LSTM-Based Time Series Prediction and Optimization of Replenishment Pricing Model

Yifan Wang *, Lanyi Meng, Zonghan Liu

School of Electrical and Automation Engineering, Hefei University of Technology Xuancheng Campus, Xuancheng, China, 242000

* Corresponding Author Email: wyf1356972977@163.com

Abstract. Historical sales data play a significant role in the formulation of future replenishment and pricing schemes. In this paper, on the one hand, the correlation relationship between the sales volume of various categories and individual vegetables is investigated by data visualization and calculation of correlation coefficients in different time dimensions, and their "complementary" or "synchronous" relationships are obtained; on the other hand, a more reasonable vegetable replenishment volume and pricing scheme are predicted and obtained through the LSTM time-series prediction model and linear programming model, which are both accurate and accurate at the same time. On the other hand, the LSTM time series prediction model and linear programming model are used to predict and obtain a more reasonable total vegetable replenishment and pricing scheme, which is accurate and flexible at the same time.

Keywords: Time dimension, Spearman correlation coefficient, LSTM, Time series forecasting.

1. Introduction

In the fresh food superstore, the freshness period of vegetables is relatively short, and the quality deteriorates with the sales time, so it needs to be replenished every day. Supermarkets based on the historical sales and demand for commodities to carry out vegetable purchase transactions, but merchants do not know exactly the specific single product and purchase price, the need to make the restocking decision for each vegetable category in this case, when a reliable market demand analysis is critical to the restocking and pricing decisions. (Data in this article are from http://www.mcm.edu.cn/)

2. Research on the correlation relationship between different categories of vegetable products or different single products

2.1. Reflections on associative relationships

Correlation can be divided into many categories, but according to the background of this paper, it should be analyzed from different categories and different single products first.[1] For the category level, it can be studied in terms of the interrelationship between category sales volume, etc. For the single product level, it can be studied from the different division methods of the timeline unit for correlation; finally, it can be analyzed from the aspect of correlation and supplemented with the clustering model to make a complete explanation of the distribution of data.

2.2. Study of correlations between different categories

2.2.1 Distribution pattern of category sales volume

(1) Hourly time units

July 1, 2020 was selected as a sample, and the change of sales volume over time was plotted with hour as the time unit, as shown in Fig 1. The basic characteristics of the distribution pattern are: except for root vegetables, the curve has two peaks, i.e., the first one at about 10:00 a.m. and the second one at about 16:00-18:00 p.m. Between the two peaks and after the second peak, the sales volume of vegetable category shows a Between the two peaks and after the second peak, the sales volume of vegetable categories showed a downward trend, and the peak time was more in line with the ideal time for neighboring residents to purchase vegetables.

(2) In days

After comparing several sets of data for several weeks near different dates, the most representative data for three weeks were found and a line graph of sales volume in days was drawn.2 It can be concluded that the sales data for the five vegetable categories other than root vegetables have a more obvious periodicity in days, and the trend of the sales volume for the five vegetable categories is highly consistent, with the peaks occurring steadily around the double days.

(3) By month

It can be seen that the sales volume of each vegetable category increases significantly around August, and the trend of the sales volume of the six vegetable categories is relatively consistent. Therefore, around August, the peak sales season of each vegetable variety is , i.e., the peak sales volume of each vegetable category occurs around August. The anomaly in the pattern for root vegetables is related to people's hoarding behavior,as shown in Fig 1 and 2.

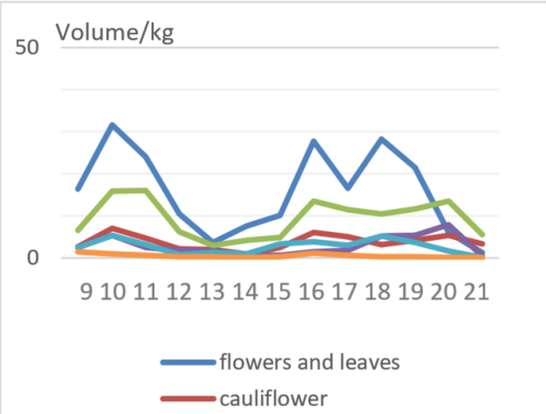

Fig. 1 Sales of each category of vegetables by time period

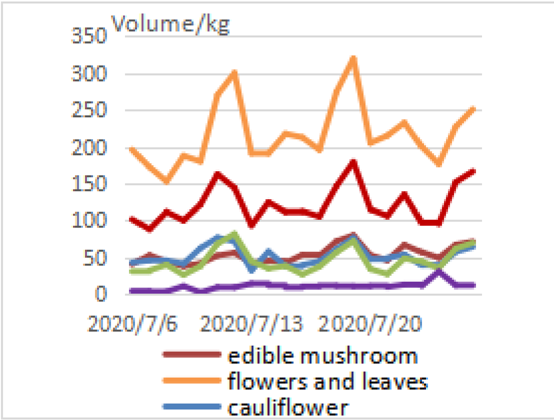

Fig. 2 Sales of each category of vegetables per day

2.2.2 Spearman correlation coefficient analysis of category sales volume

Based on the sales data of the above six vegetable categories, the non-parametric and intuitive Spearman correlation coefficient analysis model was used to analyze the correlation.

(1) Calculation of Spearman's correlation coefficient

Let Xi represent six vegetable categories (i=1,2,3,4,5,6), then the columns of the returned matrix X represent random variables consisting of the six vegetable categories, and the rows represent sales volume data observations at different time units.

its Spearman rank correlation coefficient:

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

where d_i is the rank difference between the column vectors of the X matrix, andr_s which is the spearman rank correlation coefficient (-1 $\leq r_s \leq$ 1), and n represents the number of rows of the constructed column vectors. [2]

(2) Hypothesis testing of Spearman's correlation coefficient

large sample data situation, Spearman's rank correlation coefficient of ther_s sampling distribution tends to be normal, therefore, the significance test was conducted using the normal approximation in the large sample scenario at that location. [3]

Statistics at this time

$$r_s\sqrt{n-1} \sim N(0,1) \tag{2}$$

The correlation coefficient matrix was tested for the 3 different time dimensions. When the unit is hourly, as shown in Table 2, the values of three vegetable categories, namely, foliage, pepper, and eggplant, are 0.2, 0.14, and 0.39, respectively, indicating that there is no significant correlation between their vegetable sales with edible mushroom varieties. While the rest of the sample observations are less than 0.1, it can be considered that there is a significant correlation between different vegetable categories,as shown in Table 1.

Table 1 Matrix of P-values obtained in terms of days.

1.00	0.06	0.02	0.09	0.00	0.01
0.06	1.00	0.00	0.09	0.05	0.20
0.02	0.00	1.00	0.01	0.04	0.14
0.09	0.09	0.01	1.00	0.11	0.39
0.00	0.05	0.04	0.11	1.00	0.07
0.01	0.20	0.14	0.39	0.07	1.00

Similarly, the results of hypothesis testing can be obtained when the unit of days and when the unit of months are used, i.e., when the unit of days is used, the test can be obtained that there is a strong correlation between the sales volume of different vegetable categories; when the unit of months is used, the test can be obtained that there is no significant correlation between the sales volume of cauliflower, cauliflower leaves, and chili peppers and sales data of eggplant categories. [4]

2.3. Study of correlations between single categories

In order to clarify the data relationship between items, different individual items within the same category and different individual items across categories are analyzed.

2.3.1 Distribution pattern of single product sales volume

(1) Single product analysis across categories

When exploring the sales patterns of individual items across categories, a sampling of individual items from different categories is done. It is found that the two have the law of fixed difference, which proves that the overall distribution of the category can describe the distribution law of most of the single products belonging to it.

Based on this, we can make a highly representative line graph of sales volume statistics of cabbage moss and red string pepper in terms of hours/days as shown in Fig.3,Fig.4,Fig.5,and Fig.6,which can more intuitively and accurately describe the distribution pattern of sales volume of a single product across categories.

By analyzing the sales volume statistics of cabbage moss and red pepper, it can be seen that the sales volume of a single product fluctuates greatly in different time periods. In a day at about 10:00 a.m. will appear in the single product sales volume of vegetables, so this time period for the ideal time for vegetable sales; in the week in the vicinity of the double holiday is easy to appear in the single product sales volume of vegetables, similarly, this time period for the ideal time for the sale of vegetables.

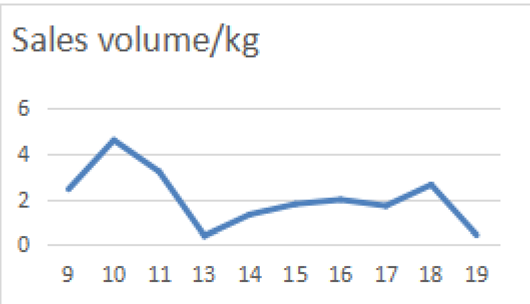

Fig 3 Cabbage Moss Sales by Hour

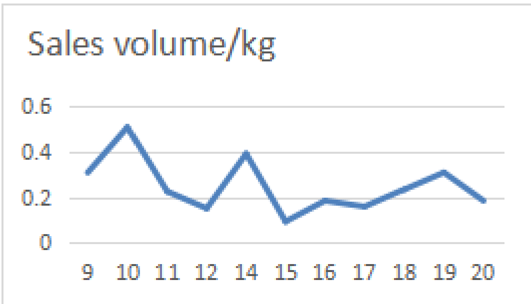

Fig 4 Red Thread Pepper Sales by Hour

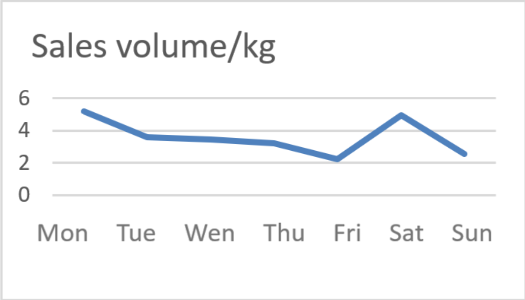

Fig 5 Cabbage Moss Sales by Days

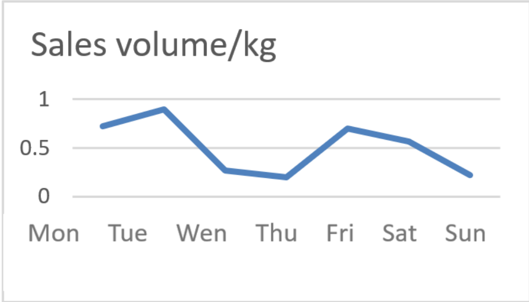

Fig 6 Red Thread Pepper Sales by Days

(2) Comparison of individual products within the same category

For different individual products within the same category, the distribution pattern should be consistent with the trend of the distribution pattern obtained from the time dimension of the category in terms of hours/days.

However, data analysis of the same kind of single product can be seen, the same kind of single product of different sources, usually in a certain period of time to show a "monopoly trend", that is, some sources of single product in a certain period of time in line with the time dimension of the law of the sales volume, but in the other time period of the sales volume of almost 0. In this case, the multi-sourcing can be handled by turning it into a single-sourcing problem.

(3) Seasonal Explanation of the Sales Patterns of Individual Vegetable Products

For the analysis of individual items, this paper abandons the time series in months. This is because the sales of single product has a short-term seasonal or long-term characteristics, some single product sales fixed in April to June each year, while some single product sales in the 12 months of existence and always maintain a more stable sales volume. Therefore, considering a variety of factors, the month-by-month study of vegetable single products is more complicated and has no obvious sales distribution pattern, so the distribution study using month as the time dimension is discarded.

2.3.2 Correlation analysis of single product sales volume

From the above analysis, it can be seen that the sales volume data of individual products of the same category show a significant trend of "this and that", i.e., there is a phenomenon of sales competition between different individual products, and there is a "complementary" relationship. In order to further verify the conclusions, this paper further analyzes the correlation between different single items based on vegetable category and pairing. Fuzzy C-means algorithm (FCM) is used for the single product sales data. Firstly, the graph is made through the elbow rule to estimate the optimal number of clusters, and it is found that the function has a large change in $x=3$ slope, i.e., the absolute value of the curve is larger when $x\leq 3$, and the absolute value of the slope of the curve is smaller when $x>3$. Therefore, 3 is taken as the number of clustering categories. [5]

After selecting the number of clustering categories, the FCM fuzzy c-mean algorithm optimized based on the Special Forces algorithm is written using matlab and optimized iteratively according to the FCM objective equation, and the clustering results are visualized as in Fig 7. [6]

Combined with the images, it can be seen that in terms of time dimension, the correlation between individual items has two kinds of correlations: "complementary" or "synchronized".

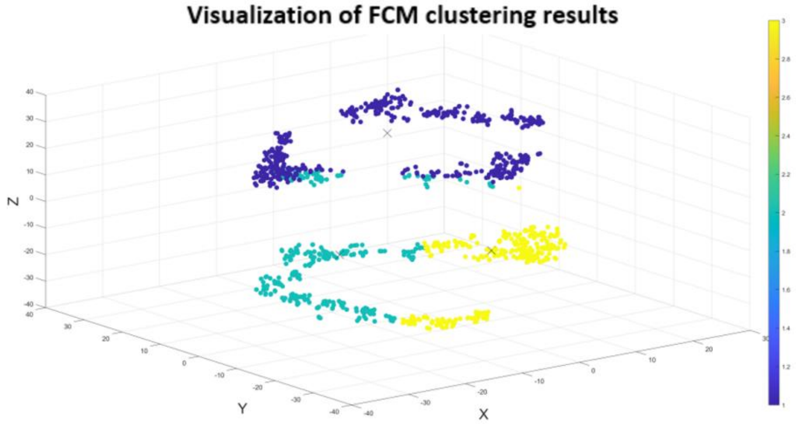

Fig. 7 Visualization of FCM clustering results

3. A study of the development of a replenishment program for superstores at the time of maximum profitability

3.1. Linear regression analysis of total sales and cost-plus pricing

In order to explore the relationship between total sales volume and cost-plus pricing, it is necessary to establish an equation to describe the relationship between total sales volume and cost-plus pricing by category. Combined with the single product sales scatterplot, we can get its corresponding regression equation:

$$y = -0.00028x + 20.015 \tag{3}$$

Combining the regression equation with the scatterplot, it can be seen that cost-plus pricing has an insignificant negative correlation with total sales, i.e., the higher the pricing, the smaller the total sales, which is consistent with the laws of life and common sense.

3.2. Total replenishment prediction based on LSTM time series forecasting model

In order to analyze the prediction of the restocking volume of each category of vegetables in the supermarket in the coming week, this paper adopts the method of LSTM time series prediction. [7]For the prediction of replenishment volume, the positive correlation between sales volume and replenishment volume can be proved firstly, and the historical data of total sales volume can be transformed into the historical data of total replenishment volume, so as to predict the replenishment volume in the next seven days by combining with the time series prediction model. For the formulation of pricing strategies, a linear programming model can be developed and solved to obtain the pricing strategy that maximizes the benefit to the superstore.

(1) MATLAB construction of training set trend graphs

In order to reveal the correlation between the factors more clearly, this paper adopts a dataset partitioning strategy, dividing the dataset into training set and test set respectively in the ratio of 9:1, and conducts multi-fold cross-validation, so as to expand the coverage of the model and improve its accuracy. [8]

In this paper, we iteratively process 251 single products, and then merge the related data of 251 single products processed with the same category, and finally derive the law when taking the category as a unit, which is not shown here.

(2) Training set predicts test set to verify model accuracy

After completing the training of the model, it is necessary to use this training set model to make predictions on the test set.

Now, we still take Amaranthus as an example to predict the test set and compare the prediction results with the real data, and the results are shown in Fig 8. And it can be seen that the time series prediction model of LSTM is more effective.

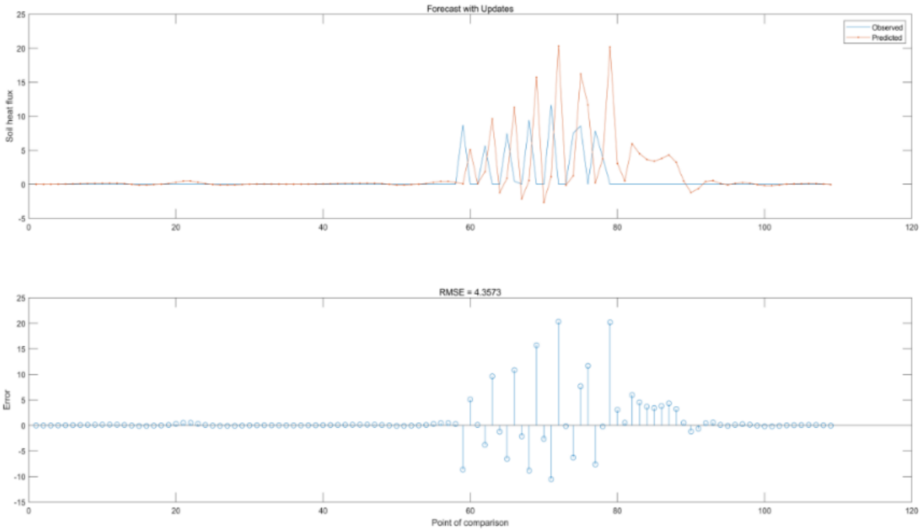

Fig. 8 Comparison of the predicted results with the real data for Marmot

(3) Forecast of results for total replenishment in the coming week

Based on the above, the predicted total daily replenishment of Amaranthus for the next 7 days can be obtained as 1.61, 1.24, 1.15, 0.29, 0.05, 0.77, and 0.54 (kg).

Similarly, the total replenishment of each vegetable category in the coming week can be predicted, and the results are shown in Table 3.

Table 3 Forecast of total replenishment of vegetable commodities by category for the coming week

dates	7.1	7.2	7.3	7.4	7.5	7.6	7.7
Foliage (kg)	5.6	3.6	3.8	3.7	4.3	5.0	4.7
Cauliflower (kg)	10.3	9.6	10.2	7.6	7.4	6.3	6.1
Roots and tubers (kg)	1.4	1.2	1.8	1.9	1.7	1.4	1.7
Eggplant (kg)	4.7	5.8	4.1	4.6	6.1	5.0	6.0
Chili pepper (kg)	2.5	3.1	3.1	3.3	3.6	4.2	4.2
Edible mushrooms (kg)	2.3	2.0	1.9	2.4	2.6	2.5	2.4

3.3. Pricing strategy development based on linear regression equations

(1) Total replenishment cost of the day for the superstore

Combined with the wholesale prices of each category of vegetables obtained in the previous section, the replenishment cost can be known, and then the total replenishment cost of a category and six vegetable categories on that day can be obtained:

$$\sum cost(i,j)=\sum rep(i,j) \times prime(i,j) \tag{4}$$

Where, i=1,2,3,4,5,6, i.e., 6 vegetable categories; j=1,2,3,4,5,6,7, i.e., Monday through Sunday;$cost(i,j)$ denotes the total replenishment cost of vegetable category i at day j;$rep(i,j)$ denotes denotes the replenishment quantity of vegetables of category i at day j;$prime(i,j)$ denotes the wholesale price of vegetable category i on day j.

(2) Total sales of the day at the superstore

As a result of the analysis, it can be seen that each vegetable category may be sold at a discount, which in turn affects revenue. [9]

The probability of discounting as well as the discount rate for each category of vegetables is obtained through statistics, which is approximated as 0.9. Then the total sales of the six vegetable categories on the jth day $W(i,j)$ is

$$\sum W(i,j) =\sum\left\{\gamma(i,j) \times (1-K(\mathrm{i,j})) \times \left[rep(i,j) \times \left(1-loss(i,j)\right)\right] \times price(i,j)\right\} \tag{5}$$

Combining the analysis of the above variables, the sales revenue of each category of vegetables on a given day is obtained, i.e., the optimization objective is

$$\max\ profit\ (i,j) = \sum W(i,j) - \sum cost(i,j) \tag{6}$$

where *profit* (i,j) denotes the profit on sales of category i vegetables on day j, and $loss(i,j)$ denotes the wastage rate of category i vegetables on day j, *the*$\gamma(i,j)$ denotes the discount rate of category i vegetables on day j, and$K(\mathrm{i},\mathrm{j})$ denotes the discount probability of category i vegetables on day j.

Next, since pricing cannot increase indefinitely, neither can the total number of sales, and using the constraints on the total number of sales and pricing one can write the constraints[10]

$$rep(i,j) \geq \mathrm{sales(i,j)} \tag{7}$$

$$-0.00028 \times \mathrm{sales(i,j)} + 20.015 = price(i,j) \tag{8}$$

Solving this linear regression equation yields the average pricing strategy for the six vegetable categories for the coming week, as shown in Table 4.

Table 4 Average pricing of six vegetables for the coming week

kind	7.1	7.2	7.3	7.4	7.5	7.6	7.7
flowers and leaves	5.0	5.6	5.1	5.0	5.3	5.8	5.8
cauliflower	11.5	8.8	12.0	10.3	10.2	11.3	11.3
capsicum	8.5	12.2	12.5	11.1	13.3	13.5	13.2
eggplant	8.5	6.4	8.3	8.2	9.3	9.0	8.7
edible mushroom	14.0	13.0	13.2	13.3	13.3	13.5	14.9
rootstock	14.8	19.6	18.3	19.3	18.3	18.2	18.0
Profit ($)	147.9	153.6	182.6	160.6	179.2	174.5	180.1

4. Conclusions

This paper focuses on automatic pricing and replenishment decisions for vegetable commodities, and the LSTM time series prediction modeland FCM fuzzy mean clustering analysis used in analyzing the problem are also applicable to the calculation of production and sales balance, inventory control and dynamic pricing research for fresh agricultural products similar to vegetable commodities in marketing, which can be widely extended to general marketing plan setting. Based on the large sample data as the training set, machine learning can be used to predict and analyze the relevant output, so it can be widely extended to the general marketing plan setting, which has certain practical significance.

References

[1] PAN Ke, ZHANG Wei, WANG Yagang. Special forces algorithm: a new metaheuristic algorithm[J]. Control and Decision Making, 2022.

[2] Li Yali. Joint replenishment and pricing collaborative decision-making considering investment in preservation technology[D]. Chongqing: Chongqing Jiaotong University, 2022.

[3] Lu Jing. Research on inventory control and dynamic pricing of fresh agricultural products [D]. Tianjin: Tianjin University, 2019.

[4] Penghui Liu. Coordination and optimization model of three-level cold chain integrated inventory [D]. Shandong: Yantai University, 2016.

[5] Zeng Minmin Research on Dynamic Pricing Strategies of a Fresh Community Supermarket Based on Time Context [D]. Southwest University of Finance and Economics, 2021.

[6] Mao Lisha. Research on Pricing Strategy and Production and Marketing Mode of Vegetable Wholesale Market under the Perspective of Supply Chain [D]. Central South Forestry University of Science and Technology, 2023.

[7] Gu Si-Hong. A study on dynamic pricing of fresh products in H-retailers considering freshness variation [D]. Donghua University, 2023.

[8] Qiao Xue. Joint replenishment pricing strategy for fresh products considering sales loss [D]. Southeast University, 2023.

[9] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.

[10] Li, N., & Wang, Q. H. (2019). Time series analysis and its R language implementation. Mechanical Industry Press.