A NEW SOYBEAN NDVI DATA-BASED PARTITIONING ALGORITHM FOR FERTILIZATION MANAGEMENT ZONING

CHEN, H. – WANG, X.* – ZHANG, W. – WANG, X. Z. – DI, X. D. – QI, L. Q.

College of Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, China (phone: +86-4596-819-216)

**Corresponding author*
e-mail: 7455663@qq.com; phone: +86-138-3696-1962

(Received 17th Nov 2020; accepted 8th Feb 2021)

Abstract. With the broad application of management zoning, various partitioning algorithms are used for this purpose. The K-means algorithm is the most widely used method with the best performance. We proposed a model-based partitioning algorithm based on the K-means algorithm that does not need to process all the data in each computational iteration and thus is able to improve the management zoning speed. We constructed a calculation model for management zone partitioning using 2000 normalized difference vegetation index (NDVI) values. This model was then used to partition the next 2000 points of NDVI data acquired, and the model was updated after each 2000 NDVI values until the management zoning computation was completed for the entire field. Based on both internal evaluation indicators (the sum of squared errors (SSE) and silhouette coefficient (SC)) and external evaluation indicators (the Rand index (RI) and homogeneity), we compared the clustering performance of the two algorithms in management zone partitioning and found that when the amount of NDVI data reaches 8000 values, the proposed method achieves a management zone partition result similar to that of the conventional K-means algorithm but is faster. This advantage becomes increasingly profound as the data volume increases.
Keywords: *K-means, model-based algorithm, SSE, silhouette coefficient, Rand index, homogeneity*

Introduction

Heilongjiang Province is a major production area of high-quality soybeans in China. With the expansion of soybean acreage, the amount of chemical fertilizer applications has increased, which has in turn raised operating costs and decreased the quality of cultivated land. A large quantity of unutilized fertilizer can cause soil compaction in the field and damage soil microbial communities, thus severely hindering the sustainable development of agriculture. With the continuous development of precision agriculture, management zoning has been extensively investigated. Management zoning refers to the division of cultivated land into different management units based on the differentiation of various features, such as soil properties, the growth of crops, the crop yield, and terrain, so that the features of interest are similar within each unit but differ markedly between them. Through management zoning, differentiated fertilization schemes can be applied, which helps improve fertilizer use efficiency, protect the environment and quality of arable land and achieve the sustainable use of agricultural resources (Lopes et al., 2012; Koenig et al., 2015). Management zoning is an important foundation for the development of precision agriculture and has been thoroughly studied. Currently, the data sources for management zoning mainly include soil fertility, soil conductivity, soil texture, terrain, the crop yield, and crop canopy spectral information (Johnson et al., 2003; Flowers et al., 2005). Zoning based on soil fertility and terrain has been the most common practice, with relatively high accuracy. However, such approaches are time consuming and labor intensive, and the timing tends to be poor. Additionally, when

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1902_13911405

making measurements, it is difficult to determine the appropriate spacing, making it impossible to use these methods to zone large areas. In addition, such methods are associated with a high risk of machine-derived damage to the crop in the measurement process. Yield-based zoning methods cannot be performed in real time, as yield data are often not readily available. Satellite remote sensing-based zoning methods can be simultaneously performed in real time over a large area and rapidly generate high-resolution images and data. However, remote sensing data is geometrically and radiometrically distorted, with poor periodicity, stability, and repeatability; thus, complex data processing is needed, and the cost of generating the images is high.

The reflections of light of specific wavelength bands from plants in different growth states differ significantly. From these reflections, crop canopy spectral information can be collected through active remote sensing to estimate the growth status of crops. This technique does not rely on external light sources and can be applied at any time to rapidly acquire data in a timely manner. The normalized difference vegetation index (NDVI) data obtained through spectral detection techniques can be used to help evaluate crop growth, nutritional status, potential yield, and the impacts of pests and diseases (Jiang et al., 2006). The NDVI data that reflect the crop growth status facilitate the zoned management of crops, so that differentiated fertilization can be applied according to the growth of the crops.

Studies in the field of precision agriculture have focused on the use of management zoning to conduct differentiated fertilization management. NDVI-based management zoning is essentially a clustering process involving NDVI data. Clustering, an important part of data mining and machine learning, is the division of datasets into multiple clusters so that the data in the same cluster are similar and the data in different clusters are different. In the case of management zoning, each cluster represents a management zone. Unlike classification, clustering is a typical unsupervised learning method that does not require the labeling of data in advance (Barlow, 1989). There are many algorithms for data clustering, such as the K-means (MacQueen, 1967), hierarchical clustering (Johnson, 1967), DBSCAN (Ester et al., 1996) and fuzzy C-means (Bezdek et al., 1984). Among them, the K-means algorithm is one of the most well-known, simplest, and best clustering algorithms (Saxena et al., 2017).

Liu and Wang (2019) applied the K-means clustering algorithm for the management zoning of corn and concluded that it was the best algorithm for the NDVI data collected by GreenSeeker. However, during each clustering iteration in the K-means algorithm, all the data have to be processed to update the centroid (mean) of the data of each class. Therefore, as the data volume increases, the computational speed declines. The current clustering algorithms are faster than the previous algorithms for management zoning applications, but with the growing data volume to be processed, clustering computations face potential limitations. NDVI data are classified as a data stream, which represents an "unbounded" sequence of observations (Lu, 2005), i.e., an open time series of data with an ever-increasing data volume. Such data require a fast clustering algorithm dedicated to processing the incoming data stream.

In this study, we aim to propose a model-based partitioning algorithm for soybean fertilization management zoning that does not perform computations on all NDVI data but updates the model with data at certain intervals. Further, we aim to use the model to process the subsequent data to improve the management zoning speed.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1902_13911405

Materials and methods

Experimental site

The experimental site (*Fig. 1*) was located at Zhaoguang Farm, a state-owned mechanized farm affiliated with the Land Reclamation Bureau of Heilongjiang Province, China. Standardized production operations, mostly with large-scale agricultural equipment, were used in the experiment. Zhaoguang Farm (126°26'-127°6' E, 47°54'-48°12' N) is located in the center of Bei'an city, to the east of Kedong County in Heilongjiang Province and at the edge of the southern foot of the Xiaoxing'an Mountains. The area is located at moderate to high latitudes and has a cold-temperature monsoon climate. The average annual temperature is 0.5 °C, the frost-free period is 120 days, annual rainfall totals 570 mm, and the average annual sunshine duration is over 2700 hours (Liu and Wang, 2019). The overall climate is characterized by dry and windy springs, hot and rainy summers, quickly cooling autumns with early frost, and long and cold winters. The region has four types of soil, namely, brown soil, black soil, meadow soil and bog soil, of which black soil covers 57.4% of the total area of arable land. This experiment was conducted at the 12-3 plot of Station 17 of the 4th Management District of Zhaoguang Farm, which is 34 ha in area. The soil type in this area is black soil. The area is planted with soybeans, with two lines per ridge (1.1 m in width). The soybean variety is Heihe 43, the amount of fertilizer applied is 105 kg/ha, and the ratio of N and P in fertilizer is 1:1.3. Plot 12-3 was unobstructed, thus supporting the accuracy requirements of the received GPS signals. The data were acquired on Jun 22, 2019.

Figure 1. *Experimental site (Zhaoguang Farm, China)*

Data collection and processing

Soybean canopy NDVI data were acquired with a ground-based remote sensing detection platform consisting of six GreenSeeker plant canopy spectrum analyzers (Trimble, U.S.), a GPS receiver (Model AG332, Trimble, U.S.) and a controller area network (CAN) data logger. The soybean canopy NDVI data were acquired with GreenSeeker plant canopy spectrum analyzers, each of which was equipped with narrow-band red light (660 nm) and near-infrared light (770 nm) LEDs as active light sources, and next-generation optical sensors from NTech were used to acquire plant

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1902_13911405

canopy reflectance spectra at the two aforementioned bands, with a measurement area of 61 cm±10 cm (width) × 1.5 cm±0.5 cm (length). During the measurement process, the probe of the spectrometer sensor was directed vertically downward and maintained at approximately 80 cm above the canopy. GreenSeeker is currently the most widely used agricultural spectrum sensor. For example, John Deere's GreenStar high-clearance adjustable sprayer uses GreenSeeker to control the application of liquid fertilizer so that smart, real-time and adjustable fertilization operations can be performed. The GreenSeeker sensor does not rely on sunlight and can work around the clock because it is unaffected by atmospheric and soil reflections. Additionally, the sensor provides a fast measurement speed and accurate data and can estimate the plant growth status in real time. Therefore, the sensor overcomes the shortcomings of poor timeliness, damage to crops and a high workload for large-scale measurements; moreover, unlike passive hyperspectral remote sensing, it does not require an external light source (Jiang et al., 2019). Applications of GreenSeeker in nutritional diagnosis, yield prediction and management zoning for many crops, such as wheat and corn, have been extensively investigated. For example, Sharp et al. (2004) demonstrated that it is feasible to determine the NDVI for the management zoning of cotton through GreenSeeker data. *Fig. 2* shows the ground-based remote sensing detection platform used for the acquisition of soybean canopy NDVI data and the corresponding control box used in this study.

(a) (b)

Figure 2. *Ground-based remote sensing detection platform for the acquisition of soybean canopy NDVI data. (a) Data acquisition site; (b) Interior of the remote sensing platform control box*

The Trimble GPS receiver (Model AG332) uses the ultimate choice technique and is the most advanced high-performance dual-frequency receiver available; it provides centimeter-level measurement accuracy and has a position information updating frequency of 1 Hz. The data acquisition frequency of GreenSeeker was also 1 Hz in this study. The acquired data were analyzed and recorded through the CAN data logger. *Table 1* shows the details of some representative data collected during the test.

In this study, we used Python and ArcGIS for data processing. We compiled the model-based partitioning algorithm program using the Scikit-learn library in Python. Scikit-learn is a machine learning tool library based on Python that includes many simple and efficient data mining and analysis tools. We used ArcGIS, which is an excellent geographic information processing software, to plot the fertilization management zones.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1902_13911405

Table 1. *Details of the acquired data (partial)*

Date	Time	Longitude	Latitude	Mean NDVI
2019-6-22	12:36:31	126.6747136	48.0562793	0.650
2019-6-22	12:36:32	126.6747353	48.0562942	0.667
2019-6-22	12:36:33	126.6747584	48.0563087	0.650
2019-6-22	12:36:34	126.6747819	48.0563236	0.637
2019-6-22	12:36:35	126.6748047	48.0563384	0.658
2019-6-22	12:36:36	126.6748265	48.0563530	0.682
2019-6-22	12:36:37	126.6748489	48.0563684	0.706
2019-6-22	12:36:38	126.6748721	48.0563838	0.674
2019-6-22	12:36:39	126.6748942	48.0563986	0.659
2019-6-22	12:36:40	126.6749157	48.0564132	0.647

Fertilization management zoning method

K-means algorithm

The K-means algorithm, which was proposed by MacQueen in 1967, is a commonly used and effective clustering algorithm classified as an unsupervised machine learning method. In the past three decades, this algorithm has been used in most clustering algorithm-related studies. Later, many new clustering algorithms were developed or modified based on this algorithm (Fahad et al., 2014). During implementation, the K-means algorithm divides the data objects into K (defined by the user) clusters, each of which has a center (also called the centroid or mean) to which the distance of each of the data objects is calculated. The data object with the shortest distance to the center of a certain cluster is then assigned to that cluster. The clustering process is iterative, and the Euclidean distances between the data points and cluster centers are continuously optimized. For the set of data points X= [x_1, ⋯, x_N], the K-means algorithm presents k divisions of dataset X. If [C_1, ⋯, C_K] represents the centers of the K clusters, then we have the following function (*Eq. 1*):

$$V = \sum_{i=1}^{K} \sum_{x_j \in Xi} \left\| x_j - C_i \right\|^2 \quad \text{(Eq.1)}$$

where K is the number of partitions, X_i is a data point in the i^{th} partition, x_j is a data point in the data point set, and C_i is the cluster center of the i^{th} partition.

The K-means algorithm is a method that constantly searches for the minimum value of the above function (Huang et al., 2006; Jing et al., 2007).

Determination of the number of clusters K using the elbow method

The K-means algorithm minimizes the squared error between a data point and data center, and this minimization process is encompassed in the objective function. The sum of squared errors (SSE) between the center of each cluster and the data points in the cluster is called the distortion. In a cluster, the lower the distortion is, the closer the cluster members are to each other; the higher the distortion is, the looser the cluster structure. As the number of classes increases, the distortion decreases, but for data with a certain degree of distinction, the distortion is greatly reduced when a certain critical point is reached. Then, the decrease becomes gradual. The critical point can be

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1902_13911405

considered a point where good clustering performance is displayed. Based on the above-described principle, the elbow method calculates the mean distortion to determine the optimum number of clusters (K). The mean distortion is calculated using the following formula (*Eq.2*):

$$MD = \frac{\sum_{p \in C_i} |p - m_i|^2}{n} \quad \text{(Eq.2)}$$

where *MD* is the mean distortion, *Ci* is the dataset of the ith partition, p represents a certain piece of data in the ith partition, m_i is the center of the C_i partition, and n is the data volume of the ith partition.

Model-based algorithm

The proposed algorithm is based on the K-means clustering algorithm. The number of clusters (K) is set by the user, and the centers of clusters are randomly set by the algorithm. Then, the distances between the data points and cluster center are calculated, and the data points are clustered based on the minimum distance. The calculation is iteratively performed to update cluster centers, and when the accuracy requirements are met and 2000 data points are processed, a clustering model is constructed based on the cluster centers. Subsequently, without updating the cluster centers, the established clustering model is used to perform the clustering calculations, and when 2000 data points have been processed, the clustering model is updated. Similarly, in data intervals of 2000 points, the clustering model is updated until all the data are clustered. The flow chart of the proposed algorithm is shown in *Fig. 3*.

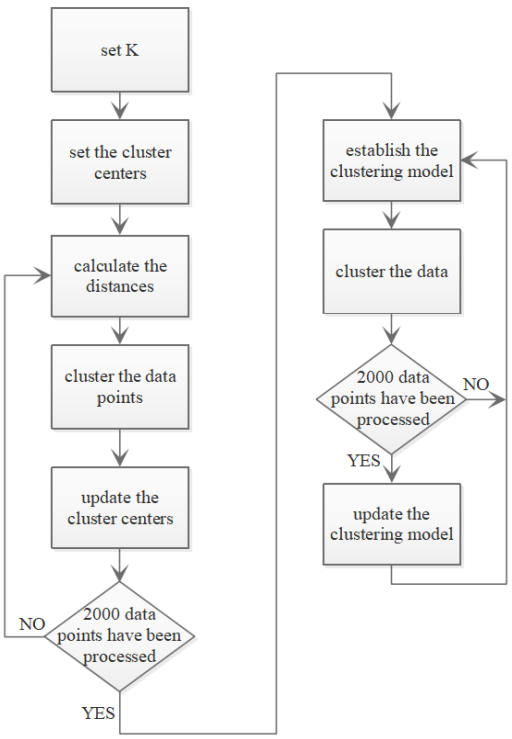


Figure 3. *Flow chart of model-based algorithm*

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1902_13911405

Evaluation of clustering performance

Two kinds of evaluation indicators are commonly used to assess clustering performance: internal evaluation indicators and external evaluation indicators. Internal evaluation indicators are able to assess clustering performance without comparing the clustering result with the ideal result, and external evaluation indicators have to rely on comparing the clustering result with the ideal clustering result (Amigó et al., 2009). To avoid the limitations of a single evaluation method and obtain accurate results, we use both internal and external evaluation indicators. Notably, the clustering result obtained through the K-means algorithm is used as the ideal clustering result to assess clustering performance with external indicators.

The SSE is used as an indicator for the evaluation of unsupervised clustering results. This parameter can be used to assess clustering performance without a comparison database (ideal clustering result). The purpose of the K-means algorithm is to find a solution that divides a dataset into K clusters, where the data in the same cluster are similar, the data from different clusters are different and each data point has a minimized SSE relative to the corresponding cluster center. The SSE is a common indicator of clustering quality. For the same number of clusters (K), the lower the SSE is, the higher the clustering quality. In terms of principles, the SSE evaluation method is similar to the elbow method. The calculation formula of the SSE is given as *Eq. 3*.

$$SSE = \sum_{i=1}^{K} \sum_{p \in C_i} \left| p - m_i \right|^2 \quad \text{(Eq.3)}$$

where *SSE* is the sum of squared errors, *K* is the number of partitions, C_i is the dataset of the i^{th} partition, *p* represents certain data points in the partition, and m_i is the center of partition C_i.

The silhouette coefficient (SC) (Rousseeuw, 1987), which was proposed by Rousseeuw in 1986, is an internal evaluation indicator for unsupervised clustering based on two factors: cohesion and separation. This variable can be used to evaluate different clustering algorithms based on a given dataset and the effects of different methods of operation of the same clustering algorithm on the partitioning result.

In the samples, the average distance from each data point (i) to the other samples in the same cluster is defined as a(i), and accordingly, the average distance from each data point (i) to the other samples in cluster C_j is defined as B_{ij}. Therefore, b(i)= min[B_{i1}, B_{i2}, ..., B_{ik}]. The lower the value of a(i) is, the stronger the association of the sample data point (i) with the corresponding cluster is, so a(i) is called the intracluster dissimilarity for sample data point (i). The higher the value of b(i) is, the weaker the association of sample data point (i) with other clusters is, so b(i) is called the intercluster dissimilarity for sample data point (i). The SC-based partitioning algorithm was proposed to estimate the number of clusters (Azimi et al., 2017; Ünlü and Xanthopoulos, 2019). The definitions of a(i), b(i), and SC can be found in the following formula (*Eq. 4*):

$$SC(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \quad \text{(Eq.4)}$$

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1902_13911405

In the equation, -1 ≤ SC (i) ≤ 1; a(i) represents the compactness of the cluster to which (i) belongs, and the lower the a(i) value is, the more compact the points are in the cluster; and b(i) represents the degree of separation of (i) and the other clusters, and the higher the b(i) value is, the more separated (i) is from the other clusters (Han et al., 2011) However, when SC (i) < 0, (i) is closer to the samples in other clusters than to those in the same cluster. This situation is not ideal in actual clustering and should be avoided.

The clustering process can be regarded as a series of decision-making processes in which it is determined whether each pair of data points in the dataset should be assigned to the same cluster. In this process, the Rand index (RI) (Hubert and Arabie, 1985; Vinh et al., 2010) is used to measure the accuracy of decisions. Assuming that U is the ideal cluster set and that V is the clustering result, four statistical measures are established as follows: a is the number of datapoint pairs that fall in the same cluster in both U and V; b is the number of datapoint pairs that fall in the same cluster in U but not in V; c is the number of datapoint pairs that fall in the same cluster in V but not in U; and d is the number of datapoint pairs that do not fall in the same cluster in U and V. Then, RI is defined according to *Eq. 5*.

$$RI = \frac{a+d}{a+b+c+d} \quad \text{(Eq.5)}$$

The *RI* values are in the range of [0,1]. When the clustering result perfectly matches the ideal clustering result, *RI* is 1. However, the *RI* value cannot guarantee a value of 0 in the case of random clustering, so the adjusted Rand index (ARI) (Yeung and Ruzzo, 2001; Steinley, 2004) is introduced to address this issue and is defined in *Eq. 6*.

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \quad \text{(Eq.6)}$$

where *ARI* is the adjusted Rand index, *RI* is the Rand index, *E*[*RI*] is the expected value of the *RI*, and max(*RI*) is the maximum value of the *RI*.

Homogeneity is an external evaluation indicator used to assess clustering performance based on conditional entropy analysis. Specifically, data from the same cluster should be from the same class (Hirschberg and Rosenberg, 2007). This parameter is defined in *Eq. 7*.

$$H = 1 - \frac{\sum_{c,k} n_{c,k} \log \frac{n_{c,k}}{n_k}}{\sum_c n_c \log \frac{n_c}{D}} \quad \text{(Eq.7)}$$

where *H* is the homogeneity, n_c is the number of data points in class *c*, n_k is the number of data points in partition *k*, $n_{c,k}$ is the number of data points present in class c and cluster *K* at the same time, and *D* is the number of data points in the dataset.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1902_13911405

Results and discussion

Calculation of the number of management zones K

Using the elbow method, we calculate the mean distortion and plot the changes. As shown in *Fig. 4*, for the soybean NDVI data, when the number of partitions K is greater than 2, the mean distortion does not decrease significantly and is characterized by a smooth curve, indicating that a number greater than 2 is reasonable for the number of management zones based on the acquired soybean NDVI data. As the number of management zones increases, the requirements for working machinery that can provide variable-rate fertilization increase, which inevitably increases the related costs; thus, many methods and machines are impractical in actual agricultural production activities. Therefore, it is necessary to choose a suitable number of management zones with the optimal zoning accuracy and best comprehensive economic benefits. In this study, we determined that the number of fertilization management zones should be 4 based on the fertilization site conditions and experience.

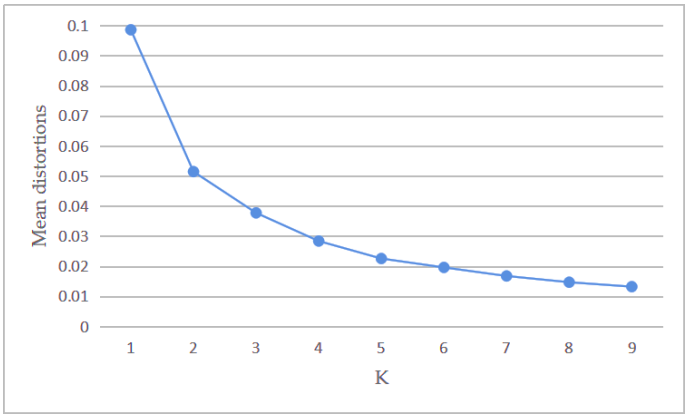


Figure 4. *Mean distortions*

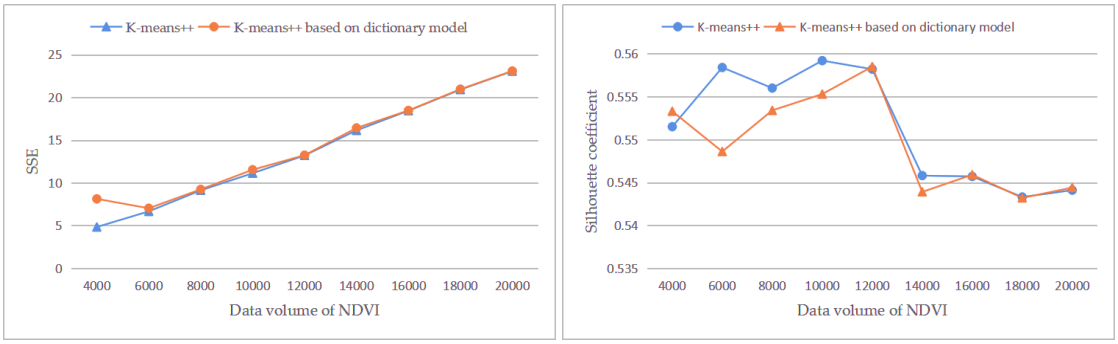
Comparison of the K-means algorithm and model-based algorithm

We compiled the K-means algorithm and the model-based algorithm programs through the Scikit-learn library and obtained the management zoning results from the two algorithms. We also compiled the program to evaluate the clustering performance of the two algorithms, and based on the results of the evaluation of clustering performance, we plotted the changes in the internal evaluation indicators (SSE and SC) for different data volumes, as shown in *Fig. 5*.

The SSE is a commonly used internal evaluation index of the clustering effect and reflects the degree of aggregation of the data points in a cluster to the clustering centroid and thus the clustering performance, as demonstrated in previous studies (Chayangkoon and Srivihok, 2016; Zhang and Zhou, 2018). A comparison of the SSE values of the K-means algorithm and the model-based algorithm shows that when 4000 data points are processed, the SSE values of the K-means algorithm and the model-based algorithm are 4.83 and 8.12, respectively, indicating that the K-means algorithm outperforms the model-based algorithm. When the volume of processed data reaches 6000 data points, the SSE values of the K-means algorithm and model-based algorithm are 6.65 and 7.02, respectively, indicating that the two algorithms perform similarly. When the volume of

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1902_13911405

processed data surpasses 6000 data points, the SSE value of the model-based algorithm using partial data is as low as that of the K-means clustering algorithm using all the data; i.e., the model-based algorithm, using less data, can perform as well as the K-means algorithm using more data.

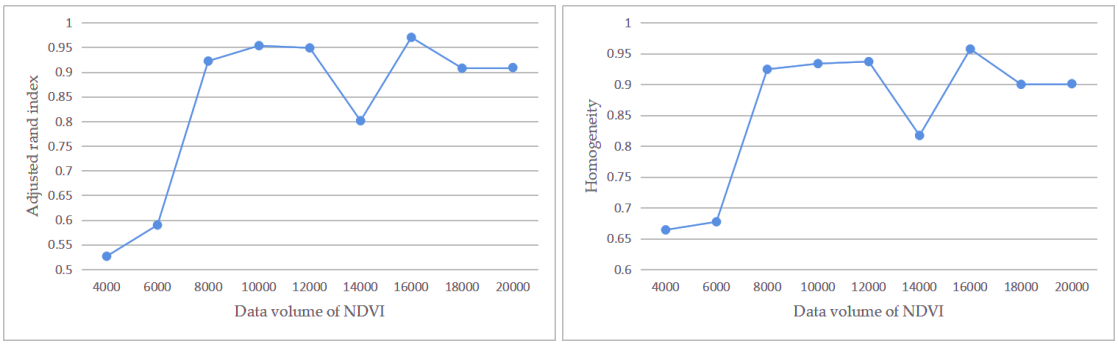


(a) (b)

Figure 5. *Changes in the internal evaluation indicators with the data volume. (a) Comparison of the SSE values of the two algorithms; (b) Comparison of the SC values of the two algorithms*

The SC is another commonly used internal evaluation indicator of clustering performance, and it reflects the degree of closeness of data points to the center of the cluster to which they belong and the degree of remoteness of data points to centers of other clusters. To obtain good clustering results, the points within a cluster should be close to each other but distant from the data points in other clusters. In previous studies (Hasanzadeh-Mofrad and Rezvanian, 2018; Dinh et al., 2019), the SC value was used to assess the quality of different clustering methods. A comparison of the SC values of the K-means algorithm and the model-based algorithm shows that when the volume of processed data reaches 8000 data points, the SC values of the two algorithms are very similar. This result is similar to that for the SSE, indicating that the clustering performance of the model-based algorithm using partial data is similar to that of the K-means algorithm using all the data.

A comparison of the changes in the external evaluation indicators (ARI and homogeneity) of the clustering performance with the data volume is shown in *Fig. 6*.



(a) (b)

Figure 6. *Changes in the external evaluation indicators with the data volume. (a) ARI changes; (b) Homogeneity changes*

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1902_13911405

The ARI is an improved external evaluation indicator of the RI indicator for clustering performance, and it reflects the correctness of the clustering result through the similarity of two different clustering results. Previously, the ARI evaluation method was used to assess the stability of five clustering methods in evaluating the clustering of single-cell RNA-seq data, and good results were achieved (Duò et al., 2018). In this study, in calculating the ARI, we treat the clustering result of the K-means algorithm as the ideal result for the model-based algorithm and find that when the data volume reaches 8000 data points, the ARI value is 0.92 and stabilizes at approximately 0.95 as the volume of data increases. Thus, after the data volume reaches 8000 data points, the clustering performance results of the two algorithms are similar. When the data volume reaches 14,000 data points, the ARI decreases to 0.80, indicating that the accuracy of the clustering model is unstable and fluctuates, which will be examined further in future studies.

In previous studies (SanJuan and Ibekwe-SanJuan, 2006; Mohammadrezapour et al., 2020), homogeneity was used to evaluate the clustering performance and proven to be effective. In this study, we also use the homogeneity indicator to evaluate the clustering results and find that when the data volume reaches 8000 data points, the homogeneity value is 0.93. Thus, when the data volume reaches 8000 data points, high clustering accuracy is achieved. When the data volumes are 4000 and 6000 data points, the homogeneity values are 0.66 and 0.67, respectively. Therefore, for these data volumes, the clustering accuracy is not high. When the data volume reaches 14,000 data points, the homogeneity value decreases to 0.82; this trend is similar to that for the ARI, indicating that the model accuracy is not stable.

We used ArcGIS to plot the fertilization management zones. We interpolated the NDVI data for crop growth based on geographical information (latitude and longitude) and generated fertilization management zones based on the above-described clustering results for the soybean NDVI data. Then, we plotted the results, as shown in *Fig. 7*.

A comparison of the management zones obtained using the K-means algorithm and the model-based algorithm shows that for a data volume of 4000 data points, the two zoning methods differ significantly, and the SSE values of the two methods differ significantly. These findings are consistent with the results of actual management zoning. The difference in SC is slight. The values of the two external evaluation indicators, RI and homogeneity, are 0.53 and 0.67, respectively, indicating that the results based on the two methods differ significantly and that the zoning accuracy of the model-based algorithm is poor. When the data volume reaches 6000 data points, the two methods differ slightly in terms of management zoning, with slight differences in the internal evaluation indicators (SSE and SC). Additionally, the values of the two external evaluation indicators, RI and homogeneity, are 0.59 and 0.68, respectively. When the data volume reaches 8000 data points or more, the two methods show no difference in terms of management zoning, with minimal differences in the internal evaluation indicators (SSE and SC). In this case, the values of the two external evaluation indicators, RI and homogeneity, are 0.92 and 0.93, respectively. These results indicate that the zoning results based on the K-means algorithm and the model-based algorithm show minimal differences and that the model-based algorithm achieves a high accuracy when the result of the K-means algorithm is used as the ideal clustering result.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1902_13911405

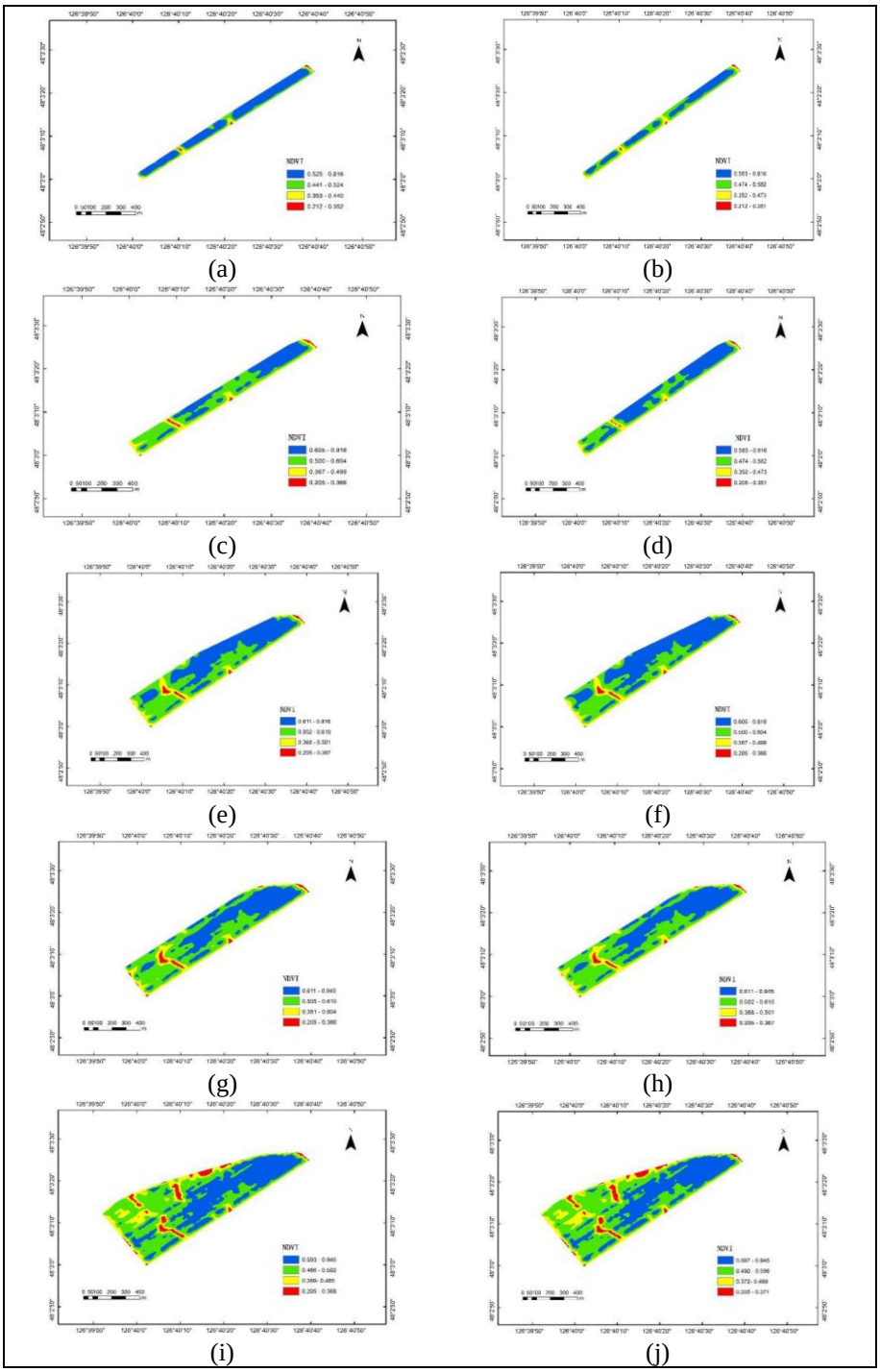


Figure 7. *Comparison of the management zones obtained using the K-means algorithm and the model-based algorithm. (a) K-means algorithm when the data volume reaches 4000 data points; (b) model-based algorithm when the data volume reaches 4000 data points; (c) K-means algorithm when the data volume reaches 6000 data points; (d) model-based algorithm when the data volume reaches 6000 data points; (e) K-means algorithm when the data volume reaches 8000 data points; (f) model-based algorithm when the data volume reaches 8000 data points; (g) K-means algorithm when the data volume reaches 10,000 data points; (h) model-based algorithm when the data volume reaches 10,000 data points; (i) K-means algorithm when the data volume reaches 18,000 data points; (j) model-based algorithm when the data volume reaches 18,000 data points*

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1902_13911405

Conclusions

In this study, we investigated partitioning algorithms for soybean fertilization management zoning by acquiring soybean NDVI data through a ground-based remote sensing detection platform and proposed a model-based algorithm based on the K-means clustering algorithm. With NDVI data acquired from an entire soybean field as the partitioning data sample and the K-means and model-based algorithms compiled in Python, we comparatively analyzed the produced fertilization management zoning charts and the corresponding internal and external evaluation indicators. When the data volume reached 8000 data points, the model-based algorithm produced SSE and SC values similar to those of the K-means algorithm. Additionally, the ARI and homogeneity values were high (approximately 0.95 for both) when the clustering result of the K-means clustering algorithm was used as the ideal clustering result, indicating that the model-based algorithm differs from the traditional K-means algorithm when the volume of sample data is below 8000 data points and does not differ from the traditional K-means algorithm when the volume of sample data is 8000 data points or more but is faster. This advantage becomes increasingly profound as the data volume increases.

Subsequent research will further improve the model-based algorithm, reduce computation time, improve the operational efficiency, establish a reasonable fertilization model based on the NDVI, calculate the recommended amount of fertilizer, and compare the difference in the recommended amount of fertilizer between the model-based algorithm and other algorithms.

Acknowledgements. This research was funded by the National Soybean Industry Technology System (No. CARS-04-PS32), the Heilongjiang Bayi Agricultural University Support Program for San Heng San Zong (ZRCPY202014), and the Heilongjiang Farms & Land Reclamation Administration Support Project (2019HKQNJTG0015).

REFERENCES

[1] Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F. (2009): A comparison of extrinsic clustering evaluation metrics based on formal constraints. – Information retrieval 12: 461-486.

[2] Azimi, R., Ghayekhloo, M., Ghofrani, M., Sajedi, H. (2017): A novel clustering algorithm based on data transformation approaches. – Expert Systems with Applications 76: 59-70.

[3] Barlow, H. B. (1989): Unsupervised learning. – Neural computation 1: 295-311.

[4] Bezdek, J. C., Ehrlich, R., Full, W. (1984): FCM: The fuzzy c-means clustering algorithm. – Computers 10: 191-203.

[5] Chayangkoon, N., Srivihok, A. (2016): Two step clustering model for K-means algorithm. – Proceedings of the Fifth International Conference on Network, Communication and Computing, pp. 213-217.

[6] Dinh, D.-T., Fujinami, T., Huynh, V.-N. (2019): Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. – International Symposium on Knowledge and Systems Sciences, Springer, pp. 1-17.

[7] Duò, A., Robinson, M. D., Soneson, C. (2018): A systematic performance evaluation of clustering methods for single-cell RNA-seq data. – FResearch 7.

[8] Ester, M., Kriegel, H.-P., Sander, J., Xu, X. (1996): A density-based algorithm for discovering clusters in large spatial databases with noise. – KDD-96 Proceedings, pp. 226-231.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1902_13911405

[9] Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S., Bouras, A. (2014): A survey of clustering algorithms for big data: Taxonomy and empirical analysis. – IEEE transactions on emerging topics in computing 2: 267-279.
[10] Flowers, M., Weisz, R., White, J. G. (2005): Yield-Based Management Zones and Grid Sampling Strategies: Describing Soil Test and Nutrient Variability. – Agronomy Journal 97: 968-982.
[11] Han, J., Pei, J., Kamber, M. (2011): Data mining: concepts and techniques. – Elsevier.
[12] Hasanzadeh-Mofrad, M., Rezvanian, A. (2018): Learning automata clustering. – Journal of computational science 24: 379-388.
[13] Hirschberg, J. B., Rosenberg, A. (2007): V-Measure: a conditional entropy-based external cluster evaluation. – Proceedings of EMNLP, Columbia University.
[14] Huang, S., Chen, Z., Yu, Y., Ma, W.-Y. (2006): Multitype features co-selection for web document clustering. – IEEE Transactions on Knowledge Data Engineering 18: 448-459.
[15] Hubert, L., Arabie, P. (1985): Comparing partitions. – Journal of classification 2: 193-218.
[16] Jiang, Z., Huete, A. R., Chen, J., Chen, Y., Li, J., Yan, G., Zhang, X. (2006): Analysis of NDVI and scaled difference vegetation index retrievals of vegetation fraction. – Remote sensing of environment 101: 366-378.
[17] Jiang, R., Wang, P., Xu, Y., Zhou, Z., Luo, X., Lan, Y. J. S. (2019): A Novel Illumination Compensation Technique for Multi-Spectral Imaging in NDVI Detection. – Sensors 19: 1859.
[18] Jing, L., Ng, M. K., Huang, J. Z. (2007): An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. – IEEE Transactions on knowledge data engineering 19: 1026-1041.
[19] Johnson, S. C. (1967): Hierarchical clustering schemes. – Psychometrika 32: 241-254.
[20] Johnson, C. K., Mortensen, D. A., Wienhold, B. J., Shanahan, J. F., Doran, J. W. (2003): Site-specific management zones based on soil electrical conductivity in a semiarid cropping system. – Agronomy journal 95: 303-315.
[21] Koenig, K., Höfle, B., Hämmerle, M., Jarmer, T., Siegmann, B., Lilienthal, H. (2015): Comparative classification analysis of post-harvest growth detection from terrestrial LiDAR point clouds in precision agriculture. – ISPRS Journal of Photogrammetry and Remote Sensing 104: 112-125.
[22] Liu, H., Wang, X. (2019): Assessing ndvi spatial pattern related to management zones. – Applied Ecology and Environmental Research 17: 6269-6285.
[23] Lopes, F. B., Silva, M. C. D., Miyagi, E. S., Fioravanti, M., Facó, O., Guimarães, R. F., de C. Júnior, O. A., McManus, C. M. (2012): Spatialization of climate, physical and socioeconomic factors that affect the dairy goat production in Brazil and their impact on animal breeding decisions. – Pesquisa Veterinária Brasileira 32: 1073-1081.
[24] Lu, Y.-H. (2005): Mining data streams using clustering. – International Conference on Machine Learning and Cybernetics, IEEE, pp. 2079-2083.
[25] Macqueen, J. (1967): Some methods for classification and analysis of multivariate observations. – Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA, pp. 281-297.
[26] Mohammadrezapour, O., Kisi, O., Pourahmad, F. (2020): Fuzzy c-means and K-means clustering with genetic algorithm for identification of homogeneous regions of groundwater quality. – Neural Computing Applications 32: 3763-3775.
[27] Rousseeuw, P. J. (1987): Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. – Journal of computational applied mathematics 20: 53-65.
[28] Sanjuan, E., Ibekwe-Sanjuan, F. (2006): Text mining without document context. – Information Processing & Management 42: 1532-1552.
[29] Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., Lin, C.-T. (2017): A review of clustering techniques and developments. – Neurocomputing 267: 664-681.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1902_13911405

[30] Sharp, T., Evans, G., Salvador, A. (2004): Weekly NDVI relationships to height, nodes and productivity index for low, medium, and high cotton productivity zones. – Proceedings of the Beltwide Cotton Conferences: 2048.
[31] Steinley, D. (2004): Properties of the Hubert-Arable Adjusted Rand Index. – Psychological methods 9: 386.
[32] Ünlü, R., Xanthopoulos, P. (2019): Estimating the number of clusters in a dataset via consensus clustering. – Expert Systems with Applications 125: 33-39.
[33] Vinh, N. X., Epps, J., Bailey, J. (2010): Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. – The Journal of Machine Learning Research 11: 2837-2854.
[34] Yeung, K. Y., Ruzzo, W. L. (2001): Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. – Bioinformatics 17: 763-774.
[35] Zhang, H., Zhou, X. (2018): A novel clustering algorithm combining niche genetic algorithm with canopy and K-means. – International Conference on Artificial Intelligence and Big Data (ICAIBD), IEEE, pp. 26-32.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1902_13911405