IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 12, 2021, accepted October 18, 2021, date of publication October 22, 2021, date of current version November 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3122112

A Principal Component Analysis-Boosted Dynamic Gaussian Mixture Clustering Model for Ignition Factors of Brazil's Rainforests

MAOFA WANG[1], GUANGDA GAO[2], HONGLIANG HUANG[3], ALI ASGHAR HEIDARI[4], QIAN ZHANG[5], HUILING CHEN[6], (Associate Member, IEEE), AND WEIYU TANG[7]

[1]Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China
[2]School of Information Engineering, China University of Geosciences (Beijing), Beijing 100000, China
[3]School of Public Foundation and Applied Statistics, Zhuhai College of Jilin University, Zhuhai 519041, China
[4]School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran 1439957131, Iran
[5]School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou 325035, China
[6]College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou 325035, China
[7]School of Computer, Zhuhai College of Jilin University, Zhuhai 519041, China

Corresponding authors: Qian Zhang (20200420@wzu.edu.cn) and Huiling Chen (chenhuiling.jlu@gmail.com)

This work was supported in part by the Youth Fund Project of National Natural Science Foundation of China under Grant 41504037, and in part by the research on key algorithms and system development of vectorization of simulated seismic monitoring waveform records (National level, presided over, done, from January 2016 to December 2018) under Grant RMB 230000.

(Maofa Wang, Guangda Gao, Hongliang Huang are co-first authors.)

ABSTRACT Analysis of Brazil's rainforest fires caused by various factors has become a hot topic nowadays,. Mining of rainforest fire data through learning unlabeled training samples can reveal inherent properties and patterns, providing a clue for fire prevention. Among commonly used mining approaches, clustering algorithms based on density estimation can relatively effectively capture the potential ignition features through probability calculation, while the Gaussian mixture model (GMM) based on Expectation-Maximum (EM) can effectively quantify fire distribution curves and decompose a fire object into different shape clustering problems based on the actual distribution characteristics of fires data, and thus cluster fires more accurately. However, when the discrimination of probability density is not apparent, the clustering effect is susceptible to both the number of parameters used in clustering and the shape of the clustering problem. Therefore, in the present paper, based on a new strategy of selecting and updating the parameters in the GMM, a new hybrid clustering model called Principal Component Analysis-boosted Dynamic Gaussian Mixture Clustering model (PCA-DGM) is developed. Specifically, Principal Component Analysis (PCA) reduces the dimension of fire samples and strengthens key ignition features. Furthermore, a new dynamic distance loss function is developed by dynamically selecting density parameters or distance parameters, whose computing value is utilized as one important parameter of the clustering shape decision of the GMM. Using the PCA-DGM, which can effectively solve clustering problems with various shapes, the causes of forest fires in Brazil are analyzed at both the temporal and geographical levels, and the experimental results demonstrate that the proposed PCA-DGM in this paper has a better clustering effect than the other traditional clustering algorithms.

INDEX TERMS Forest fire, ignition factor, PCA-DGM, principal component analysis, Gaussian mixture.

I. INTRODUCTION

Hazard analysis is one of the crucial stages of advance for developing countries with a growing population toward sustainable development [1]–[3]. Forest fires [4], [5], which usually occur in the forest, are challenging to prevent and control. Although forest fires are typically instigated by lightning, they can also be caused by human carelessness, deliberate arson, volcanic outbreak and pyroclastic clouds. Moreover, heatwaves, drought, and periodic climate change [6]–[9], such as the El Niño phenomenon, can dramatically increase mountain fire risks. Greenpeace announced in 2018 that "the total global emission of carbon dioxide from wildfires is as high as 7.7 billion metric tons per year" [10].

The associate editor coordinating the review of this manuscript and approving it for publication was Binit Lukose.

Brazil is one of the countries most seriously affected by forest fires [11]. The Amazon Forest in Brazil, which accounts for half of the world's rainforest area and 20% of the forest area, holds the world's largest and most tropical rainforest species [12]. Known as the ''lung of the earth'', it significantly influences the whole earth's environment. Due to forest fires, forest area diminished quickly in the tropical rainforests of multiple states such as Rondonia State, Maton Grosso State, and Para State [11]. Therefore, the study of forest fires' characteristics in Brazil can help to protect the country's environment [13] and reduce economic losses [14] by planning and implementing relevant policies.

In this paper, to help the government to make decisions and intervene in the occurrence of forest fires, the factors responsible for the occurrence of forest fires in Brazil are identified by studying the temporal and geographical characteristics of Brazilian states. To this end, a new, improved Gaussian mixture clustering model (GMM) called PCA-DGM is established based on Principal component analysis (PCA), the expectation-maximization (EM) algorithm, and the loss function of distance clustering.

PCA, proposed by Karl Pearson in 1901 [15], [16], is used to analyze and obtain the main components of data using eigenmatrix transformation. PCA is a simple method for analyzing multivariate statistical distribution with characteristic quantity [17]. The results can be interpreted as an explanation of the variance in the original data. In other words, PCA provides an effective way to reduce data dimension.

Expectation-maximization (EM) was developed by Arthur P. Dempster, Nan laird, and Donald Rubin in their classic paper published in 1977 [18]. The EM algorithm [19]–[21] can be employed in statistical studies to explore the maximum likelihood approximation of parameters in probability models that rely on unobservable unknown variables. In statistical calculation, the maximum expectation (EM) algorithm is used to achieve the maximum likelihood approximation or the probability model parameters' maximum posterior approximation.The probability model rests on the hidden variables that cannot be detected. The EM algorithm is often used in machine learning and data clustering of computer vision.

Cluster analysis [22], [23], also known as clustering, is widely used in many fields as a technology for statistical data analysis. In many potential applications, clustering can be a key component within the system. The notion of clustering is based on the fact that splitting similar objects into different collections or more subsets by static classification results in the member objects in the same subset having some similar characteristics, such as shorter spatial distance in a specific coordinate system. Data clustering is generally classified as unsupervised learning.

GMM [24] can be used in clustering and probability density estimation. The clustering algorithm based on density was developed for mining classes with arbitrary shape. In this algorithm, a category is regarded as an area in the dataset that is greater than a certain threshold. The advantage of GMM is that the probability of each class is obtained instead of a definite classification mark.

In this study, based on GMM, a density-based clustering algorithm, a new dimension reduction clustering algorithm is proposed. On the Brazilian government's official website, our research team obtained a dataset report of the number of forest fires in Brazil divided by states (1998-2017). Using geographic data, time information, and the number of forest fires in each state of Brazil, a Gaussian mixture model (GMM) was optimized to adjust parameters by clustering so as to obtain forest fire characteristics in different states of Brazil [25], [26]. Given that the calculation cost ratio of each iteration of GMM is based on the EM algorithm [24], it may fall into the local extreme. Therefore, the selection of the initial value is critical. In this paper, the number of GMM clustering parameters was optimized.

To sum up, the main contributions of this paper are as follows:

(1) Aiming to discover the features of forest fire using geographical and temporal data, this study proposes a dynamic clustering model framework named PCA-DGM, which is based on PCA, GMM, and a new advanced distance loss function. The innovative framework has more advantages and excellent performance in terms of clustering stability, feasibility, authenticity, accuracy, and integrity.

(2) This study designs a research method based on geographical location and time factors. We prove that the proposed research method is practical.

(3) Extensive experimental results based on synthetic and real-world datasets demonstrate that the proposed integrated clustering model is more competitive and balanced and better than other similar clustering models.

II. RELATED WORK

At present, there are few reports on the optimization of clustering results by improving clustering parameters, and there are few studies focusing on the feature extraction of rainforest fire factors. In 2009, Christos *et al.* used the results of sensitivity analysis of the BP neural network (BPN) to distinguish the influence of each variable in the development of fire risk scheme [27]. In 2011, DG Woolford *et al.* used a logistic generalized additive mixture model to study ignition factors [28]. In 2012, N. Phillip Cheney *et al.* [29] established an empirical model to predict fire behavior. In 2013, N Arndt *et al.* explored the relationship between forest ignition factors by studying independent socio-economic variables [30]. In 2014, M. Rodrigues *et al.* [31] used the logistic regression technology within the framework of the geographicallyi weighted regression model (GWR) to analyze the spatial variation of man-made wildfire explanatory factors in mainland Spain. In 2015, Bianchi *et al.* studied the effects of live fuel moisture content (LFMC) and blade ignition on forest fires [32]. In 2016, Futao Guo *et al.* [33] used Ripley's K-function and logistic regression (LR) model to predict the possibility of fire based on forest wildfires in Southeast China. In 2017, Mortimer M. m ü ller and

Harald vacik [34] studied forest fires from the perspective of lightning. In 2018, J Ruffault and F Mouillot studied the ignition factor of fire by using enhanced regression tree and a set of seven explanatory variables [35]. And what is more, Nicholas read *et al.* [36] introduced a method to decompose ignition prediction into single covariate contribution based on lightning. In 2019, Volkan Sevic *et al.* [37] introduced the Bayesian network model to predict possible forest fire causes and analyze the multilateral interaction between them. In addition, Molina J.R. *et al.* [38] found that there was a significant correlation between fire intensity and biomass consumption. In 2020, Flavio Tiago coutoa *et al.* [39] evaluated the applicability of the current meso NH electrical scheme (cells) in forest fire ignition investigation. And Neetu Verma and Dinesh Singh [40] identified climatic factors and their interrelationships that can be used to detect fires using cost-effective sensors. In 2021, Artan Hysa [41] proposed a fast and cost-free method for forest fire susceptibility assessment within the wildland urban interface (WUI) in developing metropolitan areas. And Meriame mohajaneab *et al.* [42] developed five new hybrid machine learning algorithms for a forest fire susceptibility map. It can be seen that previous workstudied the characteristics of forest fire by establishing physical mechanism model, geographic statistical model and regression prediction model, but there are also shortcomings. While the previous research results either only capture one or several ignition factors or predict and simulate forest fires, the clustering model established in this paper focuses on the characteristics of ignition factors and captures and analyzes most ignition factors based on reliable data and parameter optimization.

The main goal of cluster analysis is to collect data by classifying based on similarity to yield more critical features. The most recent clustering algorithms can be divided into three categories: distance-based clustering [22], [23], density-based clustering [24], and model-based clustering [19]–[21]. Among them, the density-based clustering method has been increasingly used because it can process data with multiple shapes at the same time. By contrast, distance clustering can only deal with spherical data. In this work, we focus on a clustering method based on density and model. Therefore, there are three main directions: dimension reduction, selection of critical data, and improvement of model parameters.

In order to solve the impact of high-dimensional data on clustering, we used PCA dimension reduction as a technology to strengthen critical data [15], [16]. Furthermore, to solve the clustering bias caused by a large amount of information, we introduced and improved the loss function of distance clustering as one of the crucial parameters of GMM clustering. Therefore, we proposed a new dynamic Gaussian mixture clustering model. In order to evaluate the performance of the model, we extracted the features of the ignition factors of Brazilian rainforests [11], and the results demonstrate the efficacy of the clustering model.

III. EXPLORATORY DATA ANALYSIS (EDA)

A. GENERAL DESCRIPTION OF THE STUDY REGION (GEOGRAPHICAL FEATURES)

The chosen country is Brazil, which has more forests in South America [43]. On the Brazilian government's official website, we obtained a dataset report on the number of Brazilian forest fires divided by 26 states per month for each year from 1998 to 2017. However, some data in this dataset is missing and has reporting errors. Therefore, through data processing and data cleaning, 22 states were selected in the data set.

To study the features of Brazil's forest fires, geopandas [44], [45] and Database of Global Administrative Areas (GADM) were used to obtain the latitude and longitude of each state in Brazil and buildgeographic charts (see Fig.1) [45]. We also used the Geographic Information System (GIS) visualization rules to display the data [46], [47]. Compared with other states, Bahia, Mato Grosso, and Sao Paulo were the three states with more forest fires between 1998 and 2017. Although Amazon state's rainforest accounts for half of the world's rainforest area, it was not the state with the largest number of forest fires between 1998 and 2017. According to historical records, most forest fires were caused by persons who live around them. To access more land for grazing or farming, people destroy rainforests and clear-out the site by scorching tree trunks, branches, and greeneries. However, whether different states, seasons, and historical factors further influence the occurrence of firesremains to be determined.

B. ANALYSIS OF OUTLIERS OF THE NUMBER OF FOREST FIRES

The study aims to find out what are the critical factors contributing to forest fires in Brazil. To this end a box diagram was established based on the dataset report on the number of Brazil's forest fires of the 26 states from 1998 to 2017 (see Fig.2). In the top pannel, the abscissa is the year, and the longitudinal axis is the number of forest fires in Brazil; In the bottom pannel, the abscissa is the state, and the longitudinal axis is the number of forest fires in Brazil;

As shown in Fig.2, some abnormal values deviate from the box. When the abscissa is the year, the number of forest fires in each year was relatively stable from 1998 to 2017. However, when the abscissa is the state, the change in the number of forest fires in Bahia, Mato Grosso, Sao Paulo, Goias, and Piaui is relatively significant, and the number of forest fires in other states changed slightly. What is more, Bahia, Mato Grosso, Sao Paulo are also high-risk areas of forest fires according to the above analysis. Whether the links between states affect forest fires is worth further studying. It makes sense to retain these data to ensure a sufficient sample size to explore the causes of forest fires in Brazil. Besides, it is necessary to continue data exploration and establish mathematical models for research.

FIGURE 1. Median number of fires.

C. DATA EXPLORATION BASED ON TEMPORAL CHARACTERISTICS

According to Brazil's geographical characteristics analyzed in section 2.1, it is known that forest fires are affected by geographical factors. Therefore, the clustering models used in this work need to be built on geographical factors. At the same time, this paper also explores the influence of time factors such as seasonal features on the number of forest fires.

Based on time data (year and month), a heatmap of forest fire numberis established with abscissa as month and ordinate as year (see Figure.3).

As shown in Fig.3, numbers of forest fires in spring and winter were much less than those in autumn and summer from the year 1998 to 2017 (see Fig.3), indicating that time is also one pivotal factor that may affect forest fires.

To explore the influence of geographical features and time factors on forest fires in Brazil in detail, a machine learning model is further established for cluster analysis in our work.

IV. RESEARCH METHOD

In this paper, an improved clustering algorithm, named Principal component analysis boosted-Dynamic Gaussian mixture clustering model, is proposed, based on principal component analysis, the dynamic Gaussian mixture model, and an improved loss function of distance clustering. Therefore, in the following, the PCA and GMM as well as their corresponding improved algorithms are introduced in detail.

A. PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is a technique for statistical analysis and simplification of datasets. Besides, it is generally used to reduce the dimension of datasets while preserving the features that contribute the most to the square difference. In brief, it utilizes an orthogonal transformation to linearly transform the experiential values of a series of perhaps correlated variables to project the values of a series of linearly uncorrelated variables. These uncorrelated variables that readers can see are named principal components. Specifically, the principal component can be regarded as a linear equation containing a series of linear coefficients to indicate the projection direction. The primary method is to decompose the covariance matrix to obtain principal components [48] (i.e., eigenvectors) and their weights (i.e., eigenvalues) through eigen decomposition of the covariance matrix. The schematic diagram of the model is as follows in Fig.4.

The basic notion of PCA is to transfer the midpoint of the coordinate axis to the center of the information (gens or data) and at that juncture rotate the axis to exploit the variance of the records on the new axis, i.e., the projection of all N data individuals in this direction is the most scattered. It means more information will be retained. In this paper, PCA will reduce the dimension of five-dimensional data to two-dimensional data, as described in the following.

$X = (X_1, \ldots, X_5)^T$ is a 5-dimensional random vector. Mean is $E(X) = \mu$, and covariance is $D(X) = \sum$. Linear transformation of X is considered as the following:

$$\begin{aligned} Z_1 &= a_1' X = a_{11}X_1 + a_{21}X_2 + \ldots + a_{51}X_5 \\ Z_2 &= a_2' X = a_{12}X_1 + a_{22}X_2 + \ldots + a_{52}X_5 \\ &\cdots \\ Z_5 &= a_5' X = a_{15}X_1 + a_{25}X_2 + \ldots + a_{55}X_5 \end{aligned} \quad (1)$$

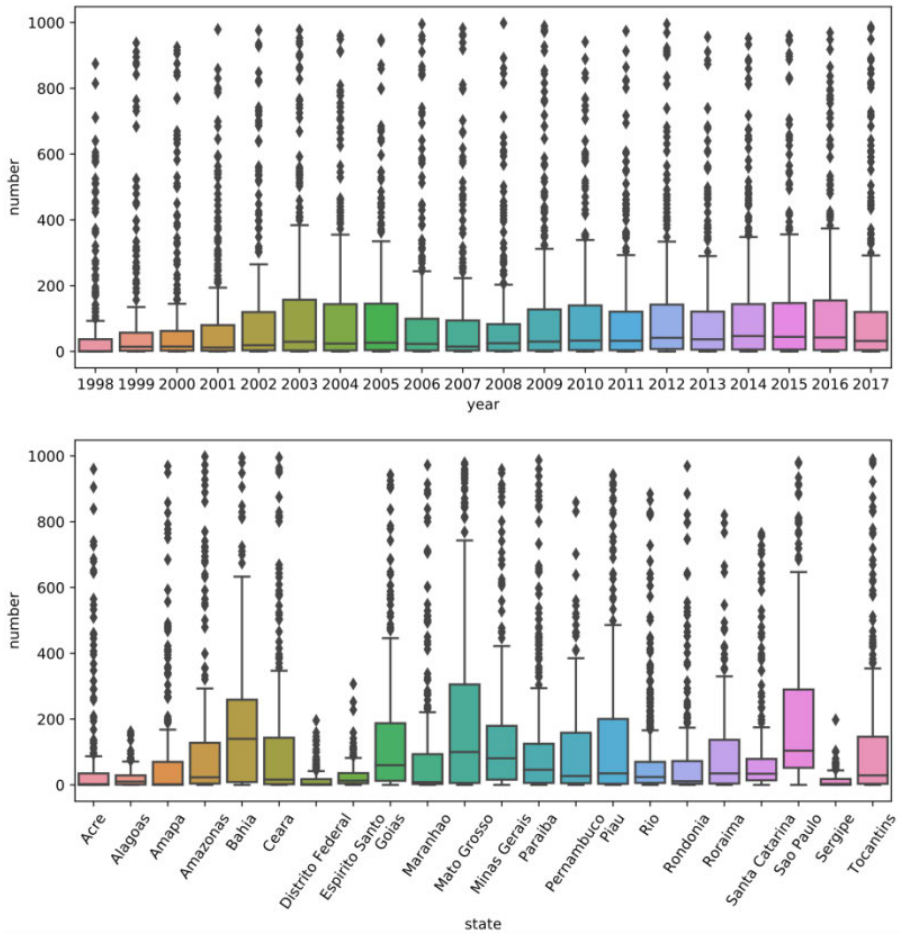


FIGURE 2. The box diagram of forest fires about year and state.

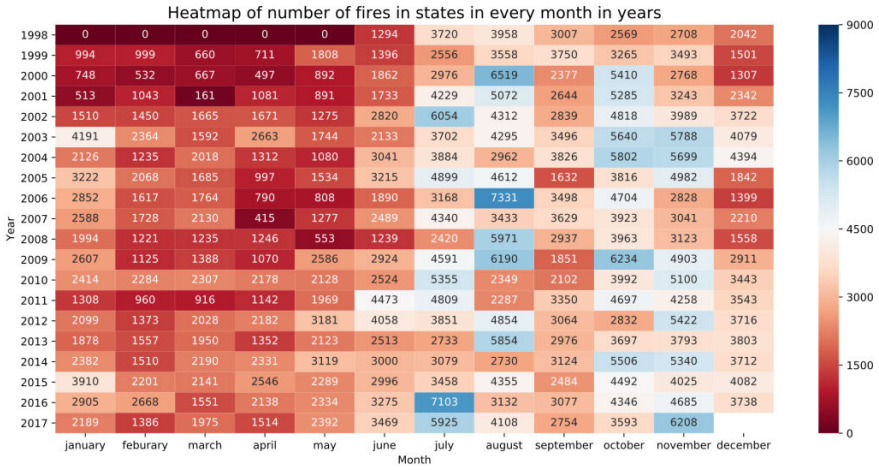


FIGURE 3. Heatmap of forest fire number about month and year.

Obviously, it is can be seen:

$$\mathrm{Var}\,(Z_i) = a_i^{'}\Sigma a_i, \quad i = 1, \ldots, 5 \tag{2}$$

$$\mathrm{Cov}\,(Z_i, Z_j) = a_i^{'}\Sigma a_j, \quad i, j = 1, \ldots, 5 \tag{3}$$

If:

(1)

$$a_i^{'}a_i = 1, \quad i = 1, \ldots, 5;$$

(2)

$$\mathrm{Cov}\,(Z_i, Z_j) = 0 (j = 1, \ldots, i-1) \text{ when } i > 1;$$

(3)

$$\mathrm{Var}\,(Z_i) = \max_{a_i a_i = 1, \mathrm{Cov}(Z_i Z_j) = 0 (j = 1, \ldots, i-1)} \mathrm{Var}\,(a'X)$$

where $Z_i = a_i'X$, i.e., the ith principal component of X.

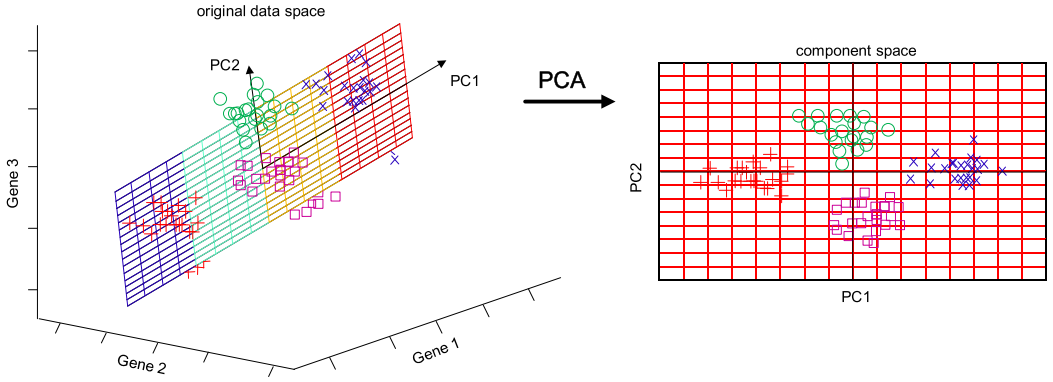


FIGURE 4. The brief general structure of Principal component analysis (PCA) in this research.

Given a sample set, it is $X_t = (x_{t1}, \ldots, x_{t5})^T$ from X.

$$X = \begin{bmatrix} x_{t1} & \ldots & x_{15} \\ \ldots & \ldots & \ldots \\ x_{n1} & \ldots & x_{n5} \end{bmatrix} = \begin{bmatrix} X_1' \\ \ldots \\ X_5' \end{bmatrix} \tag{4}$$

Therefore, the sample covariance matrix S is:

$$S = \frac{1}{n-1} \sum_{t=1}^{n} \left(X_t - \bar{X}\right)\left(X_t - \bar{X}\right)^T \stackrel{def}{=} (s_{ij})_{5\times 5} \tag{5}$$

Among:

$$\bar{X} = \frac{1}{n} \sum_{t=1}^{n} X_t \tag{6}$$

$$s_{ij} = \frac{1}{n-1} \sum_{t=1}^{n} (x_{ti} - \bar{x}_i)\left(x_{tj} - \bar{x}_j\right) \tag{7}$$

The covariance matrix $\sum Æ$ is approximately replaced by S. The eigenvalues of S are $\lambda_1 \geq \ldots \geq \lambda_5 \geq 0$, $a_i (i = 1, 2, \ldots, 5)$ is the corresponding unit orthogonal eigenvector, so that the principal component i of X is:

$$Z_i = a_i' X \tag{8}$$

B. GAUSSIAN MIXTURE MODEL (GMM)

To understand GMM, the EM algorithm needs to be introduced first. Expectation-Maximization (EM) [18], [20], [21] is a kind of maximum likelihood estimation (MLE) [49], [50]. The MLE optimization algorithm, which is usually used as an alternative to the Newton Raphson method, estimates the parameters of probability models containing latent variables or incomplete data.

The standard computing framework of the EM procedure includes the E-step and M-step. Therefore, the convergence of the EM approach can guarantee that the iteration approaches at least the local extreme extremum.

EM is an iterative technique for estimating unknown variables when some related variables are known, and its algorithm flow is as follows:

(1) Initialize distribution parameters.

(2) Repeat until convergence is achieved:

1) E Step: according to the parameters' assumed values, the unknown variables' expected estimates are given and applied to the missing values.

2) M Step: according to the estimated values of unknown variables, the maximum likelihood estimation of current parameters is given.

Based on the EM algorithm, the Gaussian mixture model is built, which is a math model composed of K number ofsingle Gaussian models. Besides, the K number of sub-

models are hidden variables of the mixture model. Therefore, much clustering information can be obtained by GMM. The schematic diagram of the algorithm is shown in Fig. 5.

Given a set of observation data generated by the Gaussian mixture model [51], [52], the following equation [53] can be satisfied.

$$p(X \mid \theta) = \sum_{c=1}^{k} \pi_c N(X|\mu_c, \sigma_c^2), \quad \sum_{c=1}^{k} \pi_c = 1 \tag{9}$$

$$\theta = \{\pi_1, \ldots, \pi_k, \mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_k^2\} \tag{10}$$

According to the dimension of the data, where $N(\mu,\sigma^2)$ shows that normal distribution with means μ and variance σ^2. π are the mixing ratio of normal distribution. k is the total number of distributions participating in mixing. The hidden variables related to the observation data are defined as $Z \rightarrow X$, and the Hidden distribution q(Z) represents the soft assignment of GMM clustering. In other words, the probability that each data comes from $c \in \{1, \ldots, k\}$ distribution. Then the hidden variable has outliers $Z = \{Z_1, \ldots, Z_k\}$.

Combined with the definition of GMM, it is brought into the calculation framework of the EM algorithm. The derivation process of the E step is as follows.

$$\begin{aligned} q^t(Z_c) &= p\left(Z_c \mid X, \theta^{(t-1)}\right) \\ &= \frac{p(X \mid Z_c, \theta)\, p(Z_c \mid \theta)}{\sum_{c=1}^{k} p(X \mid Z_c, \theta) p(Z_c \mid \theta)} \end{aligned} \tag{11}$$

For GMM contents: the calculation of the E step is as follows.

$$q^t(Z_c) = p\left(Z_c \mid X, \theta^{(t-1)}\right) = \frac{\pi_c N(X|\mu_c, \sigma_c^2)}{\sum_{c=1}^{k} \pi_c N(X|\mu_c, \sigma_c^2)} \tag{12}$$

Next, the M step calculates the model parameters' hidden variables through the E step. The calculation of the M step

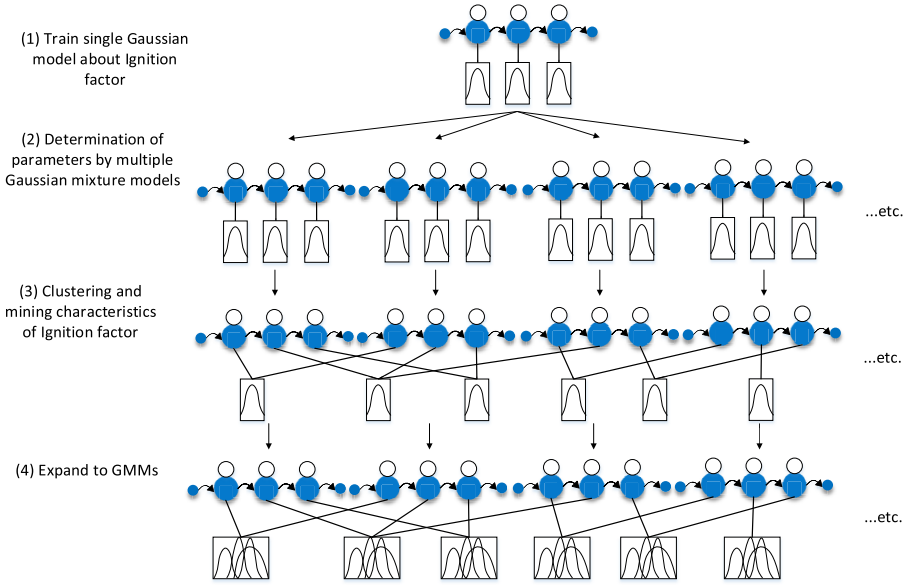


FIGURE 5. The general structure of the Gaussian mixture model (GMM) in this research. Structure of the Gaussian mixture model (GMM): First, train single Gaussian model about ignition; Second, determine parameters by multiple Gaussian mixture models; Third, do clustering and mining characteristics of ignition factor. Finally, expand to Gaussian mixture models.

solves parameter optimization. The derivation process of the M step is as follows.

$$\begin{aligned} & E_{q^{(t)}}\left[\log p\left(X, Z \mid \theta\right)\right] \\ & = \sum_{i=1}^{N} \sum_{c=1}^{k} q\left(Z_c\right) \log p\left(X_i, Z_c \mid \theta\right) \\ & = \sum_{i=1}^{N} \sum_{c=1}^{k} q\left(Z_c\right) \log \left[\pi N\left(X_i \mid \mu_c, \sigma_c^2\right)\right] \end{aligned} \tag{13}$$

It can perform the M step's computational procedure by introducing the analytic form of univariate normal distribution and gaining partial derivation of model parameters in (14), as shown at the bottom of the next page.

Through PCA dimension reduction, GMM clustering was used to study ignition factors of Brazil's rainforest. Next, the detailed improved strategies and experimental results will be described and discussed.

C. THE PROPOSED PRINCIPAL COMPONENT ANALYSIS-BOOSTED DYNAMIC GAUSSIAN MIXTURE CLUSTERING MODEL (PCA-DGM)

Firstly, according to the forest fires report in Brazil in the dataset, the records were divided into different groups by state between 1998 and 2017. Then, all the 6215 lines of records were preprocessed through fault elimination, de-duplication, and outlier analysis. Finally, 6183 lines were obtained.

Secondly, in this paper, the year and month data were used to reflect the influence of time, the longitude and latitude data were used to reflect geographical influence, and the forest fires report in Brazil was used to reflect fire severity. We integrated the data into a five-dimensional vector. By using the Brazilian states' longitudes and latitudes and the number of forest fires reported in Brazil divided by state, a five-dimensional vector U = (year, month, number, latitude, longitude) is constructed for this study. For simplicity, it uses x as latitude and y as longitude and we can get U = (year, month, number,x, y). Finally, this paper uses PCA described in section 3.1 to reduce the five-dimensional vector group into two dimensions and influences each element in the vector group on the two principal components. Thus, the two principal components can replace the vector group.

According to the results in Table 1, the following linear expression is obtained.

$$\begin{aligned} \mathrm{Z}_1 = & \ 0.\mathrm{od}1046\mathrm{year} + 0.3788\mathrm{month} \\ & + 0.5136\mathrm{number} - 0.5744x + 0.5019\mathrm{y} \end{aligned} \tag{15}$$

$$\begin{aligned} \mathrm{Z}_2 = & \ 0.1614\mathrm{year} + 0.5608\mathrm{month} \\ & + 0.4960\mathrm{number} + 0.4099x - 0.4954y \end{aligned} \tag{16}$$

It can be seen that, for the first principal component, the primary influence is latitude, longitude, and the number of fires; and for the second principal component, the primary influence is month, longitude, and the number of fires. This shows that the number of forest fires, seasonal factors, and geographical environments significantly influence the two principal components.

Next, a clustering model called Principal components analysis-boosted Dynamic Gaussian mixture clustering model (PCA-DGM) was proposed, which is the above principal component data after dimension reduction. Using dynamically selected density parameters or distance parameters, the proposed optimized GMM clustering analysis was then carried out. Generally, these records are considered to obey Gaussian distribution. Therefore, it is crucial to select

TABLE 1. Coefficient of principal component.

Index / Coefficient / Principal component	year	month	number	x	y
The first principal component	0.1046	0.3788	0.5136	-0.5744	0.5019
The second principal component	0.1614	0.5608	0.4960	0.4099	-0.4954

feasible and accurate metrics for testing and efficacy modeling [3], [54], [55].

The output result can be obtained by Eq. (17):

$$f = p(Z \mid \theta) = \sum_{c=1}^{k} \pi_c N(X|\mu_c, \sigma_c^2) \tag{17}$$

where f is the probability of each principal component data aggregating in the same class. According to f values to the cluster, the model is called the principal component analysis-Gaussian mixture clustering model (PCA-GMM), i.e., the clustering model of GMM will be performed after dimension reduction of the records.

In the paper, the loss function of the distance clustering model (such as the K-Means model [56], [57]) was introduced into our work. Its establishment process and properties are as follows.

The loss function of the K-Means clustering model is:

$$J\left(c^1, \ldots, c^m, u_1, \ldots, u_s\right) = \frac{1}{m} \sum_{i=1}^{m} \left\| Z - u_{c^i} \right\|^2 \tag{18}$$

where u_{c^i} is the nearest cluster center to Z. The objective function of distance clustering is:

$$J' = \min_{c^i \epsilon c^1, \ldots, c^m} \frac{1}{m} \sum_{i=1}^{m} \left\| Z - u_{c^i} \right\|^2 \tag{19}$$

The objective function can gain optimal c^i and u_{c^i}.

So the distance from a data point to a cluster point is:

$$Q = \left\| Z - u_{c^i} \right\|^2 \tag{20}$$

Obtaining the distance between the two principal components data is:

$$D = \left\| Z^{(k)} - Z^{(t)} \right\|^2, \quad k \neq t \tag{21}$$

where $Z^{(k)}$, $Z^{(t)}$ are principal component data.

Based on the above proposed f, the model adds an essential parameter D, so the output f is changed to F:

$$F = p(Z_i \mid \theta) \quad \text{and } D \tag{22}$$

According to F, it is stipulated that:

(1) If PCA-GMM determines that the clustering probability in one class for two samples is less than 50% and D < Q, it will output D as the determining parameter of clustering, i.e., the two data are clustered by distance.

(2) If PCA-GMM determines that the clustering probability in one class for two samples is less than 50%, and D ≥ Q, it will output $p(Z_i \mid \theta)$ as the determining parameter of clustering, i.e., the two data are clustered by probability.

(3) If PCA-GMM determines that the probability of two data clustering in one class is more than 50%, and D ≥ Q, it will output $p(Z_i \mid \theta)$ as the determining parameter of clustering, i.e., the two data are clustered by probability.

(4) If PCA-GMM determines that the probability of two data clustering in one class is more than 50%, and D < Q, it will output $p(Z_i \mid \theta)$ as the determining parameter of clustering, i.e., the two data are clustered by probability.

$$\begin{aligned}
\max_{\theta} E_{q^{(t)}}\left[\log p(X, Z \mid \theta)\right] &\Longleftrightarrow \max_{\theta} \sum_{c=1}^{k} \sum_{i=1}^{N} \left[q(Z_c) \log \pi_c - \frac{X_i - \mu_c}{2\sigma_c^2}\right] \\
&\Rightarrow \frac{\partial}{\partial \mu_c} \sum_{i=1}^{N} \left[q(Z_c) \log \pi_c - \frac{X_i - \mu_c}{2\sigma_c^2}\right] = 0 \Rightarrow \mu_c = \frac{\sum_{i=1}^{N} q(Z_c) X_i}{\sum_{i=1}^{N} q(Z_c)} \\
&\Rightarrow \frac{\partial}{\partial \sigma_c} \sum_{i=1}^{N} \left[q(Z_c) \log \pi_c - \frac{X_i - \mu_c}{2\sigma_c^2}\right] = 0 \Rightarrow \sigma_c^2 = \frac{\sum_{i=1}^{N} q(Z_c) (X_i - \mu_c)^2}{\sum_{i=1}^{N} q(Z_c)} \\
&\Rightarrow \frac{\partial}{\partial \pi_c} \sum_{i=1}^{N} \left[q(Z_c) \log \pi_c - \frac{X_i - \mu_c}{2\sigma_c^2}\right] = 0 \Rightarrow \pi_c = \frac{\sum_{i=1}^{N} q(Z_c)}{N}
\end{aligned} \tag{14}$$

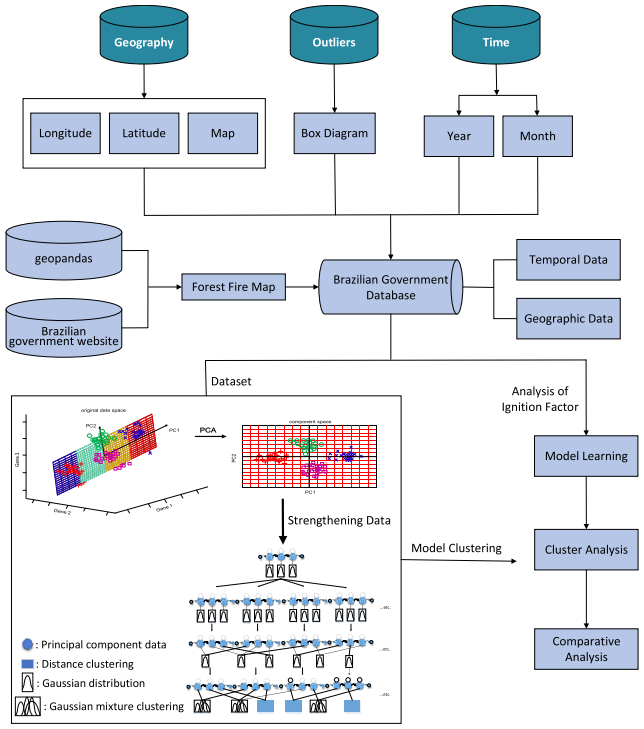


FIGURE 6. Dynamic Gaussian mixture clustering model based on PCA enhancement.

PCA-GMM is also based on EM algorithm for parameter estimation. If there is no stop threshold, the EM algorithm will infinitely optimize the cluster allocation to achieve infinite accuracy. Therefore, the theoretical running time of EM is infinite. However, once there is a stop condition, its complexity should be $O(MN\hat{}3)$, where M is the number of iterations and N is the number of parameters. Pca-dgm adds new parameters, but its computational complexity is not high. Therefore, the complexity of this method is close to that of EM algorithm.

According to the above four rules, PCA-DGM can dynamically select better parameters for clustering by comparing the output probability with the distance value. The proposed PCA-DGM model flow is shown in detail in Fig. 6.

V. EXPERIMENTS AND RESULTS

Cluster experiments were carried out on the data of principal components using PCA-DGM. The parameter selection of PCA-DGM is related to the likelihood function. In order to ensure clustering effect, Akaike Information Criterion (AIC) [58], [59] and Bayesian Information Criterion (BIC) [60] were used to determine the number of parameters of the GMM clustering model. That is, K is solved by AIC and BIC. The calculation formula of AIC and BIC are as follows.

$$\mathrm{AIC} = 2\mathrm{k} - 2\ln(\mathrm{L}) \tag{23}$$

$$\mathrm{BIC} = \mathrm{k}\ln(\mathrm{n}) - 2\ln(\mathrm{L}) \tag{24}$$

where k is the number of parameters, and L is the likelihood function.

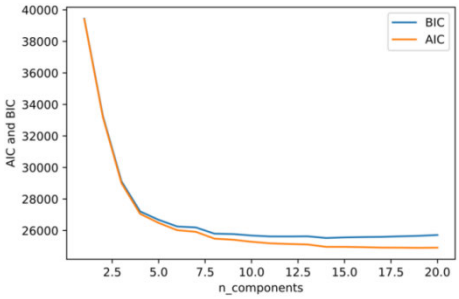


FIGURE 7. Parameter adjustment curve based on AIC and BIC.

TABLE 2. Number of clusters per class.

Class	Count
3	2374
1	2170
0	1161
2	478

Small values of AIC and BIC [58], [60] indicate that the number of clusters is better because they can gain better parameters. sklearn was used to obtain the values of AIC and BIC (see Fig. 7). As can be seen from Fig.7, when the number of clusters was greater than 2, AIC and BIC decreased. However, if the number of clusters was set to greater than 10, it led to the situation that the categories were not clear enough. The number of clusters was determined to be in the range of 4-10, and it is clear that AIC and BIC values are relatively stable in this range. In addition, It was found that AIC and BIC values of 4-10 were stable, and thus the optimal number of clusters was set as 4.

To sum up, we determine that the number of K-values of the GMM model is 4 using the EM algorithm, and that the distance value is taken as an essential parameter. Finally, PCA-DGM was established to study the ignition factors of Brazilian rainforests.

Subsequently, PCA-DGM was built for clustering analysis using the obtained optimal number of clusters. In the experiment, two principal component data were clustered into four categories (see Fig. 8 (a)). To gather the clustering categories more clearly, some evenly distributed data were randomly generated, and the cluster shape was ovalized by dot product with principal component data (see Fig. 8 (b)).

Specifically, the four clusters of PCA-DGM were called class 0, class 1, class 2, and class 3. PCA-DGM separated the four clusters. There was no data mixing. Two charts show that the clustering effect of the improved GMM is excellent and achieves the expected results.

Table 2 shows the number of clusters in each class. As can be seen from the table, class 3 ranks first, followed by class 1. Therefore, these clusters should be focused on..

Next, based on numbers of clusters, a histogram was generated with month, year, and the state as abscissa and quantity

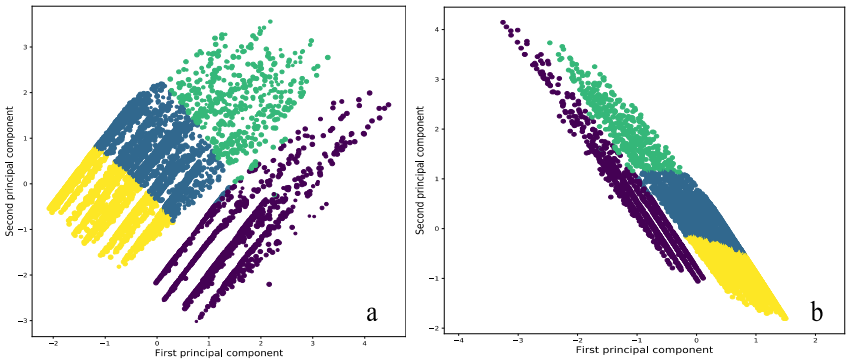


FIGURE 8. Clustering results of the PCA-DGM model: (a) Normal results of the PCA-DGM clustering model and (b) the results of the PCA-DGM clustering model are ovalized.

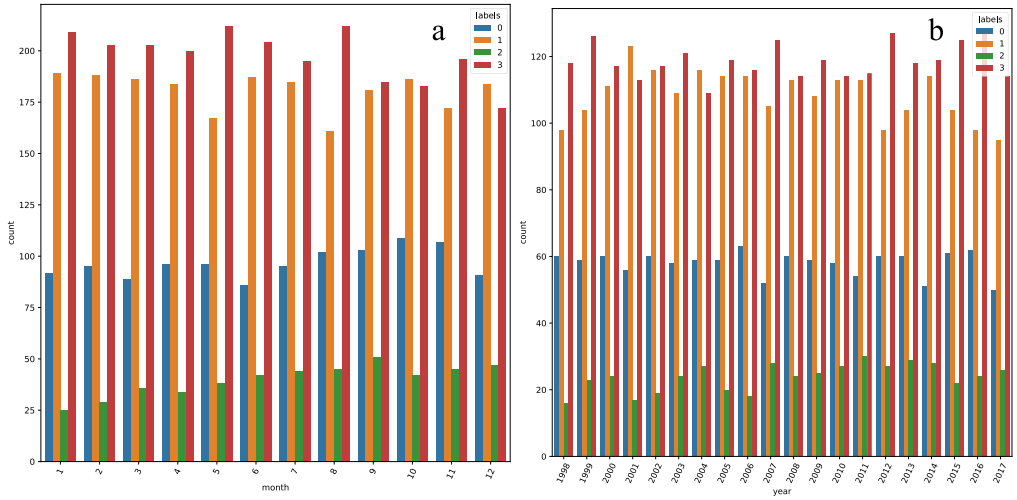


FIGURE 9. Clustering result bar charts based on principal component data: (a) month and (b) year.

as ordinate (see Figs. 9 and 10). In these bar charts, much information can be gained.

(1) In the bar chart about the month, in October and December, most of the data is in cluster 1, indicating that the states in class 1 were more likely to have forest fires during this period (October and December). Also, it showed that states in class 3 were more likely to have forest fires in other months.

(2) In the bar chart about the year, principal component data clustered in classes 1 and 3 between 1998 and 2017. Therefore, we focused on these two classes to study the relations between fires and the geographical and temporal characteristics of the states.

(3) In the bar chart about the state, what kind of class each state belongs to can be seen. For example, Acre, Alagoas, and Amapa all belong to class 3, whereas Rio de Janeiro and Rondonia belong to class 0, 1, 2. In addition, clustering results of 22 states with principal components can be calculated. The histogram (In Fig.10) can be used to study which class has the most fires in each state. That is, the clustering results can show the differences of each state. Moreover, the geographical and temporal influence on forest fires in each state can be mined through specific data.

In the following, the specific data of month, year, and state histogram were output and presented in a table (see Table 3).

(1) For the state, it can be found that in the same class, the state with more forest fires and the states closes to it also have more forest fires. For example, Bahia is near Distrito Federal, and they are both in class 1 and class 3. Amazonas is near Acre and Amapa, and most of them are in class 3. Notably, most of Tocantins' information is clustered in class 2, which indicates that the characteristics of forest fires in Tocantins are much different from those in other states, and it is less affected by other states. Therefore, the study of one state in the same category can help to study forest fires in other states and other regions to a certain extent.

(2) For the year, the principal component data focus on class 1 and class 3, so it is significant by focusing on forest fires in classes 1 and 3. However, The number of classes 0 and 2 is small, and the amount of information contained is also small. Therefore, the data of only recent 5-10 years were considered.

TABLE 3. The number of clusters in the four classes.

state	L0	L1	L2	L3	year	L0	L1	L2	L3	month	L0	L1	L2	L3
Acre	0	0	0	239	1998	60	98	16	118	1	92	189	25	209
Alagoas	0	0	0	239	1999	59	104	23	126	2	95	188	29	203
Amapa	0	0	0	239	2000	60	111	24	117	3	89	186	36	203
Amazonas	0	27	0	212	2001	56	123	17	113	4	96	184	34	200
Bahia	0	51	0	188	2002	60	116	19	117	5	96	167	38	212
Ceara	0	66	0	173	2003	58	109	24	121	6	86	187	42	204
Distrito Federal	0	63	0	176	2004	59	116	27	109	7	95	185	44	195
Espirito Santo	0	68	0	171	2005	59	114	20	119	8	102	161	45	212
Goias	0	97	0	142	2006	63	114	18	116	9	103	181	51	185
Maranhao	0	103	0	136	2007	52	105	28	125	10	109	186	42	183
Mato Grosso	0	266	0	207	2008	60	113	24	114	11	107	172	45	196
Minas Gerais	0	170	0	69	2009	59	108	25	119	12	91	184	47	172
Paraiba	5	351	0	116	2010	58	113	27	114					
Pernambuco	13	170	0	56	2011	54	113	30	115					
Piaui	29	197	2	11	2012	60	98	27	127					
Rio de Janeiro	199	442	56	0	2013	60	104	29	118					
Rondonia	142	51	46	0	2014	51	114	28	119					
Roraima	145	34	60	0	2015	61	104	22	125					
Santa Catarina	142	14	83	0	2016	62	98	24	128					
Sao Paulo	236	0	3	0	2017	50	95	26	114					
Sergipe	239	0	0	0										
Tocantins	11	0	228	0										

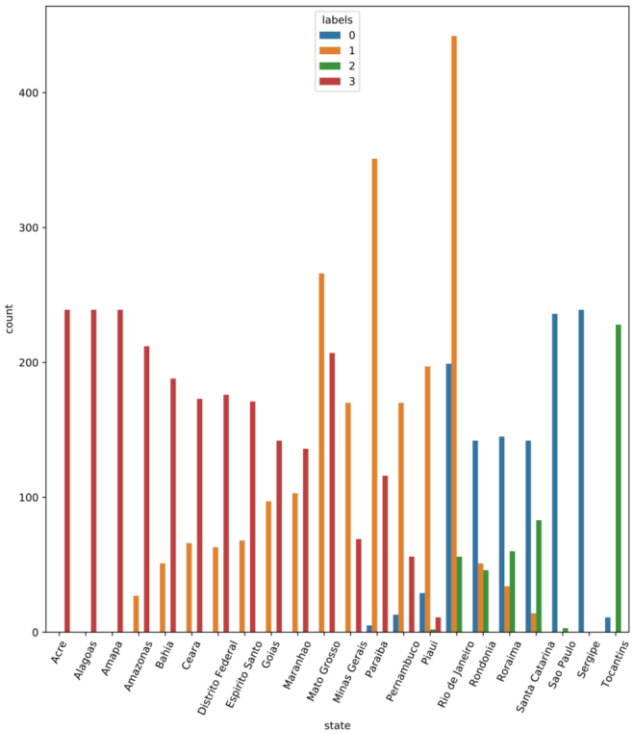


FIGURE 10. State clustering result bar chart based on principal component data.

(3) For the month, it involves seasonal factors. Therefore, according to the table's specific data, this result can focus on this category according to which category has the largest number in each month At the same time, data mining was performed with the specific data of each class. For example, January's principal component dataare concentrated in class 3 (209 in total), so researchers can study them according to the specific data (year, month, number of forest fires, longitude, and latitude).

VI. EXPERIMENTS AND RESULTS

PCA-DGM is an improved clustering model that combines the density clustering model and the loss function of distance clustering. In order to evaluate the clustering effect, the PCA-DGM model was compared with PCA-GMM, the typical density clustering models (the original GMM and DBSCAN) and the distance clustering model (K-Means [25], [56]). The results show that PCA-DGM is better than other clustering algorithms in terms of clustering effect. Moreover, PCA-DGM can obtain more forest fire characteristics in each state (see Fig. 11).

The results show that PCA-DGM is excellent at separating the four types of data, whereas K-means clustering results are overlapped. The reason is that K-means clustering is a distance-based clustering algorithm that can only deal with spherical data. DBSCAN clustering cannot solve all the clustering problems of non-spherical structure (for example, the Brazilian forest fire studied in this paper). When the clustering problem is aspheric, the clustering algorithms based on distance and DBSCAN have a poor clustering effect. Although PCA-GMM clustering achieved a better result, some discrete points were occurring for achieving a better clustering result. By contrast, PCA-DGM can process

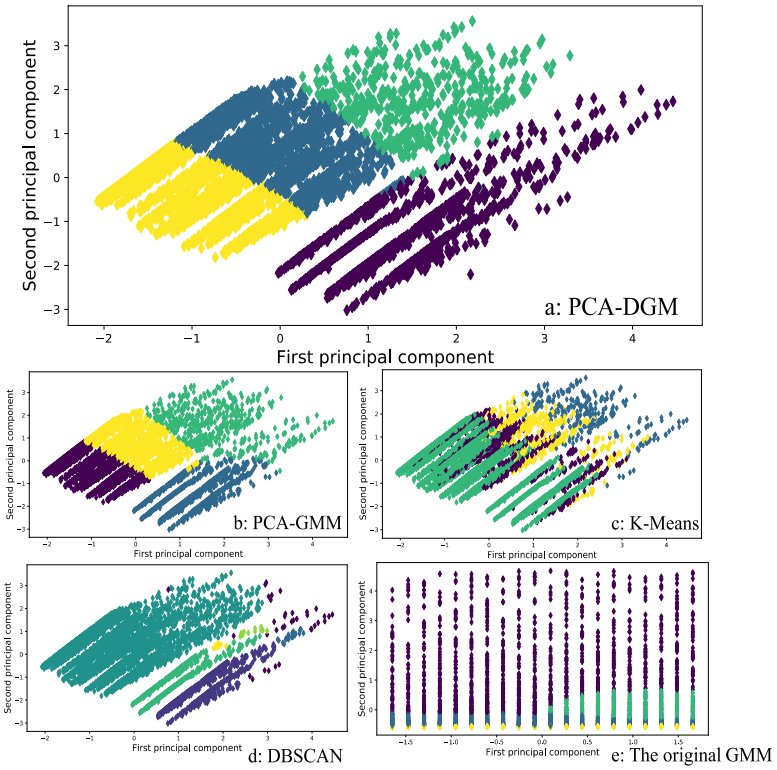


FIGURE 11. Comparison of experimental results of five clustering models: (a) PCA-DGM clustering model, (b) PCA-GGM clustering: Dimension reduction Gaussian mixture clustering model based on probability clustering, (c) K-Means clustering: Distance clustering Model, (d) DBSCAN clustering: Density clustering Model and (e) the original GMM clustering: Gaussian mixture clustering model containing only the number of years and records of forest fires in Brazil.

complex spherical data and cluster them. It perfectly solves the problem of the ignition factor data of Brazilian rainforests in this paper. The clustering results show that the reported forest fire records are between spherical and non-spherical. Because the proposed PCA-DGM algorithm is a dynamic clustering algorithm based on density and distance that can find clusters of arbitrary shapes, it is better than the distance clustering and density clustering models in terms of clustering effect. From the above experiment results, it is believed that PCA-DGM can better solve the clustering problem of forest fires in Brazil. Moreover, deep learning approaches can be introduced in the future to boost the model's performance.

VII. CONCLUSION

In recent years, Brazil's rainforests have been increasingly damaged by natural disasters due to both climate changes (such as seasonal factors) and human activities (such as deforestation). Therefore, it is a practical need to establish a set of models to study the factors contributing to forest fire occurrence in Brazil. This study proposed a new hybrid machine learning framework, which uses the PCA-enhanced GMM model to achieve this goal. First, PCA was used to strengthen the data, and it was added to the GMM structure. Numbers of forest fires in Brazil from 1998 to 2017 and the longitude and latitude data in the GADM database (shape: 6215×5, processed: 6183×5) were used to form 5-Dimensional data. Next, PCA was used to strengthen the data into 2-D principal component data. Then, GMM clustering using PCA was adjusted to improve the performance based on the EM algorithm called the PCA-DGM model.

To test the performance of the proposed PCA-DGM model, 6183 lines principal component data after data processing were used for clustering experiments, and PCA-DGM was compared with PCA-GMM, the traditional clustering algorithm K-Means clustering, the original GMM, and DBSCAN. The experimental results show that PCA-DGM is better than the K-Means benchmark model and can deal with clusters with arbitrary shapes. Therefore, the newly developed PCA-DGM can be used as a valuable tool for studying forest fires' ignition factors in Brazilian states, including geographical environment factors (interstate influence) and time characteristics (seasonal factors), and thus help local authorities to carry out forest fire prevention work more effectively. Furthermore, due to the excellent clustering effect of PCA-DGM in Brazil forest fires, the proposed enhanced machine learning model can be applied to other fields and data with multi-shape outside the forest fire area. In the future, this study's extensions will include the use of more advanced feature selection [61]–[65] and the integration of multiple clustering models to obtain more feature information. Moreover,

the recently proposed metaheuristics [66]–[69] can also be employed to optimize the model further.

DATA AND COMPUTER CODE AVAILABILITY

Brazilian government (http://dados.gov.br/dataset/sistema-nacional-de-informacoes-florestais-snif) and Database of Global Administrative Areas (https://gadm.org/download_country_v3.html). Name of code: PCA-DGM, developer: Hongliang, telephone: +8618318217999, email: huanghongliang2020@126.com, year first available: 2021, hardware required: i5 CPU and 4G RAM, the software required: win7 or win10, program language: Python, program: 1.5M. Readers can access the code by checking it through the website in GitHub (https://github.com/JacksonHuang-yoo/PCA-DGM/tree/main).

REFERENCES

[1] Z. Zhang, C. Luo, and Z. Zhao, "Application of probabilistic method in maximum tsunami height prediction considering stochastic seabed topography," *Natural Hazards*, vol. 104, no. 3, pp. 2511–2530, Dec. 2020.
[2] Q. Quan, S. Gao, Y. Shang, and B. Wang, "Assessment of the sustainability of *Gymnocypris eckloni* habitat under river damming in the source region of the Yellow River," *Sci. Total Environ.*, vol. 778, Jul. 2021, Art. no. 146312.
[3] D. Yu, Y. Mao, B. Gu, S. Nojavan, K. Jermsittiparsert, and M. Nasseri, "A new LQG optimal control strategy applied on a hybrid wind turbine/solid oxide fuel cell/in the presence of the interval uncertainties," *Sustain. Energy, Grids Netw.*, vol. 21, Mar. 2020, Art. no. 100296.
[4] A. Shenoy, J. F. Johnstone, E. S. Kasischke, and K. Kielland, "Persistent effects of fire severity on early successional forests in interior Alaska," *Forest Ecol. Manage.*, vol. 261, no. 3, pp. 381–390, Feb. 2011.
[5] A. S. Cheng and L. Dale, "Achieving adaptive governance of forest wildfire risk using competitive grants: Insights from the Colorado wildfire risk reduction grant program," *Rev. Policy Res.*, vol. 37, no. 5, pp. 657–686, Sep. 2020.
[6] J. R. Lima, V. F. Mansano, and F. S. Araújo, "Coexistence and geographical distribution of Leguminosae in an area of Atlantic forest in the semi-arid region of Brazil," *J. Syst. Evol.*, vol. 50, no. 1, pp. 25–35, Jan. 2012.
[7] K. Zhang, Y. Zhang, and J. Tao, "Predicting the potential distribution of *Paeonia veitchii* (Paeoniaceae) in China by incorporating climate change into a Maxent model," *Forests*, vol. 10, no. 2, p. 190, Feb. 2019.
[8] J. F. Johnstone, F. S. Chapin, T. N. Hollingsworth, M. C. Mack, V. Romanovsky, and M. Turetsky, "Fire, climate change, and forest resilience in interior Alaska," *Can. J. Forest Res.*, vol. 40, no. 7, pp. 1302–1312, 2010.
[9] L. He, Y. Chen, and J. Li, "A three-level framework for balancing the tradeoffs among the energy, water, and air-emission implications within the life-cycle shale gas supply chains," *Resour., Conservation Recycling*, vol. 133, pp. 206–228, Jun. 2018.
[10] J. Doyle, "Picturing the clima(c)tic: Greenpeace and the representational politics of climate change communication," *Sci. Culture*, vol. 16, no. 2, pp. 129–150, 2007.
[11] J. Ning, D. Guoming, and F. Meng, "Transformation and fragmentation of tropical rainforest landscape in Brazil," *Geograph. Res.*, vol. 30, no. 3, pp. 780–794, 2015.
[12] D. Nepstad, G. Carvalho, A. C. Barros, A. Alencar, J. P. Capobianco, J. Bishop, P. Moutinho, P. Lefebvre, U. L. Silva, and E. Prins, "Road paving, fire regime feedbacks, and the future of Amazon forests," *Forest Ecol. Manage.*, vol. 154, no. 3, pp. 395–407, Dec. 2001.
[13] J. J. Camarero, G. Sangüesa-Barreda, C. Montiel-Molina, F. Seijo, and J. A. López-Sáez, "Past growth suppressions as proxies of fire incidence in relict Mediterranean black pine forests," *Forest Ecol. Manage.*, vol. 413, pp. 9–20, Apr. 2018.
[14] F. D. V. Pegas and J. G. Castley, "Ecotourism as a conservation tool and its adoption by private protected areas in Brazil," *J. Sustain. Tourism*, vol. 22, no. 4, pp. 604–625, May 2014.
[15] F. R. S. K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *London, Edinburgh, Dublin Philosoph. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901.
[16] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, no. 6, p. 417, 1933.
[17] X.-F. Wang, P. Gao, Y.-F. Liu, H.-F. Li, and F. Lu, "Predicting thermophilic proteins by machine learning," *Current Bioinf.*, vol. 15, no. 5, pp. 493–502, Oct. 2020.
[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B, Methodol.*, vol. 39, no. 1, pp. 1–22, 1977.
[19] R. V. Hogg, J. McKean, and A. T. Craig, *Introduction to Mathematical Statistics*. London, U.K.: Pearson, 2005.
[20] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Comput.*, vol. 6, no. 2, pp. 181–214, 1994.
[21] B. Muthén and K. Shedden, "Finite mixture modeling with mixture outcomes using the EM algorithm," *Biometrics*, vol. 55, no. 2, pp. 463–469, Jun. 1999.
[22] H. Abdi, "Additive-tree representations," in *Trees and Hierarchical Structures* (Lecture Notes in Biomathematics), vol. 84. Berlin, Germany: Springer, 1990, pp. 43–59.
[23] J. Clatworthy, D. Buick, M. Hankins, J. Weinman, and R. Horne, "The use and reporting of cluster analysis in health psychology: A review," *Brit. J. Health Psychol.*, vol. 10, no. 3, pp. 329–358, Sep. 2005.
[24] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *Int. Comput. Sci. Inst.*, vol. 4, no. 510, p. 126, 1998.
[25] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *J. Roy. Stat. Soc. C, Appl. Statist.*, vol. 28, no. 1, pp. 100–108, 1979.
[26] N. Shental *et al.*, "Gaussian mixture models with equivalence constraints," 2009.
[27] C. Vasilakos, K. Kalabokidis, J. Hatzopoulos, and I. Matsinos, "Identifying wildland fire ignition factors through sensitivity analysis of a neural network," *Natural Hazards*, vol. 50, no. 1, pp. 125–143, Jul. 2009.
[28] D. G. Woolford, J. Cao, C. B. Dean, and D. L. Martell, "Erratum: Characterizing temporal changes in forest fire ignitions: Looking for climate change signals in a region of the Canadian boreal forest," *Environmetrics*, vol. 22, no. 3, p. 485, 2011.
[29] N. P. Cheney, J. S. Gould, W. L. McCaw, and W. R. Anderson, "Predicting fire behaviour in dry eucalypt forest in southern Australia," *Forest Ecol. Manage.*, vol. 280, pp. 120–131, Sep. 2012.
[30] N. Arndt, H. Vacik, V. Koch, A. Arpaci, and H. Gossow, "Modeling human-caused forest fire ignition for assessing forest fire danger in Austria," *iForest-Biogeosci. Forestry*, vol. 6, no. 6, p. 315, 2013.
[31] M. Rodrigues, J. de la Riva, and S. Fotheringham, "Modeling the spatial variation of the explanatory factors of human-caused wildfires in Spain using geographically weighted logistic regression," *Appl. Geogr.*, vol. 48, pp. 52–63, Mar. 2014.
[32] L. O. Bianchi and G. E. Defossé, "Live fuel moisture content and leaf ignition of forest species in Andean Patagonia, Argentina," *Int. J. Wildland Fire*, vol. 24, no. 3, pp. 340–348, 2015.
[33] F. Guo, Z. Su, G. Wang, L. Sun, F. Lin, and A. Liu, "Wildfire ignition in the forests of southeast China: Identifying drivers and spatial distribution to predict wildfire likelihood," *Appl. Geogr.*, vol. 66, pp. 12–21, Jan. 2016.
[34] M. M. Müller and H. Vacik, "Characteristics of lightnings igniting forest fires in Austria," *Agricult. Forest Meteorol.*, vols. 240–241, pp. 26–34, Jun. 2017.
[35] J. Ruffault and F. Mouillot, "Contribution of human and biophysical factors to the spatial distribution of forest fire ignitions and large wildfires in a French Mediterranean region," *Int. J. Wildland Fire*, vol. 26, no. 6, pp. 498–508, 2017.
[36] N. Read, T. J. Duff, and P. G. Taylor, "A lightning-caused wildfire ignition forecasting model for operational use," *Agricult. Forest Meteorol.*, vols. 253–254, pp. 233–246, May 2018.
[37] V. Sevinc, O. Kucuk, and M. Goltas, "A Bayesian network model for prediction and analysis of possible forest fire causes," *Forest Ecol. Manage.*, vol. 457, Feb. 2020, Art. no. 117723.
[38] J. R. Molina, M. A. Herrera, and F. R. Y. Silva, "Wildfire-induced reduction in the carbon storage of Mediterranean ecosystems: An application to brush and forest fires impacts assessment," *Environ. Impact Assessment Rev.*, vol. 76, pp. 88–97, May 2019.
[39] F. T. Couto, M. Iakunin, R. Salgado, P. Pinto, T. Viegas, and J.-P. Pinty, "Lightning modelling for the research of forest fire ignition in Portugal," *Atmos. Res.*, vol. 242, Sep. 2020, Art. no. 104993.

[40] N. Verma and D. Singh, "Analysis of cost-effective sensors: Data fusion approach used for forest fire application," *Mater. Today, Proc.*, vol. 24, pp. 2283–2289, Jan. 2020.

[41] A. Hysa, "Indexing the vegetated surfaces within WUI by their wildfire ignition and spreading capacity, a comparative case from developing metropolitan areas," *Int. J. Disaster Risk Reduction*, vol. 63, Sep. 2021, Art. no. 102434.

[42] M. Mohajane, R. Costache, F. Karimi, Q. B. Pham, A. Essahlaoui, H. Nguyen, G. Laneve, and F. Oudija, "Application of remote sensing and machine learning algorithms for forest fire mapping in a Mediterranean area," *Ecol. Indicators*, vol. 129, Oct. 2021, Art. no. 107869.

[43] B. Soares-Filho, R. Rajão, M. Macedo, A. Carneiro, W. Costa, M. Coe, H. Rodrigues, and A. Alencar, "Cracking Brazil's forest code," *Science*, vol. 344, no. 6182, pp. 363–364, Apr. 2014.

[44] D. Ozturk, A. Chaudhary, P. Votava, and C. Kotfila, "GeoNotebook: Browser based interactive analysis and visualization workflow for very large climate and geospatial datasets," *AGUFM*, vol. 2016, p. IN53A-1876, 2016.

[45] Y. Tang, "Big data analytics of taxi operations in New York City," *Amer. J. Oper. Res.*, vol. 9, no. 4, pp. 192–199, 2019.

[46] L. He, J. Shen, and Y. Zhang, "Ecological vulnerability assessment for ecological conservation and environmental management," *J. Environ. Manage.*, vol. 206, pp. 1115–1125, Jan. 2018.

[47] K. Zhang, Q. Wang, L. Chao, J. Ye, Z. Li, Z. Yu, T. Yang, and Q. Ju, "Ground observation-based analysis of soil moisture spatiotemporal variability across a humid to semi-humid transitional zone in China," *J. Hydrol.*, vol. 574, pp. 903–914, Jul. 2019.

[48] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

[49] P. H. Moose, "A technique for orthogonal frequency division multiplexing frequency offset correction," *IEEE Trans. Commun.*, vol. 42, no. 10, pp. 2908–2914, Oct. 1994.

[50] Q. Zhao, P. Sur, and E. J. Candès, "The asymptotic distribution of the MLE in high-dimensional logistic models: Arbitrary covariance," 2020, *arXiv:2001.09351*. [Online]. Available: http://arxiv.org/abs/2001.09351

[51] S. Hadavi, "Predicting activity noise levels in occupied classrooms by means of cluster analysis," M.S. thesis, Gina Cody School Eng. Comput. Sci., Concordia Univ., Montreal, QC, Canada, 2020.

[52] D. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*. Boston, MA, USA: Springer, 2008.

[53] L. Hang, *Statistical Learning Method*. Beijing, China: Tsinghua Univ. Press, 2012.

[54] Z. Zhang, M. Liu, M. Zhou, and J. Chen, "Dynamic reliability analysis of nonlinear structures using a Duffing-system-based equivalent nonlinear system method," *Int. J. Approx. Reasoning*, vol. 126, pp. 84–97, Nov. 2020.

[55] C. Yang, F. Gao, and M. Dong, "Energy efficiency modeling of integrated energy system in coastal areas," *J. Coastal Res.*, vol. 103, pp. 995–1001, Jul. 2020.

[56] M. Capó, A. Pérez, and J. A. Lozano, "An efficient K-means clustering algorithm for tall data," *Data Mining Knowl. Discovery*, vol. 34, pp. 776–811, Jul. 2020.

[57] W.-L. Zhao, C.-H. Deng, and C.-W. Ngo, "K-means: A revisit," *Neurocomputing*, vol. 291, pp. 195–206, May 2018.

[58] H. E. Brown, M. S. Doyle, J. Cox, R. J. Eisen, and R. S. Nasci, "The effect of spatial and temporal subsetting on *Culex tarsalis* abundance models—A design for sensible reduction of vector surveillance," *J. Amer. Mosquito Control Assoc.*, vol. 27, no. 2, pp. 120–128, Jun. 2011.

[59] Y. Li, Q. Zhang, L. Wang, and L. Liang, "An AIC-based approach to identify the most influential variables in eco-efficiency evaluation," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 113883.

[60] W. N. F. Junior, O. Resende, G. K. I. Pinheiro, L. C. D. M. Silva, and E. R. Costa, "Use of AIC and BIC in desorption isotherms of tamarind seeds (*Tamarindus indica* L.)," *Engenharia Agrícola*, vol. 40, no. 4, pp. 511–517, Aug. 2020.

[61] Y. Zhang, R. Liu, X. Wang, H. Chen, and C. Li, "Boosted binary Harris hawks optimizer and feature selection," *Eng. Comput.*, vol. 37, pp. 3741–3770, May 2020.

[62] J. Hu, H. Chen, A. A. Heidari, M. Wang, X. Zhang, Y. Chen, and Z. Pan, "Orthogonal learning covariance matrix for defects of grey wolf optimizer: Insights, balance, diversity, and feature selection," *Knowl.-Based Syst.*, vol. 213, Feb. 2021, Art. no. 106684.

[63] X. Zhang, Y. Xu, C. Yu, A. A. Heidari, S. Li, H. Chen, and C. Li, "Gaussian mutational chaotic fruit fly-built optimization and feature selection," *Expert Syst. Appl.*, vol. 141, Mar. 2020, Art. no. 112976.

[64] Q. Li, H. Chen, H. Huang, X. Zhao, Z. Cai, C. Tong, W. Liu, and X. Tian, "An enhanced grey wolf optimization based feature selection wrapped kernel extreme learning machine for medical diagnosis," *Comput. Math. Methods Med.*, vol. 2017, pp. 1–15, Jan. 2017.

[65] T. Liu, L. Hu, C. Ma, Z.-Y. Wang, and H.-L. Chen, "A fast approach for detection of erythemato-squamous diseases based on extreme learning machine with maximum relevance minimum redundancy feature selection," *Int. J. Syst. Sci.*, vol. 46, no. 5, pp. 919–931, Apr. 2015.

[66] S. Li, H. Chen, M. Wang, A. A. Heidari, and S. Mirjalili, "Slime mould algorithm: A new method for stochastic optimization," *Future Gener. Comput. Syst.*, vol. 111, pp. 300–323, Oct. 2020.

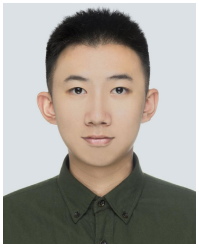
[67] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen, "Harris hawks optimization: Algorithm and applications," *Future Gener. Comput. Syst.*, vol. 97, pp. 849–872, Aug. 2019.

[68] Y. Yang, H. Chen, A. A. Heidari, and A. H. Gandomi, "Hunger games search: Visions, conception, implementation, deep analysis, perspectives, and towards performance shifts," *Expert Syst. Appl.*, vol. 177, Sep. 2021, Art. no. 114864, doi: 10.1016/j.eswa.2021.114864.

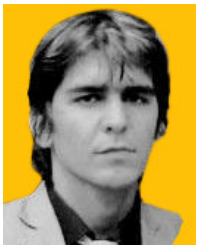
[69] I. Ahmadianfar, A. A. Heidari, A. H. Gandomi, X. Chu, and H. Chen, "RUN beyond the metaphor: An efficient optimization algorithm based on Runge Kutta method," *Expert Syst. Appl.*, vol. 181, Nov. 2021, Art. no. 115079.

MAOFA WANG received the Ph.D. degree in geo-information engineering from Jilin University, China. He is currently an Associate Professor with the School of Computer and Information Security, Guilin University of Electronic Technology, China. He has published more than 20 papers in international journals and conference proceedings, including the *Journal of Seismology*, *Computers & Geosciences*, and others. His current research interests include machine learning, deep learning, and meta-learning and their applications to address geological problems.

GUANGDA GAO is currently pursuing the Ph.D. degree in minerals exploration with the China University of Geosciences (Beijing), China. She is a Lecturer with the School of Information Engineering, China University of Geosciences (Beijing). She has published more than ten papers in international journals and conference proceedings. Her present research interests include deep learning and machine learning and their applications to address geological problems.



HONGLIANG HUANG is currently pursuing the bachelor's degree with the Zhuhai College of Jilin University. He has reviewed several SCI papers. He has seven papers accepted in international conferences and domestic journals. His current research interests include machine learning, mathematics, and their application in solving geological problems and the medical field.



ALI ASGHAR HEIDARI received the B.Sc. and M.Sc. degrees (Hons.) in geospatial engineering and information systems from the College of Engineering, University of Tehran, Tehran, Iran. He has been an Exceptionally Talented Researcher with the School of Computing, National University of Singapore (NUS) and the University of Tehran, and an Elite Researcher with the Iran's National Elites Foundation (INEF). He has authored more than 120 research articles with over 7200 citations (H-index of 48) in prestigious international journals, such as IEEE Internet of Things Journal, IEEE Transactions on Industrial Informatics, *Information Fusion*, *Information Sciences*, *Future Generation Computer Systems*, *Renewable and Sustainable Energy Reviews*, *Energy*, *Journal of Cleaner Production*, *Energy Reports*, *Energy Conversion and Management*, *Applied Soft Computing*, *Knowledge-Based Systems*, IEEE Access, and *Expert Systems with Applications*. He has been ranked globally among the top computer scientists prepared by Guide2Research. He has been ranked in the world's top 2% scientists list of Stanford University, in 2020 and 2021, with several highly cited and hot cited articles. Publons has recognized him as the top 1% peer reviewer in computer science and cross-field with more than 350 papers reviewed for highly reputed journals. His research interests include performance optimization, advanced machine learning, evolutionary computation, optimization, prediction, solar energy, information systems, and mathematical modeling. For more information, researchers can refer to his website https://aliasgharheidari.com.

QIAN ZHANG received the master's degree from the College of Computer Science and Artificial Intelligence, Wenzhou University, China. He is currently an Assistant with the Wenzhou University of Technology. His current research interests include machine learning and data mining and their applications, such as medical diagnosis and bankruptcy prediction, among others.

HUILING CHEN (Associate Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Technology, Jilin University, China. He is currently an Associate Professor with the College of Computer Science and Artificial Intelligence, Wenzhou University, China. His present research interests include evolutionary computation, machine learning, data mining, and their applications to medical diagnosis, bankruptcy prediction, and parameter extraction of the solar cell. He has published more than 200 papers in international journals and conference proceedings, including *Information Sciences*, *Pattern Recognition*, *Future Generation Computer Systems*, *Expert Systems with Applications*, *Knowledge-Based Systems*, *Neurocomputing*, PAKDD, and others. He has more than ten ESI highly cited papers and two hot cited papers. With more than 12165 citations and an H-index of 63, he is ranked worldwide among top scientists for Computer Science & Electronics prepared by Guide2Research, the best portal for computer science research (https://guide2research.com/u/huiling-chen). He has been ranked in the world's top 2% scientists list of Stanford University, in 2020 and 2021, with several highly cited and hot cited articles. He is currently serving as the Editorial Board Member for *Computers in Biology and Medicine*, *Scientific Reports*, IEEE Access, and *Computational and Mathematical Methods in Medicine*. He is also a Reviewer for many journals, such as *Applied Soft Computing*, *Artificial Intelligence in Medicine*, *Knowledge-Based Systems*, and *Future Generation Computer Systems*.

WEIYU TANG is currently pursuing the bachelor's degree with the Zhuhai College of Jilin University. He has published four papers in international conferences and domestic journals. His current research interests include machine learning, data analysis, and Java development, as well as their application in solving geological problems and transportation.