic rhizomes | eggplant | capsicum | edible mushroom |
|---|---|---|---|---|---|---|
| philodendron | 1 | 0.1 | -0.6 | 0.3 | -0.6 | -0.8 |
| cauliflower | 0.1 | 1 | -0.8 | 0.1 | 0.3 | 0.1 |
| Aquatic rhizomes | -0.6 | -0.8 | 1 | -0.1 | 0.2 | 0.5 |
| eggplant | 0.3 | 0.1 | -0.1 | 1 | -0.5 | 0.2 |
| capsicum | -0.6 | 0.3 | 0.2 | -0.5 | 1 | 0.6 |
| edible mushroom | -0.8 | 0.1 | 0.5 | 0.2 | 0.6 | 1 |

In summary, there is a certain positive correlation between the sales of leafy vegetables, cruciferous vegetables, and solanaceous vegetables. Similarly, there is also a certain positive correlation between the sales of aquatic root and stem vegetables, chili, and edible fungi. This indicates that when people buy leafy vegetables, they may also buy cruciferous vegetables and solanaceous vegetables. However, the aquatic root and stem vegetables, chili, and edible fungi that are negatively correlated with leafy vegetables are less likely to be purchased at the same time.

ISSN 2522-3488 Vol. 8, Issue 2: 37-43, DOI: 10.25236/IJNDES.2024.080206

### *2.3 Hierarchical Cluster Analysis*

Due to the large number of individual products, up to 250, the normal distribution and correlation coefficient test cannot fully analyze the interrelationships among individual products. Therefore, we use hierarchical clustering analysis[4] to cluster the data and obtain cluster groups. Variables within the same group have a stronger correlation. By visualizing the data, we obtain a complete phylogenetic tree, in which most individual products do not have mutual relationships. We extracted the parts with mutual relationships as shown in Figure 2 below:

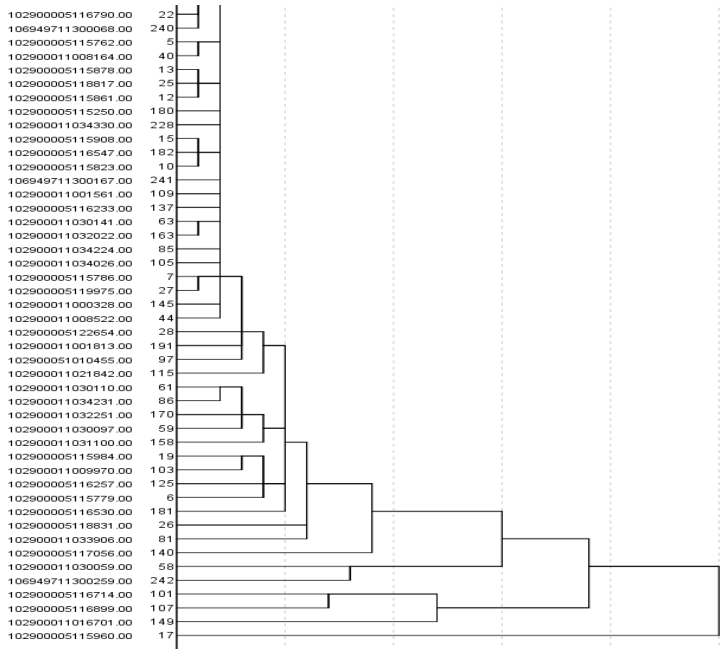


*Figure 2: Phylogenetic map of individual product cluster analysis*

From right to left on the phylogenetic tree, the first bifurcating line connects individual products that are grouped separately. Cabbage (102900005115960) has very little correlation with other vegetable products, indicating that the sales of other vegetables have minimal impact on it. However, there is a certain correlation between broccoli, lotus root, and Wuhu green peppers, suggesting that their sales may influence each other, allowing us to roughly predict the sales changes of other vegetables in the same category through changes in the sales of one vegetable. As for other vegetable products, due to their weak correlation, they can be grouped. Finally, considering that sales volume is a variable that significantly fluctuates over time, we selected days as the unit, summed the sales of each category on the same day, and drew the curve of each category's sales volume over time, as shown in Figure 3 below:

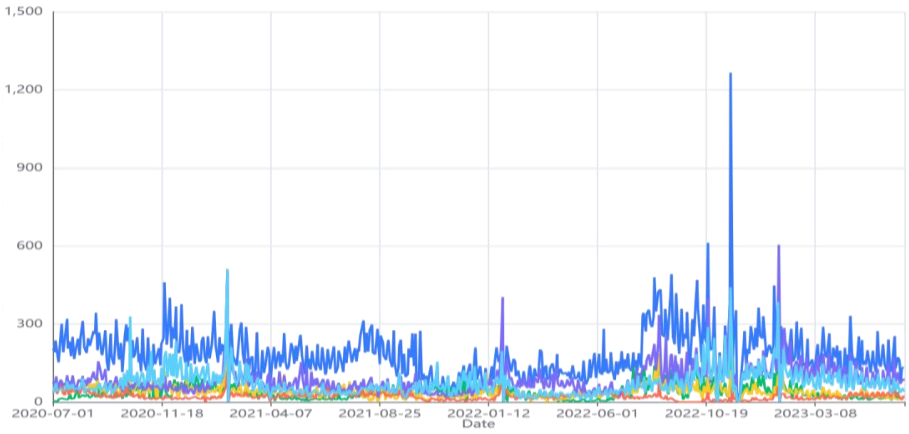


*Figure 3: Graph of sales volume by category over time*

By observing the above charts, we find that the annual sales volume of flowering and leafy vegetables is significantly higher than that of other categories, while the sales volume of eggplant vegetables is significantly lower than that of other categories, which is consistent with the previous conclusion. In

ISSN 2522-3488 Vol. 8, Issue 2: 37-43, DOI: 10.25236/IJNDES.2024.080206

addition, the charts show that there are seasonal and cyclical trends in the sales volume of vegetables. Specifically, sales of chili peppers, foliage, and mushrooms increase in the winter and decrease in the summer. 2022 The end of the year saw a significant increase in sales of all vegetables, with foliage vegetables showing particularly significant growth. We believe this may be influenced by unexpected factors such as natural disasters, epidemic outbreaks, and economic changes.

## 3. Forecast of trends in vegetable sales

### *3.1 Modeling LSTM time series*

To realize the accurate prediction of the total daily replenishment of each vegetable category in the coming week, we constructed a prediction model for the daily replenishment of vegetables based on collecting the basic data of vegetable sales, processing the predicted data of vegetable sales, and so on. In the prediction process, the long and short-term neural network processed prediction data is converted from single-column data to double-column data. For the total amount of replenishment at the moment of S, the two-column data corresponds to the replenishment data at the moment of S and the moment of S+1, thus converting the "unlabeled" data into "labeled" data. In terms of parameter settings, the prediction model has three long and short-term neural network logic units, such as the input layer, forgetting layer, and output layer, to accurately predict the total daily replenishment of vegetables under the activation function.

### *3.2 LSTM time series model[5] validation*

In the prediction comparison between the training set and the test set, we predicted the test set using the training set and verified its accuracy. The results are shown in Figure 4 below, where the prediction results of the training set and test set have a high degree of overlap. This indicates that the LSTM-based time series prediction method has better performance.

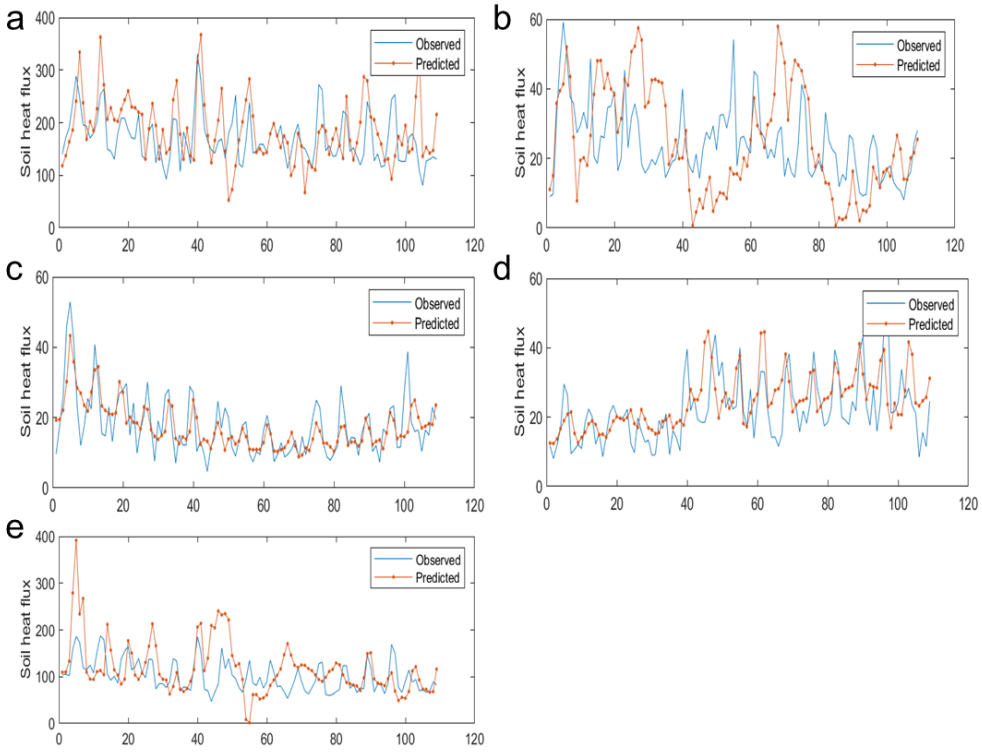


*Figure 4: Comparison of predicted and actual values for five vegetables. From a to e, they are leafy flowers, cauliflower, aquatic tubers, solanaceous vegetables, and chili peppers.*

### *3.3 LSTM time series model solving*

We fit 1095 data from 2020 to 2023 by (LSTM) and use LSTM to build a correlation between factors

ISSN 2522-3488 Vol. 8, Issue 2: 37-43, DOI: 10.25236/IJNDES.2024.080206

and train it by Matlab; 90% of the data is used to train the model and the rest is used to validate the accuracy of the model, and we get the prediction plots as shown in the following Figure 5.

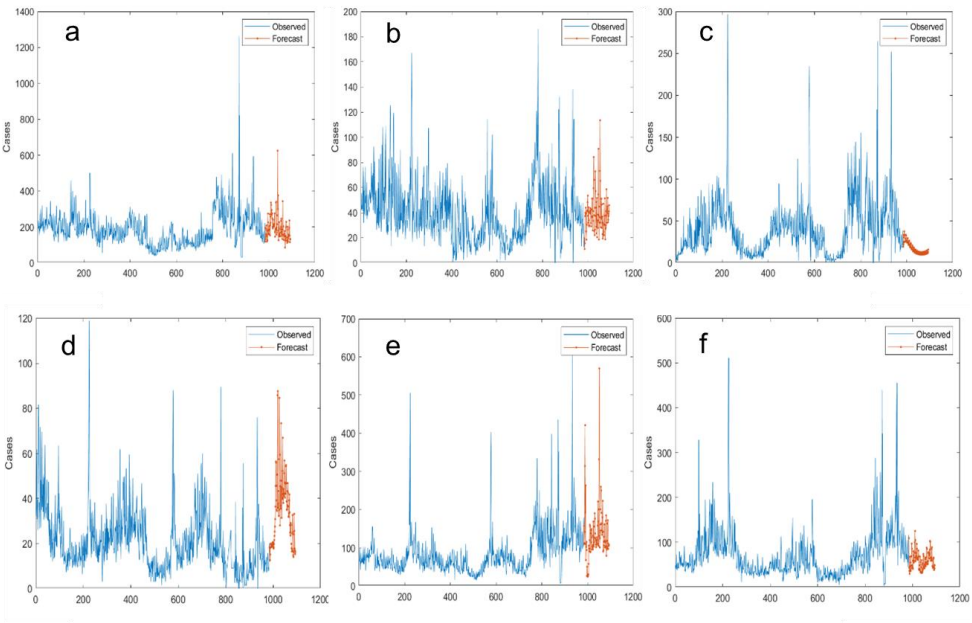


*Figure 5: Forecasts of total replenishment of foliage, cauliflower, aquatic roots and tubers, eggplants, peppers, and edible mushrooms.*

The final daily replenishment totals for each vegetable category for the coming week are obtained as shown in Table 2.

*Table 2: Projected total daily replenishment for each vegetable category for the coming week*

| | philodendron | cauliflower | Aquatic rhizomes | eggplant | capsicum | edible mushroom |
|---|---|---|---|---|---|---|
| 7-1 | 143.4391 | 19.78373 | 20.98011 | 23.8929 | 90.9523 | 56.8813 |
| 7-2 | 109.8989 | 14.36559 | 13.66348 | 18.2938 | 70.1541 | 37.9862 |
| 7-3 | 91.6328 | 12.99388 | 12.56712 | 17.6316 | 65.3719 | 30.0305 |
| 7-4 | 113.5676 | 13.56121 | 15.89001 | 20.8132 | 72.6002 | 40.2179 |
| 7-5 | 120.9082 | 17.0402 | 16.00011 | 16.1521 | 79.6609 | 42.9732 |
| 7-6 | 135.2453 | 17.21315 | 19.01663 | 20.9399 | 82.2053 | 54.1896 |
| 7-7 | 134.2231 | 17.01613 | 18.33456 | 22.7316 | 80.6871 | 50.5229 |

## 4. Achieving an optimal replenishment strategy

We developed an integer programming model[6] aimed at optimizing superstore revenue by selecting suitable replenishment quantities and pricing strategies across different vegetable categories.

To tackle this integer programming challenge, we defined an objective function to maximize and set constraints to outline the solution's boundaries[7]. These elements were integrated into a fitness function for evaluating potential solutions, represented as chromosomes through binary encoding. This approach converts each decision variable into binary form, with the chromosome length being the product of decision variables and binary digits (L = 12*m).

Our methodology[8] begins with creating an initial population of diverse solutions (chromosomes), from which we select the fittest individuals as parents for the next generation. We then apply crossover and mutation operations to introduce variability and enhance solution quality. Crossover involves swapping segments between two chromosomes to create a new solution, while mutation alters random genes within a chromosome to introduce new traits.

This iterative process progresses until a satisfactory solution emerges, encompassing steps from defining the objective and constraints, through chromosome representation, to population initialization and genetic operations. The optimal solution obtained from this process is detailed in Table 3 and Table

ISSN 2522-3488 Vol. 8, Issue 2: 37-43, DOI: 10.25236/IJNDES.2024.080206

4.

*Table 3: Part of predicted results for each category of vegetables*

| date | **cauliflower** | | **philodendron** | | **capsicum** | |
|---|---|---|---|---|---|---|
| | Total daily replenishment (kg) | Pricing strategy (%) | Total daily replenishment (kg) | Pricing strategy (%) | Total daily replenishment (kg) | Pricing strategy (%) |
| 2023-07-01 | 53.364 | 50.26% | 211.466 | 33.35% | 75.154 | 75.15% |
| 2023-07-02 | 20.123 | 25.26% | 260.557 | 27.53% | 108.967 | 108.98% |
| 2023-07-03 | 41.155 | 50.55% | 261.434 | 42.98% | 109.651 | 109.65% |
| 2023-07-04 | 39.275 | 87.46% | 164.155 | 20.48% | 43.093 | 43.10% |
| 2023-07-05 | 50.934 | 41.22% | 262.581 | 91.73% | 110.127 | 110.82% |
| 2023-07-06 | 31.55 | 39.37% | 260.573 | 38.84% | 108.937 | 108.97% |
| 2023-07-07 | 29.175 | 52.37% | 203.144 | 55.58% | 70.088 | 70.98% |

*Table 4: Another part of predicted results for each category of vegetables*

| date | **eggplant** | | **edible mushroom** | | **Aquatic rhizomes** | |
|---|---|---|---|---|---|---|
| | Total daily replenishment (kg) | Pricing strategy (%) | Total daily replenishment (kg) | Pricing strategy (%) | Total daily replenishment (kg) | Pricing strategy (%) |
| 2023-07-01 | 9.781 | 96.10% | 65.401 | 33.56% | 21.837 | 84.68% |
| 2023-07-02 | 8.742 | 41.85% | 66.972 | 100.63% | 11.688 | 44.01% |
| 2023-07-03 | 5.662 | 64.84% | 31.953 | 107.61% | 15.671 | 88.64% |
| 2023-07-04 | 10.865 | 105.21% | 49.359 | 86.88% | 9.679 | 32.94% |
| 2023-07-05 | 3.785 | 95.84% | 46.763 | 92.76% | 18.394 | 52.65% |
| 2023-07-06 | 6.454 | 109.71% | 58.61 | 91.70% | 10．924 | 33.99% |
| 2023-07-07 | 6.193 | 84.90% | 61.975 | 62.49% | 14.385 | 38.17% |

## 5. Conclusions

This paper provides a research idea and framework for solving the problem of automatic pricing and replenishment prediction of vegetables in the superstore sales domain. By exploring the relationship between vegetable sales through correlation coefficient and cluster analysis, combined with the LSTM model, integer programming, and genetic algorithm, we successfully predicted the sales trend and the total amount of replenishment of vegetables in a specified period, gave a pricing strategy to maximize the revenue of the superstore, and verified the feasibility of the method.

## References

*[1] Huang He. Research on automatic commodity pricing model based on deep learning[J]. Modern Commerce Industry, 2019, 40(09): 188-190.*
*[2] Liu Hebing, Han Jingjing, Zhong Chenhui, et al. Research on vegetable price prediction model based on multi-scale feature fusion[J]. Journal of Henan Agricultural University, 2022, 56(5): 858-867.*
*[3] Alabdallah A, Ohlsson M, Pashami S, et al.The Concordance Index decomposition -- A measure for*

ISSN 2522-3488 Vol. 8, Issue 2: 37-43, DOI: 10.25236/IJNDES.2024.080206

*a deeper understanding of survival prediction models[J]. 2022.DOI:10.48550/arXiv.2203.00144.*
*[4] Chen Liping, Xing Xiaodan, Zhang Yuting, et al. Research on Consumer Behavior in Online Shopping [J]. E-commerce Research, 2023, 14(4): 56-62. (in Chinese)*
*[5] Bekele R, Mcpherson M. A Bayesian performance prediction model for mathematics education: A prototypical approach for effective group composition[J].British Journal of Educational Technology, 2011, 42(3):395-416. DOI:10.1111/j.1467-8535.2009.01042.x.*
*[6] Lin Meina. Research on Sales forecasting based on multi-level time series model [D]. Jinan university, 2022.*
*[7] Ji Y N. Research on supply chain pricing strategy of fresh e-commerce under the background of online shopping promotion [J]. Logistics technology, 2022, (02): 140-144.*
*[8] Zhang Yan, Mou Jinjin, Wang Shuyun. Business super have some samples with price control supply chain optimization decision [J]. Journal of China management science, 2023, 31 (10): 266-275.*