

Identification of vegetable pests using spectral clustering analysis

Xia Ji'an¹, Yang Yuwang^{1*}, Cao Hongxin², Zhang Wenyu²,
Ge Sijun², Chen Guangwei³

(1. School of Computer Science and Engineering, Nanjing University of Science and Technology, 210094 China;

2. Institute of Agricultural Information, Jiangsu Academy of Agricultural Sciences, 210014 China;

3. Purdue University-West Lafayette, IN. 47906, USA)

Abstract: The broad bean, an annual leguminous plant, is widely planted in China as a food source. During its growing period, the leaves and pods of the broad bean are often damaged by pests, which reduces yield and even results in crop death. Information on crop growth can be obtained quickly and comprehensively by collecting green crop reflectance spectra followed by using cluster analysis to identify the spectra of infected crops. In this study, a spectral acquisition device was designed to collect visible-near infrared reflectance spectra for three types of broad bean leaf samples in a darkroom with supplemental lighting. The tested samples included healthy leaves, leaves with slight insect infestation, and leaves with strong insect infestation, 30 spectra were sampled from each type. Principal component analysis (PCA) was used to conduct dimension reduction of the spectrum dataset. Hierarchical cluster analysis of the sampled spectra was conducted using the single-link, complete-link, and average-link methods and the sum of squares of deviations (Ward's method). Our results showed that hierarchical clustering effectively identified the three types of spectral samples. Ward's method had the highest identification rate of sample spectra, followed by the single-link and average-link methods, whereas the complete-link method exhibited the poorest performance. The single-link method required the least amount of time, followed by the complete-link and average-link methods, whereas Ward's method required the most time.

Keywords: broad bean, pest identification, spectrum technology, clustering analysis

Citation: Xia, J. A., Y. W. Yang, H. X. Cao, W. Y. Zhang, S. J. Ge, and G. W. Chen. 2017. Identification of vegetable pests using spectral clustering analysis. *International Agricultural Engineering Journal*, 26(3): 176–183.

1 Introduction

In China, broad bean crops cover an area of approximately 3.2 million acres, with an annual output around 6.6 billion pounds. Broad beans are used as both food and nectar plants. Their flowers, leaves, and stems are used as an auxiliary treatment for constipation, hypertension, and edema, which give the crop a high economic and pharmaceutical value. Major pests that feed on broad beans include the broad bean weevil, aphid, and vegetable leaf miner; of these, the broad bean weevil is the most common. The adult bean weevil mainly infects

bean pollen and tender leaves, and the larva parasitizes the pod and eats the bean. If infestation is not detected quickly and preventive measures are not implemented, significant decreases in crop output may occur.

Precision agriculture (PA) offers a fast and accurate method of obtaining information on crops that suffer from pest infestation, allowing strategic management during the crop growing period (Gebbers et al., 2010). Pest control in modern agriculture utilizes computer-based intelligent equipment, remote sensing technology, and real-time monitoring methods to distinguish healthy and infested plants and the degree of crop damage. By comparing the spectral features of infested and healthy crops, differences in crop spectral reflectance within characteristic bands may be detected.

The spectrum analysis has a long history; the application of modern near-infrared spectrum technology

Received date: 2016-10-10 Accepted date: 2017-08-22

* Corresponding Author: Yang Yuwang, Professor of Computer Science and Engineering College, Nanjing University of Science and Technology, Nanjing, 210094, China. Email: yuwangyang@njust.edu.cn; Tel: +8613512503829.

began with quality analyses of agricultural products. Birth and Norris et al. (1957, 1958) used near-infrared spectrum analysis to examine the quality of eggs, fruits, and vegetables. Recently, analytical methods based on spectrum technology have been widely studied and applied in agriculture. Niewitetzld et al. (2010) built a model for the automatic and optimal selection of rape seeds using infrared spectrum technology. Sankaran et al. (2011) used visible-near infrared spectrum analysis to build an identification model to determine the degree of huanglongbing infection in citrus. Lin et al. (2014) selected spectra with wavelengths from 550 to 650 nm with image information to analyze cabbage leaves damaged by diamondback moths. Anna et al. (2014) used Fourier-transform infrared spectroscopy in three sample groups of coast live oak to determine its resistance to phytophthora ramorum prior to infection. Naresh et al. (2013) studied different spectral indices to analyze cotton pests and obtained 69%-74% accuracy. Zhang et al. (2012) studied spectral and digital images of crop leaf area index (LAI), and the results showed that the spectral method was more stable than the digital image method using LAI measurements.

Cluster analysis is an unsupervised multivariate statistical method. Based on characteristics of the data, it can analyze relationships between close and distant objects using similarity or difference indices. Cluster analysis has been widely applied in agriculture in recent years. Jan et al. (2015) applied cluster analysis via advanced machine learning to detect biotic stress. Li et al. (2012) used Fourier-transform infrared spectroscopy and a cluster analysis to study the characteristics of broad bean diseases and pests. Gumienna et al. (2016) used PCA alongside the nonlinear iterative partial least squares (NIPALS) algorithm to detect the bioethanol production efficiency of 258 different corn types, and the result showed that the ethanol yield of the fermentation process depended on the variety of the grain. Cheilane et al. (2016) used hierarchical classification to analyze the mineral components of breadfruit. Jin et al. (2016) used PCA and hierarchical clustering to study the oxidation resistance and color of 110 different herb teas, compared them with eight types of green tea.

At present, pattern recognition and data mining of crop spectra and images is a major focus of precision agriculture. Through cluster analysis of the spectra of specific crops, which can provide data and theoretical support for further research and application of cluster analysis in agriculture, we can effectively recognize and classify crop pests and diseases.

2 Materials and methods

2.1 Plant and pest materials

Three types of broad bean leaves were selected as the experimental subjects: healthy leaves, leaves slightly infested by pests, and leaves strongly infested by pests (Figure 1). Thirty sample spectra of each type were collected, such there were 90 experimental samples in total. Leaves and broad bean weevils were collected from the experiment field at the Jiangsu Academy of Agricultural Sciences on May 1, 2016.

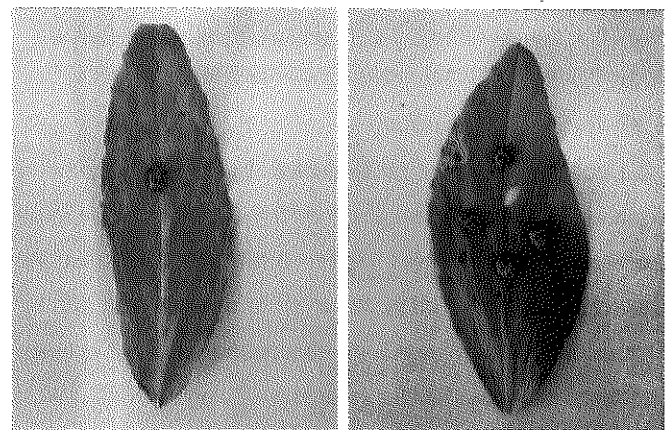


Figure 1 Experimental samples

2.2 Spectral acquisition

An Ocean Optics portable fiber optic spectrometer (USB4000) with an acquisition band from 200 to 1100 nm and an optical resolution of 1.5 nm was used to acquire crop spectra. Reflected light was transmitted to the spectrometer through probes and fiber optics (Figure 2), and the spectrometer transferred the collected spectral data to a laptop for download and analysis. It is difficult to ensure a stable and continuous source of sunlight during long experiments. For this reason, and because stray natural light can affect spectral collection, the experiment was conducted in a darkroom environment, using a 25 W UVB halogen tungsten lamp as the light source. A round BaSO₄ test whiteboard was used to

balance the spectra and adjust the integration time. During spectral acquisition, the distance between the fiber optics probe and the center of the sample surface was 3.5 cm. To minimize the impact of the incident angle of the light source on the reflectance spectrum, we set angles between the light source and the sample at 45°, 90°, and 135° for reflectance spectrum measurements; the average value of the spectra from all three directions was considered the reflectance spectrum value for the sample.

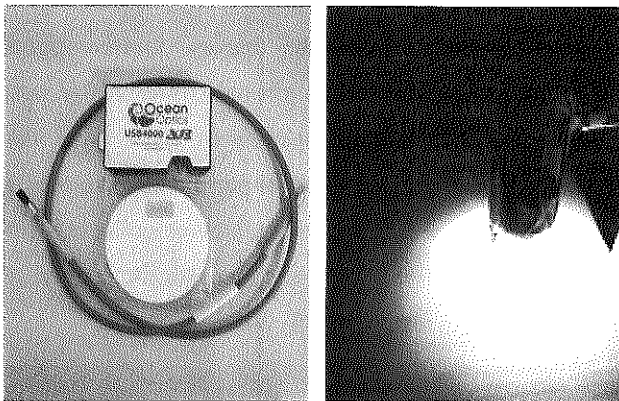


Figure 2 Spectral acquisition device and environment

2.3 Data preprocessing

Because the spectra collected by the spectrometer contained noise and baseline drifts, the Savitzky–Golay convolution method was used to smooth and de-noise the spectra. Savitzky–Golay is a filtering and smoothing technology widely used in absorption and reflectance spectrum analysis (Ruffin et al., 1999; Turton et al., 1992; Luo et al., 2005).

$$X_i^* = \frac{\sum_{j=-r}^r X_{i+j} W_j}{\sum_{j=-r}^r W_j} \quad (1)$$

The 5-point convolution smoothing method was applied to collect spectrum data considering the weighting factors of the Savitzky–Golay coefficient window:

$$X_i^* = \frac{-3X_{i-2} + 12X_{i-1} + 17X_i + 12X_{i+1} - 3X_{i+2}}{35} \quad (2)$$

Through the data pre-proceeding, the reflectance spectra of the three kinds of samples is shown in Figure 3. Chlorophyll plays a crucial role in the absorption of light by green plants. Because chlorophyll tends to strongly absorb spectra from a band range of 200 to 450 nm,

spectra within this band have a low average reflectance, generally within 10%. Because wavelengths around 550 nm are considered within the range of significant spectrum reflection for chlorophyll, the reflectance spectrum curve for spectral bands from 490 to 600 nm displays a wave peak and medium reflectivity (approximately 8%-28%) within this range. Between the visible light waveband and the near-infrared waveband (i.e., around 760 nm), reflectivity increases sharply (45%-55%) and forms a red edge, which is the most significant feature in the reflectance spectrum curve for green plants. Because pests have complicated chemical components with differing optical power absorption, the pest spectrum does not follow any rule, and further research is required to analyze the spectral characteristics of certain pests.

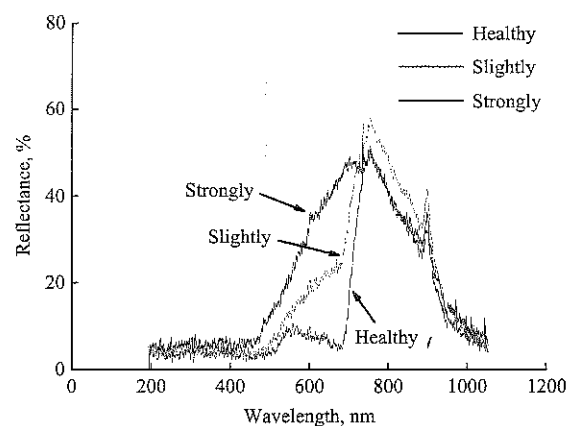


Figure 3 Reflectance spectra of the three sample types

2.4 Cluster analysis

Cluster analysis is a multivariate method of classification that uses individual characteristics to ensure that individuals within the same class will have maximum homogeneity while different classes will exhibit maximum heterogeneity. Hierarchical clustering is the most widely used method of cluster analysis and is known for its stable clustering results and straightforward process (Praisler et al., 2015). It can also be applied in high-dimensional data clustering and therefore is widely used in precise agriculture research (Mratinic et al., 2012; Balan et al., 2015; Justyna et al., 2016).

This study used hierarchical clustering to conduct cluster analyses of crop spectra. Hierarchical clustering uses the distance coefficient as its classification statistic, which requires consideration of two core factors: (I)

similarity measurement between samples and (II) similarity measurement between classes. Many methods can be used to measure the similarity between samples (Li et al., 2010; Kavitha et al., 2013). Zong et al. (2014) used Euclidean distance, cosine angle, and correlation methods to measure similarity between pest spectra of brassica chinensis, showing that Euclidean distance performed best in identifying pest spectra. There are various methods of measuring similarity between classes; the proper method should be chosen based on the nature and characteristics of the clustering object.

Table 1 Distance measurement of hierarchical clustering

Similarity measurement of hierarchical clustering			
Similarity measurement between samples		Similarity measurement between classes	
1	Similarity distance	1	Single-Link
2	Cosine angle	2	Complete-Link
3	Correlation coefficient	3	Average-Link
4	Relative error distance	4	Centroid-Link
5	Maximum dissimilarity coefficient	5	Sum of Squares of Deviations (WARD)

Four methods of the single-link were selected, namely, complete-link, average-link methods and Ward's method to measure the similarity between classes. Then the hierarchical cluster analyses of the spectra of three types of crops were conducted and the clustering results were analyzed.

(I) Single-link method

Assuming that d_{ij} is the similarity distance between sample i and sample j , and D_{pq} is the distance between class G_p and class G_q , then Formula (3) can be used to calculate the distance between different classes:

$$D_{pq} = \text{Min}_{i \in G_p, j \in G_q} d_{ij} \quad (3)$$

The single-link method defines interclass distance as the distance between the most similar samples in two classes.

(II) Complete-link method

By adopting the assumption in Formula (3), Formula (4) can be used to calculate the distance between Class G_p and Class G_q :

$$D_{pq} = \text{Max}_{i \in G_p, j \in G_q} d_{ij} \quad (4)$$

The complete-link method defines interclass distance as the longest similarity distance between two samples

from two classes.

(III) Average-link method

By adopting the assumption in Formula (3), n_p and n_q refer to the sample quantity; Formula (5) can then be used to calculate interclass distance using the average-link method:

$$D_{pq} = \frac{1}{n_p n_q} \text{Max}_{i \in G_p, j \in G_q} d_{ij} \quad (5)$$

The average-link method uses the average value of distances between samples as the measurement between classes.

(IV) Ward's method

Assuming that n samples must be divided into k classes denoted by $G_1, G_2 \dots G_k$, n_i refers to the sample quantity in Class G_i . $\bar{X}^{(i)}$ is the center of gravity of G_i , and $X_i^{(i)}$ is the i^{th} sample in i . The total sum of squares of deviations for k classes is

$$W = \sum_{i=1}^k \sum_{i=1}^{n_i} (X_i^{(i)} - \bar{X}^{(i)})'(X_i^{(i)} - \bar{X}^{(i)}) \quad (6)$$

Ward's method combines the two classes with the smallest variation in W until all samples are combined into one class.

3 Results and discussion

3.1 Hierarchical clustering of pest spectra

Matlab 2012b was used to implement the hierarchical clustering algorithm. Using the Euclidean distance for distance measurements between samples, and using four different interclass measurement methods for the hierarchical cluster analysis of the three types of samples, we obtained four different tree diagrams (Figures 4).

These tree diagrams showed that all four clustering methods first calculated the Euclidean distance between samples, then categorized samples with the smallest Euclidean distance into one class, and finally calculated the distance between new classes and other samples. Because different methods were used to calculate interclass distance, the final clustering results were different. The advantage of the hierarchical clustering method is that the principle is simple; thus, we can observe in detail the process by which small classes are aggregated into large classes, and the similarity between samples can be seen in the similarity distance.

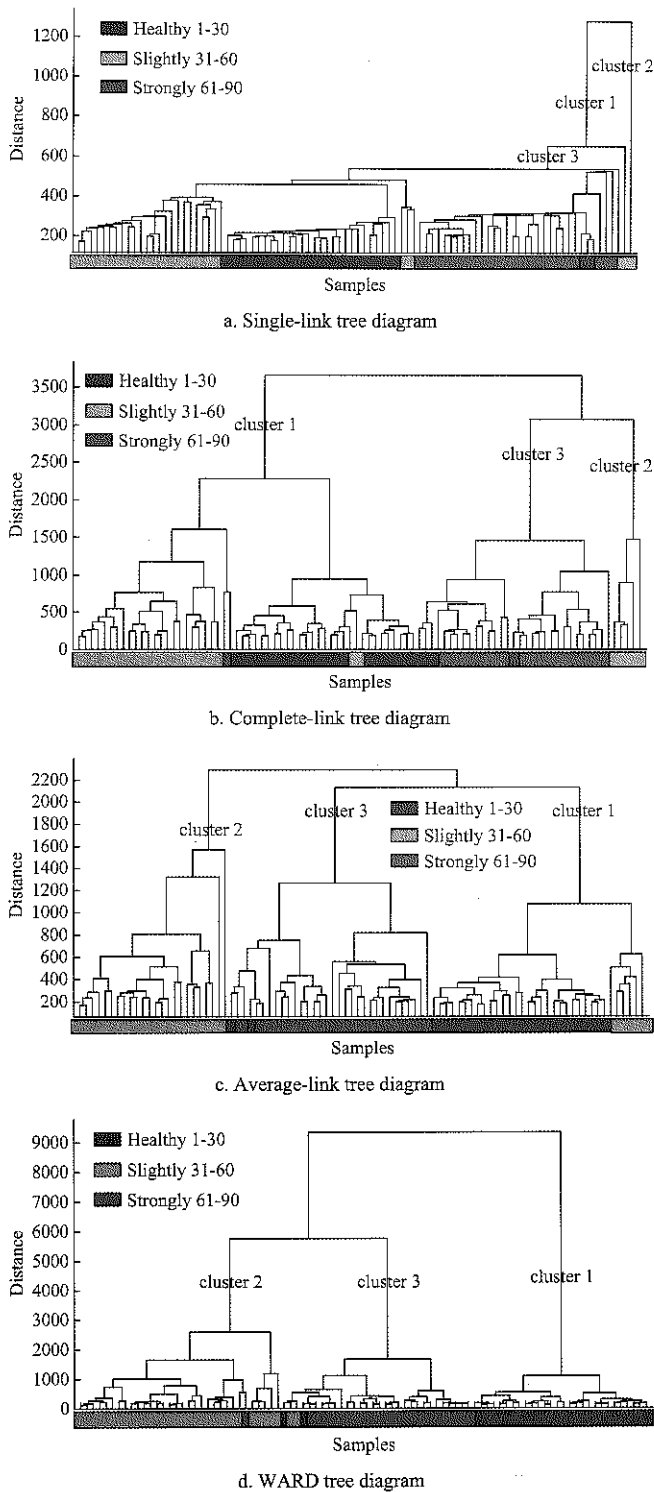


Figure 4 Hierarchical clustering results of broad bean pests' spectra

3.2 Cluster analysis of pest spectra

Because the pest spectral clustering contained high-dimensional data clustering, PCA was used to conduct dimension reduction processing of the spectral data, which can effectively evaluate the effects of hierarchical clustering.

PCA transforms multiple variables into a smaller number of comprehensive factors. Through linear

transformation, PCA can transform the original data into a set of main feature components that are linearly independent in dimension and can be used to extract data and reduce the dimensions of high-dimensional data (Maurizio et al., 2009).

PCA determines the final data dimension by calculating the contribution rate of the feature components. Generally speaking, corresponding values with contribution rates higher than 70% can be used as the data dimension (Jolliffe et al., 2016). We used Matlab 2012b to run the PCA algorithm and applied dimension reduction processing to the original spectral data. Table 2 lists the corresponding relationships between different eigenvectors and dimensions in the spectral matrix. We chose two-dimensional spectral data, and the contribution rate of the eigenvectors was 99.78%.

Table 2 Contribution rate of eigenvectors

Principal component analysis of pest spectra					
Data dimension	1	2	3	4	5
Contribution rate	0.9436	0.9978	0.9989	0.9990	0.9992

Through PCA of the sample spectra, the high-dimensional spectral data were reduced to a lower dimension for expression and analysis. It showed that the distribution diagram of the three samples in two-dimensional space following PCA in Figure 5.

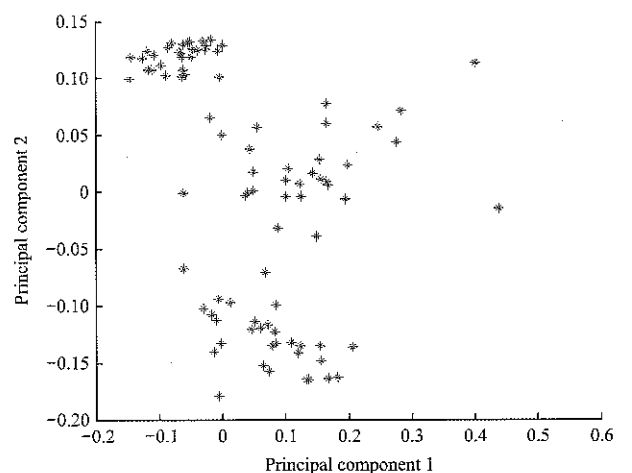


Figure 5 Pest spectral distribution

Using PCA to reduce the dimension of the high-dimensional pest spectral data, we demonstrated the distribution of sample spectra in low-dimensional space. Figure 6 shows the hierarchical clustering results of pests' spectra by using different interclass measurement methods. In the new two-dimensional space, when we

used different interclass measurement methods, the hierarchical clustering results were different. At the

boundary of the three types of samples, the clustering results were clear distinction.

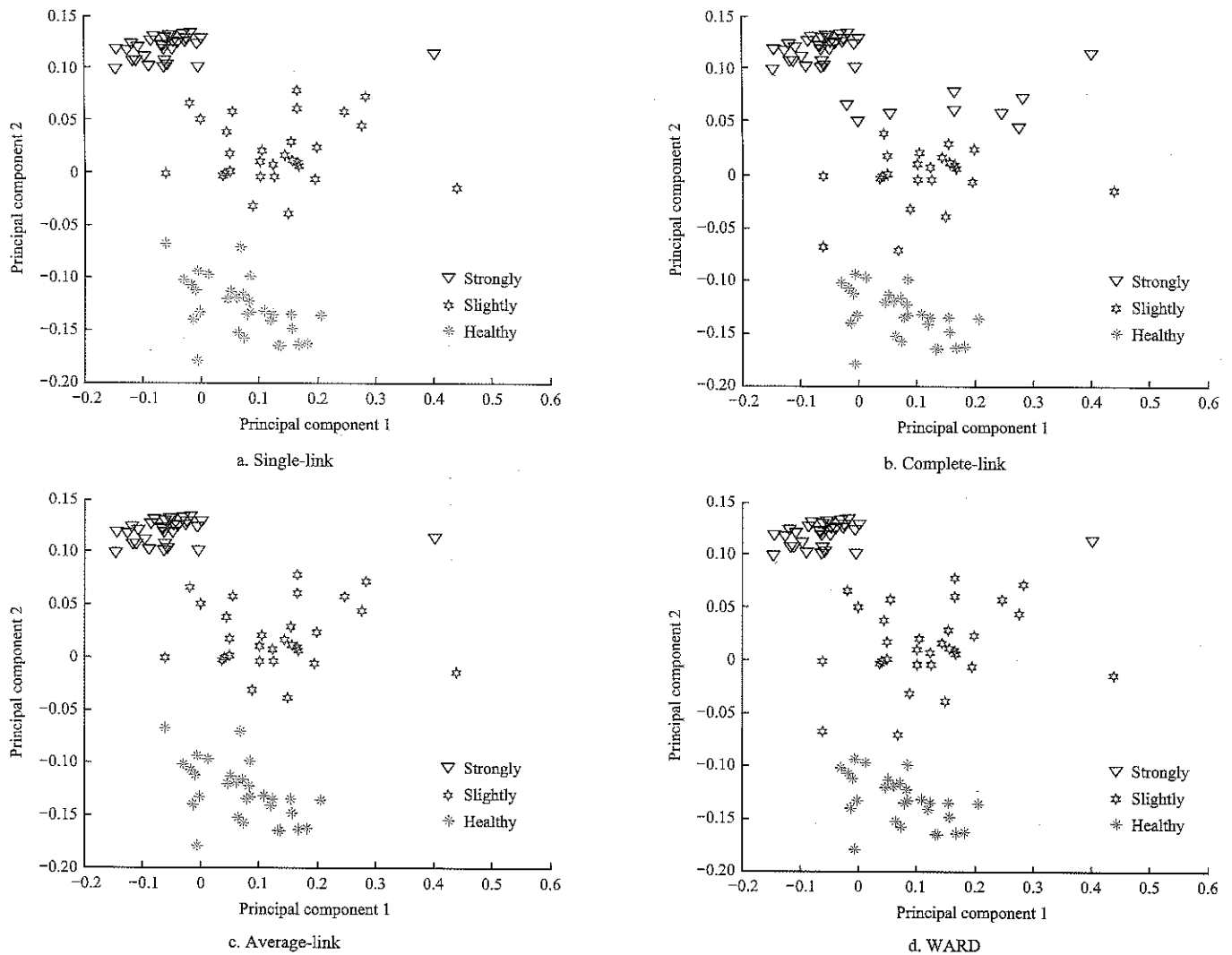


Figure 6 Principal component analysis results for four methods of analyzing pest spectral

Four different hierarchical clustering methods were used to conduct cluster analysis of three types of pest spectra; the statistical results were shown in Table 3.

Table 3 Recognition rate of hierarchical clustering

Model	Recognition rate		
	Healthy	Slightly	Strongly
Samples	30	30	30
Single	32 (93.3%)	29 (96.6%)	29 (96.6%)
Complete	30 (100%)	23 (76.6%)	37 (76.6%)
Average	32 (93.3%)	29 (96.6%)	29 (96.6%)
WARD	30 (100%)	31 (96.6%)	29 (96.6%)

In accordance with the clustering result in the experiment, the clustering effects of the four kinds of similar distance are summarized, in which, the single-link distance's identification rates of the three types of samples are 93.3%, 96.6% and 96.6% respectively; the complete-link distance's identification rates of the three

types of samples are 100%, 76.6% and 76.6% respectively. The identification rates of average-link distance are 93.3%, 96.6% and 96.6% respectively. The identification rates of WARD distance are 100%, and 96.6% and 96.6%, respectively.

It can be seen from the results that the WARD distance is the best for the identification of the pests' sample spectrum, followed by single and average distance, the identification rate of complete distance is the worst.

We executed each of the four hierarchical clustering methods 50 times and calculated the average time required by each method. We used an AMD A8 processor with a frequency of 3.6 GHz in Windows 7 and 8 GB of RAM memory. The time consumed for each clustering method was 0.52 s (Single), 0.61 s (Complete), 0.55 s (Average), and 0.65 s (WARD), respectively.

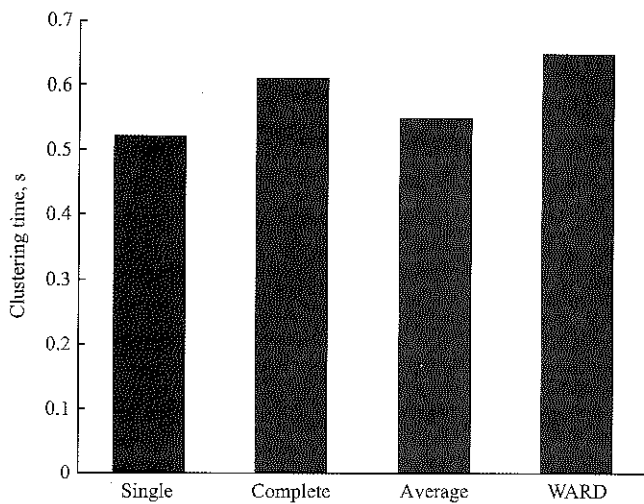


Figure 7 Clustering time consumption

4 Conclusion

The quality of crop spectral acquisition was affected by various factors, such as the light source, acquisition distance, incident light angle, pest density, and integration time of the spectrometer. To improve the accuracy of spectral acquisition, the experiment was conducted multiple times in a darkroom environment; the spectra were acquired from different angles, with the average result used as the experimental data. However, guaranteed accuracy of the acquired spectral data was not possible under these conditions. Consequently, we recommend finding a more reliable method of obtaining crop pest spectral information in a fast and accurate manner.

Some currently used clustering methods, such as K-means and fuzzy C-means, require predefined cluster numbers, and the clustering center is randomly chosen during the process, which tends to yield unstable clustering results. Hierarchical clustering avoids defining the clustering number by using a similarity matrix during clustering yet generates the same result. The direct observation of the sample clustering process and the results obtained by combining the clustering results through tree diagrams and PCA results is beneficial to research on crop spectral analysis.

Because there are many types of crops and pests that have different characteristic reflectance spectra, it is necessary to determine the appropriate cluster analysis method. This study used hierarchical clustering and PCA to analyze crop and pest spectra. In future research,

hierarchical cluster analysis can be widely used to analyze and identify crop disease and pest spectra to improve the detection of diseases and pests for precision agriculture.

Acknowledgements

The research is funded by the National Natural Science Foundation of China (No. 61671244, 61640020). The Agricultural Innovation Program of Jiangsu, China (No. CX(13)3054, CX(14) 2114 and CX(16)1006). The Key Research and Development Program of Jiangsu, (No. BE2016368-1).

[References]

- [1] Anna, O. C., L. E. Rodriguez-Saona, B. A. Mcpherson, D. L. Wood, and P. Bonello. 2014. Identification of *Quercus agrifoli* (coast live oak) resistant to the invasive pathogen *phytophthora ramorum* in native stands using Fourier-transform infrared(FT-IR) spectroscopy. *Frontiers in Plant Science*, 5(521): 1–9.
- [2] Balan, A. V., E. Toma, C. Dobre, and E. Soare. 2015. Organic farming patterns analysis based on clustering methods. *Agriculture & Agricultural Science Procedia*, 6: 639–646.
- [3] Birth, G. S., and K. H. Norris. 1958. An instrument using light transmittance for nondestructive measurement of fruit maturity. *Food Technology*, 12(11): 592–594.
- [4] Cheilane, T., S. A. R. Soares, A. F. S. Queiroz, A. M. P. D. Santos, and S. L. C. Ferreira. 2016. Determination and evaluation of the mineral composition of breadfruit (*Artocarpus altilis*) using multivariate analysis technique. *Microchemical Journal*, 128: 84–88.
- [5] Gebbers, R., and V. I. Adamchuk. 2010. Precision agriculture and food security. *Science*, 327(5967): 828–831.
- [6] Gumienna, M., A. Szwengiel, M. Lasik, K. Szambelan, D. Majchrzycki, J. Adamczyk, J. Nowak, and Z. Czarnecki. 2016. Effect of corn grain variety on the bioethanol production efficiency. *Fuel*, 164: 386–392.
- [7] Jan, B., A. K. Mahlein, T. Rumpf, C. Römer, and L. Plümer. 2015. A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. *Precision Agriculture*, 16(3): 239–260.
- [8] Jin, L., X. B. Li, D. Q. Tian, X. P. Fang, Y. M. Yu, H. Q. Zhu, Y. Y. Ge, G. Y. Ma, W. Y. Wang, W. F. Xiao, and M. Li. 2016. Antioxidant properties and color parameters of herbal teas in China. *Industrial Crops and Products*, 87: 198–209.
- [9] Jolliffe, I. T., and J. Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical*

- Transaction of the Royal Society A-Mathematical Physical and Engineering Sciences*, 374(2065): 1–16.
- [10] Justyna, B. C., G. Małgorzata, and S. Piotr. 2016. Monitoring of essential and heavy metals in green tea from different geographical origins. *Environmental Monitoring and Assessment*, 188(3): 1–11.
- [11] Kavitha, K. A., P. Mintu, and K. Lubna. 2013. Comparative analysis of similarity measures in document clustering. *The 2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE)*, 857–860. Beijing, China, 20-23 August.
- [12] Li, L. B., and H. Y. Deng. 2010. Research of similarity measurements in the clustering analysis. *The Second International Conference on Information Technology and Computer Science*, 3-6. Kiev, Ukraine., 24-25 July.
- [13] Li, Z. Y., G. Liu, L. Li, Q. H. Ou, and X. Zhao. 2012. FTIR spectroscopic study of broad bean diseased leaves. *Spectroscopy and Spectral Analysis*, 13(11): 1217–1220.
- [14] Lin, L. B., H. N. Li, P. F. Cao, F. Qin, and S. Yang. 2015. The characteristic analysis of spectral image for cabbage leaves damaged by diamondback moth pests. *The International Conference on Photonics and Optical Engineering (icPOE 2015)*, 9449, 1–6. Xi'an, China, 13-15 Oct.
- [15] Luo, J. W., K. Ying, and J. Bai. 2005. Savitzky–Golay smoothing and differentiation filter for even number data. *Signal Processing*, 85(7): 1429–1434.
- [16] Maurizio, V., and S. Gilbert. 2009. Clustering and disjoint principal component analysis. *Computational Statistics and Data Analysis*, 53(8): 3194–3208.
- [17] Mratinic, E., F. A. Milica, and R. Jovkovic. 2012. Analysis of wild sweet cherry (*Prunus avium* L.) germplasm diversity in south-east Serbia. *Genetika*, 44(2): 259–268.
- [18] Naresh, K., N. Ram, M. L. Khichar, R. K. Saini, and B. Biswas. 2013. Effect of different growing environments on population dynamics of sucking pests in relation to various spectral indices in cotton. *Journal of the Indian Society of Remote Sensing*, 41(2): 309–317.
- [19] Niewietzki, O., P. Tillmann, H. C. Becker, and C. Möllers. 2010. A new Near-Infrared reflectance spectroscopy method for high-throughput analysis of oleic acid and linolenic acid content of single seeds in oilseed rape (*Brassica napus* L.). *Journal of Agricultural and Food Chemistry*, 58(1): 94–100.
- [20] Norris, K. H., and J. Rowan. 1957. Use of the automatic green rot detector to improve the quality of liquid. *Food Technology*, 11: 374–377.
- [21] Praislser, M., S. C. Ghinita, A. Stoica, and L. Dumitriu. 2015. Hierarchical Cluster Analysis: a reliable tool allowing more detailed (regional) traceability investigations. *19th International Conference on System Theory, Control and Computing (ICSTCC)*, 157–161. Cheile Gradistei, Romania., 14-16 October.
- [22] Ruffin, C., and R. L. King. 1999. The analysis of hyperspectral data using Savitzky-Golay Filtering-Theoretical Basis (Part1). *IEEE 1999 International Geoscience and Remote Sensing Symposium*, 2: 756–758.
- [23] Chao, W. L., H. H. Su, S. Y. Chien, W. Hsu, and J. J. Ding. 2011. Visible-near infrared spectroscopy for detection of Huanglongbing in citrus orchards. *Computers and Electronics in Agriculture*, 77(2): 127–134.
- [24] Turton, B. C. H. 1992. Novel variant of the Savitzky-Golay filter for spectroscopic applications. *Measurement Science & Technology*, 3(9): 858–863.
- [25] Zhang, R. X., J. L. Ba, Y. Ma, S. Q. Wang, J. Zhang, and W. D. Li. 2012. A comparative study on wheat leaf area index by different measurement methods. *First International Conference on Agro-Geoinformatics*. 1-5. Shanghai, China, 2-4 Aug.
- [26] Zong, J. X., Y. W. Yang, L. Wang, W. Zhao, and M. He. 2014. Detection of vegetable pests identification based on spectrum. *International Agricultural Engineering Journal*, 1(23): 47–55.