Agriculture Stimulates Chinese GDP: A Machine Learning Approach

Nan Zhenghan and Omar Dib(✉)

Department of Computer Science, Wenzhou-Kean University, Wenzhou, China
odib@kean.edu

Abstract. GDP is a convincing indicator measuring comprehensive national strength. It is crucial since the industrial structure, living standards, and consumption level are closely related to GDP. In recent years, the Chinese GDP has maintained rapid growth. Admittedly, the contribution of agriculture to GPD is gradually decreasing. As the foundation of life, the structure of agricultural production still needs to be improved. Therefore, this paper applies machine learning skills to investigate how to improve the agricultural production structure to promote GDP. A total of 47 agricultural products were selected and analyzed. We extracted the production data from 1980 to 2018. K-means clustering model was used to group products into several clusters. The Holt-winters model predicts the following year's production of the different agriculture products to simulate next year's GDP. The linear regression model quantifies the relationship between clusters and GDP. Based on that relationship, we provide suggestions on stimulating GDP growth. For assessment, both linear regression and neural network models are used to simulate the GDP after considering the recommendations. Results show that the proposed approach offers relevant recommendations to stimulate the Chinese GDP based on the agriculture data.

Keywords: GDP · Agriculture · Decision making · Machine learning · K-means clustering · Holt-winters · Linear regression · Neural networks

1 Introduction

According to Yi (2021), GDP (Gross Domestic Product) is the total monetary or market value of all the finished goods and services produced within a country's borders in a specific period. It's important because it reflects a country's economic strength and international status. In recent years, the Chinese GDP has maintained rapid growth. However, the development of the Chinese economy still has some issues. The industry is still the central pillar of Chinese economic growth, but this tendency violates sustainable development. Therefore, transformation to the service industry is the ultimate goal of development. Although the influence of agriculture on Chinese GDP decreases because of advanced technology, it is vital as the foundation of life. Therefore, this paper will analyze the impact of different agricultural products on GDP. Based on the analysis, a suggestion on how to modify the structure of agricultural production will be made.

L. C. Tang and H. Wang (Eds.): BDET 2022, LNDECT 150, pp. 21–36, 2023.
https://doi.org/10.1007/978-3-031-17548-0_3

Many researchers have already studied the relationship between GDP and agriculture, education, and other factors. Thereby governments can stimulate the GDP growth based on those relationships. In (Kira et al. 2013), the authors adopted the Keynes model to observe the economic development trend and gave some suggestions on stimulating the GDP. But they focused on the macroeconomy field. Therefore, they didn't cluster different factors to explore the impact of other clusters on GDP. In Ifa & Guetat's (2018) paper, the authors applied the Auto-Regressive Distributive Lags (ARDL) approach to explore the relationship between public spending on education and GDP per capita. Nevertheless, they only described the positive and negative correlation of these two factors. They didn't find the formula to manifest the quantified relationship between education and GDP. Furthermore, with only positive and negative correlations, it isn't easy to give concrete suggestions on stimulating GDP growth.

Unlike previous papers, many others ignored other factors and only focused on the relationship between agriculture and GDP. Hussain (2011) used the linear model, Ordinary Least Square (OLS), to research the contribution of agriculture growth rate towards GDP growth rate. As can be noticed, the authors assumed linearity of the variables, which may not always be the case in practice. Indeed, the relationship between variables might either be linear or non-linear. In Azlan's (2020) paper, they compared Artificial Neural Network (ANN) and ordinary least squares (OLS) model performance on predicting the stock. The results show that ANN has a more accurate prediction because it can describe the non-linear relationship.

Predicting each cluster's output is necessary to simulate next year's GDP. Yoon (2020) applied advanced machine learning skills to predict GDP under different real-world settings. They used a gradient boosting model and a random forest model. In another research, when Shih (2019) compared machine learning skills and time series, they indicated that time series methods have the most negligible value of the error measures than machine learning skills. Thus, time series methods can give a better accuracy when predicting.

The previous papers didn't research the influence of factors on GDP in the microdomain. Even if some studied the microdomain, they didn't quantify the relationship between those factors and GDP. Some papers quantified this relationship, but they omitted the non-linear relationship. As for the time analysis, some articles didn't apply time series methods to provide better prediction accuracy.

Machine learning algorithms have also been used to extract knowledge from supervised and unsupervised data in various fields. For example, Liu (2021) investigates the roles and interactions of gene variants, using machine learning (ML) and big data analysis to discover the potential autism spectrum disorder. Dou (2021) proposed an auto-machine learning approach for predicting the risk of progression to active tuberculosis based on Its association with host genetic variations. Kong (2021) developed a model-free, machine-learning-based solution to exploit reservoir computing in the normal functioning regime with a chaotic attractor.

This paper applies K-means Xu (2019) clustering to observe the impact of different clusters in agriculture on GDP. Holt-Winters model is used to predict the output of clusters due to its reliability. After that, the linear regression methods will quantify the relationship between those clusters and GDP. Based on the relationship between

clusters and GDP, a suggestion on stimulating GDP growth is made. Finally, both linear regression and neural networks Aggarwal (2018) are used to assess the performance of the proposed approach.

2 Methodology

This section presents the methodology adopted to apply machine learning skills to explore the improvement of the agricultural production structure to stimulate GDP. We summarize the workflow of the study in Fig. 1 and elaborate on the steps as follows. Firstly, the data is retrieved from two sources. One is China's statistical yearbook (2020): The GDP data from 1980 to 2018. The other is the National Bureau of Statistics of China, Agriculture part: 1. The output of Tea and Fruits 2. The production of Major Farm Products 3. The output of Livestock Products 4. The output of Aquatic Products (From 1980 to 2018).

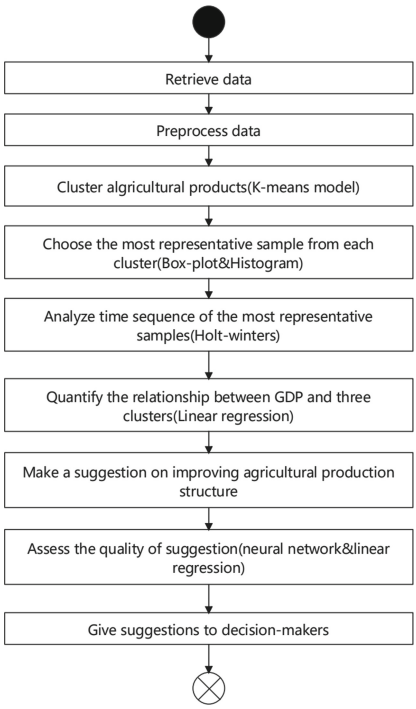

Fig. 1. Experiment flowchart.

Secondly, the data is preprocessed before applying machine learning skills. Mainly, the data is cleaned and combined. The reason is that a mass of production data is not continuous from 1980 to 2018. Many data were primarily missed before 2000 because of the backward traceability system. Thus, those products that don't have continuous data are omitted. Then the remaining 47 agricultural production data from 1980 to 2018

is combined with the GDP data from 1980 to 2018. In this way, the relationship between GDP and agricultural production can be studied. Thirdly, those agricultural products need to be clustered. In this paper, there are 47 samples, which constitutes the real data. However, this paper handles the agricultural production structure on a microscopic level. If the whole agriculture data is not divided into sub-parts, this paper can only study the macroscopic relationship between agriculture and GDP. Therefore, it's important to cluster the agricultural data so that the robustness of the internal structure of agriculture can be closely researched.

Fourthly, the most representative samples need to be selected from each cluster. The reason is that studying the impact of each agricultural product in the same cluster on GDP can't reflect the superiority of clustering. Clustering has already extracted the commonality of different products and integrated them into the same cluster based on the commonality. That's why repeatedly studying the impact of those similar products on GDP is meaningless and duplicated. Due to the commonality of various agricultural products in the same cluster, each cluster can be regarded as the minimum unit. However, each cluster is the integration of many products. Hence, it's vital to find a product representing the whole cluster to study the relationship between each cluster and GDP.

Fifthly, with the representative products obtained from the fourth step, the time sequence is analyzed. There are two aims of interpreting the time sequence of each usual product. One is that the trend of time sequence can be obtained. A recommendation is proposed to improve the agricultural production structure to stimulate the Chinese GDP. The other is to assist the simulation part, and more details will be shown in the eighth step. Sixthly, the relationship between GDP and three clusters is quantified. This step aims to find which clusters can promote the GDP and which clusters can impair the GDP. Also, the degree of influence can be studied from the weight of each cluster. By considering the effect of each cluster on GDP and the trend of each cluster obtained from step5, the suggestion can be made to improve the agricultural production structure. Seventhly, a suggestion on improving the agricultural production structure is proposed based on the influence and trend mentioned above.

Eighthly, due to the uncertainty of the suggestion's quality, it's significant to assess it. This paper assumes that the current year is 2014. Therefore, a suggestion was proposed in 2014 to improve the 2015 agricultural production structure. 2015 predicted production data would be a base for the suggestion. Based on 2015 predicted data, this paper simulates that the government implements the suggestion on 2015 predicted data. Then the expected production can either be increased or decreased. Then, the simulated GDP can be obtained based on the 2015 modified production. Also, the predicted GDP can be obtained based on the 2015 predicted GDP. Finally, the predicted 2015 GDP and simulated 2015 GDP are compared to assess the quality of the suggestion. After the evaluation, only a convincing suggestion can be applied in real-world settings. Ninthly, once the validity of the suggestions is confirmed, they are communicated with the decision-makers to improve the agricultural production structure.

3 Results and Analysis

3.1 K-Means Clustering

There is a total of 47 agricultural products. It's meaningless and impossible to research the impact of each product on the GDP. In addition, all of them belong to agriculture, so there must be some similarity among them. That's why K-means clustering is applied to first group them into various clusters. The value k should be determined before clustering; however, there is no k value. Therefore, k was set to be 2, 3, 4 and 5 roughly.

After observing the result of clustering under these four values, there can be a basic idea about the optimal value of k. Figure 2 presents the results under each case. The results can be evaluated externally from two indicators. One is the internal compactness, and the other is the external degree of separation. More specifically, compactness measures whether the sample points in a cluster are compact enough, such as the average distance to the cluster's center, variance, etc. The degree of separation measures whether the sample is far enough away from other clusters. From these two perspectives, both k = 2 and 3 gave acceptable results. To specify the value of k, these two techniques are involved in determining the k value. They are elbow plot Dinov (2018) and average silhouette scores. Figure 3 demonstrates the elbow plot. Smaller than the optimal number of clusters, the increase of k substantially increases the compactness of each cluster. Thus, the SSE

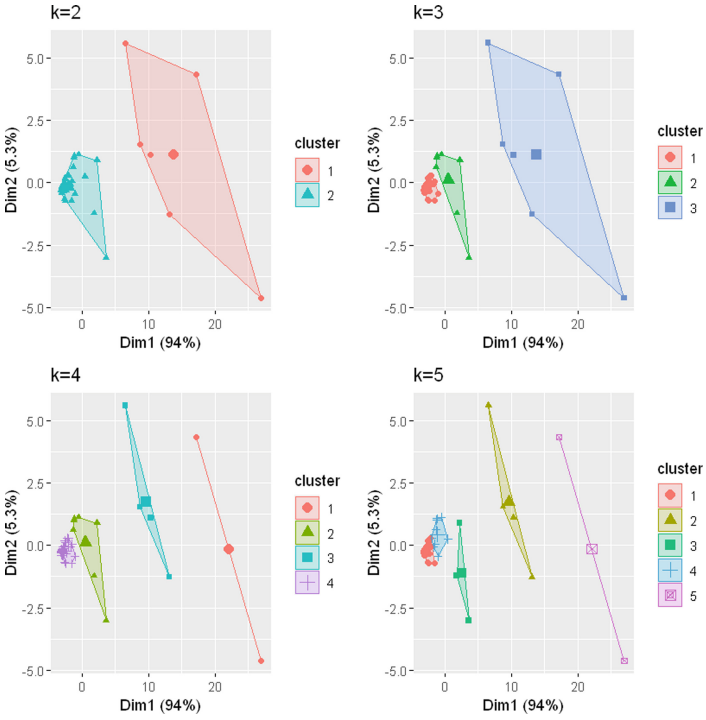

Fig. 2. Clustering under k = 2, 3, 4, 5 respectively.

drops quickly. However, when k reaches the optimal number, the expansion of k hardly impacts SSE. Therefore, the K value corresponding to this elbow is the actual clustering number.

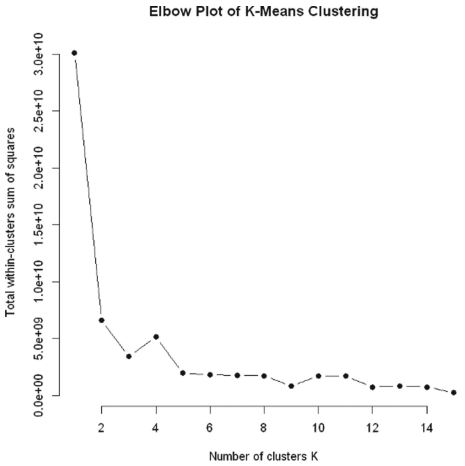

Fig. 3. Elbow plot.

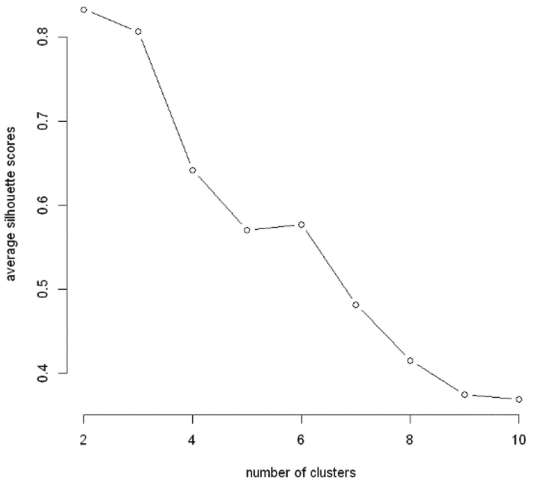

Fig. 4. Average silhouette scores.

In the elbow plot, the sample clustering is refined with the increase of cluster number K. The degree of aggregation of each cluster gradually increases, so the error square and SSE become smaller. When k is In Fig. 3, k = 2 is the actual clustering number. However, the elbow method is limited because it only considers compactness without considering the external degree of separation. Hence, Fig. 4 depicts the other approach, named average silhouette scores. The silhouette score (Shahapure, 2020) is $S = (b-a)/max(a,$

b), where a is the average distance between one sample and other samples in the same cluster, b is the average distance between the sample and all the samples in the nearest group. Here, a is the compactness above, b is the degree of separation. The smaller a is, the higher compactness is. The larger b is, the higher the degree of separation is. Therefore, k will obtain the optimal value when the S is close to 1. In Fig. 4, $k = 2$ is the optimal value, the same as the elbow method. To sum up, $k = 2$ is the real clustering number. Although the products have already been clustered wonderfully, cluster2 contains 41 samples, much more than cluster1. For the convenience of research, cluster2 was visualized to see if it could reasonably be divided into several clusters. Figure 5 demonstrates the histogram of cluster2.

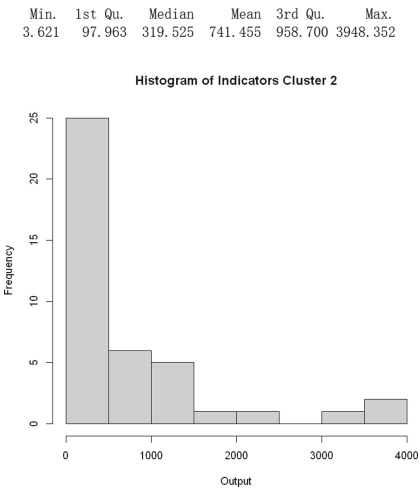

Fig. 5. Cluster2 histogram.

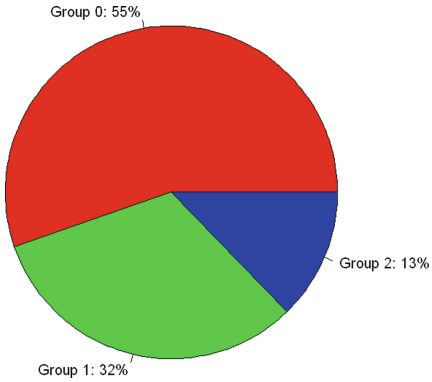

Fig. 6. Composition of agricultural products.

Here, the mean average annual output is 741.455. More than half are smaller than the mean value. Therefore, a threshold was set to separate initial cluster2 into two groups.

Those samples which are smaller than 700 will be group0. The remaining part in cluster2 is called group1. Cluster1 will be group2. Ultimately, three groups are generated for the convenience of research. Figure 6 shows the pie chart of these groups.

3.2 Clusters Analysis

After grouping all products, we select the most representative sample from each group. It is not relevant to explore the influence of each product on the GDP. Thus, the relationship between these three groups and GDP is studied. All three groups will be executed in the same manner. Here, group0 is taken as an example to demonstrate the complete process. The most representative is either the median or the mean based on the distribution of the group. Figure 7 describes the distribution of the group0. For the group0, the median was chosen as the most representative sample because an apparent right-skewed trend was found. Mode is smaller than the median, 137.526, and the median is smaller than the mean, 184.943. In this regard, the median is more reasonable to be the representative sample. Hence, the median, which is persimmon output, represents group0. Figure 8 presents the time sequence of persimmon. Notice that the abscissa begins from 0, which is 1980. With the same process, the most representative sample of group1 is apple, and garden fruit represents group2. Figures 9 and 10 display their time sequence of them, respectively.

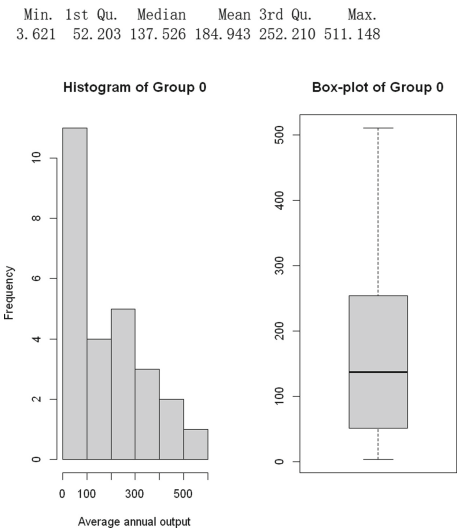

Fig. 7. Group0 distribution.

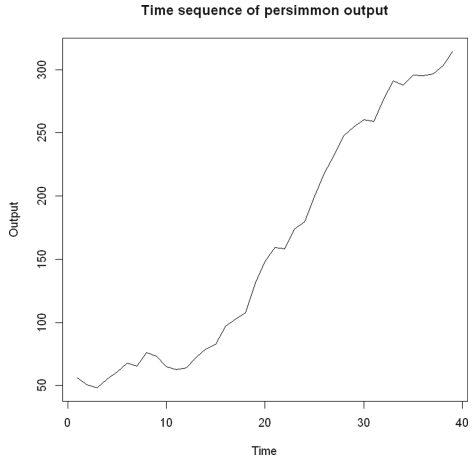

Fig. 8. Persimmon time sequence diagram.

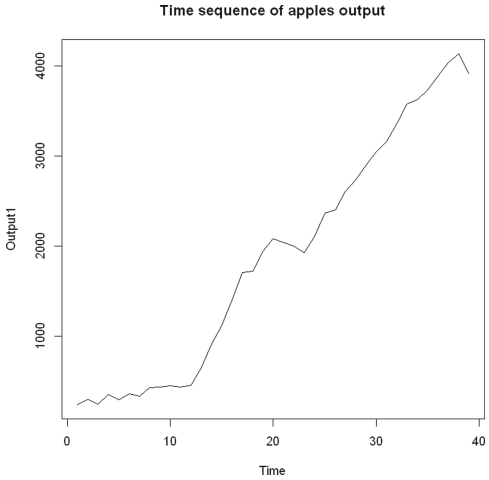

Fig. 9. Apple time sequence diagram.

3.3 Time Analysis

This paper focuses on improving the structure of agricultural production to promote GDP. To provide valid suggestions, it's essential to observe the trend of each cluster. Figures 8, 9, and 10 show the time sequence diagrams of samples. From the diagrams, the shape of the curve is quite straightforward. There is a clear upward trend and no seasonality. Thus, the holt-winters model can probably give a satisfying prediction. Actually, the holt model is competent to conduct the prediction. The holt-winters model is applied here because there might be some hidden seasonality that can be observed Canela (2019). That's why this section will apply the holt-winters model to predict the future trend of those representative samples. Take group0, for example; there are three

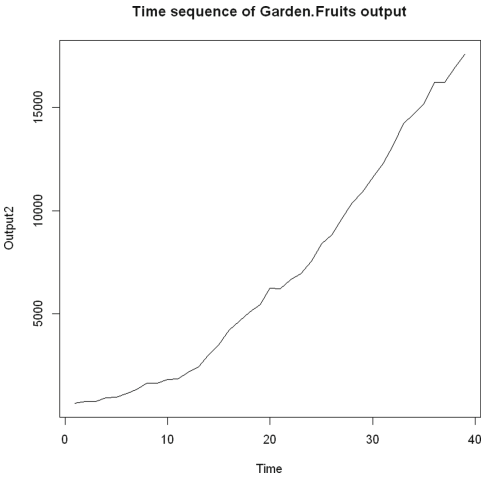

Fig. 10. Garden fruits time sequence diagram.

parameters in this model. They are alpha, beta, and gamma. Alpha is the level that is the weight of new information, beta is the trend, and gamma is the seasonality. Lower weights give less weight to recent data and vice versa. The Holt-winters function in the R language will automatically choose the best parameters by minimizing AIC and BIC values. Eventually, alpha = 1, beta = 0.2824653, gamma = FALSE. Gamma is false means that there is no seasonality. Figure 11 presents the holt-winters fitting.

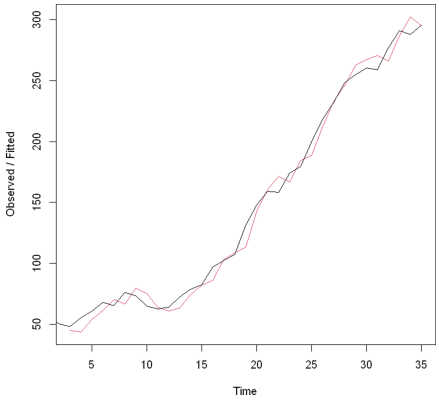

Fig. 11. Group0 holt-winters fitting.

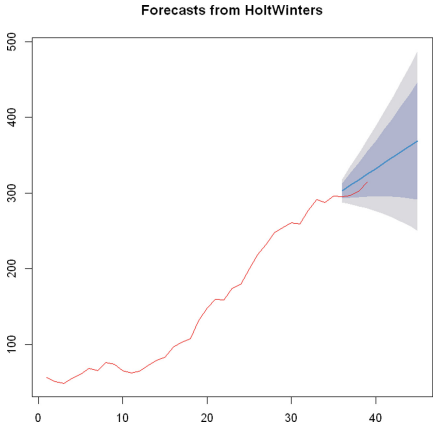

Fig. 12. Group0 forecasts from holt-winters.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	1.376459	7.941914	6.576597	2.065224	5.929615	0.7512071	0.02586935

Fig. 13. Group0 holt-winters fitness.

The red line is the fitting line, and the black line is the actual line. Figure 13 presents the fitness of the holt-winters model. Here are two significant indicators which can evaluate the fitness of the model. One is MAPE (mean absolute percentage error, which is approximately 6, meaning that the forecast is off by 6% on average. The other is the RMSE (root mean square error) which reflects the model's fitness. Here RMSE is 8. Furthermore, Fig. 11 shows that the predicted value is consistent with reality. After selecting the best parameters, Fig. 12 demonstrates the prediction. In this figure, holt-winters started prediction from 2015 to 2018. The red line is the actual value. The blue line is the predicted value. Dark grey is an 80% confidence interval, and light grey is a 95% confidence interval. In the worst case, the actual is still located in the dark grey part. Thus, it's a relevant prediction. It's not necessary to predict more than one year because the output of agricultural products is primarily determined by the annual policy and the previous year's output. Moreover, this paper aims to give governments some suggestions on improving the agricultural product structure. Incredibly, the advice can be valid for a long time because too many dynamic factors influence the output. When focusing on 2015, the actual output is 294.96, and the predicted one is 302.9005. As mentioned before, the predicted one belongs to an 80% confidence interval, which is a good prediction. After all, a slight modification of the policy can have a crucial impact on agriculture. The other two samples are executed similarly so that the result will be displayed directly. Likewise, for group1, the actual value of 2015 is 3889.9, and the predicted one is 3858.523.

3.4 Quantification Analysis

Before making the suggestion, how each group of products affects GDP should be explored. Figures 14 and 15 show the result of the fitting. The number of * represents the performance of each feature prediction in the model. Except for intercept, all the model shows a perfect performance of prediction. Besides, R-squared is 0.98, presenting a prominent fitness. Therefore, the linear regression model can explain their relationship well. Here is the concrete formula: GDP $= -2616.6$ * Group0 $- 252.4$ * Group1 $+$ 151.1 * Group2 $- 51558.2$.

```
Call:
lm(formula = GDP ~ Output.of.Persimmons.10000.tons. + Output.of.Apples.10000.tons. +
    Output.of.Garden.Fruits.10000.tons., data = df2)

Coefficients:
                          (Intercept)      Output.of.Persimmons.10000.tons.
                              51558.2                               -2616.6
       Output.of.Apples.10000.tons.   Output.of.Garden.Fruits.10000.tons.
                                -252.4                                151.1
```

Fig. 14. Linear regression fitting.

```
Call:
lm(formula = GDP ~ Output.of.Persimmons.10000.tons. + Output.of.Apples.10000.tons. +
    Output.of.Garden.Fruits.10000.tons., data = df2)

Residuals:
   Min     1Q Median     3Q    Max
-75882 -27163   7302  25507  56726

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     51558.199  18716.749   2.755  0.00926 **
Output.of.Persimmons.10000.tons.  -2616.608    400.035  -6.541 1.51e-07 ***
Output.of.Apples.10000.tons.        -252.433     30.916  -8.165 1.28e-09 ***
Output.of.Garden.Fruits.10000.tons.   151.105      8.749  17.271  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36860 on 35 degrees of freedom
Multiple R-squared:  0.9823,    Adjusted R-squared:  0.9808
F-statistic: 648.7 on 3 and 35 DF,  p-value: < 2.2e-16
```

Fig. 15. Linear regression fitting evaluation.

3.5 Decision Making

According to the formula at the end of Sect. 3.4, it was found that both group0 and group1 impair the GDP, whereas group2 can promote GDP. What's more, group0 has the most significant impact on GDP. Considering the output increase from a macro-perspective, next year's output is probably higher than the previous year. This paper will simulate 2014. Suppose this year is 2014, and the suggestion was proposed in 2014. The aim

of the suggestion is to improve the agricultural product structure in 2015. Hence, the recommendation is that governments should slow down the output growth of group0 and group1 and promote the output growth of group2 in 2015. To get the growth of output, 2013 output should be obtained. Then the development of 2014 to 2015 can be estimated from 2013 to 2014.

Table 1. 2013–2015 real, predicted, and simulated output

Year	Output of group0 (10000.tons.)	The output of group1 (10000.tons.)	The output of group2 (10000.tons.)
2013(real)	287.87	3629.81	14675.17
2014(real)–2013(real)	7.69	105.58	496.19
2014(real)	295.56	3735.39	15171.36
2015(real)	294.96	3889.9	16200.91
2015(predicted)	302.9	3858.5	15827.66
2015(simulated)–2014(real)	4.44	54.61	928.64
2015(simulated)	300	3790	16100

Table 1 shows the actual output of 2013 and 2014. The difference between 2014 and 2013 of groups 0, 1, and 2 are displayed on the third row. Based on the discrepancy and strategy above, the growth of output from 2014 to 2015 was displayed on the seventh row. Finally, the specific suggestion is that the average production of group0 should be controlled to 300, the average group1 result should be 3790, and the average group2 output should 16100.

3.6 Decisions Evaluation

In this section, the validity of the suggestion will be estimated by comparing actual, simulated, and predicted GDP. Then, both neural network and linear regression will be applied to confirm the validity. Although the linear regression in Sect. 3.4 demonstrates exciting results, there exists a possibility that the relationship between agricultural products and GDP is non-linear. Figure 16 shows the fitted neural network.

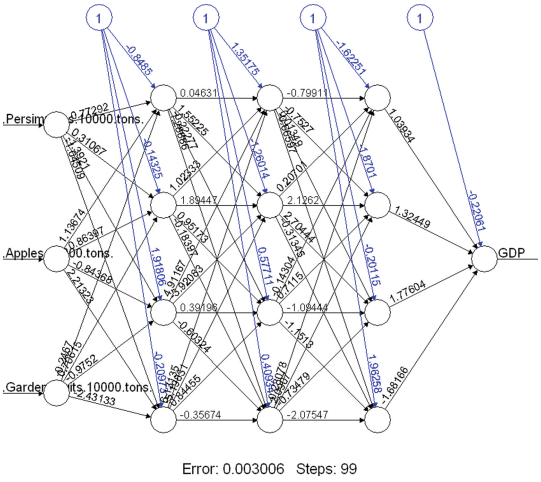

Fig. 16. Neural network

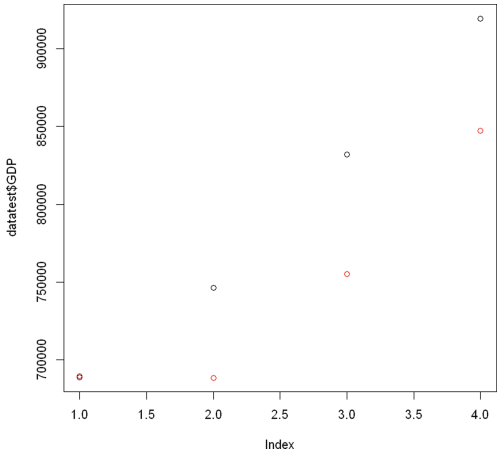

Fig. 17. Neural network prediction.

The number of hidden layers is 3, and there are four neurons in each hidden layer. The r-squared value determines the configuration. R-squared can be maximized, which is 0.5257 under this configuration. After establishing the model, Fig. 17 shows the result of the prediction. There are four years in this model which are 2015, 2016, 2017, and 2018. The red points are the predicted values, and the black points are the actual value. Although the predictions are not so accurate in the last three years, the forecast and the actual value of 2015 almost overlap. As mentioned before, this paper only needs to predict one year GDP. After predicting the 2015 GDP, the GDP after governments adopted the suggestion will be simulated. Linear regression model and neural network are applied to simulate the 2015 GDP. Table 2 shows the result.

Table 2. Simulated GDP

Year	Output of group0 (10000.tons.)	The output of group1 (10000.tons.)	Output of group2 (10000.tons.)	GDP
2015(simulated by Linear regression)	300	3790	16100	699209.5
2015(predicted by Linear regression)	302.9	3858.5	15827.66	641487.4
2015(simulated by Neural network)	300	3790	16100	705348.0
2015(predicted by Neural network)	302.9	3858.5	15827.66	694861.8
2015(real)	294.96	3889.9	16200.91	688858.2

Table 2 conveys two pieces of information. First, Table 2 reveals that the neural network performed better than linear regression while predicting the 2015 GDP. The other is that both simulated results are higher than the real one. That is to say; the suggestion is convincing and valid.

4 Conclusions and Future Works

This paper focuses on providing and verifying valid suggestions about improving agricultural production so that it can stimulate the Chinese GDP. Four machine learning skills involving K-means clustering, Holt-winters, Linear regression, and Neural network are applied throughout the whole process. The result shows that the 2015 simulated GDP is higher than the 2015 predicted GDP after adopting the suggestions, and the 2015 simulated GDP is even higher than the 2015 real GDP. Because this paper assumes that an excellent agricultural production structure can stimulate GDP, the suggestions are valid in the real world. Separately, this paper only researches the data from 1980 to 2018. The reason is that many products' output data is still being collected. 2018 is the most recent year, providing the most output data. In addition, in this work, we only applied the K-means clustering because of its simplicity and low time complexity. Other clustering algorithms can be applied to improve the accuracy. In the future, the proposed methodology in this paper can be used in other fields besides agriculture. What's more, both the time analysis and clustering models can be changed to improve the accuracy. Also, the parameters of the neural network, which are the number of neurons and layers, can be tuned to a better combination. Furthermore, the most representative products can be selected more reasonably, or the methodology can be modified to decrease the impact of a limited selection of representatives. In addition, the data can also be improved in the future. Instead of using the annual production data, the monthly data can enrich the dataset, improving the prediction accuracy.

References

1. Anwar, H., Khan, A.Q.: Relationship between agriculture and GDP growth rates in Pakistan: an econometric analysis (1961–2007). Academic Research International 1, no. 2, p. 322 (2011)
2. Azlan, A., Yusof, Y., Mohsin, M.F.M.: Univariate financial time series prediction using clonal selection algorithm. Int. J. Adv. Sci. Eng. Inf. Technol. **10**(1), 151–156 (2020)
3. Yoon, J.: Forecasting of real GDP growth using machine learning models: gradient boosting and random forest approach. Comput. Econ. **57**(1), 247–265 (2021). https://doi.org/10.1007/s10614-020-10054-w
4. Shih, H., Rajendran, S.: Comparison of time series methods and machine learning algorithms for forecasting Taiwan blood services foundation's blood supply. J. Healthcare Eng. 2019 (2019)
5. Kira, A.R.: The factors affecting Gross Domestic Product (GDP) in developing countries: the case of Tanzania (2013)
6. Ifa, A., Guetat, I.: Does public expenditure on education promote Tunisian and Moroccan GDP per capita? ARDL approach. J. Finan. Data Sci. **4**(4), 234–246 (2018)
7. Liu, Z., et al.: Machine learning approaches to investigate the relationship between genetic factors and autism spectrum disorder. In: 2021 The 4th International Conference on Machine Learning and Machine Intelligence, pp. 164–171 (2021)
8. Dou, W., et al.: An AutoML Approach for Predicting Risk of Progression to Active Tuberculosis based on Its Association with Host Genetic Variations (2021)
9. Kong, L.-W., Fan, H.-W., Grebogi, C., Lai, Y.-C.: Machine learning prediction of critical transition and system collapse. Phys. Rev. Res. **3**(1), 013090 (2021)
10. Xu, J., Lange, K.: Power k-means clustering. In: International Conference on Machine Learning, pp. 6921–6931. PMLR (2019)
11. Yi, Y.: The relationship between industrial structure and economic growth in China—an empirical study based on panel data. In: E3S Web of Conferences, vol. 275, p. 01009. EDP Sciences (2021)
12. Canela, M.Á., Alegre, I., Ibarra, A.: Holt-winters forecasting. In: Quantitative Methods for Management, pp. 121-128. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-17554-2_13
13. Aggarwal, C.C.: Neural Networks and Deep Learning. Springer, vol. 10, pp. 978–3 (2018). https://doi.org/10.1007/978-3-319-94463-0
14. Dinov, Ivo D. K-Means Clustering. In: Data Science and Predictive Analytics, pp. 443-473. Springer, Cham, 2018. https://doi.org/10.1007/978-3-319-72347-1_13
15. Shahapure, K.R., Nicholas, C.: Cluster quality analysis using silhouette score. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pp. 747–748. IEEE (2020)