Journal of Economics and Administrative Sciences
2024; 30(141), pp. 457-478

P-ISSN 2518-5764
E-ISSN 2227-703X

Journal of Economics and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq

Constructing a Hybrid Algorithm to Model the Physical and Chemical Inspection Station Data of the Shatt Al-Arab Waters*

Ahmed Husham Mohammed*
Department of Statistics
College of Administration And Economics,
University of Baghdad
Baghdad, Iraq
ahmed.albasrai@uobasrah.edu.iq
***Corresponding author**

Marwan Abdul Hameed Ashour
Department of Statistics
College of Administration And Economics,
University of Baghdad
Baghdad, Iraq

Received:13/8/2023 **Accepted:16/10/2023** **Published Online First: 30 /6/ 2024**

Abstract:

The aim of this paper is to find a hybrid method between of statistical methods that deal with non-linear high-dimensions. The data often suffer from complexity and overlap problems in their mathematical functions. It is difficult to delineate or accurately determine the effect of each variable on the other, accordingly it was constructing a hybrid model between the KPCA and FCM methods. The KPCA method aims to address the problem of high-dimensional nonlinear data and reduce it by finding a kernel matrix that depends primarily on the smoothing parameter matrix (h_j) that was estimated using the ROT method. Then, the FCM was adopted to obtain the clusters. This proposal was applied to the water sector in Basrah Governorate through a study of (8) stations for physical and chemical examination through (15) variables for the years (2019, 2020, 2021) and data were collected on a monthly basis. Through the application of this methodology, the paper was able to determine (7) Basic variables, which are (TH, Na, Cl, TDS, No3, EC, O_G). As for the stations, the overlapping stations between the clusters were identified, which are (SH1, SH2, SH3, SH4, E20, T34), as for the best degree of fuzziness it was (3.6) and the best number of clusters is (k = 3).

Paper type: Research paper

Keywords: Non-linear Data, nonparametric, KPCA, Fuzzy clustering, FCM.

1. Introduction:

High-dimensional data related to natural phenomena is characterized by complexity and overlap, and this is what makes it characterized by a non-linear state, which makes traditional statistical methods useless in treating it. To overcome this problem, a hybrid scenario was proposed through which this complexity can be addressed by adopting the method of kernel principal components analysis (KPCA) which aims to address the non-linear problem in the data by estimating the smoothing parameter matrix by using the rule of thumb high order derivative method and then constructing a matrix $k(x_i, x_k)$by adopting the gaussian kernel function that will contribute to obtaining on the kernel components and reducing the high-dimensions of the variables and then identifying the most influential variables. The other method adopted by this paper is fuzzy c-mean clustering (FCM). This method is based on fuzzy logic in forming clusters because high-dimensional data always suffer from cases of uncertainty in forming groups, as it is possible for cases to belong to two or more clusters. Therefore, the formation of these groups will be based on determining the fuzzing of exponents and the partitions matrix that determines the membership degrees to each cluster.

For achieving the purpose of this paper, it was applied to one of the important sectors, which is the water sector, which is considered one of the environmental sectors that is exposed to several risks such as pollutants, low water levels, high salt concentrations, in addition to the complexity that characterizes its data, as the environmental data for water is a phrase. It consists of physical and chemical elements linked together in complex relationships that cannot be easily separated.

1.1 literature review:

There are several contributions made in the field of addressing the problem of high-dimensions and different opinions**:**

Liu and Yang (2009) developed an approach to address the problem of high-dimensions by adopting two methods of kernel principal component analysis and the fuzzy cluster for addressing the problem of classification in non-linear data.

Ali and Salman (2012) demonstrated the importance of using statistical methods that are concerned with determining causal relationships between variables by enabling classification methods that contribute to the analysis of phenomena. This study presented its contribution through family data and demographic analysis with factor analysis and cluster analysis.

Dogruparmak et al (2014) *discussed* the possibility of reducing air-monitoring stations by analyzing the reality of stations with similar characteristics and features, which helps reduce operational costs, for achieving the goal of this paper, principal component analysis and fuzzy cluster analysis were used. The study concluded with the possibility of reducing stations and reducing pollutants.

ahmed et al (2015) analysed the health sector in Iraqi health institutions by adopting a set of variables that help create classified groups that are easy to analyses by defining a group of clusters. In order to reach clusters that describe the health reality, by used the K-means cluster.

Wang and Zhang (2015) proposed an approach to identify people through speech analysis by adopting the KPCA and thus improving the assembly performance using FCM where the paper was able to reduce errors within speech analysis.

Al-mousa et al (2015) constructed a model for improving the outcomes of the K-Means by processing multidimensional data by adopting PCA, through this approach; they improved the clustering results when using principal components analysis.

Naif and Ayoub (2016) provided large data processing (data exploration) using the K-means averaging cluster method to come up with the identification and classification of the huge amount of data for identifying large impact variables.

Mohammed and Abbod (2016) presented a comparative study between the estimators of the smoothing parameters that affect the construction of the kernel principle component analysis for non-linear data and reducing dimensions by adopting three estimators: least squares cross-validation, biased crossing valid (BCV), Smoothed Cross-validation (SCV) , direct plug-in rule (DPI).

Kaittan (2018) address the high-dimensions and improved the spatial beams of the accuracy of space images through a number of methods including analysis of the main compounds PCA.

Hamed (2019) presented a study on water quality follow-up that is a priority for surface water protection. This applied to the water of the Nile River and particularly to drinking water stations (CDWPs) in Cairo using the principal compounds analysis (PCA) and cluster analysis techniques (FCM).

Essa and Alrawi (2019) presented a comparative study between two methodologies addressing the problem of high- dimensions, one dealing with linear data (PCA) and the other with non-linear data the (KPCA) method in processing the high-dimensions of satellite images of the Shatt al-Arab and the interspersed channels in Basra governorate and the surrounding surfaces.

Hojjatinia and Lagoa (2019) analysed the electrophysiological activities of the brain neurons to explore the structure of the nervous system. Several methods of dimensional analysis proposed for the activities were adopted PCA and KPCA, and then the K-means and FCM methods were adopted to obtain the clusters.

Fawzi and Alkanani (2020) they presented a comparative study between the traditional clustering (K-Means) algorithm and the fuzzy C-Means clustering algorithm and applied this methodology to Baghdad's regions are classified according to water inertia with the adoption of the hazy circle and conventional or hard circle algorithms to indicate which are the most efficient in identifying the less ultra-orthodox areas.

mohammed and Muhamed (2020) provided processing of high-dimensional non-linear data by adopting the methodology of analysis of key pulp compounds to reduce dimensions and identify the most influential variables.

El Fattahi and Sbai (2021) presented an important approach in the processing of non-linear data and analysis of key pulp components KPCA and the use of inertial model with KPCA entropy to arrive at the component-processing algorithm KEPCA that achieved more accurate results in the formation of kernel non-linear compounds.

Mushtaq et al (2023) proposed an approach to the Naive bayes algorithm with a gaussian distribution and then performed a dimensionality reduction analysis using KPCA. This algorithm was applied to a group of people with breast cancer with the aim of identifying cancerous malignant and non-malignant cells. The researchers were able to achieve high accuracy in identifying cancer cells using this algorithm compared to other algorithms.

Liu and Shao (2023) used a method that combination kernel principal components analysis to reduce dimensions with the Gaussian mixture model in order to analyze neutron rays that are used in the fields of radiation protection. KPCA was used to reduce the dimensions of characteristic values, while GMM was used to to cluster the dimensionality reduction of KPCA outputs. This methodology was compared with a group of methods. It has proven to be highly accurate compared to other methods.

Through this introduction, we can place our contribution to this paper by estimating the matrix of the boot parameter in a multiplicity of variables. The other contribution is to develop the fuzzy cluster algorithm by improving the distance scale when incorporating cluster centre variation.

Therefore, we can put the research problem in how to find a hybrid algorithm that can address the problem of higher dimensions and reducing dimensions.

The aim of this research attempted to provide a contribution to address the problem of high-dimensions and reducing dimensions to find the most influential variables and then benefit from them in forming homogeneous clusters.

2.Material and Methods:

In this section, this section will illustrate the statistical methods used in this paper, namely analysis of the main pulp compounds KPCA and the contract of FCM fog averages and the important developments proposed in improving the functioning of the FCM algorithm.

2.1 Kernel Principal Component Analysis (KPCA) :

The overlap and complexity in interpreting relationships between variables with the high- dimensions of those variables makes the data conduct nonlinear behavior (Fawzi and Jaber, 2021). In these cases, it is difficult to build a model through which the phenomenon can be simulated and for the purpose of addressing the problem of vascular dimensions and reducing them. (Blanchard and et al, 2006).

2.2 Kernel Density Estimator:

Probability functions are essentially the basics of parametric distributions, but they vary depending on the nature of the style, including parametric and non-parametric. The non-parametric approach is based on the principle of not knowing the function form, so the estimator of the kernel density $\hat{g}_h(x)$ will depend on the data and the Kernel Function (Chacon and Duong, 2018) and knowledge according to the equation:

$$\hat{g}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \qquad (1)$$

The Eq. (1) represents a one-dimensional kernel density estimator, but if we assume that we have a vector $X = \left[X_1, X_2, \dots, X_q\right]^T$ that has (q) dimensions and that for each dimension (n) of observations $x = [x_1, x_2, \dots, x_n]^T$, then we can get a kernel density estimator suitable for multivariate case (Wand and Jones, 1994):

$$\hat{g}_H(x) = \frac{1}{n|H|}\sum_{i=1}^{n}\left\{\prod_{j=1}^{q} K\left(\frac{x_j-X_{ij}}{h_j}\right)\right\} \qquad (2)$$

$$\hat{g}_H(x) = \frac{1}{n|H|}\sum_{i=1}^{n} K\left(\frac{x_1-X_{i1}}{h_1}, \frac{x_2-X_{i2}}{h_2}, \dots, \frac{x_q-X_{iq}}{h_q}\right) \qquad (3)$$

|H|: The determinant of the smoothing parameter matrix, the H is a positive definite square matrix.

In order to reach an efficient intensity, it must be determined whether it is consistent by adopting a consistency criterion, since the methods of preparation depend primarily on the estimate's proximity to the original function (MSR, MISR, and AMISR), (Hmood, 2005).

2.3 Adjusting Smoothing Parameter:

Adjusting the smoothing parameter contributes to determining the kernel density estimator. In addition, it contributes to creating a state of balance between variance and bias. Then, it is possible to build a covariance matrix, and the smoothing parameter will be estimated using the full and fuzzy method (normal distribution) with higher-order derivatives.

The use of high-derivatives in determining the appropriate formulas to adjust the smoothing parameter is one of the procedural to address the problem of high-dimensions of variables, which was diagnosed by (Bellman) in 1961, which helps to reduce the mathematical difficulties facing the formulas, and thus good smoothing parameters are reached (Chacon and et al, 2011), if we assume that we have (r) of the orders of the higher derivatives and (v) that represents the order of Taylor's expansion, then we can reach a formula for the kernel density estimator through which we can adjust the smoothing parameter matrix By adopting the AMISE scale, as follows:

$$\hat{g}_H^r(x) = \frac{1}{n\prod_{j=1}^{q} h_j}\sum_{i=1}^{n}\left\{\prod_{j=1}^{q} K_{v,h}^{rj}\left(\frac{x_j - X_{ij}}{h_j}\right)\right\}$$

$$AMISE\left(\hat{g}_H^r(x)\right) = \int Var\left(\hat{g}_H^r(x)\right) dx + \int [Bais(\hat{g}_H^r(x))]^2\, dx$$

$$Bais\left(\hat{g}_H^r(x)\right) = E(\hat{g}_H^r(x)) \quad - g^r(x)$$

$$= E\left(\frac{1}{n\prod_{j=1}^{q} h_j}\sum_{i=1}^{n}\left\{\prod_{j=1}^{q} K_{v,h}^{rj}\left(\frac{x_j - X_{ij}}{h_j}\right)\right\}\right) - g^r(x)$$

$$\frac{1}{\prod_{j=1}^{q} h_j}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\prod_{j=1}^{q} g^{(r)}\,(X_{ij} - h_j u_j) K_{v,h}^{rj}(u_j\,)(-h_j)du_j\right) - g^r(x_j)$$

$$= -\frac{\prod_{j=1}^{q} h_j}{\prod_{j=1}^{q} h_j}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\prod_{j=1}^{q}[g^{(r)}(X_{ij} - h_j u_j)]K_{v,h}^{rj}(u_j\,)(-h_j)du_j - g^r(x_j)\right)$$

Using the rule of Taylor expansion $g^{(r)}(X_{ij} - h_j u_j)$

$$= \left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\sum_{j=1}^{q}\left[g^{(r)}(x_j) + \frac{1}{v_j}\nabla g^{(r+v)}(x_j)(h_j u_j)^v\right]K_{v,h}^{rj}(u_j\,)du_j\right) - g^r(x_j)$$

By referring to the properties of the kernel density, function

$$\int k(u)du = 1\,;\ \int u\,k(u)du = 0\,;\ \int u^2\,k(u)du = \mu_2(k)$$

After distributing the amount $\mathrm{K}_{\mathrm{v,h}}^{\mathrm{rj}}(\mathrm{u_j}\,)$ on Taylor's to a bracket Taylor expansion, we can get the result:

$$= g^{(r)}(x_j) + \frac{\mu_v^{(r)}}{v!}\sum_{j=1}^{q}\nabla g^{(r+v)}(x_j)(h_j)^v - g^r(x_j)$$

$$Bais\left(\hat{g}_H^r(x)\right) = \frac{\mu_v^r(K)}{v!}\sum_{j=1}^{q}\nabla g^{(r+v)}(x_j)(h_j)^v + o(nh)^{r+v} \qquad (4)$$

By following the same derivative method above, we can get the variability formula for the high order derivatives

$$var\left(\hat{g}_H^r(x)\right) = E(\hat{g}_H^r(x))^2 - [E(g^r(x))]^2$$

$$=$$

$$E\left[\frac{1}{n\prod_{j=1}^{q} h_j^{r+1}}\sum_{i=1}^{n}\left\{\prod_{j=1}^{q} K_{v,h}^{rj}\left(\frac{x_j - X_{ij}}{h_j}\right)\right\}\right]^2 - \left[E\left(\frac{1}{n\prod_{j=1}^{q} h_j^{r+1}}\sum_{i=1}^{n}\left\{\prod_{j=1}^{q} K_{v,h}^{rj}\left(\frac{x_j - X_{ij}}{h_j}\right)\right\}\right)\right]^2$$

$$= \frac{n}{\left(n\prod_{j=1}^{q} h_j^{r+1}\right)^2}\left(E\left(\prod_{j=1}^{q} K_{v,h}^{rj}\left(\frac{x_j - X_{ij}}{h_j}\right)^2\right) - \left[E\left(\prod_{j=1}^{q} K_{v,h}^{rj}\left(\frac{x_j - X_{ij}}{h_j}\right)\right)\right]^2\right)$$

When the transformation to simplify the formula with the properties of the Kernel function that is achieved $\int \mathrm{u}\,\mathrm{k(u)du} = 0$ with the use of Taylor's on $\mathrm{g}^{(\mathrm{r})}(\mathrm{X_{ij}} - \mathrm{h_j u_j})$at the first order, then we get:

$$var\left(\hat{g}_H^r(x)\right) = \frac{1}{n\prod_{j=1}^{q} h_j^{2r+2}}\left(\left[g^{(r)}(x_j)\right]\prod_{j=1}^{q} R(K_{v,h}^{rj})\right) \qquad (5)$$

By taking the derivative$\frac{\partial \mathrm{AMISE}\left(\hat{\mathrm{g}}_{\mathrm{H}}^{\mathrm{r}}(\mathrm{x})\right)}{\partial \mathrm{h}}$, we get the formula$\left(\mathrm{h}_{\mathrm{j}}^{\mathrm{opt}}\right)$:

$$h_j^{opt} = \left[\frac{v!(q+2r)\prod_{j=1}^{q} R(K_{v,h}^{rj})}{(2v)\mu_v^{2+r}(K)\left[\int\left(\nabla g^{(r+v)}(x_j)\right)^2 dx\right]}\right]^{\frac{1}{2v+2r+q}} n^{\frac{-1}{2v+2r+q}} \qquad (6)$$

(Henderson and Parmeter, 2012)

The Eq. $\left(\mathrm{h}_{\mathrm{j}}^{\mathrm{opt}}\right)$ represents the formula for obtaining the optimal smoothing parameter and addressing the amount of $\int\left(\nabla g^{(r+v)}(x_j)\right)^2 dx$unknown using natural distribution and taking advantage of some of its characteristics. It has multiple infinite boundaries at the higher levels of derivatives. Hermite Polynomial, where between Henderson and others in 2012 could find a formula to estimate the smoothing parameter as follows:

$$h_j^{opt} = \left[\frac{\pi^{\frac{q}{2}} 2^{v+q+1}(v!)^2 R(K_v)^q}{(2v)\mu_v^2(K_v)[(2v!!+(q-1)(v!!)^2]}\right]^{\frac{1}{2v+q}} n^{\frac{-1}{2v+q}} \qquad (7)$$

v: High order derivative

q: Numbers of diminutions

The multivariate gaussian kernel function, which is consistent with the ROT (Wand and Jones, 1995) method, is be adopted, and in compensation for the results of this $\mu_2(K)^q$, $R(K)^q$ in the Eq. (7) we get the ROT formula created by Scott as explained in the following(Scott, 1992):

$$h_j^{ROT} = \left[\frac{4}{2+q}\right]^{\frac{1}{4+q}} \hat{\sigma}_j \quad n^{-\frac{1}{4+q}} \qquad (8)$$

Where:

$\hat{\sigma}_j$: It represents the standard deviation for each of the dimensions

In this paper, (H) diagonal matrix defined by (h_j) will be adopted for each of the dimensions, its elements defined as follows (Duong, 2004):

$$H = \begin{pmatrix} h_1^2 & 0 & \dots & 0 \\ 0 & h_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & h_q^2 \end{pmatrix} \qquad (9)$$

For obtaining the kernel matrix, the Gaussian Kernel function will be used in its traditional state, which is shown below:

$$k(x_i, x_k) = exp\left(-\frac{\left\|x_{ij} - x_{sj}\right\|^2}{{h_j}^2}\right) \qquad (10)$$

2.4 Statistical Tests:

2.4.1 Fitting Data Test:

(1) Kaiser Meyer Olkin test (KMO):

Kaiser presented in 1970 the KMO test, which aims to determine the suitability of the sample size, whereas, value falls between the two values **(0 , 1)** , it has calculated according to the following formula:

$$\boldsymbol{KMO = \frac{\sum_{i\neq j}^{n} r_{ij}^2}{\sum_{i\neq j}^{n} r_{ij}^2 + \sum_{i\neq j}^{n} \ell_{ij}^2}\ ; i\ \&\ j = 1, 2, \dots, q} \qquad (11)$$

Where:

$\boldsymbol{\sum_{i\neq j}^{n} r_{ij}^2}$ is the sum of the squares elements of the correlation coefficient matrix outside the main diagonal.

$\boldsymbol{\sum_{i\neq j}^{n} \ell_{ij}^2}$ is the sum of the squares elements of the partial correlation coefficients matrix.

Whenever its value is greater than 85%, this is evidence that the sample is valid, but if it is less than 50%, this is evidence that it is not valid.

(2) Bartlett Shpericity Test:

Bartlett Shpericity has calculated according to the following formula (Dziuban & Shirkiy, 1974):

$$\boldsymbol{\chi^2_{cl} = -\left(n - \frac{1}{6}[2q+5]\right) ln|R|} \qquad (\mathbf{12})$$

If the value of $\chi^2_{cl} > \chi^2_{tab\left(\frac{1}{2}q(q-1)\right)}$, the null hypothesis H_0 is rejected.

2.4.2 Identification number of the KPCA:

(1) Jolliffe criteria:

Jolliffe modified the part value criterion in determining the number of main vehicles according to the ratio (70%-90%), (Jolliffe, 2002).

(2) Average Eigen Values Criterion:

It has calculated according to the following formula (Rencher & Christensen, 2012):

$$\boldsymbol{\bar{\lambda} = \frac{\sum_{j=1}^{q}\lambda_j}{q} \; ; if \; \lambda_j > \bar{\lambda} \; to \; select \; Z_j} \qquad (\mathbf{13})$$

$\boldsymbol{\lambda_j}$: eigenvalues

(3) Scree Test:

This criterion depends on the intrinsic $\boldsymbol{\lambda_j}$ in determining the number of important PC. It can be determined by drawing the $\boldsymbol{\lambda_j}$ and determining the "**elbow**" point. All component that are after this point represent the important compounds, while compounds that form before are unimportant and can be neglected (Cattell, 1966)**.**

2.5 Fuzzy C-Means Clustering:

Clustering did considered one of the important statistical methods that are used to form aggregates and classify them to obtain homogeneous samples for a group of data, especially when that data is characterized by large patterns and trends that cannot be controlled. Therefore, technological development had a role in developing clustering methods to shift from traditional methods to more ones that are realistic. In line with the real reality, so the fuzzy theory proposed by Zadeh in 1965 had a shift in the development of many methods, including the clustering technique to show what was the called fuzzy C-Means method (FCM) (Sreenivasarao and Vidyavathi, 2010), (Ashour and Jawad, 2017).

The FCM algorithm aims to cluster large data in the form of new, more homogeneous clusters based on determining the fuzzing exponents to the target cases. (Dunn, Bezdek) developed algorithm in 1974 by developing the partitions matrix in the K-Means clustering algorithm with specifying the degree of fuzziness (El-Zaghmouri and Abu-Zanona, 2012), then we get the objective function described in the following formula (Javadi and et al, 2018):

$$obj(X,P,V) = \sum_{k=1}^{c}\sum_{i=1}^{n} p_{ik}^{m}\, \varphi(x_i \;, v_k) \qquad (14)$$

Where:

P : It is a matrix of order (k×n) and represents the degrees of belonging to each element within the clusters.

$\boldsymbol{m}$ **:** Fuzziness coefficient (fuzziness exponent), where its value is defined within the period $1 \le \boldsymbol{m} < \infty$**.**

x_i **:**It represents the original data collected**.**

v_k**:** Cluster center**.**

$\varphi(x_i; v_k)$ **:** It is a measure of similarity or difference depending on the center of the cluster$\boldsymbol{v_k}$.

In order for us to achieve the objective function, it is necessary to determine the center of the fuzzy cluster $\boldsymbol{v_k}$and to define the matrix of fuzzy divisions that includes the membership degrees to each case (p_{ik}) and the fuzzing exponents (m) and its formulas to be define as follows:

$$v_{kj} = \frac{\sum_{i=1}^{n} p_{ik}^{m} x_{ij}}{\sum_{i=1}^{n} p_{ik}^{m}} \quad \forall\, k = 1,2, \dots, c \qquad (15)$$

Accordingly, we obtain formula (15), which represents the estimation of the cluster center (Goyal and et al, 2016). Therefore, the cluster center here that become weighted by membership degrees (p_{ik})which also depends on the measure of similarity or difference, and accordingly a matrix can be estimated Degrees of belonging according to the following formula (Oliveira and Pedrycz, 2007)**:**

$$p_{ik} = \frac{h_k \varphi(x_i, v_k)}{\sum_{k=1}^{c} h_k \varphi(x_i, v_k)} \ ; \ \forall\, i = 1,2,\dots,n \qquad (16)$$

Where:

$h_k = \frac{n_k}{n}$: It represents the ratio (size of the sampling) i.e. Number of elements in each cluster $\mathrm{n_k}$ to the number of total elements n, and the condition is met $\sum_{\mathrm{k=1}}^{\mathrm{c}} \mathrm{h_k} = 1$

$\varphi(x_i, v_k)$: The measure of distance scale.

2.6 Distance Metrics:

The measure of distance or similarity $\boldsymbol{\varphi(x_i, v_k)}$ is a measure based on the formation of a similarity matrix for (n) cases and (q) variables. Accordingly, the degree of convergence between the points of each variable can be determined according to the cases to form a matrix of convergence. Proximate Matrix, if we have a vector Variables $X = [X_1 \quad X_2 \quad \cdots \quad X_q]$, the information matrix is of degree$(n \times q)$, and therefore the convergence matrix can be obtained as shown below (Wierzchon and Kłopotek, 2018)**:**

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1q} \\ d_{21} & d_{22} & \cdots & d_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nq} \end{bmatrix}$$

There are several types of measures of similarity and difference to determine the distance, but the Euclidean Distance and Square Euclidean Distance (SED) will be displayed as follows:

1- Euclidean Distance (ED)

$$d_{ik} = \sqrt{\sum_{i=1}^{n} (x_{ik} - v_k)^2} \ \ \forall\, k = 1,2,\dots,c; \ \ w_k^2 = 1 \qquad (17)$$

2- Square Euclidean Distance (SED)

$$d_{ik} = \left(\sqrt{\sum_{i=1}^{n} (x_{ik} - v_k)^2}\right)^2 \ \ \forall\, k = 1,2,\dots,c \qquad (18)$$

2.7 proposal to develop a FCM algorithm:

The K-means algorithm was hybridized with the FCM algorithm for arriving at the partitions matrix to determine the initial membership degrees by adopting the original data instead of generating it randomly. This was done in two stages**:**

The first stage: It was represented by the stage of preparing the membership degrees by adopting the K-Means algorithm approach according to the following steps:

1- Determine the number of clusters.

2- Determine the centres of initial clusters.

3- Calculate the distance according to Eq. (18).

4- Formation of primary clusters.

5- Return to step (2) and get the centres of the clusters according to the clusters achieved in step (4).

6- Formation of new clusters.

7- Calculating the division's matrix according to the membership function Eq. (16) to obtain the membership degree$\boldsymbol{p_{ik}}$**.**

The second stage: entering the matrix of belonging scores resulting from step (7) to initialize the FCM algorithm:

8- Calculation of cluster centres according to Eq. (15).

9- Calculate the objective function according to Eq. (14).

10- Condition check$|Obj^{I+1} - Obj^{I}| < \varepsilon$.

11- If the condition is met, it stops, but if the condition is not met, the $\boldsymbol{p_{ik}}$ matrix was updated according to the equation.

$$\boldsymbol{p_{ik}} = \frac{\mathbf{1}}{\sum_{k=1}^{c}\left(\frac{d^2(x_i\,,v_r)}{d^2(x_i\,,v_k)}\right)^{\frac{1}{m-1}}}\ ;\ \forall \boldsymbol{i}\,, \boldsymbol{r} = \boldsymbol{i} \qquad (\mathbf{19})$$

Repeat steps (8-10), and the process of updating $\boldsymbol{p_{ik}}$ elements continues according to step (14) until the stopping condition (10) is true.

2.8 Identification the validity of fuzzy cluster:

(1) Partition Coefficient (P.C.):

The partition coefficient scale was proposed to test the validity of fuzzy clustering by Bezdek (1974), It is calculated using the following formula (Bezdek, 1974):

$$\boldsymbol{P.C.} = \frac{\sum_{k=1}^{c}\sum_{i=1}^{n} p_{ij}^2}{n}\ ; \qquad \frac{1}{c} \le \boldsymbol{P.C.} \le \mathbf{1} \qquad (\mathbf{20})$$

The higher the value of P.C. close (1), then the fuzzy cluster has a crisp cluster, and when it is close to or less than the limit (1/c), the fuzzy cluster is invalid cluster. However, if the value of P.C. Between the two values, the fuzzy cluster is good (valid).

(2) Xie-Beni criteria:

The cluster validity criterion (δ_{XB})verifies the validity of the cluster structure based on the objective function and the fuzzing exponents. Accordingly, it is considered better than the partitions coefficient criterion, which depends only on the membership degrees of the partitions matrix. The best cluster structure is determined by determining the lowest value achieved for this criterion, and it is calculated to the following formula (Höppner, Klawonn, Kruse, & Runkler, 1999):

$$\boldsymbol{\delta_{XB}(X, P, V)} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{c} p_{ij}^{m}\left\|x_{ij}-v_k\right\|^2}{n\left(\min_{i,k}\left(\left\|v_{rk}-v_{jk}\right\|^2\right)\right)} \qquad (\mathbf{21})$$

Where:

$\boldsymbol{n}$:number of observation; $i = 1,2, \dots, n$

$\boldsymbol{c}$**:** Number of clusters**;** $k = 1,2, \dots, c$; $\boldsymbol{v_k}$ **:** cluster center.

$\boldsymbol{p_{ij}}$ **:** Elements of the membership degree matrix

$\boldsymbol{min\left\|v_{rk} - v_{jk}\right\|}$**:** Minimum distance between clusters

2.9 Description of the application sector:

The environmental sector, especially the water sector, is considered one of the complex sectors in building its models. This complexity is related to the two types of its data, which are physical and chemical elements. It is also known that these elements are linked to each other through complex relationships and are characterized by nonlinearity. Due to these characteristics, this data was processed according to the methods proposed in this article. The paper, therefore, the water sector in Basrah Governorate was targeted as a result of the environmental problems that it suffers from, especially in this sector. so we prepared this paper for the purpose of determining the structure of homogeneous stations through a set of physical and chemical variables, as information was collected from (8) stations in Basrah Governorate, which are Shatt Al-Arab stations (H1, H2, H2 B, H3, H4), station Qurna for the Tigris River T34, the two stations of the city for the Euphrates River (E20, E21), and it was represented by the data of the physical and chemical examination of water on a monthly basis from 2019 to 2021, which included (15) variables, which are shown in the table below:

Table (1) shows environmental variables (physical and chemical properties of water)

environmental variables	Variable	Climate variables	Variable
Potential for Hydrogen (PH)	X_1	Nitrogen (Na)	X_9
Dissolved Oxygen (DO_2)	X_2	Sulphate (SO4)	X_{10}
Phosphorous (PO_4)	X_3	**chlorine** (Cl)	X_{11}
Nitrates (NO_3)	X_4	Total dissolved solids (TDS)	X_{12}
Calcium (Ca)	X_5	**Electrical conductivity** (EC)	X_{13}
Magnesium (Mg)	X_6	Total Alkalinity (ALK)	X_{14}
Total Hardness (TH)	X_7	Oil and Grease (O_G)	X_{15}
Potassium (K)	X_8		

Source: Ministry of Environment - Technical Section

3- Discussion of Results:

3.1 Discuss the results of the KPCA:

First: test the suitability of the data for analysis:

The suitability of the environmental data obtained from the water test was tested in Basrah Governorate and when conducting it, the results were as shown in Table (1).

Table (2) Testing the suitability of the physical and chemical examination data for the waters of Basrah Governorate

Sector	q	N	KMO	Decision	Chi-Sq.	χ^2_{TAB}	
						P-value	Decision
environmental	15	145	0.9223	Great fit	2365.293	0.0000	Sig.

It is clear from the results of the table (2) that the data was highly suitable for the procedure of e kernel principal components analysis, as the (KMO) scale recorded an appropriate rate of (0.9223), and this was confirmed by the Bartlett Shpericity test, as it showed the Chi-Sq. statistic. The calculated value (2365.293) is significant because the P-value was less than (5%), then the relevance and accuracy of the researched data is accepted.

Second: Identify the kernel principal component contributing:

In this paragraph, the results of the analysis of the kernel components of the environmental data will be discussed about the smoothing parameter matrices (H)were estimated using the (ROT_{hod}) method, and then the gaussian kernel function was adopted in calculating the kernel covariance matrix $K(x,x)$. The first stage included testing the criterion for determining the cut-off value by applying (3) criteria as discussed in the experimental side (Average Eigen value, Jolliffe, Scree graph), and it was obtained the results are shown in Table (3 and 4):

Table 3: Shows the eigenvalues and contribution Proportion variance of the KPCA

Number PC.	GK – H		
	Eigen value	Proportion Variance	Cumulative Proportion Variance
Z_1	293.606	0.83567	0.83567
Z_2	31.654	0.09009	0.92576
Z_3	19.022	0.05414	0.97990
Z_4	2.805	0.00798	0.98789
Z_5	2.663	0.00758	0.99547
Z_6	0.699	0.00199	0.99746
Z_7	0.545	0.00155	0.99901
Z_8	0.264	0.00075	0.99976
Z_9	0.049	0.00014	0.99990
Z_{10}	0.019	0.00005	0.99995
Z_{11}	0.008	0.00002	0.99998
Z_{12}	0.004	0.00001	0.99999
Z_{13}	0.002	0.00001	1.00000
Z_{14}	0.001	0.00000	1.00000
Z_{15}	0.000	0.00000	1.00000

Table (3) shows the number of KPC resulting when estimating the covariance matrix using the gaussian kernel (GK) method, which showed through the eigenvalues and the cumulative contribution proportion to the variance achieved that there are (5) PC out of (15) compounds, and here the scope for reduction becomes clear the number of component, but this does not determine who are the most influential and contributing component and therefore the contributing vehicles will be determined through a number of criteria, which are shown in table (4).

Table 4: Comparison between the criteria for selecting kernel principal compounds, determining the cut-off value, and the cumulative contribution percentages of the smoothing parameter matrix estimated according to the (ROT_{hod}) method.

Sector	Dime. (q)	S. size (n)	Actual KPC.	N. KPC by criterion					
				$\bar{\lambda}$	Cum. pro. variance	Jollffie	Cum. pro. variance	S.G.	Cum. pro. Variance
Environmental	15	145	5	2	0.92576	1	0.83567	1	0.83567

As the results of Table (4) show that the number of compounds achieved was (5) compounds, while the criteria for selecting the most compounds were able to determine the number of contributing compounds.

$(\bar{\lambda})$ the cut-off value was at the component (Z_2) with a contribution ratio of (0.92576), meaning that the number of the main compounds are $(Z_2\ ,Z_1)$, while the Jollffie criterion specified the value of the cut-off at the vehicle (Z_1) with a cumulative contribution rate of (0.83567). The Scree Graph criterion came with results similar to the Jollffie criterion, and due to what was achieved by the criterion$(\bar{\lambda})$, what was achieved can be adopted in determining the most influential variable.

Table 5: Shows the saturation degrees representing the degree of binding of variables to the KPC

K.F Var.	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	Z_{11}	Z_{12}	Z_{13}	Z_{14}	Z_{15}
X_1	0.01	0.02	0.07	-0.04	0.04	0.30	0.31	-0.40	0.29	-0.53	0.48	0.07	0.16	-0.11	0.03
X_2	-0.12	-0.26	0.08	0.30	0.76	-0.38	-0.24	-0.05	0.01	-0.17	0.10	0.02	0.00	-0.02	0.03
X_3	0.12	-0.03	0.07	0.09	0.13	-0.32	0.88	0.24	-0.13	0.05	-0.04	0.00	0.02	0.06	-0.01
X_4	-0.03	0.82	-0.11	0.54	0.12	0.08	-0.01	0.02	0.01	0.03	0.02	0.01	0.00	0.00	0.00
X_5	0.23	-0.07	0.03	-0.07	0.44	0.50	0.12	-0.16	-0.02	0.58	0.04	0.17	-0.12	-0.25	-0.08
X_6	0.32	0.03	0.01	-0.02	-0.06	-0.21	-0.10	0.43	0.73	0.12	0.24	0.08	-0.13	-0.17	-0.05
X_7	0.36	-0.01	0.01	-0.02	0.18	0.16	-0.02	-0.08	0.23	0.05	-0.11	-0.49	0.18	0.66	0.17
X_8	0.31	0.05	0.00	0.00	-0.16	-0.31	-0.13	-0.16	-0.39	0.27	0.69	-0.05	0.12	0.11	-0.10
X_9	0.35	0.05	0.01	0.00	0.00	-0.11	-0.01	-0.21	-0.07	-0.23	-0.20	-0.52	-0.30	-0.36	-0.48
X_{10}	0.28	-0.07	-0.08	0.01	0.13	0.39	-0.11	0.64	-0.38	-0.39	0.16	0.02	0.00	-0.01	-0.01
X_{11}	0.36	0.05	-0.01	-0.02	-0.04	-0.13	-0.03	-0.14	-0.11	-0.08	-0.11	-0.01	-0.15	-0.33	0.82
X_{12}	0.35	0.05	-0.02	-0.03	-0.01	-0.09	-0.03	-0.21	-0.03	-0.21	-0.16	0.61	-0.44	0.42	-0.15
X_{13}	0.35	0.03	-0.02	0.00	0.01	-0.11	-0.09	-0.07	0.00	-0.05	-0.33	0.29	0.77	-0.19	-0.17
X_{14}	0.08	-0.16	0.83	0.44	-0.24	0.14	-0.05	0.02	-0.02	0.03	-0.04	0.02	-0.02	0.01	0.00
X_{15}	0.09	-0.46	-0.53	0.63	-0.26	0.12	0.06	-0.10	0.07	0.05	0.00	0.00	-0.01	-0.01	0.00

Table (5) results illustrated a summary of the degree of correlation of the basic water environmental variables represented by the results of physical and chemical examinations of the waters of Basrah Governorate are associated with the resulting KPC (Z1 and Z2), the most influential variables were diagnosed where the component (Z1) showed each of the variants Na, X11 = (Cl), X12 = TDS, X13 = EC, while the compound (Z_2) showed that the variable X4 = No3 and the variable X15 = O_G**.**

3.2 Discussion of the results FCM:

After KPCA results have determined the most influential fundamental variables in the aquatic environment, clusters will be formed from inspection stations in Basra governorate in accordance with the FCM cluster algorithm. In this algorithm, restricted simulation of real data is adopted for the purpose of determining the degree of overlap and then conducting the cluster stations according to the variables. (O_G, EC, TDS, Cl, Na, TH, and No3), the following results were reached:

3-2-1 Discussing the results of 2019:

1. Determine the number of clusters (k) appropriate and the fuzzing exponents (m) for the year 2019

Table 6: Shows the results of the test by adopting the P.C. standard and δ_{XB} to determine the fuzzing exponents (m) and the number of clusters k for the year 2019

Clustering Algorithm		HFCM			
Comparison		Obj.	Iteration	G. fit cluster	
				P.C.	δ_{XB}
k	m				
2	1.2	925585.52	12	0.9967	378.3792
2	2	747413.11	15	0.9119	386.1952
2	2.8	527603.09	15	0.8024	239.9643
2	3.6	345639.62	18	0.7135	185.1575
3	1.2	279377.64	9	0.9987	176.3731
3	2	233255.92	35	0.8635	212.4914
3	2.8	141635.51	24	0.7004	104.8721
3	3.6	73437.70	27	0.5790	58.0576

Discuss the results of the table (6) as follows:

When **k = 2,** it is clear from the results of table () that the best degree of fuzzing is when (m = 3.6), as it achieved the lowest value of the objective function (Obj_Fun = 345639.62), compared to the other degrees of fuzzing.

According to **P.C.** standard. It showed the validity of the fuzzy cluster at the fuzzing exponent (m = 2.8), achieving a percentage of (0.8024), and the validity of the fuzzy cluster was achieved at the fuzzing exponent (m = 3.6), achieving a percentage of (0.7135).

According to the $\boldsymbol{\delta_{XB}}$ criterion, the validity of the fuzzy cluster structure was demonstrated at the fuzzing exponent (m = 3.6), as it achieved (185.1575), which is the lowest value compared to the other values.

When **k = 3,** it is clear from the results of table (6) that the best degree of fuzzing is when (m = 3.6), as it achieved the lowest value of the objective function (Obj_Fun = 73437.70), compared to the other degrees of fuzzing.

According to **P.C.** standard. It showed the validity of the fuzzy cluster at the fuzzing exponent (m = 2.8), achieving a percentage of (0.7004).

According to the $\boldsymbol{\delta_{XB}}$ criterion, the validity of the fuzzy cluster structure was demonstrated at the fuzzing exponent (m = 3.6), as it achieved (58.0576), which is the lowest value compared to the other values

We can also determine the level of overlapping points by determining the Average maximum membership Index AverageMax As shown in the figure below:

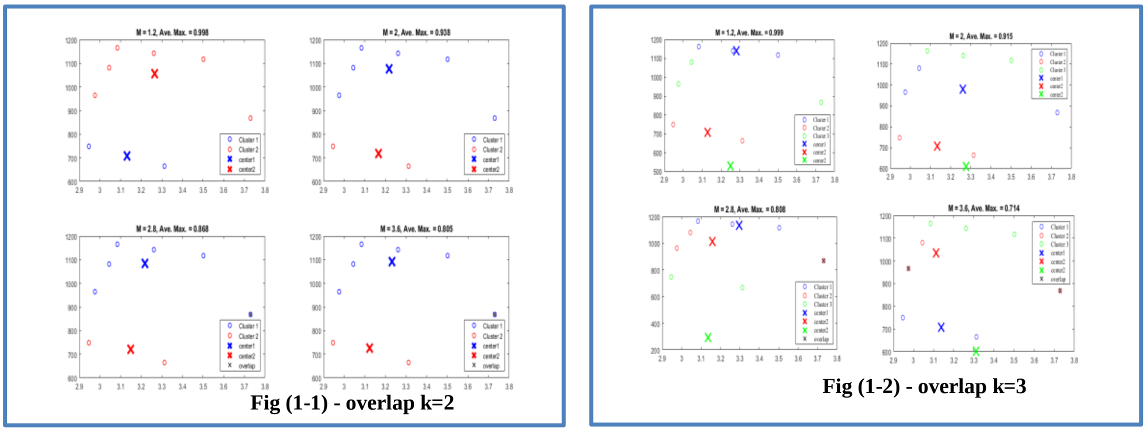

Fig (1-1) - overlap k=2

Fig (1-2) - overlap k=3

Figure 1: Shows instances of selection of fuzziness exponent according to the AverageMax standard in case of k = 2,3 for 2019

Figure (1) shows overlaps by degrees of fuzziness. In (k = 2)figure (1-1) shows, overlaps that show increased uncertainty. One station appeared at (m = 2.8) with an average maximum membership (AverageMax = 0.8684), a station (SH4) that belonged to the first cluster to the membership degrees (0.5570) and the second cluster to the degree of belonging (0.4430). In the case of the (m = 3.6) and a maximum average membership (AverageMax = 0.8047), the interference case appeared in one station, station (SH4), which belonged to the first cluster with a membership degree (0.5310) and to the second cluster with a membership degree (0.4690).

In the case of k = 3, Figure (1-2) shows overlaps showing an increase in uncertainty, one station appeared at the (m = 2.8) and with an average maximum membership (AverageMax = 0.8080), which is station (SH4) that belonged to the first cluster to the degree of membership (0.5746), and to the third cluster with a degree of belonging (0.2201), and to the second cluster with a degree of belonging (0.2053). In the case of the (m = 3.6) and with a maximum average membership (AverageMax = 0.808), two stations showed, namely (SH3), which belonged to the first cluster with a degree of membership (0.5561), and to the third cluster with a degree of affiliation (0.3326), and to the second cluster with a degree of membership (0.3326), and (0.1113) degree to the second cluster , and the station (SH4) belonged to the first cluster with a degree of membership (0.4831), to the third cluster with a degree of affiliation (0.2656), and to the second cluster with a degree of membership (0.2513).

2.The clusters are therefore formed according to the specific cases:

Table 7: Shows the fuzzy cluster formation of the physical and chemical examination stations in the case of k = 2, 3 for the year 2019

Clustering Algorithm		**HFCM**					
Comparison		**Stations Order**					
M	**Clusters**						
3.6	C_1	E20	T34				
	C_2	SH1	SH2B	SH2	SH3	SH4	E21
3.6	C_1	SH1	SH3	SH4			
	C_2	E20	T34				
	C_3	SH2B	SH2	E21			

The results of table (7) show the formation of clusters. In the case of k = 2, the cluster (2) contained (6) stations. This is evidence of the uniformity of the resulting inspection information. While cluster (1) contained two stations (E20, T34), in the case of k = 3, three clusters with the first and third contained three stations, while the second contained two stations.

3-2-2 Discussing the results of 2020

1. Determine the number of clusters (k) appropriate and the fuzzing exponents (m) for 2020

Table 8: Shows test results by adopting P.C. and δ_{XB} standard for determining fuzzing exponents and number cluster k for 2020

Clustering Algorithm		**HFCM**			
Comparison		**Obj.**	**Iteration**	**G. fit cluster**	
				P.C.	$\boldsymbol{\delta_{XB}}$
k	M				
2	1.2	5429289.87	5	1	1929.5
2	2	5371372.23	8	0.9927	1415.9
2	2.8	4747416.49	13	0.9393	1932.6
2	3.6	3635337.68	13	0.8639	791.1
3	1.2	1785074.75	10	0.9965	1318.7
3	2	1457903.63	44	0.8720	1088.7
3	2.8	962884.21	53	0.7449	779.95
3	3.6	543718.11	43	0.6568	328.32

Discuss the results of the table (8) as follows:
When **k = 2,** it is clear from the results of table (8) that the best degree of fuzzing is when (m = 3.6), as it achieved the lowest value of the objective function (Obj_Fun = 3635337.68), compared to the other fuzzing exponents.
According to P.C. standard. It was shown that the fuzzy cluster is invalid at any fuzzing exponents because it achieved a percentage higher than (85%), when the fuzzy cluster is crisp cluster.

According to the $\boldsymbol{\delta_{XB}}$ criterion, the validity of the fuzzy cluster structure was demonstrated at the fuzzing exponent (m = 3.6), as it achieved (791.1), which is the lowest value compared to the other values.
When **k = 3,** it is clear from the results of table (6) that the best degree of fuzzing is when (m = 3.6), as it achieved the lowest value of the objective function (Obj_Fun = 543718.11), compared to the other fuzzing exponents.
According to **P.C.** standard. It showed the validity of the fuzzy cluster at the fuzzing exponent (m = 2.8), achieving a percentage of (0.7449), and the validity of the fuzzy cluster was achieved at the fuzzing exponent (m = 3.6), achieving a percentage of (0.6568).

According to the $\boldsymbol{\delta_{XB}}$ criterion, the validity of the fuzzy cluster structure was demonstrated at the fuzzing exponent (m = 3.6), as it achieved (328.32), which is the lowest value compared to the other values
We can also determine the level of overlapping points by determining the Average maximum membership Index AverageMax As shown in the figure below:

we can also determine the level of overlapping points by determining the Index AverageMax As shown in the figure below:

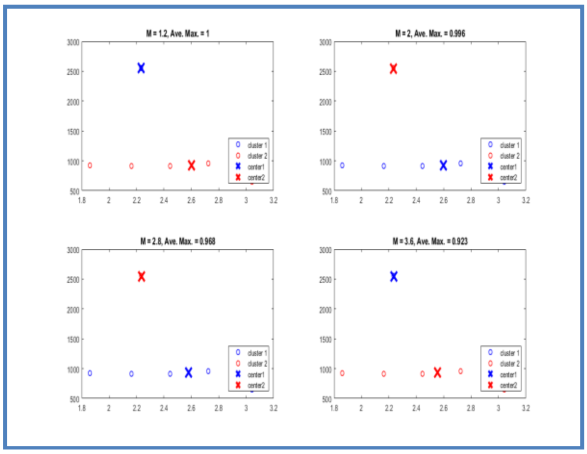

Fig (2-1) - overlap k=2

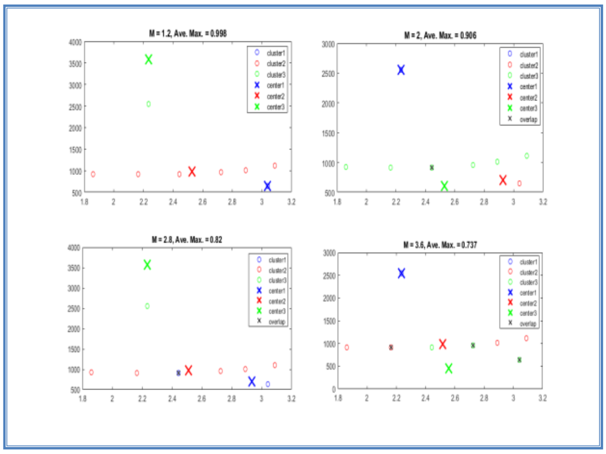

Fig (2-2) - overlap k=3

Figure 2: Shows instances of selection of fuzziness degree according to the AverageMax standard for
k = 2 and 3 for 2020

Figure (2), which shows the overlap cases according to the degrees of fuzziness, in the case of k = 2, Figure (2-1) shows that there is no overlap between the clusters, as AverageMax recorded the average maximum membership degree between [1, 0.5], which indicates a low state of uncertainty that is, the membership of the stations in the cluster are greater than (82%) and may reach one, offset by a very weak degree of a membership to the other cluster.

In the case of k = 3, Figure (2-2) shows the interference cases that show the increase in uncertainty. One station appeared at the fuzzing exponent (m = 2.8) and with an AverageMax = 0.82, which is station (SH1), which belonged to the third cluster with a membership degree(0.4903), and to the first cluster with a membership degree (0.4829), and to the second cluster with a membership degree(0.0268). In the case of the fuzzing exponent (m = 3.6) and an AverageMax = 0.737, three stations were shown: SH2, which belonged to the second cluster with a membership degree (0.4797), to the third cluster with a degree of membership (0.4740), and to the first cluster with a membership degree (0.0463).), and station (E20) belonged to the third cluster with a degree of membership (0.5241) and to the second cluster with a degree of membership (0.4304) and to the third cluster with a degree of membership (0.0455), and station (T34) belonged to the third cluster with a degree of membership (0.5844) and to the second cluster with a degree of membership (0.3332) and to the third cluster with a degree of membership (0.0824).

2.The clusters are therefore formed according to k = 3, m = 3.6

Table (9) shows the fuzzy cluster formation of the physical and chemical examination stations for the year 2020

Clustering Algorithm		**HFCM**		
Comparison		**Stations Order**		
M	Clusters			
3.6	C_1	SH1	SH3	SH4
	C_2	E20	T34	
	C_3	SH2B	SH2	E21

3-2-3 Discussing the results of 2021:

1.Determination of the appropriate number of clusters (k) and the fuzzing exponents (m) for the year 2021

Table 10: Shows the results of the test by adopting the P.C. standard andδ_{XB} to determine the fuzziness exponents and the number of clusters k for the year 2021

Clustering Algorithm		**HFCM**			
Comparison		**Obj.**	**Iteration**	**G. fit cluster**	
				P.C.	$\boldsymbol{\delta_{XB}}$
k	M				
2	1.2	27905165.91	6	1	5765.8
2	2	27566905.48	8	0.9922	3850.7
2	2.8	24297219.56	11	0.9389	3432.5
2	3.6	18624752.69	15	0.8626	2824.3
3	1.2	7843449.13	11	0.9997	3341.3
3	2	6914895.76	37	0.8601	3108.7
3	2.8	4664858.61	100	0.7234	2230.2
3	3.6	2603088.23	47	0.6401	683.84

Discuss the results of the table (10) as follows:

When **k = 2,** it is clear from the results of table (10) that the best degree of fuzzing is when (m = 3.6), as it achieved the lowest value of the objective function (Obj_Fun = 18624752.69), compared to the other fuzzing exponents.

According to P.C. standard. It was shown that the fuzzy cluster is invalid at any fuzzing exponents because it achieved a percentage higher than (85%), the fuzzy cluster has crisp cluster.

According to the $\boldsymbol{\delta_{XB}}$ criterion, the validity of the fuzzy cluster structure was demonstrated at the fuzzing exponent (m = 3.6), as it achieved (2824.3), which is the lowest value compared to the other values.

When **k = 3,** it is clear from the results of table (6) that the best degree of fuzzing is when (m = 3.6), as it achieved the lowest value of the objective function (Obj_Fun = 2603088.23), compared to the other fuzzing exponents.

According to **P.C.** standard. It showed the validity of the fuzzy cluster at the fuzzing exponent (m = 2.8), achieving a percentage of (0.7234), and the validity of the fuzzy cluster was achieved at the fuzzing exponent (m = 3.6), achieving a percentage of (0.6401).

According to the $\boldsymbol{\delta_{XB}}$ criterion, the validity of the fuzzy cluster structure was demonstrated at the fuzzing exponent (m = 3.6), as it achieved (683.84), which is the lowest value compared to the other values

We can also determine the level of overlapping points by determining the Average maximum membership Index AverageMax As shown in the figure below:

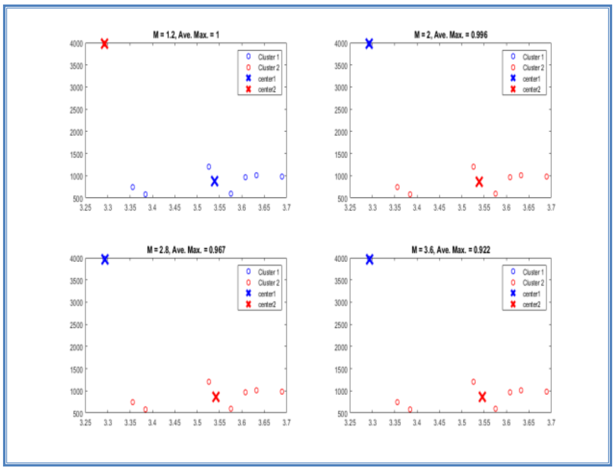

Fig (3-1) - overlap k=2

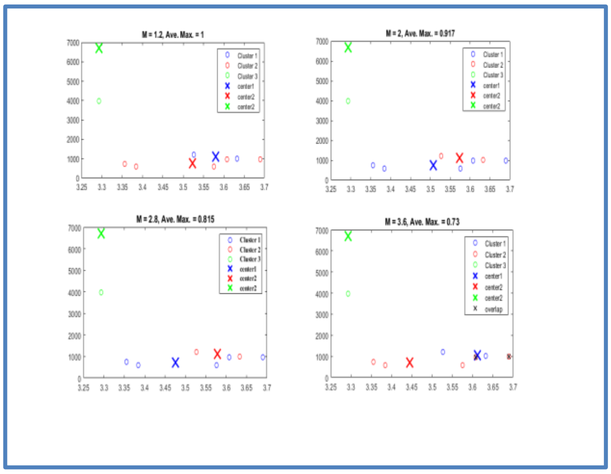

Fig (3-2) - overlap k=3

Figure 3: Shows cases of selecting degrees of fuzzing according to the AverageMax criterion in the case of
k = 2 and 3 for the year 2021

Figure (3), which shows the overlap cases according to the fuzzing exponents, in the case of k = 2, Figure (3-1) shows that there is no overlap between the clusters, as AverageMax recorded the membership degrees between [0.5, 1], which indicates a low state of uncertainty that is membership degree of the stations in the cluster have exceeded (61%), and it may reach one, and it is offset by a very weak membership degree to the other cluster.

In the case of k = 3, Figure (3-2) shows the interference cases that show the increase in uncertainty at the fuzzing exponent (m = 3.6) and with an average maximum affiliation 0.7296, which is station (SH2) that belonged to the second cluster with a membership degree (0.5158).) and to the first cluster with a degree of affiliation (0.4356) and to the third cluster with a membership degree (0.0486). As for the degrees of fuzziness (2.1, 2, 2.8), they did not achieve any overlap between the clusters, as they achieved an average maximum membership degree (1, 0.917, 0.815), respectively.

1. The clusters are therefore formed according to k = 3, m = 3.6

Table 11: Shows the fuzzy cluster formation of the physical and chemical examination stations for the year 2021

Clustering Algorithm		HFCM			
Comparison		Stations Order			
M	Clusters				
3.6	C_1	SH3			
	C_2	SH2	E20	E21	T34
	C_3	SH3			

3-3 Conclusions and recommendations:

Through the findings of the paper, we conclude that the method of the main core components contributes to treating the higher dimensions and reduces complexity. We also concluded the importance of the **Jollffie** criterion in determining the cut-off value, as it gives a wider space in the process of determining the contributing components. We also conclude that the use of the hybrid scenario between KPCA and FCM methods help to improve the clustering process under fuzzy logic, The study concluded that the PC segmentation coefficient showed weakness in determining the validity of the fuzzy cluster, while the objective function and the $\boldsymbol{\delta_{XB}}$ criterion were the best because of their characteristics. By applying this paper to the water sector in Basrah Governorate, the paper was able to from determining the most influential physical and chemical variables by studying (8) stations for examination and then clustering these stations according to fuzzy logic, where the paper reached to determine the best degree of fuzziness was m = 3.6 and the best number of clusters was k = 3 As for the stations that caused The cluster overlap status was SH1, SH2, SH3, SH4, E20, and T34.

Therefore, we recommend those interested in the water sector to study the importance of the variables achieved, as well as the stations resulting from clustering, and the development of techniques, statistical algorithms, and artificial intelligence in addressing these phenomena.

Authors Declaration:

Conflicts of Interest: None

-We Hereby Confirm That All The Figures and Tables In The Manuscript Are Mine and Ours. Besides, The Figures and Images, Which are Not Mine, Have Been Permitted Republication and Attached to The Manuscript.

- Ethical Clearance: The Research Was Approved By The Local Ethical Committee in The University.

4.References :

1- Ahmed, A. D., Abboud, S. N. and Mahdi, D. I. (2015), "Classification The Iraq Provinces according to some variants of the health sector", AL-Qadisiyah Journal for Administrative and Economic Sciences, Vol.17, No.3, pp. 271-285.

2- Ali, O. A. and Salman, F. H. (2012), "Using the Statistical Analysis for deduction the childhood status in Iraq during 2006-2010", Journal of Economics and Administrative Sciences,Vol.18, No.66, pp. 306-331.

3- Al-mousa, Y., Al-Jasem, A. and Dahhand, M. L. (2015), "Improve the Result of K-Means Algorithms using Factor Analysis". Rescerch Journal of Aleppo University, Vol.15, pp. 1-22.

4- Ashour, M. A. H. and Jawad, M. A. (2017), "Using Fuzzy Games Theory to Determine the optimal Strategy for The Mobile Phone Networks in The Baghdad and Basra governorates", Journal of Economics and Administrative Sciences; Vol.23, No. 95, pp. 399-427.

5- Bezdek, J. C. (1974). “Cluster Validity with Fuzzy Sets”. Journal of Cybernetics, Vol.3, No.3, pp. 58-73.
6- Blanchard, G., Bousquet, O. and Zwald, L. (2006), "Statistical properties of kernel principal component analysis", Mach Learn, Vol. 66 , No. 2-3 , p. 259–294.
7- Chacon, J. E. and Duong, T. (2018). "Multivariate Kernel Smoothing and Its Applications", (1ed), Chapman and Hall/CRC, pp. 1-248.
8- Chacon, J. E. E., Duong, T. and Wand, M. P. (2011), "Asymptotics for general multivariate kernel density derivative estimators", Statistica Sinica, Vol. 21, No. 2 , pp. 807-840.
9- Dogruparmak, S. C., Keskin, G. A., Yaman, S. and Alkan, A. (2014), "Using principal component analysis and fuzzy c–means clustering for the assessment of air quality monitoring", Atmospheric Pollution Research, Vol. 5, No. 4, pp. 656-663.
10- Duong, T. (2004). "Bandwidth selectors for multivariate kernel density estimation", Ph.D thesis, University of Western Australia, School of Mathematics and Statistics.
11- Dziuban, C. D., and Shirkiy, E. C. (1974). “When is a correlation matrix appropriate for factor analysis? Some decision rules”. Psychological Bulletin, Vol.81, No.6, pp.358-361.
12- El Fattahi, L. and Sbai , E. H. (2021), "Fault Detection and Identification for Nonlinear Process Based on Inertia-Based KEPCA and a New Combined Monitoring Index", Journal of Electrical and Computer Engineering, Vol.2021, pp. 1-10.
13- El-Zaghmouri, B. M. and Abu-Zanona, M. A. (2012), "Fuzzy C-Mean Clustering Algorithm Modification and Adaptation for Applications and Adaptation for Applications", World of Computer Science and Information Technology Journal, Vol. 2, No. 1, pp. 42-45.
14- Essa, A. M. and Alrawi, A. G. (2019), "Comparison Between The Method of Principal Component Analysis And Principal Component Analysis Kernel For Imaging Dimensionality Reduction", Iraqi Journal of Statistical Sciences, Vol. 16, No. 2, pp. 11-24.
15- Fawzi, H. M. and Jaber, A. G. (2021), "Comparison of Some Non-Parametric Quality Control Methods", Journal of Economics and Administrative Sciences; Vol. 27, No. 130, pp. 197-209.
16- Fawzi, R. M. and Alkanani, I. H. (2020), "Turbid of Water By Using Fuzzy C- Means and Hard K- Means". Baghdad Science Journal, Vol. 17, No. 3, pp. 988-993.
17- Goyal, L. M., Mittal, M. and Sethi , J. K. (2016), "Fuzzy model generation using Subtractive and Fuzzy C-Means clustering", CSI Transactions on ICT, Vol. 4, pp. 129–133.
18- Hamed, M. A. R. (2019), "Application of Surface Water Quality Classification Models Using Principal Components Analysis and Cluster Analysis", Journal of Geoscience and Environment Protection, Vol. 7, pp. 26-41.
19- Henderson, D. J. and Parmeter, C. F. (2012), "Normal reference bandwidths for the general order, multivariate kernel density derivative estimator", Statistics and Probability Letters, 9. Vol. 82, No.12, pp. 2198-2205.
20- Hmood, M. Y. (2005), "Comparing Nonparametric Estimators For Probability Density Estimation", Ph.D Thesis, University of Baghdad, College of Administration and Economics.
21- Hojjatinia, S. and Lagoa, C. M. (2019), "Comparison of Different Spike Sorting Subtechniques Based on Rat Brain Basolateral Amygdala Neuronal Activity", United States; San Diego, PublisherInstitute of Electrical and Electronics Engineers Inc., pp. 2251-2258.
22- Höppner, F., Klawonn, F., Kruse, R., and Runkler, T. (1999). “Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition”, (2ed.), John Wiley & Sons, LTD Chichester, pp.1-304.
23- Javadi, S., Rameez, M., Dahl, M. and Pettersson, M. I. (2018), "Vehicle Classification Based on Multiple Fuzzy C-Means Clustering Using Dimensions and Speed Features", Procedia Computer Science, Vol.126, pp. 1344–1350.
24- Jolliffe, I. (2002). “Principal Component Analysis”, (2 ed.). Springer New York, NY, pp.1-478.

25- Kaittan, M. Q. (2018), "Improve the Spatial Resolution of Multispectral satellite Image using Different Image Sharpening Techniques", Iraqi Journal of Science, Vol. 59, No. A1, pp. 227-232.

26- Liu, L. and Shao, H. (2023), "Study on neutron–gamma discrimination method based on the KPCA-GMM. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers", Detectors and Associated Equipment, November , Vol.1056, 168604.

27- Liu, X. and Yang, C. (2009), "greedy kernel PCA for training data reduction and nonlinear feature extraction in classification", Yichang, China, Society of Photo-Optical Instrumentation Engineers (SPIE),Vol. 7495, pp. 1-9.

28- mohammed, H. Y. and Muhamed, L. A. (2020), "on kernel principal component analysis", Journal of Administration and Economics, Vol. 123, pp. 376-394.

29- Mohammed, L. A. and Abbod, A. A. (2016), "comparing bandwidth estimatore (smooting parameter) by Using Of Kernel function in the principal Component Analysis", Journal Of AL-Turath University College, No. 20, pp. 412-436.

30- Mushtaq, Z., Qureshi, M. F., Abbass, M. J. and Al-Fakih, S. M. Q. (2023), "Effective kernel-principal component analysis based approach for wisconsin breast cancer diagnosis", Electronics Letters, Vol. 59, No. 2, pp. 1-4.

31- Naif, Q. N. and Ayoub, M. K. (2016), "user (K-Means) for clustering in Data Mining with application", Journal of Economics and Administrative Sciences, Vol. 22, No. 91, pp. 389-406.

32- Oliveira, J. V. and Pedrycz, W. (2007), "Advances in Fuzzy Clustering and its Applications", (1ed),John Wiley and Sons Ltd, The Atrium, Southern Gate, Chichester, pp. 1-454.

33- Rencher, A. C., & Christensen, W. F. (2012). “Methods of Multivariate Analysis”, (3ed.), Wily, a John Wily and Sons, INC, pp. 1-768.

34- Scott, D. W. (1992), "Multivariate Density Estimation: Theory, Practice, and Visualization", 1ed., A Wiley-Interscience Publication, John Wiley and SONS, INC, pp. 1-317.

35- Sreenivasarao, V. and Vidyavathi, S. (2010), "Comparative Analysis of Fuzzy C- Mean and Modified Fuzzy Possibilistic C -Mean Algorithms in Data Mining", International Journal of Computer Science and Technology, Vol. 1, No. 1, pp. 104-106.

36- Wand, M. P. and Jones, M. C. (1994), "Multivariate plug-in bandwidth selection", Computational Statistical, No. 9, pp. 97-116.

37- Wand, M. P. and Jones, M. C. (1995), "Kernel Smoothing", Chapman and Hall/CRC; New York, pp. 1-224.

38- Wang, J. and Zhang, Y. (2015), "Speaker Recognition Based on KPCA and KFCM", International Conference on Mechatronics, Electronic, Industrial and Control Engineering (MEIC 2015), Atlantis Press, pp. 181-184.

39- Wierzchon, S. T. and Kłopotek, M. A. (2018), "Modern Algorithms of Cluster Analysis", 1ed. Springer, Cham, pp. 1-417.

Journal of Economics and Administrative Sciences
2024; 30(141), pp. 457-478

P-ISSN 2518-5764
E-ISSN 2227-703X

بناء خوارزمية هجينة لنمذجة بيانات محطاة الفحص الفيزيائي والكيميائي لمياه شط العرب*

احمد هشام محمد
جامعة بغداد/ كلية الإدارة والاقتصاد/ قسم الاحصاء
بغداد، العراق
ahmed.albasrai@uobasrah.edu.iq

مروان عبد الحميد عاشور
جامعة بغداد/ كلية الإدارة والاقتصاد/ قسم الاحصاء
بغداد، العراق

Received:13/8/2023 Accepted:16/10/2023 Published Online First: 30 /6/ 2024

مستخلص البحث:

ان الهدف الاساسي من هذه الورقة هو ايجاد اسلوب هجين بين عدد من الاساليب الاحصائية التي تتعامل مع الابعاد العليا غير الخطية اذ ان البيانات غالباً ما تعاني من مشكلات التعقيد والتداخل في دوالها الرياضية فيكون من الصعب فصل او تحديد بدقة اثر كل متغير على الاخر، وعليه تم بناء نموذج هجين بين اسلوبي المركبات الرئيسية اللبية KPCA والعنقدة الضبابية FCM، اذ يهدف اسلوب KPCA الى معالجة مشكلة البيانات غير الخطية عالية الابعاد وتخفيضها من خلال ايجاد مصفوفة كيرنل Kernel matrix التي تعتمد بشكل اساسي على مصفوفة معلمة التمهيد (h_j) التي تم تقديرها باعتماد طريقة ROT ومن ثم اعتماد اسلوب عنقدة المتوسطات الضبابية FCM للحصول على العناقيد. تم تطبيق هذه المنهجية على قطاع المياه في محافظة البصرة من خلال دراسة (8) محطات للفحص الفيزيائي والكيميائي من خلال (15) متغيراً للسنوات (2019، 2020، 2021) وتم جمع البيانات بشكل شهري، ومن خلال تطبيق هذه المنهجية تمكنت الورقة من تحديد (7) متغيرات اساسية وهي (**TH، Na، Cl، TDS، No3، EC، G_O**) اما بالنسبة للمحطات فقد تم تحديد المحطات المتداخلة بين العناقيد وهي (**SH1، SH2، SH3، SH4، E20، T34**)، اما بالنسبة للافضل درجة تضبيب كانت (**3.6**) و وافضل عدد عناقيد هي (k=3) .

نوع البحث: ورقة بحثية

المصطلحات الرئيسة للبحث: البيانات غير الخطية، اللامعلمية، KPCA، العنقدة الضبابية، FCM.

*بحث مستل من اطروحة دكتوراه