Clustering Algorithms: An Application for Adsorption Kinetic Curves

Eréndira Rendón, Roberto Alejo and José L. García-Rivas

***Abstract*—Clustering algorithms have been used in different areas of knowledge with different goals such as noise detection, outliers, and descriptive tasks. The adsorption kinetics is a curve that describes the rate retention to the adsorbate on the adsorbent at time, which is represents as a two-dimensional graph. In this paper, we present a computational application to determine the experimental conditions that influence when equilibrium point is reached into adsorption kinetics curve using the K-means clustering algorithm and, Parallel Coordinates concept, in order to prove our method we used adsorption kinetic curves Q-PVA . Results obtained were compared with two designs of experiments (three-stage nested design and hierarchical design with crossed factors).**

***Index Terms*—Adsorption kinetic curves, chitosan, K-means algorithm, parallel coordinates.**

I. Introducción

Las técnicas de agrupamiento (clustering) son técnicas de aprendizaje no supervisado, que se han utilizado en diferentes áreas de conocimiento, como el aprendizaje automático, minería de datos, reconocimiento de patrones, procesamiento de imágenes, bioinformática [1], ciencia de materiales [2] y recientemente en "Materials Informatics" [3]. El agrupamiento es una tarea descriptiva que tiene como objetivo encontrar grupos en un conjunto de datos y explicarlos mediante sus características [4]. Específicamente, las técnicas de agrupamiento encuentran particiones en conjuntos de datos, tal que los objetos de cada partición son lo más similar posible entre ellos y disimilares a los objetos de otras particiones [4].

En años recientes ha emergido un área, que en el idioma inglés se denomina "Materials Informatics", la cual ha proporcionado una alternativa en la Ciencia de Materiales para la predicción de propiedades y diseño de nuevos materiales [5], [6]. Tradicionalmente, en la Ciencia de Materiales, los experimentos son realizados directamente en el laboratorio, para después realizar un análisis de los resultados por parte del científico o investigador [7]. Actualmente, existen varios ejemplos donde se han diseñado nuevos materiales utilizando algún software, los cuales posteriormente son desarrollados en el laboratorio para su comprobación [8], [9], [10], [11], [12], [13]. Por ejemplo, Varde [7] realizó un estudio del proceso de tratamiento térmico de los materiales, el cual involucra el calentamiento controlado y el enfriamiento rápido de un material para lograr propiedades térmicas y mecánicas deseadas, para lo cual se utilizó el algoritmo k-means para agrupar las gráficas resultantes de los experimentos del temple a partir de las condiciones iniciales de dicho experimento, también se utilizaron árboles de decisión para realizar el proceso de manera inversa, es decir, a partir de las gráficas obtener las condiciones iniciales del experimento. Además se obtuvo una eficiencia de 90%-95% en la estimación de nuevos materiales. Saad et al., [14] analizó las propiedades de los componentes atómicos utilizando una técnica de aprendizaje no supervisado, que les permitió separar 67 componentes octetos en varias clases de acuerdo a su estructura cristalina; de igual manera utilizaron técnicas de aprendizaje supervisado, donde encontraron la estructura cristalina correcta de 55 compuestos logrando una exactitud del 95% y finalmente, utilizando algoritmos de regresión lograron predecir el punto de fusión de 44 sub-octetos con un error relativo promedio de 12.8%. Tinoco et al., [15] utilizaron redes neuronales artificiales, máquinas de vectores de soporte y redes funcionales de base radial (NN, SVM y RBF, por sus siglas en inglés, respectivamente) de manera exitosa para un mejor entendimiento de las propiedades mecánicas de los experimentos del laboratorio *Jet Grouting*, donde se calculó la tangente del módulo elástico de Young, aplicando una tensión máxima del 50%. El conocimiento obtenido en este estudio permitió un mejor entendimiento del comportamiento de las propiedades de las mezclas suelo-cemento a través del tiempo. Suh et al., [16] utilizaron análisis de componentes principales (PCA, por sus siglas en inglés) para reducir el número de descriptores (número medio de electrones de valencia, electronegatividad, diferencia de radios, concentración, estequiometría, fracción de energía de cohesión y energía de ionización) de arreglos multidimensionales que describen materiales superconductores a altas temperaturas, para posteriormente realizar la clasificación de los materiales en superconductores y no-superconductores. Morgan [17] propuso un método llamado DMQC (por sus siglas en inglés, Data Mining of Quantum Calculations), cuyo objetivo es reducir el tiempo de predicción de estructuras cristalinas con métodos ab initio. El método DMQC utiliza la técnica de análisis de componentes principales para reducir el número de descriptores, y encontrar las correlaciones lineales entre las energías de las estructuras cristalinas de los materiales de la base de datos, y así reducir el tiempo de predicción. Ortiz y Eriksson [18] utilizaron algoritmos de minería de datos para encontrar nuevos materiales con propiedades capaces de detectar radiaciones nucleares, para lo cual emplearon una base de datos de 22,000 compuestos inorgánicos que

Eréndira Rendón, Tecnológico Nacional de México / IT Toluca, Metepec, Estado de México, C.P. 52149, México, e-mail: erendonl@toluca.tecnm.mx.

Roberto Alejo, Tecnológico Nacional de México / IT Toluca, Metepec, Estado de México, C.P. 52149, México, e-mail: ralejoe@toluca.tecnm.mx. *Autor para correspondencia: Roberto Alejo.*

José L. García-Rivas, Tecnológico Nacional de México / IT Toluca, Metepec, Estado de México, C.P. 52149, México, e-mail: jgarciar@toluca.tecnm.mx.

fueron tomados de la base de datos de estructuras cristalinas ICSD. Los materiales fueron descritos por sus propiedades electrónicas; además, utilizaron algoritmos de minería de datos para encontrar 136 nuevos y prometedores materiales los cuales pueden ser utilizados potencialmente para detectar radiaciones. Apostolakis [19] proporciona una introducción a las técnicas de minería de datos y su potencial uso en la cristalografía. También se presenta una visión general de las tareas principales de la minería de datos: descriptiva y predictiva, mostrando como se han aplicado estas técnicas a la cristalografía inorgánica, donde se han utilizado técnicas como el análisis de componentes principales (PCA). Cavas et al. [20] realizaron una comparación del modelo de Thomas [21] y las redes neuronales artificiales para describir el comportamiento del proceso de adsorción mediante $P.oceanica$. El modelo de la red neuronal propuesto dio mejores resultados que el modelo de Thomas (ver Ref. [21]), con lo cual se pudo determinar que las hojas secas de $P.oceanica$ pueden ser utilizadas para la eliminación de colorantes de las aguas residuales de la industria textil. Por su parte, Meden [22] presenta un amplio estudio de cómo la minería de datos puede ayudar a predecir nuevas estructuras cristalográficas de materiales inorgánicos.

La capacidad de adsorción de los adsorbentes comúnmente se representan en gráficas en dos dimensiones, estas gráficas son una excelente herramienta visual para su análisis y comparación de los procesos de adsorción [23]. En este trabajo se propone una solución computacional que permite determinar las condiciones experimentales (variables dependientes) que influyen en el punto de equilibrio de las curvas de cinética de adsorción. Para lo cual se utilizó el algoritmo de agrupamiento k-means [24] y el concepto de coordenadas paralelas [25]. Para probar nuestra propuesta se utilizaron curvas de cinéticas de adsorción mediante un hidrogel a base de quitosano y alcohol polivinílico (Q-PVA). Las curvas de cinética de adsorción de un hidrogel base quitosano, se obtuvieron directamente en el laboratorio, es decir los experimentos se realizaron modificando variables como: temperatura, tipo de colorante y tamaño de esfera o perla, para determinar la cantidad adsorbida de colorante Amarillo 5 (Ama5) o Amarillo 6 (Ama6) por las perlas del hidrogel base quitosano. La velocidad de agitación, pH y presión se mantuvieron constantes.

II. Cinética de Adsorción

En general, los estudios de adsorción, equilibrio y cinética se realizan con procedimientos estándar que consisten en mezclar un volumen fijo de solución de colorante con una cantidad conocida de quitosano en condiciones controladas de tiempo de contacto, velocidad de agitación, temperatura y pH. En algunas ocasiones la concentración residual del colorante es determinada por espectrofotometría en el máximo de longitud de onda de absorción por radiación UV [26]. La cinética describe la velocidad de adsorción del adsorbato en el adsorbente y determina el tiempo en que se alcanza el equilibrio. La adsorción implica la concentración de uno o más componentes de un líquido en la superficie de un sólido. El sólido se denomina adsorbente y las moléculas adsorbidas en la superficie del sólido, con mayor concentración que en la fase fluida, se conocen como adsorbato. La adsorción se establece debido a las fuerzas de atracción entre las moléculas de fluido y la superficie sólida. Si las fuerzas son de tipo Van der Waals [27], conllevan una fisisorción sobre la superficie del adsorbente, resultado de interacciones intermoleculares débiles entre el sólido y el fluido. La adsorción activada o quimisorción ocurre cuando se forman enlaces químicos entre las moléculas de fluido y la superficie adsorbente. La energía de adsorción en fisisorción es muy inferior a la que se implica en un enlace químico, y por tanto, la reversibilidad del proceso se obtiene sometiendo al sistema a un calentamiento, o bien al vacío, de forma que disminuya la presión del adsorbato [23].

El conocimiento del equilibrio de adsorción para un determinado sistema adsorbato-adsorbente posibilita el diseño de las condiciones de operación, presión y temperatura de trabajo para desarrollar proyectos a mayor escala. Además, a través de la información que se obtiene de los datos de equilibrio de adsorción, es posible establecer las características del adsorbente; por tanto, en ocasiones el equilibrio de adsorción se utiliza para caracterizar materiales para su posterior uso como adsorbente o catalizador.

III. Agrupamiento

Agrupamiento es una técnica de aprendizaje no supervisada cuyo objetivo es encontrar o descubrir grupos o particiones en conjuntos de datos u objetos [1]. A estas particiones normalmente se les denomina grupos, tal que los objetos que pertenezcan a un mismo grupo sean similares entre ellos y disimilares a los objetos de los otros grupos [28]. El agrupamiento es una de las tareas más importantes en el análisis y la minería de datos [29], y ha sido ampliamente utilizada en la detección de anomalías e identificación de características sobresalientes en conjuntos de datos en diferentes áreas del conocimiento como la biología, la antropología, ciencia de materiales, medicina, estadística y matemáticas por mencionar algunas [30], [31]. Se han desarrollado una gran diversidad de métodos de agrupamiento desde sus inicios en la década de 1950 [24], [4], los cuales han sido divididos en dos grupos: divisorios y jerárquicos. Uno de los algoritmos de agrupamiento divisorios o de partición más conocido y utilizado es el k-means [32], inclusive, hoy en día. La mayoría de los algoritmos de partición tienen como base la optimización de una función criterio [1], generalmente, para k-means esta función es representada por E (Ec. 1), y su valor depende de las particiones o grupos (C_i) en el conjunto de datos $\mathbf{X}$.

$$E = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mathbf{m}_i\|^2, \tag{1}$$

donde E es la suma del error cuadrático de todos los objetos en el conjunto de datos $\mathbf{X}$, y los centros o medias $\mathbf{m}_i$ (Ec. 2) del grupo C_i, $\mathbf{x}$ es un punto en el espacio que representa a un objeto dado en un espacio multidimensional [28], y

$$\mathbf{m}_i = \frac{1}{\|C_i\|} \sum_{j}^{\|C_i\|} \mathbf{x}_j. \tag{2}$$

Básicamente, el funcionamiento del algoritmo k-means inicia seleccionando o calculando k centros o medias iniciales $\mathbf{m}_i^0$,

dependiendo del criterio de selección, comúnmente se toman aleatoriamente k objetos de $\mathbf{X}$; A continuación, se asigna cada objeto $\mathbf{x}_j \in \mathbf{X}$ a su centro mas cercano $\mathbf{m}_i$. Posteriormente, se calculan los nuevos centros o medias $\mathbf{m}_i$ (Ec. 2) hasta que el algoritmo converge a un valor mínimo de E (Ec. 1), o hasta un máximo de repeticiones Q, las cuales se establecen antes de iniciar el procedimiento, i.e., repetir este proceso hasta que $\|E^{(q)} - E^{(q-1)}\| < \Delta$ o $q = Q$, donde $q = \{1, 2, ..., Q\}$ y corresponde al número de repetición en ese momento. En el algoritmo 1, se explica a detalle este proceso.

Algorithm 1 Pseudo-código del algoritmo k-means

Entrada: Conjunto de datos $\mathbf{X}$, número de grupos (k) ;
Salida: Grupos obtenidos $\{C_1, C_2, ..., C_k\}$;

1: Establecer el criterio de convergencia: Error mínimo ($\Delta = 0.0001$) y máximo número de repeticiones ($Q = 1000$);
2: Asignar aleatoriamente k objetos de $\mathbf{X}$ como centros o medias iniciales ($\mathbf{m}_i$);
3: $q = 1$;
4: **repeat**
5: **for** $i = 1$ **to** k **do**
6: $\mathbf{x}^* \leftarrow$ Min $dist(\mathbf{x}_j, \mathbf{m}_i)$; // $\mathbf{x}_j \in \mathbf{X}$.
// $\boldsymbol{x}^*$ es el subconjunto de los vecinos más cercanos $\boldsymbol{x}_j$ a $\boldsymbol{m}_i$
7: $C_i \leftarrow \mathbf{x}^*$; *// Asignar los objetos $\boldsymbol{x}^*$ a su centro mas cercano*
8: **end for**
9: **for** $i = 1$ **to** k **do**
10: $s = \|C_i\|$; *// Calcular los nuevos centros*
11: $\mathbf{m}_i = \frac{1}{s}\sum_j^s \mathbf{x}_j$; // $\mathbf{x}_j \in C_i$
12: **end for**
13: $q++$;
14: **until** $\{(\|E^{(q)} - E^{(q-1)}\| < \Delta) \,||\, (q = Q)\}$;

A. Coordenadas Paralelas

Inselberg [33] define a un objeto multidimensional como un grupo de variables que están asociadas a una misma instancia. Estas variables pueden representar diferentes características de un objeto o presentar la misma característica en diferentes condiciones. Existen diferentes formas de representar a un objeto, por ejemplo, puede ser representado como un sistema de coordenadas paralelas, donde cada eje vertical (ordenadas) representa una característica o atributo del objeto (dimensión), el cual puede ser continuo o categórico. Cada uno de los ejes verticales de un sistema de coordenadas paralelas puede tener su propia escala [34]. Las coordenadas paralelas son un método que se ha utilizado para visualizar un plano n-dimensional en una representación 2D y han sido de mucha utilidad para visualizar patrones de los datos o para percibir relaciones entre características. Formalmente, las coordenadas paralelas se definen sobre el plano cartesiano [35], donde se realizan n-copias del eje Y, equidistantes y perpendiculares al eje X en los puntos que son representados como p_j, donde cada eje representa una coordenada del punto $\mathbf{P} = \{p_1, p_2, ..., p_n\}$, en la n-ésima dimensión.

Considerando las definiciones anteriores, una curva de cinética de adsorción se puede ver como un punto $\mathbf{P}_i$ en un espacio n-dimensional, que está representado por una línea polígonal (aquella que está formada solo por segmentos de recta unidos) con n vértices para los valores p_j, donde (en este trabajo) p_j son las cantidades adsorbida y los vértices j el tiempo en el cual fue adsorbido p_j. Por ejemplo, la Fig. 1 muestra dos curvas de cinética de adsorción representadas por los vectores $\mathbf{P}_1 = \{247, 339, 491, 546, 632, 691, 750, 800, 847, 877, 877\}$, $\mathbf{P}_2 = \{134, 174, 363, 467, 541, 606, 647, 769, 771, 774, 774\}$, en la cual el eje X representa el tiempo de adsorción del colorante, y el eje Y el valor de adsorción. En particular, en la Fig. 1 se observa que para el punto $\mathbf{P}_1$ se absorbió 247 $(q(mg/g))$ en 0.5 horas, 339 $(q(mg/g))$ en 1 hora, y finalmente 877 $(q(mg/g))$ en 72 horas. El mecanismo de coordenadas paralelas, como se ve en la Fig. 1, permite comparar gráficamente objetos multidimensionales diferentes [25], en este caso, los puntos $\mathbf{P}_1$ y $\mathbf{P}_2$, representan dos experimentos de cinética de adsorción en condiciones diferentes. $\mathbf{P}_1$ y $\mathbf{P}_2$ corresponden a una cinética de adsorción de Ama5 a $50C$ y a $10C$, con esfera chica y mediana, respectivamente.

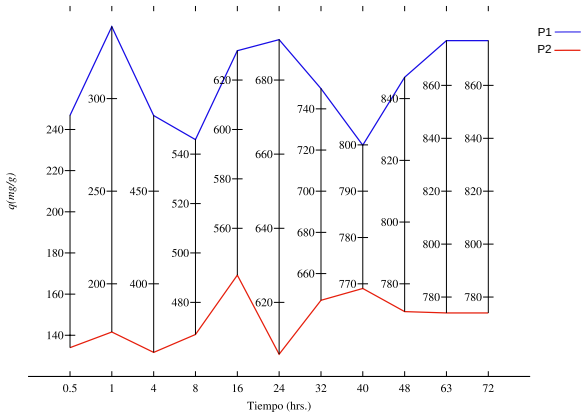

Fig. 1. Ejemplo de la cinética de adsorción del colorante amarillo 5 (Ama5) con esfera chica y mediana a $50C$ y a $10C$, $\mathbf{P}_1$ y $\mathbf{P}_1$, respectivamente.

IV. Materiales y Métodos

En esta sección se presentan las principales características experimentales de este trabajo, como los materiales y métodos empleados. Primeramente, se discute el procedimiento seguido para la generación de las muestras, las cuales corresponden a experimentos de laboratorio independientes con características distintas para la remoción de Ama5 y Ama6. Posteriormente, se detalla el proceso de transformación de cada experimento a una curva de cinética de adsorción, la cual es visualizada en un sistema de coordenadas paralelas (sección III-A). Más adelante, se muestran las características técnicas de la aplicación del algoritmo de agrupamiento k-means, seguidamente, de la interpretación de los grupos obtenidos, y finalmente, el diseño de experimentos utilizado para dar mayor soporte a las conclusiones presentadas.

La preparación de cada muestra o *experimento* consiste en mezclar un volumen fijo de solución de colorante con una cantidad conocida de quitosano en condiciones controladas de tiempo de contacto, velocidad de agitación, temperatura y pH. La capacidad de adsorción se expresa en mg de colorante adsorbido por gramos de adsorbente $(q(mg/g))$. Se trabajó con dos colorantes, amarillo 5 y 6 (Ama5 y Ama6, respectivamente) en solución acuosa mediante un hidrogel de quitosano-alcohol polivinílico, en el cual se realizaron experimentos con perlas de hidrogel en tres tamaños $(2, 2.6$ y $3.1mm)$ a tres diferentes temperaturas $(10, 30$ y $50C)$ con muestras tomadas en los tiempos: 0.5, 1, 4, 8, 16, 24, 32, 40, 48, 63

y 72 horas. Los resultados de estos experimentos permitieron conformar una base de datos de gráficas o curvas de cinética de adsorción [23] de hidrogel base quitosano independientes, las cuales fueron organizadas en dos conjuntos de experimentos. El primer grupo consistió en ejecutar el procedimiento descrito previamente, una sola vez, a los cuales se nombraron como Exp_i, donde $i = 1, 2, ..., 18$ (véase la Tabla I). En el segundo, cada experimento fue ejecutado tres veces, por ejemplo, la combinación {Ama5, a $10C$, esfera chica} corresponde a tres experimentos diferentes (1, 2 y 3, ver la Tabla I), por esa razón se denominaron ExpRep_j, para $j = 1, 2, ..., 54$, o en otras palabras experimentos con réplica. La Tabla I, detalla las principales características experimentales en cada conjunto de experimentos asociados a sus respectivas curvas de cinética de adsorción. Asimismo, debido a que los resultados de cada

TABLA I
CARACTERÍSTICAS DE LOS EXPERIMENTOS ESTUDIADOS

Tipo de Colorante	Temperatura (ºC)	Tamaño de la perla	Experimento
	10	Chica	Exp_1
	30	Chica	Exp_2
	50	Chica	Exp_3
	10	Mediana	Exp_4
Ama5	30	Mediana	Exp_5
	50	Mediana	Exp_6
	10	Grande	Exp_7
	30	Grande	Exp_8
	50	Grande	Exp_9
	10	Chica	Exp_{10}
	30	Chica	Exp_{11}
	50	Chica	Exp_{12}
	10	Mediana	Exp_{13}
Ama6	30	Mediana	Exp_{14}
	50	Mediana	Exp_{15}
	10	Grande	Exp_{16}
	30	Grande	Exp_{17}
	50	Grande	Exp_{18}
		Con Réplica	
	10	Chica	ExpRep_1, ExpRep_2, ExpRep_3
	30	Chica	ExpRep_4, ExpRep_5, ExpRep_6
	50	Chica	ExpRep_7, ExpRep_8, ExpRep_9
	10	Mediana	ExpRep_{10}, ExpRep_{11}, ExpRep_{12}
Ama5	30	Mediana	ExpRep_{13}, ExpRep_{14}, ExpRep_{15}
	50	Mediana	ExpRep_{16}, ExpRep_{17}, ExpRep_{18}
	10	Grande	ExpRep_{19}, ExpRep_{20}, ExpRep_{21}
	30	Grande	ExpRep_{22}, ExpRep_{23}, ExpRep_{24}
	50	Grande	ExpRep_{25}, ExpRep_{26}, ExpRep_{27}
	10	Chica	ExpRep_{28}, ExpRep_{29}, ExpRep_{30}
	30	Chica	ExpRep_{31}, ExpRep_{32}, ExpRep_{33}
	50	Chica	ExpRep_{34}, ExpRep_{35}, ExpRep_{36}
	10	Mediana	ExpRep_{37}, ExpRep_{38}, ExpRep_{39}
Ama6	30	Mediana	ExpRep_{40}, ExpRep_{41}, ExpRep_{42}
	50	Mediana	ExpRep_{43}, ExpRep_{44}, ExpRep_{45}
	10	Grande	ExpRep_{46}, ExpRep_{47}, ExpRep_{48}
	30	Grande	ExpRep_{49}, ExpRep_{50}, ExpRep_{51}
	50	Grande	ExpRep_{52}, ExpRep_{53}, ExpRep_{54}

experimento (Exp_i y ExpRep_j) corresponden a condiciones diferentes, como la temperatura, el tamaño de la esfera (chica, mediana o grande), el tipo de colorante (Ama5 o Ama6), entre otras; y a la necesidad de que puedan compararse entre sí, sus respectivas curvas de cinética de adsorción, es fundamental el uso de un mecanismo de visualización especializado, como el concepto de coordenadas paralelas (ver sección III-A), en el que cada experimento Exp_i o ExpRep_j se transforma en un vector $\mathbf{P}= \{p_1, p_2, p_3, ..., p_j, ..., p_n\}$, donde p_j es la cantidad de colorante removido en cada intervalo de tiempo j; $j = 1, 2, ..., n$, para más detalle véase ejemplo de la Fig. 1. En este trabajo los valores de tiempo son fijos, i.e., $n = 11$.

La parte central de este trabajo es la aplicación del algoritmo k-means para encontrar similitudes entre las curvas de cinética de adsorción de colorantes Ama5 y Ama6, en diferentes experimentos realizados en un laboratorio bajo distintas condiciones. El procedimiento seguido fue primeramente ejecutar o aplicar el algoritmo k-medias, 100 veces, con el propósito de reducir la posibilidad de caer en un mínimo local (la principal debilidad del algoritmo k-means [1]) y encontrar valores de inicialización más apropiados. Al finalizar la aplicación de varias pruebas (100) de del algoritmo k-medias se eligieron los valores que obtuvieron mejores resultados. El mecanismo de evaluación de la calidad del agrupamiento fue el error cuadrático (Ec. 1), y se escogió la inicialización del algoritmo que dio un valor mínimo de E.

La interpretación de los agrupamientos cosiste en analizar la estructura u organización de las curvas de cinética de adsorción correspondientes a condiciones y características diferentes y encontrar las similitudes o diferencias fundamentales, que puedan permitir establecer condiciones iniciales $apriori$ antes de que un experimento en el laboratorio sobre este tema sea llevado a acabo. Por ejemplo, hasta qué punto el tamaño de la perla es crítico, o inclusive la temperatura. Esto permitirá reducir costos de experimentación en laboratorio al tener información $apriori$ sobre el comportamiento del experimento a realizar. Para esta interpretación en este trabajo se uso el concepto de coordenadas paralelas, el cual permite encontrar patrones o regularidades en datos multidimensionales en un espacio bidimensional [34], [25]. Para determinar si los resultados obtenidos con el algoritmo k-means, son estadísticamente significantes se utilizaron dos tipos de diseño de experimentos: jerárquico con factores cruzados y un diseño jerárquico anidado en tres etapas [36].

V. PRUEBAS Y RESULTADOS EXPERIMENTALES

Esta sección esta dividida en tres partes, en la primera se presentan los resultados experimentales generados en el primer grupo de experimentos, es decir, sin réplicas (Exp_1 al Exp_{18}); y la segunda los resultados obtenidos con réplicas: ExpRep_1 al ExpRep_{54} (ver Tabla I). Finalmente, para dar mayor soporte a los conclusiones obtenidas en este trabajo los resultados del análisis estadístico son discutidos.

A. Sin Réplica

El algoritmo k-medias fue aplicado 100 veces al conjunto de datos que representa los experimentos del 1 al 18 (Exp_1 al Exp_18), obteniéndose que el valor de $k = 3$ es el que generó el menor error cuadrático E (Ec. 1). Así los resultados presentados en esta sección corresponden a la aplicación del algoritmo k-medias con $k = 3$. Los grupos obtenidos por el algoritmo k-means cuando se utilizaron los experimentos sin réplica, se presentan en la Tabla II, donde la primera columna representa el número de grupo al que fue asignado cada Exp_m después de aplicarse el algoritmo; y la segunda identifica el número de experimento (m) en ese grupo. Para analizar globalmente los resultados de la Tabla II, se realizaron gráficas independientes de cada grupo (Fig. 2), las cuales corresponden a la representación de la cinética de adsorción

TABLA II
AGRUPACIÓN DE LAS CURVAS DE CINÉTICA DE ADSORCIÓN

No. de grupo	No. de experimento (m) en cada grupo
1	1, 2, 13, 14, 15, 16
2	10, 11, 12, 17, 18
3	3, 4, 5, 6, 7, 8, 9

por medio de coordenadas paralelas. En la Fig. 2, el eje X es el tiempo transcurrido en el experimento y el Y la cantidad de colorante removido (para más detalle véase la sección III-A). La Fig. 2 da evidencia de la utilidad del mecanismo de coordenadas paralelas para encontrar tendencias o patrones en datos multidimensionales, es este caso, en curvas de cinéticas de adsorción generadas en condiciones diferentes. Las curvas de cinéticas de adsorción de los experimentos 1, 2, 13, 14, 15, 16, es decir, del primer grupo, son mostradas en la Fig. 2a, dónde se observa que tienen un comportamiento muy semejante, en otras palabras, el agrupamiento realizado por el algoritmo k-means da evidencia de la fuerte relación o semejanza en la estructura que existe en estos experimentos. La Tabla III, grupo 1, exhibe que se puede llegar al punto de equilibrio en la adsorción de Ama6, si el tamaño de la perla es mediana y que la variación de temperatura es un factor que interviene en menor medida. En la Fig. 2b se puede ver que las

TABLA III
CARACTERÍSTICAS DE LOS EXPERIMENTOS AGRUPADOS

Grupo	Exp.	Colorante	Esfera	Temp. (ºC)	$q(mg/g)$
	1	Ama5	Chica	10	953
	2	Ama5	Chica	30	935
	13	Ama6	Mediana	10	914
1	14	Ama6	Mediana	30	954
	15	Ama6	Mediana	50	966
	16	Ama6	Grande	10	1017
	10	Ama6	Chica	10	1208
	11	Ama6	Chica	30	1222
2	12	Ama6	Chica	50	1204
	17	Ama6	Grande	30	1060
	18	Ama6	Grande	50	1047
	3	Ama5	Chica	50	877
	4	Ama5	Mediana	10	774
	5	Ama5	Mediana	30	825
3	6	Ama5	Mediana	50	763
	7	Ama5	Grande	10	814
	8	Ama5	Grande	30	817
	9	Ama5	Grande	50	784

curvas de cinéticas de adsorción de los experimentos 10, 11, 12, 17, 18 (grupo 2), los cuales presentan un comportamiento similar al del grupo 1, es decir, son muy semejantes entre ellas (las curvas del mismo grupo). La Tabla III, grupo 2, muestra que el colorante es igual para todos los experimentos y el tamaño de la esfera o perla es chica o grande. No obstante, los valores de remoción mayores ($q(mg/g)$), están relacionados con un tamaño chico de la perla y no hay evidencia (en la Tabla III) de que la temperatura sea determinante en ello. La Fig. 2c visualiza las curvas de cinética de adsorción del grupo 3, cuyos experimentos son: 3, 4, 5, 6, 7, 8, 9, y en la Tabla III sobresale, que en este grupo, solo se remueve colorante Ama5; y que el tamaño de la perla puede ser grande, mediano o chico a diferentes temperaturas. Sin embargo, se aprecia una

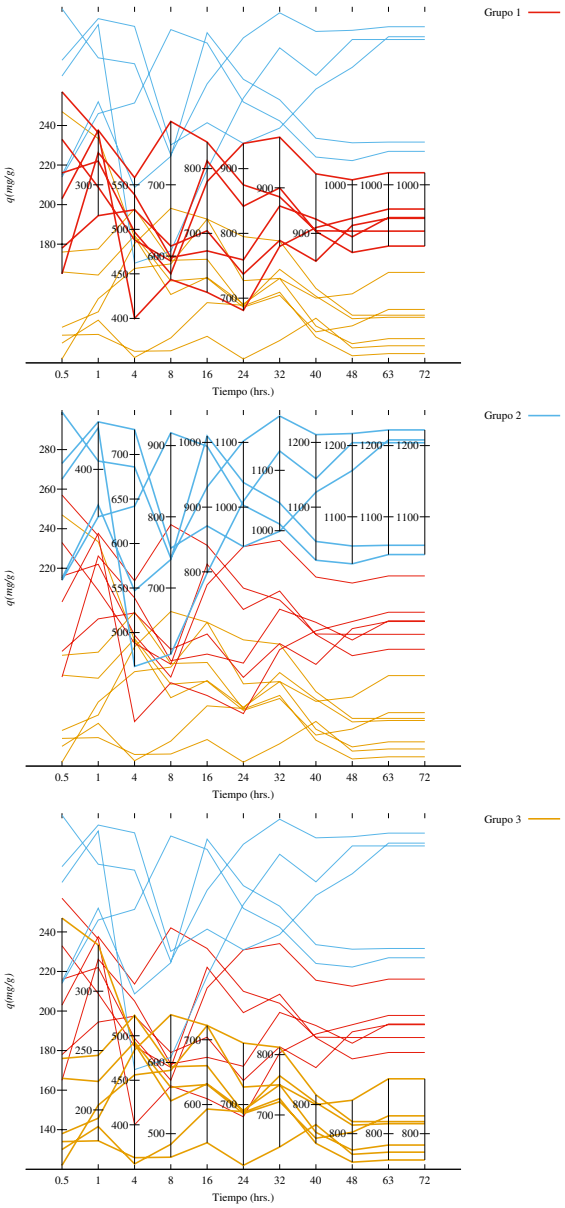

Fig. 2. Curvas de cinética de adsorción representadas en coordenadas paralelas.

tendencia en la remoción de colorante dependiendo del tamaño de la perla, en otras palabras, el promedio de remociones son aproximadamente 787 y de 805 $q(mg/g)$, con perlas medianas y grandes, respectivamente; y al parecer la temperatura no define claramente estos niveles de remoción. Resumiendo los resultados de la Fig. 2 y la Tabla III, se puede destacar que las curvas que forman el grupo 1, alcanzan su punto de equilibrio entre 914 y 1017 $q(mg/g)$, cuyas curvas representan los experimentos con tamaños de esferas chica, mediana o grande a temperaturas de 10, 30 o 50 ºC, y colorante Ama5 o Ama6. En el grupo 2, sus curvas alcanzaron su punto de equilibrio entre 1047 y 1222 $q(mg/g)$, y describen experimentos con tamaño de esfera chica y grande, a temperaturas de 10, 30 o 50 C, y colorante Ama6. Se visualiza que el grupo 3 alcanza el punto de equilibrio entre 774 y 877 $q(mg/g)$ con tamaños

de las esferas chica, mediana y grande, con el colorante Ama5.

Finalmente, se observa en los resultados presentados en esta sección, que existe un patrón de comportamiento en estos experimentos, el cual destaca al tamaño de la esfera como característica crítica en la remoción de colorante, ya sea Ama5 o Ama6. Asimismo, es notable que los grupos se definen desde las primeras horas o minutos y convergen en mayor medida en los tiempos finales.

B. Con Réplica

Para confirmar los hallazgos presentados en la sección anterior, se procedió a realizar una experimentación más exhaustiva, para ello se amplió la etapa de pruebas, utilizando experimentos replicados, es decir, se usaron tres réplicas de cada experimento inicial, para tener un total de 54 experimentos, para mayor detalle véase la Tabla I. La Tabla IV, muestra el agrupamiento de los experimentos después de la aplicación del algoritmo k-means. Los valores corresponden a los números de experimento (curvas de cinética de adsorción) asignados a cada grupo. Asimismo, al igual que la sección V-A, se probaron diferentes valores de k, y de manera semejante el valor más óptimo de k fue igual a 3, por tal motivo solo se muestran tres grupos. La Fig. 3 corresponde a las curvas de

TABLA IV
GRUPOS OBTENIDOS (CON RÉPLICA)

No. de grupo	No. de experimento en cada grupo
1	1, 2, 3, 4, 5, 6,7, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48
2	8, 9, 10, 11, 12, 13, 14, 15,16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27
3	28, 29, 30, 31, 32, 33, 34, 35, 36, 49, 50, 51, 52, 53, 54

cinética de adsorción para cada uno de los grupos obtenidos y presentados en la Tabla IV (grupo 1, 2 y 3, respectivamente). Las características de los experimentos contenidos en esos grupos se resume en la Tabla V. La Fig. 3a muestra las curvas de la cinética de adsorción del grupo 1, cuyos valores de sus características son: colorante (Ama5 y Ama6), tamaño de la perla (chica, mediana y grande), temperatura (10, 30 y 50 C) y alcanzan su punto de equilibrio entre 911 y 1040 $q(mg/g)$. La Fig. 3b presenta las curvas de la cinética de adsorción del grupo 2, en la cual los valores de sus características son: colorante (Ama5), tamaño de la perla (chica, mediana y grande), temperatura (10, 30 y 50 C) y alcanzan su punto de equilibrio entre 753 y 874 $q(mg/g)$. Por su parte, la Fig. 3c exhibe las curvas de la cinética de adsorción del grupo 3, con valores de sus características: colorante (Ama6), tamaño de la perla (chica y grande), temperatura (10, 30 y 50 C) y alcanzaron su punto de equilibrio entre 1036 y 1248 $q(mg/g)$. En la Fig. 3, se observa que el algoritmo de agrupamiento k-means permite la organización de experimentos ejecutados con diferentes condiciones, en grupos con características o estructuras similares. Esto se evidencia al analizar las curvas de cinética de adsorción de cada grupo, donde es notable la similitud entre ellas. Lo que es benéfico a la hora de llevar a la práctica experimentos de laboratorio a mayor escala, y de esta

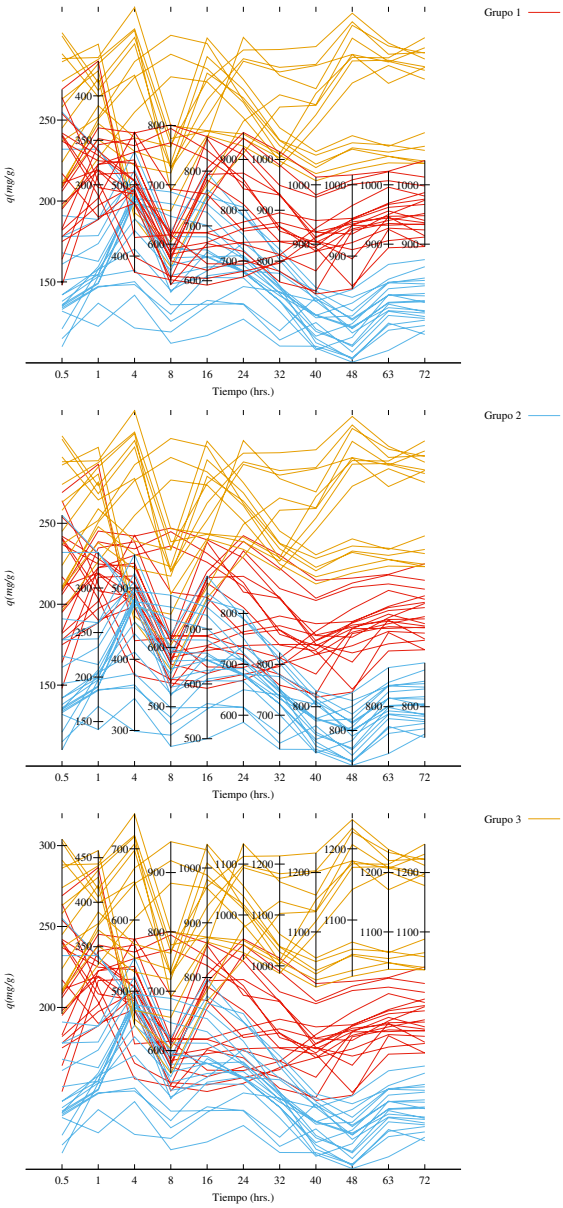

Fig. 3. Curvas de cinética de adsorción para los experimentos con réplica.

forma reducir costos de experimentación. En particular, en este trabajo, el agrupamiento favorece el conocimiento acerca de qué características y condiciones de experimentación son más favorables para el diseño y configuración del experimento de cinética de adsorción. La Tabla V contiene el detalle de la configuración de los experimentos correspondientes a cada grupo. En ella se muestra que, al igual que en la experimentación anterior, la característica que determinan en mayor medida alcanzar el punto de equilibrio en la cinética de adsorción es el tamaño de la esfera o perla, esto es más evidente en los grupos 2 y 3, en los cuales hay una tendencia en el tamaño de la perla. Para el Ama5, los tamaños ideales son mediana y grande, y para el Ama6, la esfera chica, pero sobre todo grande es la más apropiada. En el caso de la temperatura esta tendencia no es tan clara, lo que se traduce en que alcanzar el punto de equilibrio esta más estrechamente relacionado al tamaño

de la perla y no a la temperatura. Esta sección muestra cómo

TABLA V
CARACTERÍSTICAS DE LOS EXPERIMENTOS DE CADA GRUPO

Grupo	Exp.	Colorante	Esfera	Temp. (ºC)	$q(mg/g)$
	1	Ama5	Chica	10	986
	2	Ama5	Chica	10	939
	3	Ama5	Chica	10	933
	4	Ama5	Chica	30	947
	5	Ama5	Chica	30	934
	6	Ama5	Chica	30	924
	7	Ama5	Chica	50	895
	37	Ama6	Mediana	10	935
	38	Ama6	Mediana	10	911
1	39	Ama6	Mediana	10	895
	40	Ama6	Mediana	30	975
	41	Ama6	Mediana	30	973
	42	Ama6	Mediana	30	914
	43	Ama6	Mediana	50	980
	44	Ama6	Mediana	50	966
	45	Ama6	Mediana	50	951
	46	Ama6	Grande	10	1040
	47	Ama6	Grande	10	1013
	48	Ama6	Grande	10	998
	8	Ama5	Chica	50	874
	9	Ama5	Chica	50	861
	10	Ama5	Mediana	10	784
	11	Ama5	Mediana	10	774
	12	Ama5	Mediana	10	763
	13	Ama5	Mediana	30	843
	14	Ama5	Mediana	30	829
	15	Ama5	Mediana	30	801
	16	Ama5	Mediana	50	785
	17	Ama5	Mediana	50	748
2	18	Ama5	Mediana	50	753
	19	Ama5	Grande	10	838
	20	Ama5	Grande	10	816
	21	Ama5	Grande	10	787
	22	Ama5	Grande	30	832
	23	Ama5	Grande	30	813
	24	Ama5	Grande	30	804
	25	Ama5	Grande	50	802
	26	Ama5	Grande	50	778
	27	Ama5	Grande	50	771
	28	Ama6	Chica	10	1248
	29	Ama6	Chica	10	1198
	30	Ama6	Chica	10	1178
	31	Ama6	Chica	30	1211
	32	Ama6	Chica	30	1223
	33	Ama6	Chica	30	1230
	34	Ama6	Grande	50	1221
3	35	Ama6	Grande	50	1192
	36	Ama6	Grande	50	1198
	49	Ama6	Grande	30	1087
	50	Ama6	Grande	30	1057
	51	Ama6	Grande	30	1035
	52	Ama6	Grande	50	1064
	53	Ama6	Grande	50	1039
	54	Ama6	Grande	50	1035

el algoritmo k-means permite agrupar experimentos distintos de acuerdo con los resultados experimentales de cinética de adsorción (en once intervalos fijos de tiempo), y de esta forma analizar las características individuales de cada experimento que conforman los grupos obtenidos, y así ayudar a determinar las condiciones de operación óptimas para el futuro diseño de experimentos de cinética de adsorción. Para confrontar estos resultados, se realizó un diseño de experimento jerárquico con factores cruzados y un diseño jerárquico anidado en tres etapas. Para ambos experimentos se trabajó con un nivel de confianza del 95% y 97.5% ($\alpha = 0.05$ y $\alpha = 0.025$). Con respecto al análisis de varianza para el diseño de experimentos jerárquico con factores cruzados, la hipótesis nula de igualdad de la temperatura dentro del colorante no se rechaza, lo que significa que no existe diferencia significativa en la temperatura. También la hipótesis nula de la interacción tamaño de esfera-temperatura dentro del colorante, no se rechaza a una significancia de $\alpha = 0.05$ y $\alpha = 0.025$. Mas explícitamente, no hay evidencia de una diferencia significativa entre la interacción del tamaño de la esfera y la temperatura. Además la hipótesis nula de la igualdad del tipo de colorante se rechaza, lo que significa que sí existe diferencia significativa en la adsorción colorante Ama5 y Ama6.

Por otro lado, también se realizó un diseño de experimentos jerárquico anidado con tres etapas, concluyendo que existe diferencia significativa en el tamaño de la esfera de quitosano a un nivel de significancia $\alpha = 0.05$. Además en la interacción temperatura-tamaño de esfera, no existe diferencia significativa a un valor $\alpha = 0.025$. La hipótesis nula de tipo de colorante se rechaza, lo que significa que existe diferencia significativa en el tipo de colorante en la cantidad absorbida por las perlas de quitosano. La hipótesis nula (igualdad del tipo de colorante) es rechazada en concordancia a los diseños de experimentos realizados, lo cual puede ser justificado con los resultados del algoritmo k-means, es decir, el algoritmo obtuvo tres grupos de los cuales en dos de ellos hace perfectamente la separación con respecto al tipo de colorante (grupos 2 y 3). Finalmente, se observa en los resultados, lo cual coincide con este diseño de experimentos, las esferas o perlas de hidrogel (chica y grande) adsorben más el colorante Ama6 que el Ama5, además la temperatura tiene poca influencia.

VI. CONCLUSIÓN

En este trabajo se presenta una aplicación de los algoritmos de agrupamiento (k-means), utilizando curvas de cinética de adsorción base quitosano, para lo cual se utilizó el concepto de coordenadas paralelas para representar las curvas de cinética de adsorción. Para validar los resultados obtenidos del agrupamiento, se utilizaron dos diseños de experimentos: jerárquico con factores cruzados y jerárquico anidado en tres etapas. Los resultados experimentales presentados en este trabajo dan evidencia de que el tamaño de la perla influye en mayor medida en la remoción de colorante, y que la temperatura presenta una menor incidencia en los niveles de eliminación del colorante. Asimismo, este trabajo refleja la utilidad de las técnicas de aprendizaje automático para reducir costos de experimentación, en áreas como la química. Actualmente estamos profundizando en el estudio de la cinética de adsorción a través de redes neuronales artificiales y algoritmos genéticos.

AGRADECIEMIENTOS

Este trabajo ha sido parcialmente financiado por el proyecto 5046/2020CIC de la UAEMex y el 8239.20-P del TecNM.

REFERENCIAS

[1] P. M. de Sá, *Pattern Recognition: Concepts, Methods and Applications*. Springer-Verlag Berlin Heidelberg, 2001.

[2] W. D. Callister, Jr. and D. Rethwish, *Materials Science and Engineering An introduction*. John Wiley & Sons, 2014.
[3] K. Rajan, *Data-driven Discovery for Accelerated Experimentation and Application*. Oxford: Butterworth-Heinemann, Elsevier, 2013.
[4] L. Kaufman and P. Rousseeuw, *Finding groups in data: an introduction to cluster analysis.* Wiley Inter-Science, 1990.
[5] A. Agrawal and A. Chouhary, "Perspective: Materials informatics and big data: Realization of "fourth paradigm" of science in materials science," *APL materials*, vol. 4, no. 5, 2016.
[6] A. Jain, G. Hautier, S. P. Ong, and K. Persson, "New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships," *Journal of Materials Research*, vol. 31, no. 8, pp. 977–994, 2016.
[7] A. S. Varde, "Graphical data mining for computational estimation in materials science applications," Ph.D. dissertation, Worcestur Polytechnic Institute, 2006.
[8] K. Rajan, "Materials informatics," *Materials Today*, vol. 8, no. 10, pp. 38 – 45, 2005.
[9] R. Ramprasad, G. R. Batra, A. Pilana, Mannodi-Kanakkithodi, and C. Kim, "Machine learning in materials informatics: recent applications and prospects," *Natural partner journal "Computational materials"*, vol. 3, no. 54, 2017.
[10] R. Hrubiak, L. George, S. K. Saxena, K. Rajan *et al.*, "A materials database for exploring material properties," *Journal of Materials (JOM)*, vol. (61), pp. 59–62, 2009.
[11] A. Agrawal, P. D. Deshpande, A. Cecen, G. P. Basavarsu, A. N. Choudhary, and S. R. Kalidindi, "Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters," *Integrating Materials and Manufacturing Innovation*, vol. 3, no. 1, 2014.
[12] K. Rajan, *Data Mining and Inorganic Crystallography*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 59–87.
[13] J. Pacheco-Sanchez, R. Alejo, H. Cruz-Reyes, and F. Alvarez-Ramirez, "Neural networks to fit potential energy curves from asphaltene-asphaltene interaction data," *Fuel*, vol. 236, pp. 1117 – 1127, 2019.
[14] Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, "Data mining for materials: Computational experiments with *ab* compounds," *Phys. Rev. B*, vol. 85, p. 104104, Mar 2012.
[15] J. Tinoco, A. G. Correia, and P. Cortez, "Application of data mining techniques in the estimation of mechanical properties of jet grouting laboratory formulations over time," in *Soft Computing in Industrial Applications*, A. Gaspar-Cunha, R. Takahashi, G. Schaefer, and L. Costa, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 283–292.
[16] C. Suh, A. Rajagopalan, X. Li, and K. Rajan, "The application of principal component analysis to materials science data," *Data Science Journal*, vol. 1, no. 1, pp. 19–26, 2002.
[17] D. Morgan, G. Ceder, and S. Curtarolo, "Data mining approach to ab-initio prediction of crystal structure," *MRS Proceedings*, vol. 804, p. JJ9.25, 2003.
[18] C. Ortiz, O. Eriksson, and M. Klintenberg, "Data mining and accelerated electronic structure theory as a tool in the search for new functional materials," *Computational Materials Science*, vol. 44, no. 4, pp. 1042 – 1049, 2009.
[19] J. Apostolakis, *An introduction to data mining*. German: Springer Berlin Heidelberg, 2009, pp. 1–35.
[20] L. Cavas, Z. Karabay, H. Alyuruk, H. Dogan, and G. K. Demic, "Thomas and artificial neural network models for the fixed-bed adsorption of methylene blue by beach waste posidonia oceanica (l.) dead leaves," *Chemical Engineering Journal*, vol. 171, pp. 557–562, 2011.
[21] H. Thomas, "Heterogeneous ion exchange in flowing system," *J. Am. Chem. Soc*, vol. 66, no. 10, pp. 1664–1666, 1944.
[22] A. Meden, "Inorganic crystal structure prediction- a dream coming true?" *Acta Chim.Slov*, vol. 53, no. 2, pp. 148–152, 2006.
[23] G. Crini, "Application of chitosan, a natural aminopolysaccharide, for dye removal from aqueous solutions by adsorption proposes using batch studies: a review of recent literature," *Progress in Polymer Science*, vol. 33, no. 4, pp. 399–477, 2008.
[24] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. Cam and J. Neyman, Eds., vol. 1. University of California Press, 1967, pp. 281–297.
[25] A. Inselberg, "Visual analytics for high dimensional data," in *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, ser. AVI'18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3206505.3206608
[26] M. Rinaudo, "Chitin and chitosan: properties and applications," *Progress in Polymer Science*, vol. 31, no. 7, pp. 603–632, 2006.
[27] V. Parsegian, *Van der Waals Forces: A Handbook for Biologists, Chemists, Engineers, and Physicists*. Cambridge University Press, 2005. [Online]. Available: https://books.google.com.mx/books?id=K4F7uHJVEOoC
[28] J. Han, M. Kamber, and J. Pei, *Data mining concepts and techniques*, 2nd ed. Waltham, Mass.: Morgan Kaufmann Publishers, 2006.
[29] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.
[30] M. S. Sánchez, R. M. Valdovinos, A. Trueba, E. Rendón, R. Alejo, and E. López, "Applicability of cluster validation indexes for large data sets," in *2013 12th Mexican International Conference on Artificial Intelligence*, 2013, pp. 187–193.
[31] A. Ghosal, A. Nandy, A. K. Das, S. Goswami, and M. Panday, "A short review on different clustering techniques and their applications," in *Emerging Technology in Modelling and Graphics*, J. K. Mandal and D. Bhattacharya, Eds. Singapore: Springer Singapore, 2020, pp. 69–83.
[32] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651 – 666, 2010, award winning papers from the 19th International Conference on Pattern Recognition (ICPR). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865509002323
[33] A. Inselberg, *Parallel Coordinates: Visual Multidimensional Geometry and its Applications*, 1st ed. New York. U.S.A: Springer-Verlag New York, 2009.
[34] ——, *Parallel Coordinates*. Boston, MA: Springer US, 2009, pp. 2018–2024.
[35] A. Bryden, K. Rajan, and R. LeSar, "Chapter 7 - visualization in materials research: Rendering strategies of large data sets," in *Informatics for Materials Science and Engineering*, K. Rajan, Ed. Oxford: Butterworth-Heinemann, 2013, pp. 121 – 146.
[36] D. Mongomery, *Diseño y Análisis de Experimentos*. Limusa Wiley, México, 2004.

Eréndira Rendón Doctora en Ciencias Computacionales por el Instituto Tecnológico de Toluca. Se desempeñan como profesora-Investigadora en la División de Estudios de Posgrado e Investigación del Tecnológico Nacional de México, campus Toluca. Sus principales intereses en académicos se centran en la Minería de datos y recientemente en "Materials Informatics".

Roberto Alejo Doctor en Sistemas Informáticos Avanzados por la Universitat Jaume I, España (2011), adscrito a la División de Estudios de Posgrado e Investigación del Tecnológico Nacional de México, campus Toluca, con un profundo interés científico en la aplicación de la inteligencia artificial a la solución de problemas reales. Asimismo es especialista en redes neuronales artificiales, aprendizaje automático y minería de datos.

José L. García-Rivas Dr. en Ciencias en Ingeniería Química. Instituto Tecnológico de Ciudad Madero (2012). Actualmente es profesor investigador en la División de Estudios de Posgrado e Investigación del Tecnológico Nacional de México, campus Toluca. Su áreas de estudio se centran en la remoción de colorante amarillo 5 y 6 en flujo continuo con perlas de quitosano para la solución de problemas ambientales.