Integrating ARIMA and GA for Profit Maximization in Supermarket Pricing Strategy

Chenghao Zhang [1, #], Haoyang Zhang [1, #], Yimiao Guo [2, #]

[1] Hua Lookeng Honors College, Changzhou University, Changzhou, China, 213000

[2] Wu Jinglian College of Economics, Changzhou University, Changzhou, China, 213000

* Corresponding Author Email: eternitypolaris@163.com

[#]These authors contributed equally

Abstract. Supermarkets need to consider a variety of factors when making vegetable replenishment and pricing decisions to meet market demand while maximising revenue. In this paper, firstly, the descriptive statistics of weekday and non-workday sales data are carried out separately and regressed to illustrate the correlation between the pricing and sales volume of each category, secondly, the ARIMA time series prediction of cost and sales volume are carried out separately, and with the goal of maximising the economic benefits in the next 7 days, the final mark-up rate and the maximum revenue are obtained by substituting into the GA model under the optimal daily replenishment volume and pricing strategy.

Keywords: Replenishment programme; pricing strategy; correlation Analysis; ARIMA; Genetic Algorithm.

1. Introduction

Advanced forecasting models and algorithms are utilised to predict future sales volumes and thus make more accurate decisions when developing replenishment plans and pricing strategies.Wang et al. conducted a comparison of single pricing, two-stage pricing and dynamic pricing [1].Ghare and Schrader first developed a dynamic model of perishable inventory, which takes into account the fact that, in the event of a stockout, the decline in the perishable inventory level rate is equal to the sum of market demand rate and inventory decay rate [2]. Cao Hong and Zhou Jiang's research shows that static pricing strategies are no longer effective in dealing with unstable seasonal product markets, and dynamic pricing strategies can better solve such problems [3]. In this paper, we focus on how to build a suitable optimisation model to make the total daily replenishment and pricing strategy for the coming week in order to get the highest economic efficiency when supermarkets make replenishment plans on a category-by-category basis.

2. Replenishment and Pricing Modelling

2.1. Flowchart for solving the optimal pricing model

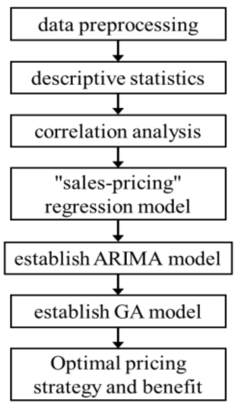

Figure 1 Overall flow of the thesis.

This study analyses supermarket sales data through descriptive statistics and regression models to investigate the association between pricing and sales volume. The ARIMA model is used to predict the cost and sales volume, and then the GA model is used to synthesise the "pricing-volume" relationship to optimise the replenishment and pricing decisions to maximise revenue, as shown in Figure 1.

2.2. Data preprocessing

The data were obtained from the actual operational records of a superstore for the period from 1 July 2020 to 30 June 2023, including sales flow details, wholesale prices, and merchandise wastage rate data, reflecting the sales and wastage of different vegetable categories.

Considering the difference between sales on weekdays and non-workdays, this paper extracts the sales volume data of vegetables on weekdays and rest days from the sales data separately, and makes a graph of the difference between sales on weekdays and non-workdays for each category as shown in Figure 2:

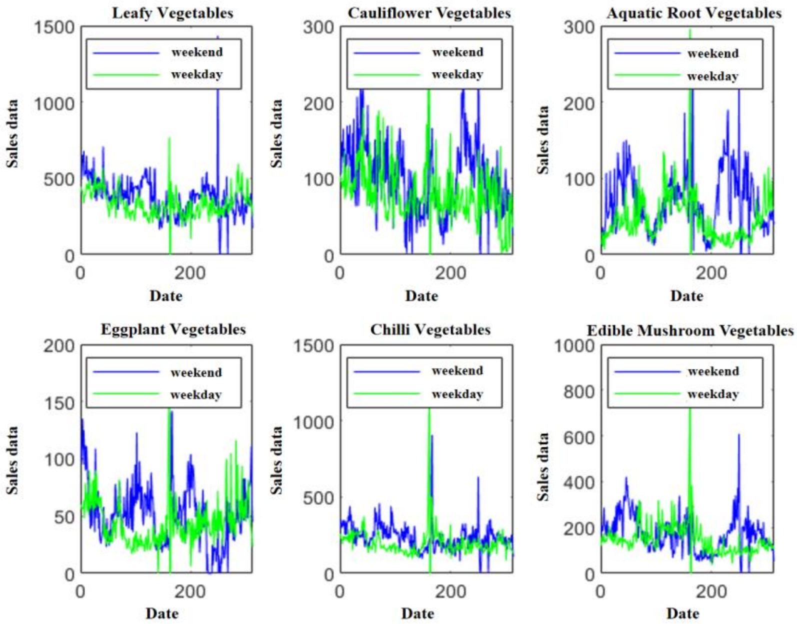

Figure 2 Sales by category on weekdays and non-working days.

By analysing the graphs, it can be observed that there is a significant difference in the volume of vegetable sales between weekdays and rest days. Specifically, the volume of vegetable sales on weekdays is significantly lower than on rest days. Therefore, in the discussion of this paper, both cases of weekdays and rest days will be explored separately in order to understand the pattern of variation in sales volume more comprehensively.

2.3. Establish a "pricing-sales" regression equation and conduct correlation analyses.

Firstly, in order to investigate the relationship between sales volume and pricing of vegetables, this study plotted a heat map of "pricing-volume" for each category of vegetables over a period of three years [10-11], as shown in Figure 3.

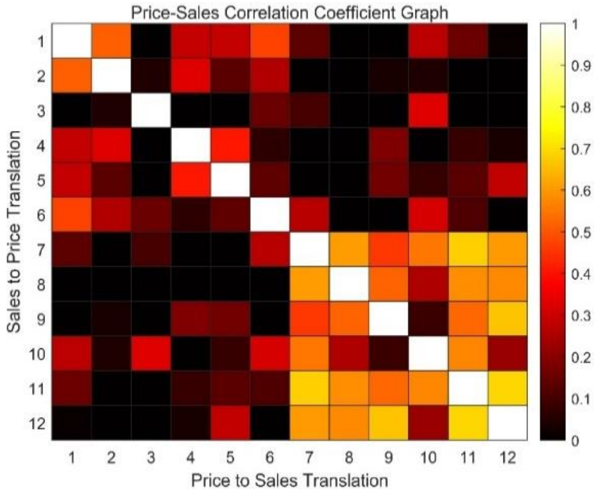

Figure 3 "Price - Sales" Heat Map

In the heat map, different combinations of daily sales volume and pricing are represented by using different colours. Such a visualisation allows for an intuitive look at the potential correlation between sales volume and pricing, and we read the following representative information from the graph:

(a) There appears to be a strong positive correlation between the Foliage category and its sales figures, with a correlation coefficient of roughly 0.7 or more.

(b) The correlation between the cauliflower category and its sales data is weak, showing a moderate to low correlation.

(c) There is a degree of positive correlation between the cost of aquatic rootstocks and eggplants and their sales data.

(d) The correlation between the chilli category and its sales data is relatively weak.

(e) The edible mushroom category showed a more significant positive correlation with its sales figures.

In order to further investigate the correlation between pricing and sales volume of each category, this study makes the "pricing-sales" regression equations for weekdays and days off respectively [8-9]. Due to space constraints, only the regressions for weekdays are shown in Table 1-2:

Table 1 Regression coefficients of sales by category on weekdays

	X_1	X_2	X_3	X_4	X_5	X_6
a constant (math.)	96.893	33.904	-9.597	-10.579	-71.124	51.579
C_1	5.355	-0.655	-4.663	1.714	4.320	-1.625
C_2	-0.086	-2.069	2.193	0.018	-3.696	1.256
C_3	-0.321	0.650	-1.092	0.937	1.486	-1.671
C_4	-11.901	0.294	2.241	-1.593	9.249	-3.652
C_5	-2.268	-0.800	-0.078	0.747	-2.658	4.299
C_6	9.015	-0.545	0.740	0.560	-1.294	-3.863
X_1	/	0.132	0.049	0.055	0.132	0.154
X_2	0.492	/	0.142	-0.047	0.251	0.080
X_3	0.327	0.255	/	-0.049	0.180	0.465
X_4	0.832	-0.189	-0.110	/	1.488	-0.550
X_5	0.218	0.111	0.045	0.163	/	0.421
X_6	0.389	0.054	0.176	-0.092	0.644	/

Table 2 Goodness of fit of the regression equation for each category on working days

mathematical equation	1	2	3	4	5	6
R^2	0.7147	0.5074	0.5584	0.5762	0.7895	0.7640

Where X_{1-6} denotes the sales of Flower leaf, Florescent vegetables, Aquatic rhizomes, Solanaceae, Chili peppers, Edible fungi and C_{1-6} denotes their corresponding costs.

Based on the data analysed in the table, the following results can be observed: the weighted unit price of eggplant vegetables on weekdays shows a significant negative correlation with the sales volume of cauliflower vegetables. Meanwhile, there is a strong positive correlation between the weighted unit price of eggplant vegetables and the sales volume of chilli vegetables as well as between the weighted unit price of edible mushroom vegetables and the sales volume of foliar vegetables. In addition, strong positive correlations were also observed between weighted unit price and sales volume of foliar vegetables, weighted unit price of foliar vegetables and sales volume of chilli vegetables, and weighted unit price of chilli vegetables and sales volume of edible mushroom vegetables. In addition, there was a strong negative correlation between the weighted unit price of foliar vegetables and the sales volume of aquatic root vegetables.

2.4. Forecasting wholesale prices and sales volume for each category for the next 7 days based on ARIMA

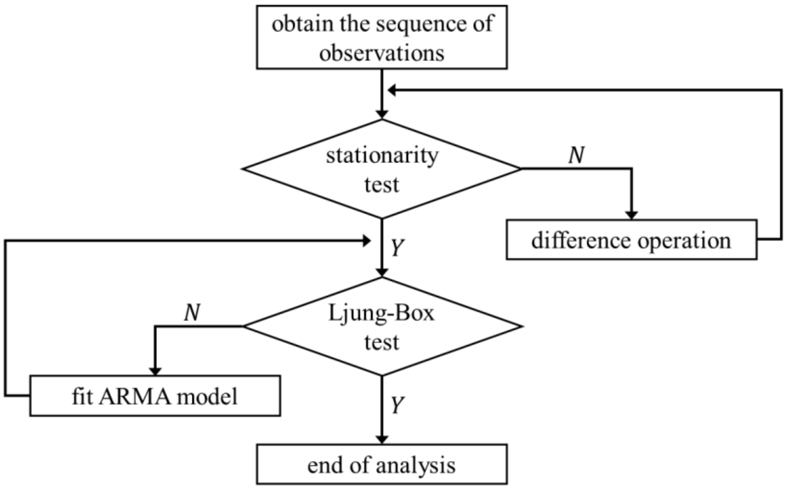

Figure 4 ARIMA model building steps.

In performing ARIMA model analysis [4], the sequence of observations is first obtained and then the sequence is tested for smoothness. If the series is not smooth, a differencing operation is executed to achieve smoothness. Subsequently, the Ljung-Box test is used to determine whether autocorrelation exists in the series. Once the data meets the requirements, the ARIMA model is fitted and the analysis is concluded as shown in Figure 4.

(1) Projected wholesale prices

Due to space issues, only the wholesale price forecasts for the cauliflower category are shown here as shown in Table 5.

Table 5 ARIMA (1,1,2) model test

ARIMA model (1,1,2) test table		
term (in a mathematical formula)	notation	(be) worth
	Df Residuals	1080
sample size	N	1085
Q statistic	Q_6 (P value)	0 (0.990)
	Q_{12} (P value)	9.491 (0.148)
	Q_{18} (P value)	26.359 (0.010***)
	Q_{24} (P value)	37.886 (0.004***)
	Q_{30} (P value)	50.024 (0.001***)
Information guidelines	AIC	1320.912
	BIC	1345.854
goodness of fit	R^2	0.93
Note: ***, **, * respectively represent 1 percent, 5 percent and 10 percent significance levels.		

Time series analyses show that the ARIMA (1,1,2) model exhibits a good fit on several measures. While the Q unity measure reveals the presence of autocorrelation in longer lags (18th, 24th, and 30th order), which may point to additional seasonal or cyclical components that are not captured by the model, the non-significant P values at short-term lag orders indicate the independence of the model residuals. In addition, the relatively low AIC value (1320.912) and BIC value (1345.854) imply that the model still has good forecasting accuracy after accounting for the number of parameters. More importantly, the R^2 value of 0.93 indicates that the model is able to explain the vast majority of the variability, pointing to a strong model fit. Together, these metrics support the ARIMA (1,1,2) model as a valid representation of time series data. The fit effect is shown in Figure 5.

Figure 5 ARIMA (1,1,2) fit

For each of the other categories, we also make predictions and get the final results as shown in Table 6:

Table 6 Predicted replenishment volume for each category (unit: kg)

Date	Flower leaf	florescent vegetables	Aquatic rhizomes	Solanaceae	Chili peppers	Edible fungi
7.1	141.1111	25.1717	23.2357	24.2619	91.1841	57.7820
7.2	130.6818	22.4852	23.2806	22.3435	92.7116	45.8051
7.3	133.7565	21.2209	23.7051	18.9382	93.2528	35.1966
7.4	136.6295	20.6167	24.1186	18.5798	93.3621	47.5023
7.5	140.8881	20.3190	24.5215	17.8745	93.3692	56.2354
7.6	142.3948	20.1635	24.9139	19.7397	93.3714	48.5030
7.7	141.2547	20.0740	25.2962	20.8012	93.3812	49.5133

(2) Sales volume of each category for the next 7 days

When determining the replenishment volume, we need to take into account the attrition rate [12], which affects the actual quantity that can be sold. The replenishment quantity should be calculated based on the expected sales volume and the wastage rate.

We have determined that the calculation of the replenishment volume calculation is as follows:

$$RQ_i = \frac{TS_i}{1-\gamma_i} \tag{1}$$

Where RQ_i and γ_i denote the replenishment volume and wastage rate of the i category respectively, and TS_i denotes the projected sales volume of the i category.

The data collected in this study centrally incorporated information on the loss rate of each category of vegetables. In order to quantitatively analyse the average loss of each category, we executed statistical processing on the wastage rate data and calculated the average wastage rate of each category as shown in Table 7.

Table 7 Wastage rates by category

Classification name	philodendron	cauliflower (Brassica oleracea var. botrytis)	Aquatic rhizomes	eggplant	chilli	edible mushroom
Loss rate (%)	11.158	9.168	9.441	6.751	8.004	8.297

For reasons of space, only the forecast wholesale price of eggplant is shown here as Table 8.

Table 8 ARIMA (5,1,1) model test

ARIMA model (5,1,1) test table		
term (in a mathematical formula)	notation	(be) worth
	Df Residuals	1077
sample size	N	1085
Q statistic	Q_6 (P value)	0.028 (0.868)
	Q_{12} (P value)	2.513 (0.867)
	Q_{18} (P value)	40.233 (0.000***)
	Q_{24} (P value)	58.434 (0.000***)
	Q_{30} (P value)	72.062 (0.000***)
Information guidelines	AIC	7868.739
	BIC	7908.646
goodness of fit	R^2	0.548
Note: ***, **, * respectively represent 1 per cent, 5 per cent and 10 per cent significance levels .		

Firstly, the goodness of fit of the model (R^2) is 0.548, indicating that the model explains about 54.8% of the variability in the cigar sales data. Although the goodness of fit is not very high, in time series analysis, the actual data are usually complex and noisy, so a lower goodness of fit is acceptable. In addition, the information criterion metrics of the model AIC is 7868.739 and BIC is 7908.646. The lower values of AIC and BIC indicate that the model relatively well balances the trade-off between goodness-of-fit and model complexity, which supports the rationality of choosing the ARIMA(5,1,1) model, and the fitting effect graph is shown in Figure 6:

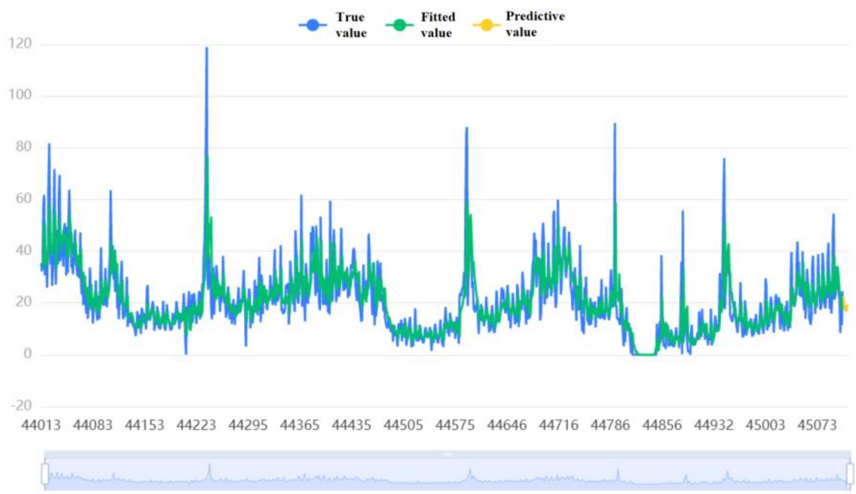

Figure 6 ARIMA (5,1,1) fit

The same forecast was done for the remaining vegetable categories for the next 7 days and the forecasts are shown in Table 9.

Table 9 Predicted wholesale prices for each category (yuan/kg)

Date	Flower leaf	florescent vegetables	Aquatic rhizomes	Solanaceae	Chili peppers	Edible fungi
7.1	3.6869	7.8021	11.6559	4.4434	5.9474	4.6861
7.2	3.6758	7.8770	11.5614	4.4440	5.9464	4.6922
7.3	3.6784	7.8199	11.4402	4.4447	5.9453	4.6908
7.4	3.6754	7.8643	11.3234	4.4453	5.9442	4.6883
7.5	3.6747	7.8306	11.2106	4.4460	5.9432	4.6856
7.6	3.6731	7.8569	11.1018	4.4466	5.9421	4.6829
7.7	3.6718	7.8371	10.9969	4.4473	5.9411	4.6803

2.5. GA-based analysis of optimal additive rates and returns over the next 7 days

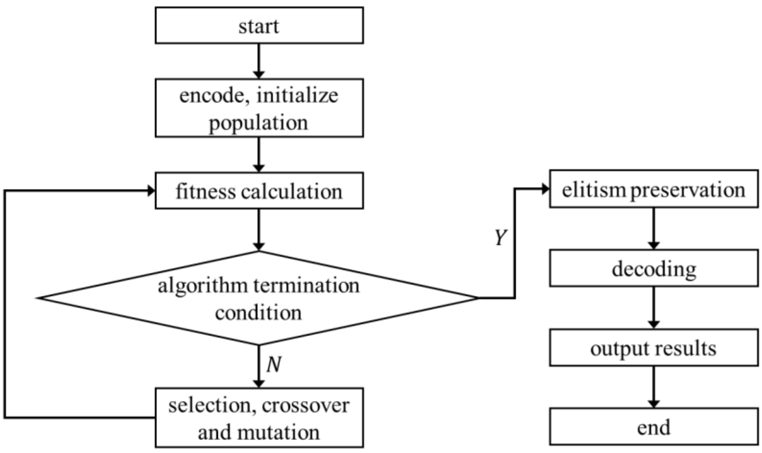

Figure 7 GA modelling steps

Genetic Algorithm [5-7] is an optimisation technique inspired by natural selection and genetics. It represents solutions as chromosomes, creates new solutions through genetic operations, and evaluates their performance using a fitness function. It represents solutions as chromosomes, creates new solutions through genetic operations, and evaluates their performance using a fitness function. The best chromosomes are selected for reproduction, passing genetic information to the next generation. The best chromosomes are selected for reproduction, passing genetic information to the next generation. GA offers advantages such as searching multiple solutions simultaneously, handling continuous and discrete variables, and being applicable to complex optimisation problems, as shown in Figure 7.

Assuming that W is the total benefit of the superstore for the next 7 days, $C_{i,j}$ denotes the wholesale price of the category in j on day i, and $B_{i,j}$ denotes the cost-plus rate of the category in j on day i, the following planning equation is obtained:

$$\max W = \sum_{i=1}^{7}\sum_{j=1}^{6} TS_{i,j}\cdot\left(\mathrm{C}_{i,j}\cdot\left(1+B_{i,j}\right)\right)-\mathrm{C}_{i,j}\cdot RQ_{i,j} \tag{2}$$

Which needs to satisfy the relationship between sales volume and replenishment mentioned in the above article, and the additive rate is greater than 0. Otherwise, it will result in the merchant losing money:

$$\begin{cases} RQ_j = \dfrac{TS_j}{1-\gamma_j},\ j=1,2,\ldots,6 \\ B_{i,j} > 0 \end{cases} \tag{3}$$

3. Results

3.1. Non-working day "pricing-volume" regression equation

Based on the analysis of the data in the table, it can be observed that there is an extremely strong negative correlation between the weighted unit price of eggplant vegetables and the sales volume of cauliflower and leafy vegetables on rest days. Further, there is an extremely strong positive correlation between the weighted unit price of eggplant vegetables and the sales volume of chilli vegetables and the weighted unit price of edible mushroom vegetables and the sales volume of foliar vegetables. Further, a strong positive correlation was also observed between weighted unit price of cauliflower vegetables and sales volume, weighted unit price of cauliflower vegetables and sales volume of chilli vegetables, weighted unit price of cauliflower vegetables and sales volume of edible mushroom vegetables, and weighted unit price of chilli vegetables and sales volume of edible mushroom vegetables.

3.2. GA model planning

Substituting this model into the GA model yields the final optimal daily markup rate and the 7-day maximum return for each category as shown in Table 10.

Table 10 GA planning results

	Date	Flower leaf	florescent vegetables	Aquatic rhizomes	Solanaceae	Chili peppers	Edible fungi
optimal bid increment	7.1	1.805564	0.544176	0.078567	1.250531	0.681406	1.460159
	7.2	1.814026	0.529485	0.087388	1.250205	0.681704	1.457619
	7.3	1.812013	0.540666	0.098901	1.249879	0.682002	1.457694
	7.4	1.814309	0.53197	0.110242	1.249553	0.6823	1.45901
	7.5	1.814836	0.538559	0.121409	1.249227	0.682598	1.4604
	7.6	1.816092	0.533402	0.132397	1.248901	0.682896	1.461802
	7.7	1.817049	0.537277	0.143205	1.248576	0.683194	1.463207
Profit		5931	1501	899	34366	31132	4384

4. Conclusions

In fresh food superstores, vegetable commodities are characterised by a short freshness period and easy deterioration of their quality, so superstores need to replenish daily based on the historical sales and demand of each commodity. In this study, based on the sales volume data of each category, a decision model is established based on ARIMA time series forecasting and genetic algorithm, which provides a category-based replenishment plan for the superstore, and develops a reasonable daily replenishment volume and pricing strategy to achieve the highest economic efficiency.

References

[1] Wang Y, Zhang J X, Tang S. Dynamic pricing for non-instantaneous deteriorating items [J]. Journal of Intelligent Manufacturing, 2015, 26(4): 629-640.

[2] Ghare P M. A model for an exponentially decaying inventory [J]. Journal of Industrial Engineering, 1963, 14 (5): 238-243.

[3] Cao Hong, Zhou Jiang. Research on dynamic pricing of seasonal products [J]. Research on Quantitative and Technical Economics, 2004, 21(3): 122-125.92.

[4] ZHAO Guoshun. Research on stock price trend forecasting based on time series analysis [D]. Xiamen University, 2009.

[5] Li Bin. Research on multi-flexible job shop scheduling problem based on improved genetic algorithm[D]. Minnan Normal University, 2023. DOI: 10.27726/d.cnki.gzzsf.2023.000599.

[6] Chen Chao. Improvement research of adaptive genetic algorithm and its application[D]. South China University of Technology,2011.

[7] Lei, Y.J., Zhang, S.W., Li, X.W., et al. MATLAB Genetic Algorithm Toolbox and Applications [M]. Xi'an: Xi'an University of Electronic Science and Technology Press, 2005: 2-4

[8] He Xiuli. Multivariate linear modelling and ridge regression analysis [D]. Huazhong University of Science and Technology, 2006.

[9] Xia Li. Research on regional electricity consumption forecasting method based on ARIMA model and regression analysis[D]. Nanjing University of Science and Technology,2013.

[10] Li Jing. Research on standardisation method of land evaluation indicators [D]. Gansu Agricultural University, 2013.

[11] Zhang Qi. Research on distribution and pricing strategy of fresh produce supply chain [D]. Southeast University, 2022.DOI: 10.27014/d.cnki.gdnau.2021.002135.

[12] Lai, Gengen. Research on ordering and loss control strategy of fresh products in Company A[D]. Donghua University, 2020.DOI: 10.27012/d.cnki.gdhuu.2019.000183.