Wheat harvester convoys spatiotemporal patterns mining using a recursive search-based DBSCAN algorithm

Weixin Zhai[1,2], Ruijing Han[1,2], Jiawen Pan[1,2], Caicong Wu[1,2*]

(1. *College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China*;
2. *Key Laboratory of Agricultural Machinery Monitoring and Big Data Applications, Ministry of Agriculture and Rural Affairs, Beijing 100083, China*)

Abstract: Due to varying crop maturity periods and uneven distribution of agricultural machinery, China has developed a unique service model known as cross-regional agricultural machinery operations. Currently, China's comprehensive mechanization rate for grain crops is relatively high, creating a substantial market for cross-regional agricultural machinery operations. Research on the behavioral patterns of cross-regional agricultural machinery migration is both urgent and significant. Considering the actual rules of cross-regional migration during the wheat harvest and the characteristics of the trajectory data, this paper proposes a trajectory mining method using a recursive search-based DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm. One representative finding of this study is that by mining the trajectory data of wheat harvesters within 25 d of peak harvest period, 131 cross-regional trajectories were identified, consisting of 11 633 harvesters. Three main routes of wheat harvester cross-regional migration were identified, along with several smaller routes outside their range. The overall spatiotemporal pattern aligns with observed realities in China. This study can provide valuable references for operators to optimize cross-regional routes, for agricultural machinery manufacturers to develop location-based services, and for relevant government departments to formulate policies.

Keywords: trajectory data mining, cross-regional convoy, cross-regional agricultural machinery operations, wheat harvester, spatiotemporal patterns

DOI: 10.25165/j.ijabe.20251806.9793

Citation: Zhai W X, Han R J, Pan J W, Wu C C. Wheat harvester convoys spatiotemporal patterns mining using a recursive search-based DBSCAN algorithm. Int J Agric & Biol Eng, 2025; 18(6): 221–229.

1 Introduction

Wheat is one of China's primary grain crops. Its maturation and harvesting are highly affected by weather conditions, leading to a short operational window and placing high demands on harvesting efficiency. As of 2024, the comprehensive mechanization rate for crop cultivation and harvesting in China has exceeded 75%, with the mechanized wheat harvesting rate reaching as high as 98%. Due to the large north-south span of China's wheat-producing areas, the time differences in crop maturity across regions, and the imbalance in the number of agricultural machines, China has developed a unique social service model for wheat cross-regional harvesting[1]. During the 2021 wheat harvest season, approximately 75% of the harvesters in major wheat-producing areas participated in cross-regional operations, covering around 84% of the total harvested wheat area. The median cross-regional migration distance (measured as straight-line distance) was approximately 597 km, and about 69% of the harvesters traveled more than 300 km[2]. Based on daily national agricultural machinery operation data, the national wheat harvesting center migration route was analyzed, which showed an alternating trend from east to west and gradually advancing northward, consistent with the wheat maturity period in various regions. The cross-regional operation of agricultural machinery is one of the important means to improve the technical efficiency of grain production. Studying the cross-regional movement of wheat harvesters can provide a reference for optimizing decisions for operators, agricultural machinery manufacturers, and relevant government departments, improving overall production efficiency, and ensuring national food security. It has great economic and social significance[3].

Based on practical production patterns and operational requirements, analyzing the cross-regional migration patterns of agricultural machinery through trajectory data mining demonstrates significant practical applications across multiple domains including agricultural production, management, and policy formulation. The trajectory analysis enables identification of typical migration routes, operational timeframes, and regional demand peaks for agricultural machinery. For manufacturers, this analytical approach facilitates optimization of warehouse facility distribution and enhancement of supporting services through systematic investigation of machinery utilization patterns and demand characteristics. Operators can leverage these insights to refine cross-regional harvesting routes, thereby reducing operational costs and improving resource utilization efficiency. From a policymaking perspective, the derived data patterns provide data-driven support for governmental guidance in agricultural machinery industry development. Furthermore, integration with crop distribution data enables the establishment of precision agricultural service platforms and collaborative machinery sharing/leasing systems, which collectively enhance service efficiency and optimize supply-demand

Received date: 2025-03-18 **Accepted date:** 2025-08-05

Biographies: Weixin Zhai, PhD, Associate Professor, research interest: big data of spatio temporal, big data mining of agricultural machinery, cartography and geographic information system, Email: zhaiweixin@cau.edu.cn; **Ruijing Han**, Master, research interest: big data mining of agricultural machinery, Email: hanrj2@cau.edu.cn; **Jiawen Pan**, PhD, research interest: computational intelligence, computer vision. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China. Email: cau_panjiawen@cau.edu.cn

***Corresponding author: Caicong Wu**, PhD, Professor, research interest: the navigation and big data mining of agricultural machinery. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China. Tel: +86-13810521813, Email: wucc@cau.edu.cn.

coordination in agricultural equipment allocation.

The cross-regional operation of agricultural machinery has attracted attention from a wide range of research perspectives. There have been many studies on its impact on production efficiency[4-6], economic benefits[1], path scheduling[7-9], etc. However, due to problems such as small sample sizes or incomplete data, large-scale quantitative analysis is still difficult.

In recent years, the installation of Beidou high-precision positioning terminals on agricultural machinery and the construction of the agricultural machinery operation big data system have enabled researchers to obtain a large amount of high-quality agricultural machinery trajectory data. These high-frequency, information-rich data can accurately correspond to each piece of agricultural machinery, facilitating research on large-scale agricultural machinery movements[2]. Zhang et al.[10] used social network analysis to conduct cluster analysis from the aspects of grouping and cohesive subgroups, exploring the network structure and internal characteristics of wheat combine harvesters across regions; Li et al.[11] counted the agricultural machinery flow in provinces, cities, and counties, analyzing the degree of dependence on harvesters in various locations.

Previous studies have used regions formed by administrative divisions as research objects and studied the flow between different regions. Using the center of gravity shift map obtained by Li et al.[11] as an example, this study used cities as units, first mapping a small number of trajectories to each administrative district, and then studying the characteristics and laws of agricultural machinery movements across regions between different cities. This article considers each agricultural machine as an independent research object and conducts research based on the original movement trajectory of the agricultural machinery. It is not restricted by administrative divisions, resulting in more thorough data utilization, finer granularity, and more accurate results.

In the past, the analysis of trajectory data movement patterns usually focused on cars, ships, pedestrians, animals, etc., and also had certain applications in meteorological fields such as vortices, ocean hurricanes, and red tides[12-15]. To identify groups of vehicles moving together over time, numerous classic movement patterns and mining algorithms have been proposed, such as Flock[16], Convoy[17], Swarm[18], and Platoon[19]. However, different patterns define the spatiotemporal constraints of accompanying vehicles differently. Regarding time constraints, Flock and Convoy require continuous time points for accompanying vehicles; Swarm does not require time point continuity; Platoon requires that vehicle groups move together within a specific period. Regarding spatial constraints, Flock requires vehicles to be in a disk-shaped geographical area, while other studies relax this constraint to a density-reachable area. Numerous studies have built upon these classic patterns, improving them according to different data characteristics. Li et al.[20] proposed a more relaxed mobility model and used grid partitioning for efficient clustering. Liu et al.[21] proposed an optimized mining algorithm based on the BP model, using a divide-and-conquer approach to optimize mining efficiency in the spatial dimension. Previous studies often impose strict temporal and spatial constraints when determining whether agricultural machinery belongs to the same convoy. Since real-world harvester convoys tend to be loosely organized, this study adopts more relaxed constraints in defining convoys. In addition, agricultural machinery trajectories typically exhibit a “field-road” binary semantic characteristic, with field trajectories accounting for a relatively high proportion, while road network features are indistinct and highly dispersed. Traditional movement pattern extraction methods tend to generate a large number of highly similar convoy results under such conditions.

Trajectory clustering methods enable the grouping of similar trajectories, where the representation of time series and the calculation of similarity measures are critically important. Some sub-trajectory clustering algorithms can identify frequently traveled common paths[22-24], but these methods compromise trajectory integrity and are therefore unsuitable for extracting agricultural machinery convoys. Park et al.[25] proposed a new trajectory spatiotemporal similarity measure based on graph theory to mine travel patterns. Most of these studies search based on the evolving trajectory of moving objects, which often incurs high computational costs and requires efficiency optimization: Zhou et al.[26] generated grid indexes, extracted grid sequence features of the ship’s spatiotemporal trajectory for mining, and improved model confidence. Dutta et al.[27] proposed a new evolutionary clustering algorithm based on multi-objective criteria, using the search function of archived multi-objective simulated annealing (AMOSA) to cluster data sets. However, harvester trajectories vary greatly in temporal duration and spatial coverage, making similarity measurement not only difficult to define precisely but also computationally expensive. Moreover, the resulting clusters may not accurately align with our definition of convoys.

The main research content of this paper is to first utilize the DBSCAN clustering algorithm to spatially cluster the trajectory data at each time interval, and then merge these results to obtain the spatiotemporal distribution of wheat harvesters across all periods. Subsequently, based on this, recursive search is performed to identify harvester cross-regional migration convoys. Finally, the cross-regional movement patterns of wheat harvesters nationwide are analyzed to explore the patterns and laws of cross-regional migration.

2 Materials and methods

2.1 Data acquisition and preprocessing

The research period of this paper covers the large-scale mechanized harvesting in China’s main wheat-producing areas in 2022, specifically 25 d from May 28 to June 21, based on statistics from the Agricultural Mechanization Management Department of the Ministry of Agriculture and Rural Affairs of the People’s Republic of China[28]. The research data consist of GNSS trajectory data of grain combine harvesters during operation. The data come from the agricultural machinery operation big data system, which collects real-time position data from agricultural machinery of various brands such as Lovol, World, Zoomlion, and YTO. Currently, the platform has connected about 700 000 pieces of agricultural machinery, each equipped with a Beidou terminal with a positioning accuracy of about 5 m. The BeiDou/GNSS terminal is a mobile device mounted on the harvester, which continuously collects the machine’s location data in real time during operation. Each piece of agricultural machinery has a unique identification number, and the parameters of each trajectory data include timestamp, longitude, latitude coordinates, etc.

In this experiment, the data sampling interval is set to one day, with each harvester recording the first valid data after 8:00 am daily. A total of 1 091 392 pieces of valid data were obtained from 80 294 harvesters over the entire time period. The daily number of harvesters is shown in Figure 1. Harvesters usually upload data only during the startup period. Due to factors such as weather and long-distance migration, valid trajectory data are not uploaded daily. To

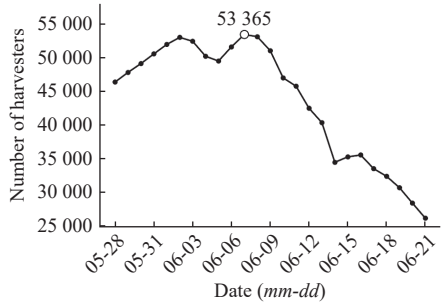

Figure 1 Daily number of harvesters

exclude harvesters that have not operated across a large area, a total of 36 904 harvesters that uploaded data for at least 15 d were selected, as shown in Figure 2. This accounted for 46% of all harvesters, with a total of 769 397 pieces, representing 70.5% of all trajectory data. Data cleaning involved removing abnormal harvesters, drift, duplication, and abnormal trajectory points. The specific geographic distribution of the data used in this study is shown in Figure 3. The provinces with the largest number of records include Henan, Jiangsu, Shandong, Anhui, and Hebei. This dataset effectively covers China's major wheat-producing regions.

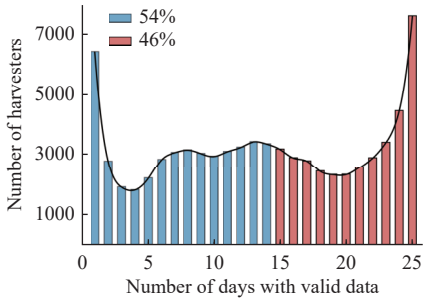

Figure 2 Frequency distribution of trajectory data

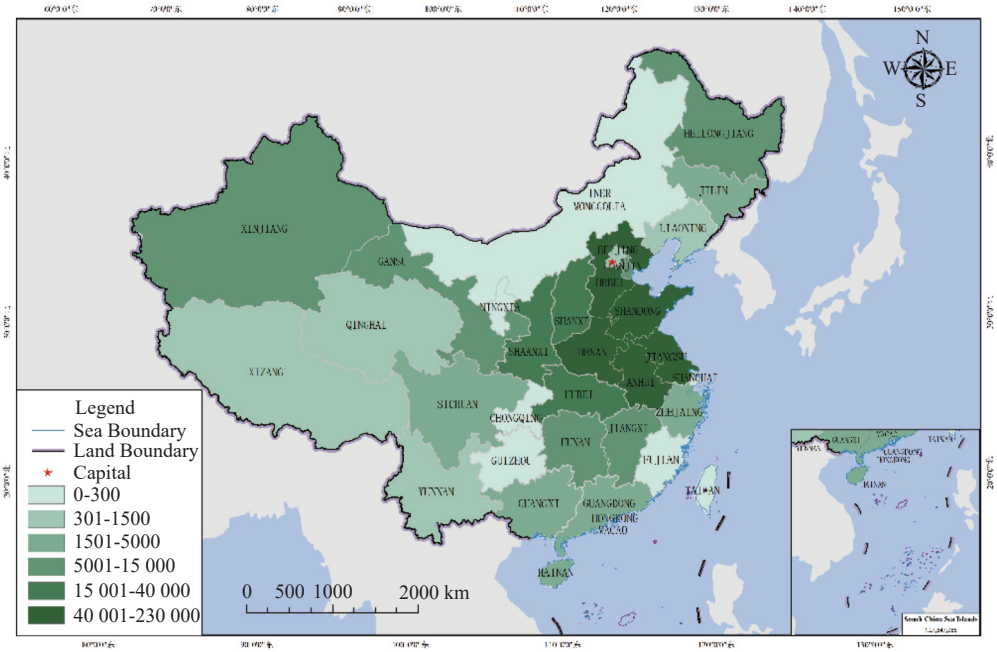

Figure 3 Spatial distribution of the trajectory data

2.2 Definition of cross-regional convoy

To define the pattern of cross-regional convoys of wheat harvesters, a series of field investigations targeting real-world convoys was first conducted, and communication with multiple harvester operators through questionnaires and interviews was engaged in. Through this process, two main types of actual convoys were identified: small-scale convoys primarily composed of individual operators, and large-scale professional service teams. The former are typically smaller in size—sometimes consisting of as few as two machines—and are formed primarily for mutual support and cost reduction. These convoys tend to be highly flexible, with their composition and routes frequently changing in response to factors such as order sources, overall harvesting progress, operational costs, weather conditions, and agricultural policies. The latter, although more structured, also adjust routes dynamically based on service orders and harvesting schedules. Their members may occasionally split and regroup depending on operational demands. Figure 4 illustrates a possible movement pattern of such machinery. O_1, O_2, and O_3 represent the trajectories of three harvesters, with dots indicating their locations at different time points. While the paths are not completely identical, there are moments—such as at timestamps t_2, t_4, and t_7—where full or partial convergence among the machines occurs.

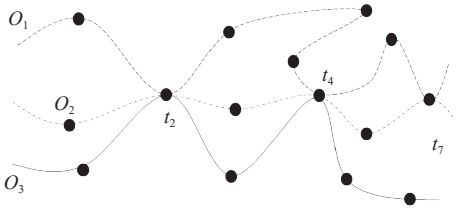

Figure 4 Illustration of cross-regional convoys during wheat harvesting

This paper examines the primary cross-regional mobility patterns of harvester through the lens of cross-regional convoys. The movement pattern of cross-regional harvester convoys is defined in this study as follows: each cross-regional convoy consists of at least $MinSize$ harvesters, and any two harvesters s_i, s_j (where $i, j \leq MinSize$) must belong to the same cluster for at least $MinT$ time interval. It should be noted that the definition of "convoy" in this study does not strictly correspond to the narrow, real-world convoys formed by harvester operators during cross-regional wheat harvesting; rather, it aims to capture various migration patterns and

regularities from a big-data perspective. Specifically, each harvester is constrained to belong to only one convoy, differing from many existing studies. This decision is motivated by the inherently loose structure of harvester convoys and relaxed spatial-distance constraints; without this restriction, a large number of highly similar convoy results would be generated. Moreover, to extract the dominant spatiotemporal patterns of cross-regional harvester convoys, our method favors convoys that are larger in scale and longer in duration.

2.3 Overview

Based on data mining principles, the actual cross-regional migration trajectories of harvesters were analyzed and the concept of cross-regional harvester convoys was proposed. By identifying common patterns and laws in the cross-regional wheat harvest process, the study searches for harvester convoys with high spatiotemporal similarity in their trajectories and explores the dominant cross-regional movement patterns during the wheat harvest. First, the DBSCAN clustering algorithm is employed to cluster and summarize the spatial data at each time interval, resulting in a series of clusters that are temporally and spatially proximate. After the valid data are filtered, the clustering results for the entire period are recursively searched individually, the appropriate parameters are selected, and each convoy that meets the criteria is finally identified.

Figure 5 below presents a schematic diagram of the technical approach of this study. The first step involves data acquisition and preprocessing. After obtaining the wheat harvester trajectory data from the agricultural machinery operation big data system, the data are screened, cleaned, and sampled to create a wheat harvester trajectory dataset for a specific interval ΔT (1 d). The second step involves clustering. The data are spatially clustered at each time interval to obtain the spatial clustering results for each time period. The DBSCAN clustering parameters are Eps = 3 and Minpts = 2. The harvesters in each cluster have high temporal and spatial similarities. All results are combined to obtain a clustering dataset of wheat harvester operation trajectories at specific intervals over the entire time period. The purpose of the third step is to obtain a convoy with the largest possible scale and duration under the premise that each harvester belongs to only one cross-regional convoy. The fourth step involves recursive searching for cross-regional convoys, screening harvesters with data upload times greater than 15 d, and then comparing the clustering results for the entire time period individually to obtain a convoy with a scale greater than 20 and a duration greater than 15 d. On this basis, the mainstream movement patterns of wheat harvesters during cross-regional migration across the country can be analyzed, and the patterns and laws of cross-regional migration can be explored.

The relevant parameters involved in the entire process are presented in Table 1. During the initial stage of our research, extensive and in-depth communication was conducted with numerous wheat harvester operators, and on-site field research was performed to gather information related to cross-regional wheat harvesting. The values of each parameter were determined based on the actual conditions.

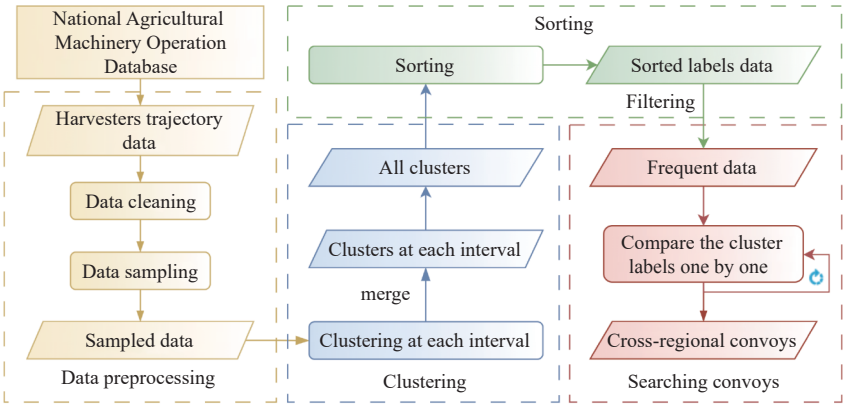

Figure 5 Technical approach

Table 1 Description and value of parameters involved in the whole process

Module	Parameter name	Parameter variable	Parameter meaning	Current value
Preprocessing	Start and end time	Tstart Tend	Start and end time of original data	Tstart=May 28th Tend=June 21st
	Sampling interval	ΔT	Sampling the data at this interval	1 d
DBSCAN cluster analysis	Neighborhood distance threshold	Eps	Core object neighborhood radius	3 km
	Number of samples threshold	Minpts	The minimum number of samples in the field	2
Cross-regional convoy recursive searching	Valid data number threshold	Vmin	Only data greater than this threshold could participate in the determination of the partner group.	15
	Minimum size of partner group	MinSize	The minimum number of harvesters at each encounter	20
	Minimum number of encounters	MinT	In each partner group, the minimum number of encounters between any two harvesters, that is, the minimum number of consistent days when each label is compared	15

2.4 DBSCAN clustering

There are numerous clustering algorithms, some of which, such as K-means, hierarchical clustering, and Gaussian Mixture Models (GMM), require the number of clusters to be specified in advance, and the optimal parameters must be determined through iterative comparison. However, due to the daily variations in the scale and structure of data during wheat harvesting, the best parameters are not universal. Searching for optimal parameters for each day's data would lead to different criteria for evaluating the fleet each day, making these algorithms unsuitable for the task. Among algorithms that do not require specifying the number of clusters, Affinity Propagation automatically determines cluster count and centers via

message passing, iteratively updating responsibility and availability values based on the similarity matrix to select exemplars and assign points[29]; Mean Shift clustering identifies local density peaks through density-gradient ascent, moving points toward higher-density regions until convergence[30]. However, both algorithms incur high computational complexity and are unsuitable for large-scale datasets. Two additional high-efficiency clustering approaches, CDC (Clustering using local Direction Centrality)[31] and HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)[32], are also tried for further verification. CDC distinguishes between interior and boundary points by measuring the uniformity of each point's K-nearest neighbor (KNN) distribution. Boundary points can form an enclosing "cage" that restricts the connection of internal points, thereby effectively separating weakly connected clusters and mitigating the impact of density heterogeneity on cluster identification. HDBSCAN, a density-based algorithm extending DBSCAN, builds a hierarchy to extract stable clusters and is well-suited to complex data distributions. These two algorithms mitigate the overlapping-cluster merging issue in DBSCAN that can produce excessively large clusters. In the results produced by CDC, the maximum cluster size remained relatively small. By adjusting its two key parameters—the number of nearest neighbors K, and T_{DCM} (a parameter related to the distinction between internal and boundary points)—the maximum cluster size could be reduced to fewer than 1,000. With HDBSCAN, given our definition of cross-regional convoys, the permissible parameter range was narrow, resulting in clusters of no more than 30 members. However, since the objective of this study is to identify cross-regional convoys, there are practical requirements regarding cluster size, and most clusters produced by CDC and HDBSCAN are too small to effectively support convoy detection. Therefore, after careful consideration, DBSCAN is selected as the most suitable clustering algorithm.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a widely used density-based clustering algorithm that can identify clusters of arbitrary shape. The core idea behind DBSCAN is that a cluster primarily consists of core points, which have high point densities (see Definitions 1 and 2). The algorithm defines a cluster as the largest set of density-connected points. By identifying high-density regions separated by low-density areas, DBSCAN separates these regions into distinct clusters. The DBSCAN algorithm requires two parameters: the neighborhood radius (Eps), which defines the range of the circular neighborhood around a given point p, and the minimum number of points ($Minpts$) required within the Eps-neighborhood of p, which corresponds to the number of harvesters.

Definition 1. (Eps-neighborhood of a point): The Eps-neighborhood of point p, is defined by $N_{Eps}(p) = \{q \in L | dis(p,q) \leq \text{Eps}$, where L is a set of points representing the given location coordinates, and $dis(p,q)$ represents the Euclidean distance between points p and q.

Definition 2. (Core point): If the number of points in the Eps-neighborhood of point p is larger than $Minpts$, it is a core point.

DBSCAN accomplishes the clustering process by extracting clusters sequentially. Starting from an arbitrary point p, its Eps-neighborhood is checked. If p is a core point, a new cluster C is created with the points in $N_{Eps}(p)$. Then, for each point q in C whose Eps-neighborhood has not yet been checked, if q is a core point, the points in $N_{Eps}(p)$ which are not already contained in C are added to the cluster. The expansion of cluster C is repeated until no new point can be added to the cluster. The clustering process terminates when no new cluster is created.

Selecting appropriate parameters is essential to obtaining reasonable clustering results. The Eps value directly affects the number and size of clusters. If the Eps value is too large, most of the trajectory points will be clustered together, resulting in fewer but larger clusters. Conversely, if the Eps value is too small, it will cause the clusters to split, generate more noise points, and reduce the number of clusters. The Minpts value is relatively less sensitive and should be carefully selected based on experience and data distribution. This study selects parameters based on interviews with actual cross-regional convoy drivers during the wheat harvest.

2.5 Sorting cluster data

It is conceivable that the same harvester may appear in multiple convoys of similar size and duration. This paper prioritizes identifying convoys with larger size and longer duration, necessitating the sorting of cluster data. First, all data are summarized into a table, with each row representing a harvester, each column representing a time interval, and each value indicating the corresponding cluster label, as listed in Table 2. Additionally, the numbers in each column are independent, with no special relationship between identical numbers. For instance, the label value of the harvester corresponding to DID "LOV230" on June 19 and June 20 is 7. This does not imply that the two clusters are completely consistent; it only indicates that they are both the seventh cluster at their respective time intervals.

Table 2 Cluster label data after integration and sorting (example)

DID	5-28	5-29	…	6-19	6-20	6-21
LOV227	0	0	…	0	0	0
LOV228	0	0	…	0	0	0
LOV229	0	0	…	0	0	0
LOV230	0	0	…	7	7	0
LOV231	0	0	…	14	0	0
LOV232	0	0	…	0	0	0
LOV233	–1	0	…	15	21	2
LOV234	0	0	…	2	0	0
LOV235	0	–1	…	2	5	0
LOV236	0	0	…	36	47	37
LOV237	0	0	…	29	60	48
LOV238	0	0	…	67	Na	Na
LOV239	0	0	…	2	Na	Na
LOV240	0	0	…	2	Na	Na
LOV241	0	0	…	0	Na	Na

Note: –1 represents noise, Na means no data on that day.

The sorting steps are:

(1) For the clustering results of each time interval, mark outliers with a label of –1, and sort the remaining data in reverse order based on cluster size.

(2) After summarizing the results of each interval, sort each harvester in reverse order based on the number of valid data (uploaded and non-noise).

(3) For harvesters with the same number of valid data, sort them by cluster size from the previous step, according to the time interval order.

Finally, all data are summarized into a table, with each row representing a harvester, each column representing a time interval, and each value indicating the corresponding cluster label.

2.6 Recursive search for cross-regional convoys

Based on the sorted clusters, the search for cross-regional

convoys is conducted. The specific process is as follows:

(1) Compare all rows with the first row. If the number of columns consistent with the first row is greater than MinT, retain the row:

• If the number of rows that meet the requirements is less than MinSize, delete the first row and proceed to step 2 with the remaining rows.

• If the number of rows that meet the requirements is greater than MinSize, compare these rows with the second row, and so on, until all rows have been compared. If the number of rows is still greater than MinSize, a convoy is identified.

(2) Delete the rows included in the convoy, and continue with step 1 for the remaining rows until the number of remaining rows is less than MinSize, ending the search.

2.7 Detailed analysis of cross-regional convoys

For the searched convoys, their average trajectories are calculated, and various mainstream movement patterns in the cross-regional migration of harvesters across the country can be obtained. Their starting point, scale, internal segmentation mode, etc. are analyzed in detail to explore more internal laws of cross-regional migration.

3 Results and discussion

3.1 Shift of wheat harvest center of gravity

Due to the extensive north-south span of wheat-producing regions in China and the associated time differences in crop maturity across these areas, coupled with the uneven distribution of agricultural machinery, China has developed a unique social service model for cross-regional mechanized wheat harvesting. Operators migrate based on wheat maturity, moving to various locations for harvesting. Figure 6 shows a national wheat harvest center of gravity shift map based on the analysis of daily national agricultural machinery operation data. The daily operation center G is calculated based on the weighted average of the harvester locations (x_i, y_i) for the day, according to the harvested area s_i (as shown in Equation (1)). The map indicates that the harvest center of gravity alternates between east and west and gradually advances northward (Li et al. 2023). However, this study uses administrative divisions as the research unit, which limits the accuracy and detail of the results. Therefore, this paper investigates the laws of cross-regional migration of harvesters during the wheat harvest based on the specific trajectories of individual harvesters.

$$G = \frac{\sum_{i=0}^{n} (x_i, y_i) \times s_i}{n \times \sum_{i=0}^{n} s_i} \quad (1)$$

3.2 DBSCAN clustering

Based on communications with drivers in the cross-regional convoy, it was learned that there are many small "convoys" consisting of only two harvesters. Therefore, the Minpts parameter of DBSCAN was set to 2. While the general movement direction of the convoy is consistent, the daily operation locations are not necessarily close, often reuniting only during rest periods. Thus, the Eps value was set to 3 km. Figure 7 shows the number of clusters of various sizes per day, with most clusters being small.

Although there are fewer large clusters, the harvesters contained within them account for a significant proportion. This is due to the nature of the DBSCAN algorithm. In areas with higher density, very large clusters are formed, while in areas with sparse

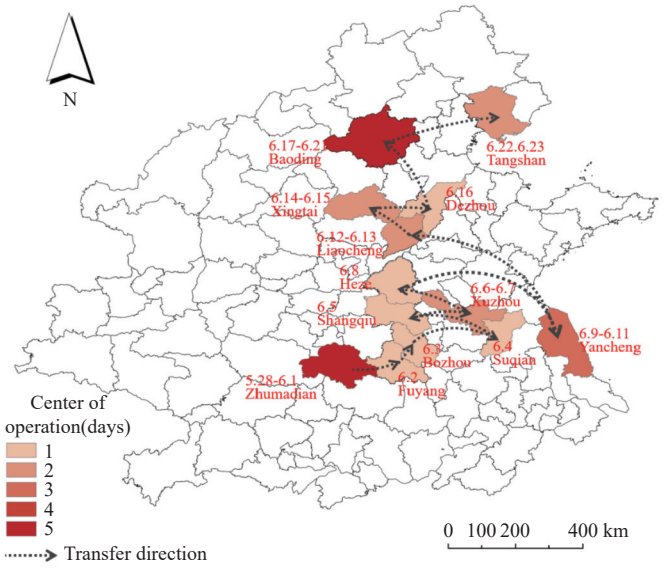

Figure 6 2022 National wheat harvest center shift map

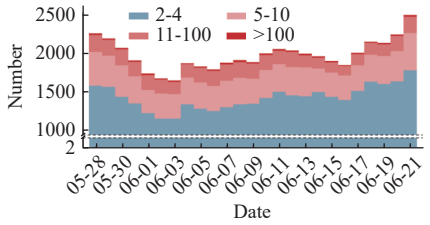

Figure 7 Distribution of daily cluster size: The vertical axis represents the number of clusters of different sizes per day

density, the clusters obtained are often smaller and most are considered noise. Figure 8 shows the clustering results for the peak day during the wheat harvest (53 365 harvesters). A total of 2942 clusters and 5192 noise points (not shown) were obtained. In Figure 8a, it can be seen that most of the harvesters are concentrated in the top few clusters. In Figure 8b, each color represents a cluster. The two largest clusters have 15,126 and 13,804 points, respectively, accounting for 61.43% of the harvesters on that day.

Applying Swarm pattern mining to the clustering results yielded 595 367 893 convoy records, a number that far exceeds the total number of harvesters. This is primarily due to the definition of the Swarm movement pattern, which allows a single harvester to be included in multiple convoys. In fact, this characteristic is shared by many existing movement pattern mining methods. Combined with our relatively permissive DBSCAN parameter settings, this resulted in numerous highly similar convoys. For example, the following are two convoy examples extracted from the same clustering result:

• Convoy A: [69, 28 186, 36 270, 43 664, 84 967] active on timestamps [0, 1, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24];

• Convoy B: [69, 28 186, 36 270, 84 967] active on timestamps [0, 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24].

In Convoy B, the inclusion of timestamp 9 led to the exclusion of harvester 43 664. The timestamp overlap between the two convoys was 95.65%, and the membership overlap was 80%. Many similar cases were observed, indicating a high level of redundancy in the results. Such redundant and highly similar convoy records do not meaningfully contribute to the identification of dominant cross-regional migration patterns of wheat harvesters, which is the primary goal of this study.

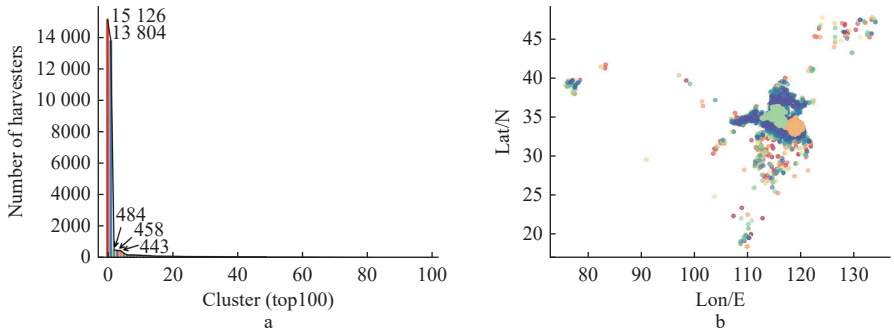

Figure 8 Clustering results for the peak day during the wheat harvest

3.3 Parameter selection

Convoy numbers and sizes are influenced by many parameters. Table 3 presents results under different parameter combinations. The parameters were determined as follows. Eps was chosen as 3 km, informed by field investigations into the dispersion of real harvester convoys. The table also includes a result for Eps = 9.5 km: such a large Eps causes over-merged clusters, yielding very few clusters and convoys. Although the overall spatial flow aligns with reality, too many details are lost. For MinT, as shown in Figure 3, it is observed that most harvesters operate for over 5 d; and because cross-regional transfer patterns over the entire peak harvest period are focused on, setting MinT below 10 d is considered too short. Setting MinT above 20 d is deemed impractical, as harvester operations are commonly interrupted by weather or long-distance transfers during actual harvesting. Results around 15 d show little variation, so 15 d was selected as a representative middle value. The selection strategy for the MinSize parameter is guided by similar principles, as small variations in this parameter are likewise found to have minimal impact on the results.

Table 3 Results obtained under different parameter settings

Eps	Minpts	MinSize	MinT	Number of Convoy	Number of Harvester
9.5	2	20	15	27	16 989
3	2	18	13	185	17 666
3	2	18	14	159	14 433
3	2	18	15	152	11 877
3	2	18	16	131	9738
3	2	18	17	129	7992
3	2	19	13	170	17 465
3	2	19	14	151	14 296
3	2	19	15	143	11 768
3	2	19	16	127	9708
3	2	20	13	162	17 316
3	2	20	14	145	14 210
3	2	20	15	131	11 633
3	2	20	16	120	9614
3	2	20	17	105	7616
3	2	21	13	153	17 166
3	2	21	14	138	14 109
3	2	21	15	131	11 671
3	2	21	16	117	11 674
3	2	21	17	104	7608
3	2	22	13	153	17 166
3	2	22	14	130	13 977
3	2	22	15	126	11 608
3	2	22	16	113	9453
3	2	22	17	96	7534
2	2	10	15	111	1630
2	2	20	15	29	847

3.4 Recursive searching cross-regional convoys

To ensure the size and duration of the convoy, and considering the actual wheat maturity period and the basic requirements for cross-regional migration, Eps = 3, Vmin = 15, MinSize = 20, and MinT = 15 are used. This resulted in a total of 131 cross-regional convoys, comprising 11 633 harvesters. Among these, two large convoys had more than 1000 harvesters, while most convoys had fewer than 100 harvesters, as shown in Figure 9. The visualization results are shown in Figure 10. Each colored polyline in the figure represents the average trajectory of a convoy. The width of the polyline corresponds to the number of harvesters. The green triangle marks the starting point, and the red dot marks the end point. The results reveal three primary modes of harvester cross-regional migration, all originating from southern Henan. The largest group operates in the main wheat-producing area, moving northeastward along the northern Henan-Hebei central route, which is also the region with the highest wheat yield and greatest harvesting pressure. The second group follows the Jiangsu-Shandong eastern route, initially moving east and then turning north. The third group moves northwest along the Shaanxi-Gansu western route. Additionally, there are smaller convoys operating across other wheat-producing areas. The overall spatial flow direction and scale align with the actual conditions of cross-regional mechanized wheat harvesting in China.

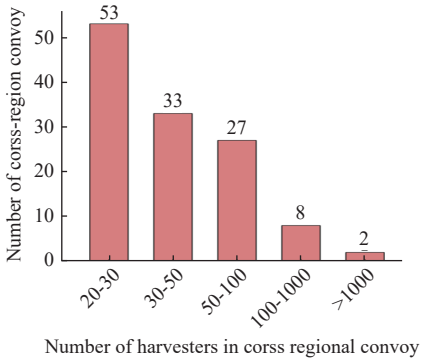

Figure 9 Distribution of cross-regional convoy sizes

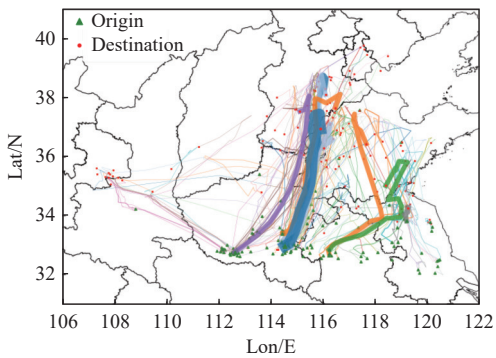

Figure 10 Average trajectory of cross-regional convoys

The largest cross-regional convoy contains 2878 harvesters, with an average trajectory starting from Pingyu County, Zhumadian City, Henan Province, moving to Zaoqiang County, Hengshui City, Hebei Province, and then turning south to Xiajin County, Dezhou City, Shandong Province. To observe the movement patterns within the convoy more deeply, the specific trajectories of each harvester are presented individually, as shown in Figure 11. The trajectories mainly cover most of Henan and Hebei provinces, as well as the western region of Shandong. It can be seen that the trajectories of most harvesters follow the average trajectory trend, carrying out cross-regional migration from south to north in the main wheat-producing areas. Compared with previous studies, this result is based on actual trajectories, is not restricted by administrative division frameworks, is more specific and intuitive, and conforms to actual crop maturity laws and cross-regional migration modes.

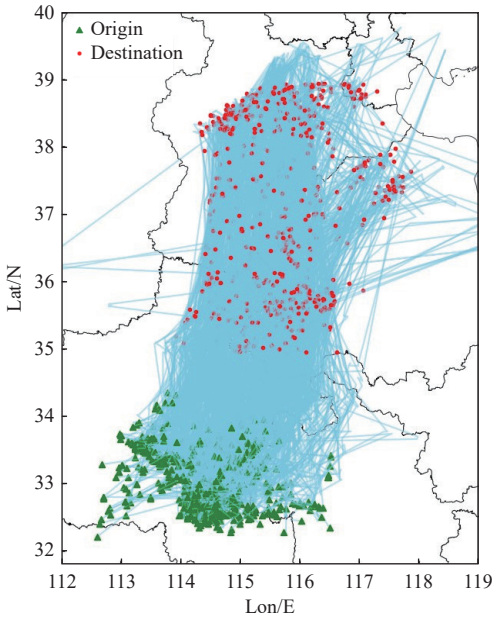

Figure 11 Detailed trajectory of harvesters

3.5 Limitations

Due to the characteristics of DBSCAN algorithm, DBSCAN uses fixed parameters to identify clusters and requires data density decrease to detect cluster boundaries. If multiple clusters overlap without data density decrease, they may be grouped into a single cluster. Therefore, large clusters will appear in high-density regions, and the distribution of all cluster sizes will show a long tail.

The Ht-index quantifies the fractal or scaling structure of geographic features[33]. The data is divided according to the average: if the tail is larger than the head, the head is further divided. A geographic feature has an Ht-index of h if the pattern of far more small things than large ones recurs (h – 1) times at different scales. If the Ht-index is 4, it indicates three instances where the number of small clusters is larger than the number of large clusters. As shown in Figure 8a, the head and tail of the day's clustering results were highly unbalanced, with the tail being much longer than the head, demonstrating clear characteristics of a long-tail distribution. The data volume of the head after the first division was only 5.11%, while the data volume of the head after the final division was 2 units, accounting for just 0.11% of the total. Figure 12 illustrates the Ht-index of daily clustering results and the proportion of head data volume after the first division. This indicates that the distribution of daily cluster sizes resembles that of the observed day, exhibiting long-tail distribution characteristics. Consequently, the search convoys also exhibit similar characteristics, ultimately resulting in two convoys being significantly larger than the others.

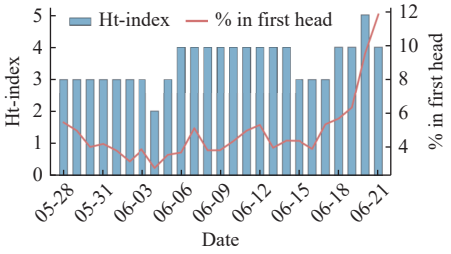

Figure 12 Daily Ht-index

In future work, improvements to clustering algorithms are planned to be explored. For example, results from DBSCAN under multiple parameter settings may be fused, or multi-level clustering may be performed to further subdivide large clusters. Additionally, how other advanced clustering algorithms can be adapted and improved for applications in this type of research will be investigated.

4 Conclusions

In response to regular differences in crop maturity across regions and the imbalance in agricultural machinery ownership, China has developed a socialized service model for cross-regional agricultural machinery operations. The agricultural machinery operation big data system based on Beidou offers a substantial amount of high-quality trajectory positioning data of agricultural machinery. This data enables a comprehensive quantitative analysis of harvester movement patterns. Research has yielded significant findings on the cross-regional migration patterns of agricultural machinery during wheat harvesting, including the development of shift maps for the center of gravity. Based on previous studies of cross-regional agricultural machinery laws, this study addresses the characteristics of harvester trajectory data dispersion and the high proportion of field trajectories, along with field survey results of the cross-regional convoy migration mode. Starting from the trajectory of each harvester, this study proposes a trajectory mining method using a recursive search-based DBSCAN algorithm. After sampling and slicing the wheat harvester trajectory data during the peak harvest period, spatial clustering is performed separately, followed by a recursive search to identify multiple cross-regional migration convoys. From 1 091 392 pieces of location data of 80 294 harvesters, 131 convoys with a scale greater than 20 and a duration greater than 15 d were searched, comprising a total of 11 633 harvesters. Except for the two largest convoys of more than 1000 harvesters, most of the convoy sizes are less than 100 harvesters. The average trajectory of the harvester within each convoy is calculated to determine its cross-regional movement paths. The results identify three primary modes of cross- regional migration: western, middle, and eastern routes. The middle route, the largest in scale, begins in southern Henan and moves northeast to Hebei, covering the main wheat-producing area. The eastern route generally starts eastward and then moves north, encompassing provinces such as Anhui, Jiangsu, and Shandong. The western route moves northwest towards Shaanxi and Gansu. Additionally, smaller convoys operate across other wheat-producing areas. The overall spatial flow direction and scale align with the observed cross-regional migration pattern during the wheat harvest. This research can provide references for operators, agricultural machinery manufacturers, relevant government departments, and other

stakeholders to optimize cross-regional migration routes, configure spare parts and maintenance resources, rationally allocate agricultural machinery, and implement location- and community-based services. This approach improves the efficiency of comprehensive information utilization and ultimately enhances the overall efficiency of cross-regional operations, ensuring bumper harvests and national food security.

While this research has made important contributions, several limitations remain. Different parameter combinations can lead to varying results, and different crops, time periods, study regions, or movement patterns may impose different requirements on the clustering outcomes. This introduces greater flexibility and complexity in parameter selection. Future work could explore improvements in parameter selection, such as the fusion of multi-parameter results, the development of dynamic parameter models, or the adaptation and enhancement of other advanced clustering algorithms for similar research applications. Moreover, this research has the potential for application in other domains. For example, identifying wheat harvester convoys with similar movement patterns could help optimize machinery scheduling and supply-demand matching, thereby improving the efficiency of cross-regional operations. It could also support the prediction of machinery maintenance and service needs to optimize resource allocation, as well as the analysis of cross-regional operation costs and their economic benefits to farmers.

Acknowledgements

This research was financially supported by the National Natural Science Foundation of China (Grant No. 32301691), the National Key R&D Program of China (Grant No. 2025YFE0103600), the China Scholarship Council (202306350108), and the 2115 Talent Development Program of China Agricultural University.

[References]

[1] Zhang X B, Yang J, Thomas R. Mechanization outsourcing clusters and division of labor in Chinese agriculture. China Economic Review, 2017; 43: 184–195.

[2] Wu C C, Li D, Zhang X Q, Pan J W, Quan L, Yang L L, et al. China's agricultural machinery operation big data system. Computers and Electronics in Agriculture, 2023; 205: 107594.

[3] Ma W Q, Liu T X, Li W Q, Yang H. The role of agricultural machinery in improving green grain productivity in China: Towards trans-regional operation and low-carbon practices. Heliyon, 2023; 9(10): e20279.

[4] Pan J, Wu C, Zhai W. A hybrid genetic slime mould algorithm for parameter optimization of field-road trajectory segmentation models. Information Processing in Agriculture, 2024; 11(4): 590–602.

[5] Zhai W, Zhang X, Liu J, et al. A dual-level interactive cascade network for agricultural machinery trajectory operation mode identification. Computers and Electronics in Agriculture, 2025; 238: 110788.

[6] Zhai W, Xu Z, Pan J, et al. A general image classification model for agricultural machinery trajectory mode recognition. Computers and Electronics in Agriculture, 2024; 227: 109629.

[7] Zhai W, Guo Z, Pan J, et al. Addressing local sparsity in massive agricultural machinery trajectories: A BiLSTM-GRU approach. Computers and Electronics in Agriculture, 2024; 226: 109376.

[8] Zhai W, Kuang X, Cheng X, et al. Reconstruction of missing points in agricultural machinery trajectory based on bidirectional adjacent information. Computers and Electronics in Agriculture, 2024; 220: 108920.

[9] Zhai W, Mo G, Xiao Y, et al. GAN-BiLSTM network for field-road classification on imbalanced GNSS recordings. Computers and Electronics in Agriculture, 2024; 216: 108457.

[10] Zhang Z, Qi S, Zhang M. Spatial flow law of cross-regional operation of wheat combine harvesters in China. Transactions of the CSAE, 2021; 37(23): 19–27.

[11] Li D, Liu X, Zhou K, Sun R Z, Wang C T, Zhai W X, et al. Discovering spatiotemporal characteristics of the trans-regional harvesting operation using big data of GNSS trajectories in China. Computers and Electronics in Agriculture, 2023; 211: 108003.

[12] Li H H, Lam J S L, Yang Z L, Liu J X, Liu R W, Liang M H, et al. Unsupervised hierarchical methodology of maritime traffic pattern extraction for knowledge discovery. Transportation Research Part C-Emerging Technologies, 2022; 143: 103856.

[13] Park S, Xu Y, Jiang L, Chen Z L, Huang S Y. Spatial structures of tourism destinations: A trajectory data mining approach leveraging mobile big data. Annals of Tourism Research, 2020; 84: 102973.

[14] Xiao Y, He X, Yang C, Liu H, Liu Y. Dynamic graph computing: A method of finding companion vehicles from traffic streaming data. Information Sciences, 2022; 591: 128–141.

[15] Zhang A S, Shi W Z, Liu Z W, Zhou X L, Geo-SigSPM: mining geographically interesting and significant sequential patterns from trajectories. International Journal of Geographical Information Science, 2024 Feb 2024, doi: 10.1080/13658816.2024.2320149

[16] Gudmundsson J, Kreveld M V. Computing longest duration flocks in trajectory data. Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems, Arlington, Virginia, USA, 2006. Available: https://doi.org/10.1145/1183471.1183479.

[17] Jeung H, Yiu M L, Zhou X, Jensen C S, Shen H T. Discovery of convoys in trajectory databases. Proc. VLDB Endow., 2008; 1(1): 1068–1080.

[18] Li Z, Ding B, Han J, Kays R. Swarm: mining relaxed temporal moving object clusters. Proc. VLDB Endow., 2010; 3(1-2): 723–734.

[19] Li Y X, Bailey J, Kulik L. Efficient mining of platoon patterns in trajectory databases. Data & Knowledge Engineering, Editorial Material, 2015; 100: 167–187.

[20] Li K, Wang H Y, Chen Z W, Chen L S. Relaxed group pattern detection over massive-scale trajectories. Future Generation Computer Systems-the International Journal of Escience, 2023; 144: 131–139.

[21] Liu Y Y, Dai H, Li J W, Chen Y, Yang G, Wang J. BP-Model-based convoy mining algorithms for moving objects. Expert Systems with Applications, 2022; 213: 118860.

[22] Ansari M Y, Mainuddin, Ahmad A, Bhushan G. Spatiotemporal trajectory clustering: A clustering algorithm for spatiotemporal data. Expert Systems with Applications, 2021; 178: 115048.

[23] Niu X Z, Zheng Y H, Fournier-Viger P, Wang B. Parallel grid-based density peak clustering of big trajectory data. Applied Intelligence, 2022; 52(15): 17042–17057.

[24] Tang C H, Chen M Y, Zhao J H, Liu T, Liu K, Yan H Y, et al. A novel ship trajectory clustering method for Finding Overall and Local Features Of Ship Trajectories. Ocean Engineering, 2021; 241: 110108.

[25] Park S, Yuan Y Q, Choe Y. Application of graph theory to mining the similarity of travel trajectories. Tourism Management, 2021; 87: 104391.

[26] Zhou C H, Liu G Y, Huang L, Wen Y Q. Spatiotemporal companion pattern (STCP) mining of ships based on trajectory features. Journal of Marine Science and Engineering, 2023; 11(3): 528.

[27] Dutta S, Das A, Patra B K. CLUSTMOSA: Clustering for GPS trajectory data based on multi-objective simulated annealing to develop mobility application. Applied Soft Computing, 2022; 130: 109655.

[28] MOA. The national "three summer" large-scale wheat harvest basically ended. http://www.moa.gov.cn/xw/bmdt/202206/t20220623_6403139.htm Accessed on [2022-06-23].

[29] Frey B J, Dueck D. Clustering by passing messages between data points. Science, 2007; 315(5814): 972–976.

[30] Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002; 24(5): 603–619.

[31] Beer A, Draganov A, Hohma E, Jahn P, Frey C M M, Assent I. Connecting the dots-density-connectivity distance unifies DBSCAN, k-center and spectral clustering. 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Long Beach, CA, 2023, Aug 06-10. pp.80–92. doi: 10.1145/3580305.3599283.

[32] Campello R J G B, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In Pei J, Tseng V S, Cao L, Motoda H, Xu G, Eds. Advances in knowledge discovery and data mining, Springer Berlin Heidelberg, 2013; pp.160–172.

[33] Jiang B, Yin J J. Ht-index for quantifying the fractal or scaling structure of geographic features. Annals of the Association of American Geographers, 2014; 104(3): 530–541.