Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 2, No. 2, 2024

Research on the Optimal Pricing Strategy of Supermarket Vegetables Based on Elastic Network Regression Model

Zhiyi Shen, Ke Huang, Ningyuan Lu, Qinglong Zhang*

Business School, University of Shanghai for Science and Technology, Shanghai, China, 200093

* **Corresponding author**: Qinglong Zhang (Email: zhangqinglong@usst.edu.cn)

Abstract: This study develops optimal restocking and pricing strategies for vegetables in supermarkets to reduce significant post-harvest losses in China, which currently result in losses exceeding 100 billion yuan annually. By employing cost-plus pricing, ridge regression, and support vector regression (SVR), this research analyzed the relationship between pricing and sales volumes for various vegetable categories. Predictive models for the week of July 1-7, 2023, indicated a consistent markup rate of about 58%, confirming the reliability of the models. For single-item analysis, both a greedy algorithm and elastic net regression were used. The greedy algorithm corrected loss rates and sales quantities to maximize profit, while the elastic net regression addressed multicollinearity and overfitting issues, leading to more accurate sales predictions. Integrating these approaches effectively predicted sales and optimized supermarket restocking and pricing strategies. The findings showed significant improvements in expected profits, validating the effectiveness of these combined models. This research demonstrates the potential for advanced analytical techniques to enhance supermarket operations, reduce waste, and better meet consumer demand.

Keywords: Cost-plus Pricing, Elastic Network Regression Model, Optimal Pricing Strategy of Supermarket Vegetables.

1. Introduction

1.1. Background introduction

China has a vast territory and a large population. Fresh agricultural products such as vegetables and fruits are consumed in China, but the management level is relatively low. The average loss rate of vegetables and fruits after harvesting in China is as high as 25% to 30%, resulting in losses exceeding 100 billion yuan per year. This is not conducive to the value-added of fresh agricultural products such as vegetables and fruits in circulation, nor is it conducive to increasing farmers' income [1]. Moreover, the freshness of vegetables and fruits will also decrease with the extension of storage time. Many varieties have a freshness period of only one day, so supermarkets replenish them every day based on vegetable retail data and demand.

1.2. Reasonable pricing for vegetables in supermarkets

Due to the different categories and origins of vegetables, and the fact that the purchase time is usually earlier than the retail time, merchants need to complete the same day replenishment and pricing decisions without determining the individual items and costs. The cost plus pricing method is a method of setting product prices based on the unit cost of the product plus a certain proportion of profit, which is applicable to vegetable pricing. At the same time, the retail volume of different categories of vegetables is related to various factors such as time, season, place of origin, and loss rate[2], and is limited by factors such as supermarket retail space. Therefore, predicting and formulating corresponding purchasing and pricing strategies based on complex environments is of crucial significance for reducing vegetable and fruit waste, meeting consumer needs, maximizing resource utilization, and maximizing supermarket profits.

1.3. Problem analysis

To explore the relationship between cost markup pricing and total sales volume by category, and to develop a daily replenishment volume and pricing strategy for the next week, this study employs cost-plus pricing, markup methods, ridge regression, and support vector regression (SVR) for analyzing vegetable item pricing and sales volumes across different categories. Based on the analysis of known data, the elastic net regression method is selected for its accuracy in regression analysis. Both SVR and elastic net regression methods are used to ensure model selection accuracy and regression result effectiveness.

2. Develop daily replenishment volume and pricing strategy

2.1. Establishing the model

Based on known information, the cost plus pricing method is adopted to determine the product price based on the unit cost of the product plus a certain proportion of profit [3]. The calculation formulas are shown in equations (1), (2), and (3):

$$\text{Price} = \text{Unit Cost} * (1 + \text{Markup Rate}) \tag{1}$$

$$\text{Unit Cost} = (\text{Total Fixed Cost} + \text{Total Variable Cost}) \tag{2}$$

$$\text{Markup Rate} = (\text{Total Fixed Cost})/(\text{Sales Volume}) * (1 + (\text{Selling Price} - \text{Purchase Price})/(\text{Purchase Price})) \tag{3}$$

Thus, using the known information on selling price, purchase price, and sales volume, the markup rate for each vegetable product can be calculated for sales pricing. Vegetables are categorized into six groups: flower and leaf, cauliflower, water root and stem, eggplant, chili, and edible mushroom. Sales pricing is the core explanatory variable, and sales volume is the dependent variable. Regression analysis is

conducted on each category after selecting a suitable model through data analysis and testing.

Traditional time series analysis methods include ARIMA, Box Jenkins method, and the EM algorithm [4]. These methods rely heavily on correct parameter model selection for accurate predictions [5]. Ridge regression, proposed by Hoerl and Kennard in 1970, is an improved least squares method that introduces a regularization term for more accurate results. Geng Juan et al. noted that ridge regression combines qualitative and quantitative analysis effectively [6]. Therefore, ridge regression is used to fit the model, with the final objective function incorporating a regularization term for enhanced accuracy.

The ridge regression method, proposed by Hoerl and Kennard in 1970, improves upon least squares by providing a more realistic fit. Sun Hailing et al. argue that ridge regression incorporates a degree of human subjectivity, blending qualitative and quantitative analysis [6]. Additionally, ridge regression introduces a regularization term to enhance the accuracy of the model's results. Thus, for the second problem, we plan to use ridge regression with a regularization term to fit the model, as described in the final objective function below:

$$J(\beta) = [\![\sum(y - X\beta)]\!]^2 + [\![\sum\lambda\beta]\!]^2 \quad (4)$$

Among them, β is the regression coefficient, and λ is the coefficient of the squared regularization term, known as the penalty coefficient. Then take the derivative of equation (4) to obtain the final ridge regression model as shown in equation (5):

$$\beta = [\![(X]\!]^T X + [\![\lambda I)]\!]^{(-1)} [\![X]\!]^T y \quad (5)$$

To verify the rationality of model selection and the accuracy of regression results, this article refers to the Support Vector Regression (SVR) method with SARIMA method used by Qian Mingjun et al. [7]. SVR addresses issues like small sample size, nonlinearity, overfitting, dimensionality disaster, and local minima by minimizing risk. From July 1, 2020, to June 30, 2023, the prices and sales volumes of vegetable items were recorded, forming a sample set (x_i^0, y_i^0), i = 1, · · ·, n, given a sample set T = {(x1, y1) , · · ·, (xn, yn)}, xi ∈ Rm , yi ∈ R , i = 1, · · ·, n. It can be seen from the above that the sample set meets the condition of linear regression problem, that is, a linear function $y = f(x) = (w \cdot x) + b$ on Rm can be used to infer the y value corresponding to any mode x. Therefore, the solution to the linear regression problem will be transformed into an optimization problem solving the following equation (6):

$$\begin{cases} \min \frac{1}{2}|(|w|)|^2 + C\frac{1}{n}\sum_1^n (\xi_i + \xi_i^*), s.t. \\ (w \cdot x_i) + b) - y_i \le \varepsilon + \xi_i, i = 1, \dots, n \\ y_i - ((w \cdot x_i) + b) \le \varepsilon + \xi_i^*, i = 1, \dots, n \\ \xi_i, \xi_i^* \ge 0, i = 1, \dots, n \end{cases} \quad (6)$$

In equation (6), C is the penalty parameter; ξ_i , ξ_i^* is the relaxation variable; ε is the threshold of the insensitive loss function.

2.2. Solving the model

Firstly, visualize the data and draw a time series chart of the cost markup ratio for each category of vegetables (Figure 1) and a box chart of the cost markup ratio for each category of vegetables (Figure 2).

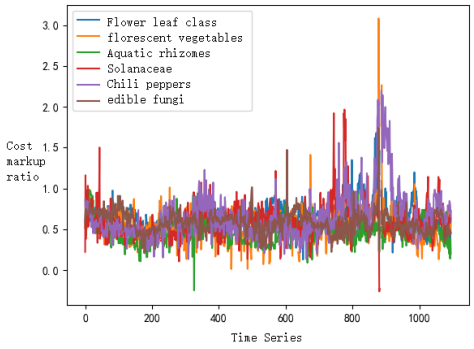

Figure 1. Cost Plus Proportional Time Series Diagram

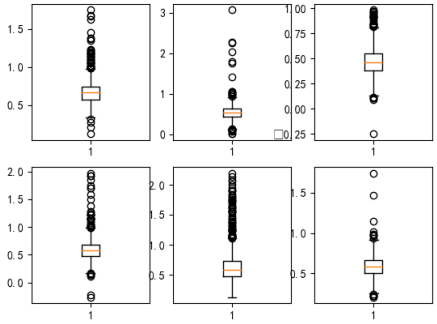

Figure 2. Cost Plus Proportional Box Plot

By traversing the penalty coefficient λ, the optimal penalty coefficient was selected for ridge regression and the model was tested. The regression results passed t-tests and F-tests, and the coefficients were significant at the 1% level, indicating a good fit of the model.

Determine the markup rate of a product based on the markup rule, that is, the higher the sales volume of a single vegetable, the less the markup; The lower the sales volume of a single vegetable product, the more the price increase, which is in line with the principle of "small profit but quick turnover". At the same time, based on the determined product markup rate and pricing strategy, the daily replenishment total and pricing of the vegetable category for the next week (July 1-7, 2023) are predicted. The predicted results are shown in Table 1:

Table.1. Ridge Regression Markup Rate Prediction Results

Date	Flower Leaf Class	Florescent Vegetables	Aquatic Rhizomes	Solanaceae	Chili Peppers	Edible Fungi
2023.07.01	64.75%	49.47%	53.63%	59.37%	58.28%	63.78%
2023.07.02	68.3%	53.22%	51.21%	51.07%	58.2%	61.82%
2023.07.03	62.36%	55.26%	51.3%	53.83%	61.96%	54.07%
2023.07.04	58.35%	48.75%	50.03%	63.26%	62.58%	57.11%
2023.07.05	62.35%	50.99%	49.74%	65.39%	57.82%	59.5%
2023.07.06	61.62%	50.99%	50.58%	58.05%	57.87%	58.65%
2023.07.07	66.33%	48.9%	56.37%	56.22%	58.35%	63.82%
Average Markup Rate	63.4371%	51.0829%	51.8371%	58.17%	59.2943%	59.8214%

Before using the Support Vector Regression (SVR) method, the data is first standardized using StandardScaler to ensure that the mean of each feature is 0 and the variance is 1. Then, the support vector regression model, i.e. equation (4), is used to solve and obtain the optimal parameters. Using the

obtained parameters, the daily replenishment total and pricing of the vegetable category for the next week (July 1-7, 2023) are predicted, and an image is generated as shown in Figure 3 to achieve visual analysis:

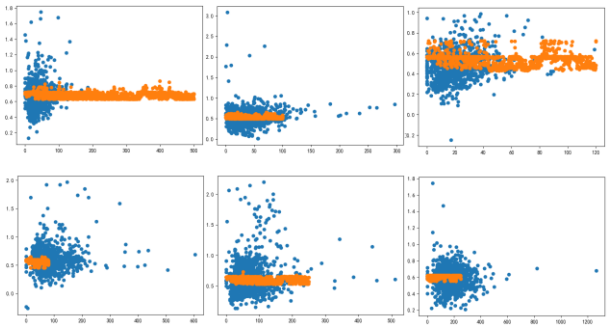

Figure 3. Scatter Plot of SVR Results

Comparing the results with Table 1, it was found that there was no significant difference, which verified the accuracy of the model construction and the reliability of the prediction results. Therefore, the model will continue to be used to predict the replenishment volume of supermarkets in the next week, and the predicted results are shown in Table 2:

Table.2. SVR Restocking Volume Prediction Results

Date	Flower Leaf Class (Kg)	Florescent Vegetables (Kg)	Aquatic Rhizomes (Kg)	Solanaceae (Kg)	Chili Peppers (Kg)	Edible Fungi (Kg)
2023.07.01	85.2985	16.2421	19.172	27.7699	73.0825	31.9687
2023.07.02	104.5199	23.2069	37.0349	27.8695	43.3721	46.8788
2023.07.03	104.9815	13.9241	19.2561	27.8695	61.4472	18.358
2023.07.04	116.0023	16.5113	14.9128	9.9683	69.8128	34.063
2023.07.05	83.2179	25.9055	32.0049	27.3447	51.7891	30.5202
2023.07.06	136.6486	21.1978	36.2253	7.474	78.6967	27.3774
2023.07.07	172.2375	23.6662	24.3932	22.6415	88.0119	26.506

3. Develop a single item replenishment volume and pricing strategy to maximize revenue

3.1. Establishing the model

On the basis of the previous model, the basic greedy algorithm is first used to correct the loss rate and the sales quantity of vegetable items. Based on the corrected data, the corresponding profit of vegetable items is calculated for screening and strategy prediction, achieving the maximization of supermarket benefits. The specific steps of the greedy algorithm are as follows:

Step 1: Select an initial solution and start from it;

Step 2: Continuously iterate, and when one step can be taken towards the goal, obtain a partial decomposition based on the local optimal strategy to reduce the problem size;

Step 3: Combine all the solutions obtained above.

Due to the limitations of greedy algorithms, which focus solely on profit without considering expected sales volume, a more comprehensive approach is necessary. Referring to the process used for problem two, a cost-plus pricing model (equations 1, 2, and 3) calculates sales pricing for different vegetable items. Sales pricing for six categories—flower and leaf, cauliflower, aquatic rhizome, eggplant, chili, and edible mushroom—serves as the core explanatory variable, with sales volume as the dependent variable. Additional variables include wholesale price, sales unit price, unit loss rate, and discount status (1 for discounted, 0 for not).

To ensure prediction accuracy, an optimized regression model is constructed. Traditional dimensionality reduction methods like clustering, partial least squares, and principal component regression have strong limitations. Xu Yunjuan et al. noted that clustering-based models are sensitive to algorithms, partial least squares and principal component regression can be biased, and ridge regression, while compressing coefficients, introduces only one regularization term [8].

Compared to traditional methods, elastic network regression combines ridge and Lasso regression, simultaneously penalizing absolute and squared values [9]. This approach addresses issues such as incorrect solutions of non-full rank matrices, multicollinearity, and overfitting in linear regression more effectively. Elastic network regression introduces two regularization terms, enhancing model accuracy.Given the large dataset and numerous explanatory variables, with a VIF value generally greater than 10 indicating severe multicollinearity, this study adopts elastic network regression for analyzing the pricing and sales volume of vegetable items across different categories in supermarkets. Referring to the method by Liu Bali et al., the model is structured as follows [10].

Assume the number of predictive variables in a linear regression model is p and the sample size is N, then:

$$\{\blacksquare(Y\_i = \beta\_0 + \beta\_1 x\_i1 + \ldots + \beta\_p x\_ip + \varepsilon\_i@\varepsilon\_i \sim N(0,\sigma\hat{}2), i = 1,2,\ldots,N\&)\dashv \quad (7)$$

In equation (7), β_0 is a constant term, β_p is the regression coefficient of supermarket vegetable single item pricing with respect to supermarket vegetable single item sales volume, and ε is the error term. It is usually assumed to follow a normal distribution with a mean of 0 and a variance of $\sigma\hat{}2$. The linear regression model is represented by a matrix as shown in equation (8):

$$\{\blacksquare(Y = X\beta + \varepsilon@\varepsilon \sim N\_N(0,\sigma\hat{}2\, I\_N)\ \&)\dashv \quad (8)$$

Therefore, the minimum two of the regression coefficients is:

$$\beta\hat{}\hat{}LS = (〚X\hat{}T\ X)〛\hat{}((-1))\ 〚X〛\hat{}T\ Y \quad (9)$$

For the pricing of single vegetable items in supermarkets, the linear regression model for the sales volume of single vegetable items in supermarkets, and the definition of the elastic network regression model is:

$$\beta\hat{}\hat{}Elastic = \arg\min\{\sum\_(i=1)\hat{}N▒〚(y\_i - \beta\_0 - \sum\_(j=1)\hat{}p▒〚x\_ij\,\beta\_j)\hat{}2 + 〛\ \lambda\_1\ 〛\ \sum\_(j=1)\hat{}p▒〚|\beta\_j| + \lambda\_2\sum\_(j=1)\hat{}p▒〚\beta\_j〛\hat{}2\ 〛\} \quad (10)$$

In equation (10), λ_1 and λ_2 are penalty parameters. When λ_1 =0 and λ_2 =0, the regression model is least squares regression. When λ_1=0 and λ_2>0, the regression model is Ridge regression; When λ_1>0 and λ_2=0, the regression model Lasso regression; When λ_1>0 and λ_2>0, the regression model is elastic network regression.

3.2. Solving the model

According to the specific steps of the greedy algorithm in model construction, combined with the visual analysis of the unit profit distribution and loss rate of sellable vegetable varieties during the period of June 23-24, 2023, as shown in the following figures (Figure 4 and Figure 5).

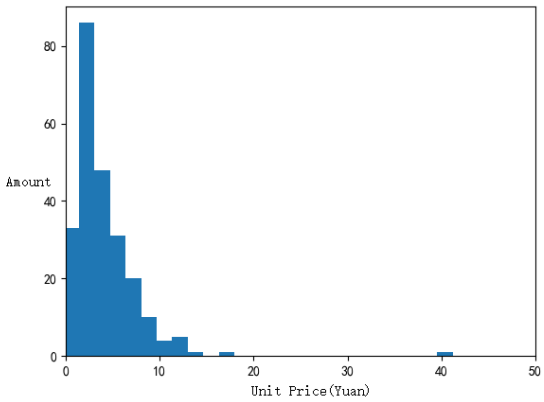

Figure 4. Unit Profit Distribution Chart

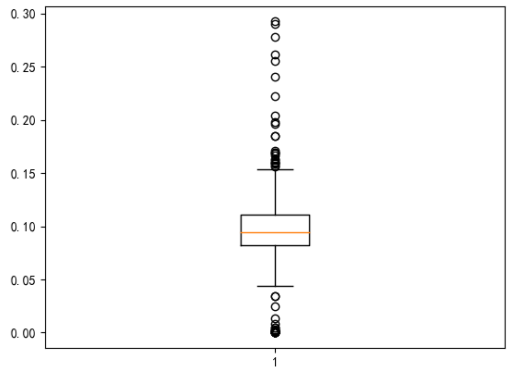

Figure 5. Box Plot of Loss Rate

Analyzing the data, it was found that the number of vegetable items with a minimum shelving quantity of 2.5kg was relatively small. Therefore, each item was sorted in descending order according to the total profit size. Based on known information, the minimum shelving quantity of 2.5kg was set as the threshold. If the number of items greater than the threshold is greater than the specified upper limit of 33 available items, the top 33 items are selected from high to low; If the number of single items greater than the threshold is within the specified range of the total number of sellable items, that is, the number of single items is between 27-33, then select the corresponding quantity of vegetable items that meet the minimum shelving quantity from high to low; If the number of items greater than the threshold is less than the specified lower limit of the total number of sellable items by 27, select from the items below the threshold from high to low until the number of sellable items reaches the specified lower limit of the total number of sellable items.

The final prediction result obtained from the greedy algorithm is shown in the table below(taking edible fungi as an example):

Table.3. Greedy Algorithm for Predicting Single Product Replenishment Volume and Profit on July 1st (Using Edible Fungi as an Example)

Type	Unit	Revise estimated sales volume（Kg）	Total profit (Yuan)
Edible fungi	Xixia flower mushroom (1)	2.57	19.16
	Apricot abalone mushroom (1)	3.22	9.03
	Golden needle mushroom (1)	7.02	14.08
	White jade mushroom (Bag)	3.48	7.22
	Double spore mushroom (box)	7.64	7.18

By traversing the penalty coefficients λ 1 and λ 2, the optimal penalty coefficient was selected for elastic network regression and the model was tested. The coefficients were significant at the 1% level, and the regression results passed t-tests and F-tests, indicating a good fit of the model. Based on the known available varieties for sale from June 24th to 30th, 2023 and the prediction of restocking volume using an elastic network regression model, the vegetable items with expected sales not meeting the standard were removed. The predicted results are shown in the following table (Table 4):

Table.4. Optimized Elastic Network Regression Model for Predicting Single Product Replenishment Volume and Profit on July 1st (Using Edible Fungi as an Example)

Type	Unit	Fixed price（Yuan/Kg）	Expected sales volume (Kg)	Total profit (Yuan)
Edible fungi	White jade mushroom (Bag)	6.06	0.93	2.14
	Crab flavored mushroom and white jade mushroom combo (Box)	11.42	1.81	7.81
	Cordyceps flowers (Portion)	3.68	1.91	2.72
	Seafood mushroom (Bun)	4.29	6.15	9.94
	Xixia flower mushroom (1)	22.54	6.83	58.36
	Double spore mushroom (Box)	5.78	8.49	18.51
	Golden needle mushroom (Box)	4.14	13.47	21.07

By comparing the results predicted by the two algorithms in Tables 3 and 4, it was found that the optimized elastic network regression model had a significant difference in regression results compared to the sales and total profit predictions obtained by the greedy algorithm. It is speculated that this may be due to the greedy algorithm not considering the predicted sales and longer time periods, only aiming to achieve the highest profit, and ignoring other factors that affect the sales of single vegetable products in supermarkets.

4. Conclusion

Based on the analysis using cost-plus pricing, ridge regression, and support vector regression (SVR), this article develops effective restocking and pricing strategies for fresh vegetables in a supermarket. By analyzing various vegetable categories and selecting appropriate regression models, the

study employs cost-plus pricing as an explanatory variable and individual vegetable sales as the response variable. The ridge regression model fits the time-series data well, and significance tests validate the results. Predictive models for the upcoming week (July 1-7, 2023) indicate a consistent markup rate of approximately 58%, confirmed by both ridge regression and SVR, demonstrating model reliability.

For single-item analysis, with constraints on total saleable items (27-33) and minimum order quantity (2.5 kg), the study devises a strategy for July 1 based on the previous week's data (June 24-30, 2023). Using a greedy algorithm, the article aims to maximize profit, resulting in a combination of 33 vegetable items with an expected profit of 695.22 yuan. Enhancing the ridge regression model with additional variables and applying elastic net regression, the study achieves more accurate sales predictions and a higher expected profit of 917.94 yuan.

In conclusion, the integration of cost-plus pricing, ridge regression, SVR, and elastic net regression models effectively predicts sales and optimizes restocking and pricing strategies, significantly improving expected profits and validating the approach's effectiveness. Future research could explore incorporating machine learning techniques and real-time data analysis to further enhance the accuracy and responsiveness of restocking and pricing strategies, ensuring sustained profitability and customer satisfaction in a dynamic market environment.

References

[1] Dan Bin, Jiang Xiaoling, Wang Fengquan. Evolution of Fresh E-commerce Circulation Model and Service Value Creation: A Dual Case Study of Hema Fresh and JD Fresh [J]. Business Economics and Management, 2024 (01): 20-36

[2] Zhang Qianqian Research on Vegetable Demand Prediction System Based on SVM [D]. Beijing Jiaotong University, 2015.

[3] An Qi. Research on Pricing of Science and Technology Novelty Search Services Based on Cost Plus Pricing Method [J]. Library Research and Work, 2021 (10): 25-31+24

[4] Du Rang Statistical Inference of Time Series Quantile Regression Model Based on EM Algorithm [D]. Changchun University of Technology, 2024. DOI: 10.27805

[5] Mao Yuanhong, Sun Chenchen, Xu Luyu, et al. Review of Time Series Prediction Methods Based on Deep Learning [J]. Microelectronics and Computer Science, 2023,40 (04): 8-17

[6] Geng Juan, Nie Wenqian. Analysis of Factors Influencing Grain Yield in Henan Province Based on Ridge Regression and LASSO Regression [J]. Shanxi Agricultural Economics, 2023 (23): 7-10. DOI: 10.16675/j.cnki.cn14-1065/f.2023.23.002

[7] Qian Mingjun, Li Minggui, Huang Xin. Railway Freight Volume Prediction Method Based on SARIMA-SVR Model [J/OL]. Railway Transportation and Economy: 1-12 [2024-06-21] http://kns.cnki.net/kcms/detail/11.1949.U.20240612.0908.004.html.

[8] Xu Yunjuan, Luo Youxi. Principal Component Lasso Dimension Reduction Algorithm and Simulation Based on Variable Clustering [J]. Statistics and Decision Making, 2021, 37 (04): 31-36. DOI: 10.13546/j.cnki. tjyjc. 2021.04.007

[9] ZOU H, HASTIE T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005,67(2):301. DOI: 10.1111/J.1467-9868.2005.00527.x

[10] Liu Bali, Hu Jinjun, Xie Lili. Seismic motion parameter sorting and comparison based on elastic network regression [J]. Journal of Harbin Institute of Technology, 2024,56 (01): 54-62