Capítulo 3

Análisis clúster o conglomerados: identificación de dos temporalidades ambientales en la zona marino-costera de Guasave

Graciano-Obeso Adalid
Alzate-Espinoza Juan Héctor
Bojórquez-Delgado Gilberto

[4] https://doi.org/10.61728/AE24250047

Introducción

Los métodos multivariados se aplican en procesos medioambientales desde principios del siglo XX, pero han tenido una enorme difusión en los últimos años, debido a la gran cantidad de información acumulada en las bases de datos y al enorme progreso de la tecnología computacional que comenzó en la década de 1960 (Palacio et al., 2020). Específicamente para la conformación de distintos conglomerados o clústeres, se aplica la técnica estadística dentro del análisis multivariado llamada "análisis de clúster o conglomerados", este método permite la conformación de grupos homogéneos hacia el interior y lo más heterogéneos entre los distintos grupos (Gutiérrez y Ciancio, 2023).

Cuando los investigadores están interesados en identificar grupos de individuos, la técnica más adecuada del análisis multivariado es el análisis de conglomerados o análisis de clúster, esta técnica se basa en el análisis y la interpretación de la asociación observada entre los individuos, de modo que el cálculo de su distancia o proximidad sirve para conformar grupos homogéneos en relación con las características seleccionadas que, a la vez, sean tan heterogéneos entre ellos como sea posible (Meneses, 2019; Rodríguez et al., 2019). En esta técnica se considera a cada unidad de análisis como un conglomerado, y posteriormente va uniendo los conglomerados de acuerdo con su homogeneidad hasta que queda un conglomerado, entendiendo por grupo homogéneo aquel cuyos miembros difieren significativamente de los de cualquier otro (Starstedt y Mooi, 2014; Tussel, 2023).

Por otro lado, esta técnica ha sido utilizada en diversas investigaciones en el mundo para agrupar elementos con características similares o dividir por zonas (Burbano et al., 2018; Price et al., 2006; Pérez et al., 2004). Además, se ha demostrado que el análisis jerárquico, facilita determinar el nivel de correlacionamiento entre las diferentes variables, permitiendo generar grupos de asociación, lo cual redunda en una mejor segmentación de los elementos que influencian el aprovechamiento de los activos medioambientales (Blanco et al., 2017).

El Análisis Clúster (AC) es un método estándar del análisis multivariado que puede reducir una compleja cantidad de información en pequeños grupos o clústers, donde los miembros de cada uno de ellos comparten características similares (Lin y Chen, 2006). El AC se considera una técnica eminentemente exploratoria que no utiliza ningún tipo de modelo estadístico para llevar a cabo el proceso de clasificación (Hair et al., 1999) y, por ello, se le podría calificar como una técnica de aprendizaje no supervisado, es decir, una técnica muy

adecuada para extraer información de un conjunto de datos sin imponer restricciones previas en forma de modelos estadísticos (Peterson, 2002).

El AC tiene por objeto formar grupos o clústers homogéneos en función de las similitudes o similaridades entre ellos (Peña, 2002). Los grupos se forman de tal manera que cada objeto es parecido a los que hay dentro del clúster con respecto a algún criterio de selección predeterminado (Rao y Srinivas, 2006; Hair et al.). Las técnicas de agrupamiento en el AC se pueden clasificar en dos categorías: el clúster jerárquico y el no jerárquico. Los procedimientos jerárquicos consisten en la construcción de una estructura en forma de árbol.

Existen dos tipos de procedimientos de obtención de clústers jerárquicos: los de aglomeración y los divisivos. Dentro de los métodos jerárquicos aglomerativos se tienen: (i) método de encadenamiento simple, (ii) métodos de encadenamiento completo, (iii) método de encadenamiento medio, (iv) método de Ward, y (v) método del centroide (Hair et al.). Estos procedimientos difieren en la forma como se calcula la distancia entre los conglomerados, entre los que se encuentran la DEC, Manhattan, coeficiente de correlación de Pearson, Chevichev y Cosine.

El clúster por medio de técnicas no jerárquicas no requiere de procesos de construcción de árboles; en su lugar, asignan los objetos a clústers una vez que el número de grupos a formar esté especificado. Los procedimientos de aglomeración no jerárquicos se denominan frecuentemente agrupaciones de k–medias, k–medianas y k–modas. Una desventaja con respecto a la técnica jerárquica consiste en que debe conocerse apriori el número de clústers a obtener, lo que implica un grado de subjetividad en el proceso (Peterson, 2002).

A pesar de lo anterior, se considera un método dinámico en el sentido en que los objetos dentro de los clústers se pueden mover de un clúster a otro, minimizando la distancia entre objetos dentro de un mismo clúster (Rao y Srinivas, 2006). Pese a las ventajas del método de aglomeración no jerárquico, en este artículo se presenta la aplicación del método jerárquico dado el interés de no querer asignar a priori el número de grupos a formar. A continuación, se describen las técnicas empleadas en el análisis clúster y el método de similitud utilizado.

- Encadenamiento medio entre grupos. Mide la proximidad entre dos grupos calculando la media de las distancias entre objetos de ambos grupos o las medias de las similitudes entre objetos de ambos grupos. Algunos autores, como Hair et al., afirman que el método está sesgado a formar conglomerados con aproximadamente la misma varianza.

• Método de Ward. Este proceso de aglomeración tiene como objetivo establecer grupos de tal forma que la suma de los cuadrados de las desviaciones con respecto a la media de cada variable es mínima para todas las estaciones al mismo tiempo. Este procedimiento tiende a combinar los conglomerados con un número reducido de observaciones y a formar grupos con aproximadamente el mismo número de grupos (Rao y Srinivas, 2006).
• Método del centroide. En este método la distancia entre los grupos se define como la distancia entre sus centroides. El centroide de cada grupo o clúster es a su vez el promedio de las posiciones de todos los puntos dentro del clúster. En este método, cada vez que se agrupa a los individuos se calcula nuevamente el centroide; así el centroide cambia a medida que se fusionan los grupos (Hair et al.).
• La distancia euclídea al cuadrado (DEC). Ese cuadrado de la suma de las diferencias al cuadrado de dos elementos en la variable o variables consideradas (Castellarin et al., 2001).

Con el objetivo analizar con técnicas estadísticas multivariadas por conglomerados o cluster las temporalidades de la zona marino-costera adyacente al litoral del municipio de Guasave, y semejanzas de la variable de temperatura superficial del mar (tsm), así como identificar la influencia de las surgencias de la zona marino-costera de Guasave.

Metodología

Área de estudio

El golfo de California está ubicado en el noroeste de México y está delimitado por la península de Baja California y la costa continental de los estados de Sonora, Sinaloa y Nayarit. Se localiza entre los 20° y 32° de latitud norte y los 105.5° y 114.5° de longitud oeste en el Pacifico Oriental, y se encuentra orientado en una dirección Noroeste (NO) – Sureste (SE) (Fig. 1).

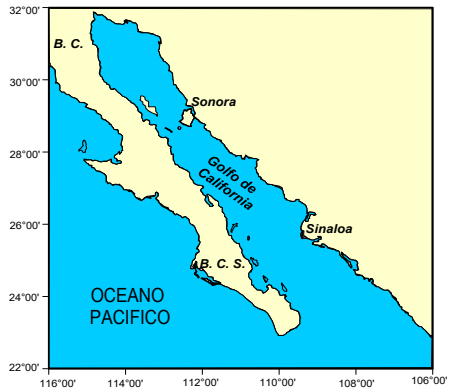


Figura 1. Ubicación geográfica del Golfo de California, México.

La presente investigación se realizó en la zona marino-costera adyacente al litoral del municipio de Guasave, Sinaloa, donde se realizaron ocho recorridos en una red de 15 estaciones ubicadas 50 km mar adentro (Fig. 2) durante febrero, abril, julio, octubre y diciembre de 2012 y febrero, marzo y abril de 2013. Para realizar los recorridos se consultaron las condiciones ambientales, específicamente del viento, el cual es un factor determinante para poder realizar los recorridos.

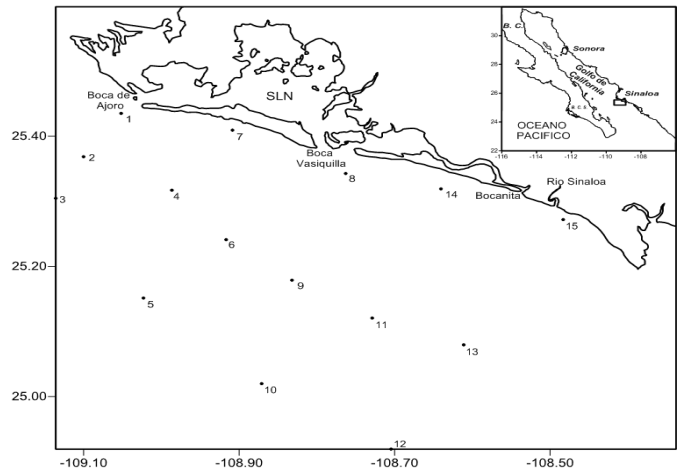
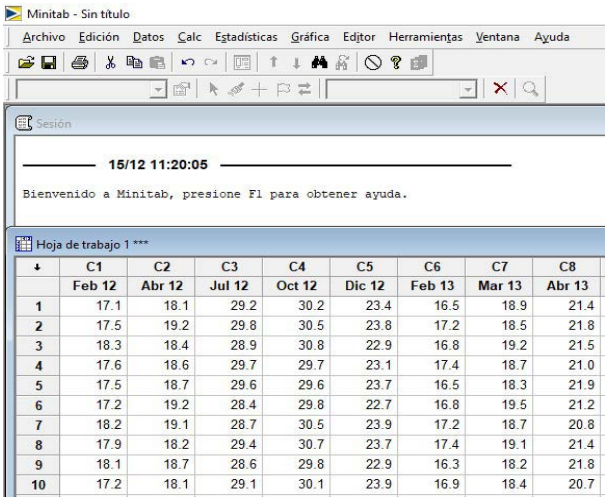


Figura 2. Delimitación del área de estudio y ubicación de las estaciones de muestreo en la zona marino-costera del municipio de Guasave, Sinaloa, México.

En cada estación se registró in situ la temperatura superficial del mar (tsm), donde se realizaron 10 mediciones por punto de muestreo.

Análisis clúster o de conglomerados con el Software estadístico Minitab versión 15

Ingresar los datos



	C1	C2	C3	C4	C5	C6	C7	C8
	Feb 12	Abr 12	Jul 12	Oct 12	Dic 12	Feb 13	Mar 13	Abr 13
1	17.1	18.1	29.2	30.2	23.4	16.5	18.9	21.4
2	17.5	19.2	29.8	30.5	23.8	17.2	18.5	21.8
3	18.3	18.4	28.9	30.8	22.9	16.8	19.2	21.5
4	17.6	18.6	29.7	29.7	23.1	17.4	18.7	21.0
5	17.5	18.7	29.6	29.6	23.7	16.5	18.3	21.9
6	17.2	19.2	28.4	29.8	22.7	16.8	19.5	21.2
7	18.2	19.1	28.7	30.5	23.9	17.2	18.7	20.8
8	17.9	18.2	29.4	30.7	23.7	17.4	19.1	21.4
9	18.1	18.7	28.6	29.8	22.9	16.3	18.2	21.8
10	17.2	18.1	29.1	30.1	23.9	16.9	18.4	20.7

Figura 3. Captura de datos de las mediciones en software minitab.

Se ingresaron los datos de las mediciones en la hoja de trabajo de la pantalla principal del software, para este caso práctico fueron las temperaturas que se tomaron durante los meses de febrero, abril, julio, octubre y diciembre de 2012, así como los meses de febrero, marzo y abril del año 2013 (Fig. 3).

Estandarizar los datos

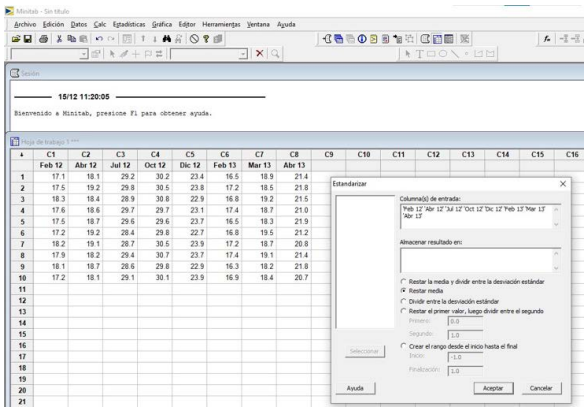


Figura 4. Estandarizar los datos de los resultados obtenidos en los puntos de muestreo.

Una vez que tenemos registrados los datos, se estandarizan los datos con el objetivo de que el software Minitab pondere todas las variables de igual manera, en la barra de herramientas seleccionamos Calc y posteriormente estandarizar (Fig. 4).

Especificar el método de enlace

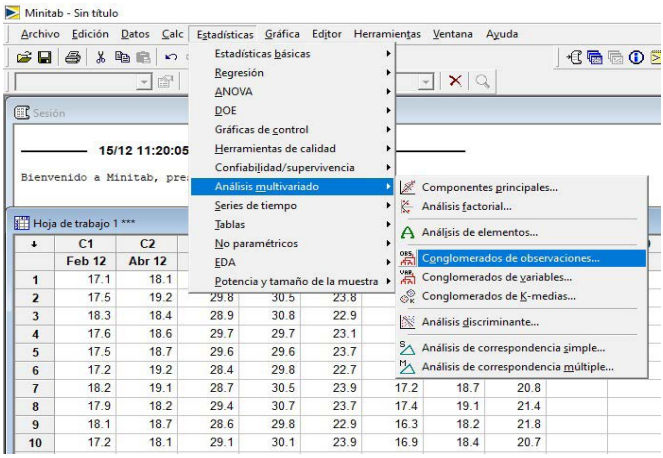


Figura 5. Elegir el método de enlace entre los conglomerados.

Una vez que se tienen los datos están estandarizados por el software, seleccionamos el método de enlace para especificar cómo se define la distancia entre los conglomerados, para este caso práctico elegimos el método completo, también conocido como el método del vecino más lejano (Fig. 5).

Especificar el método de vinculación, medición de la distancia y mostrar dendograma

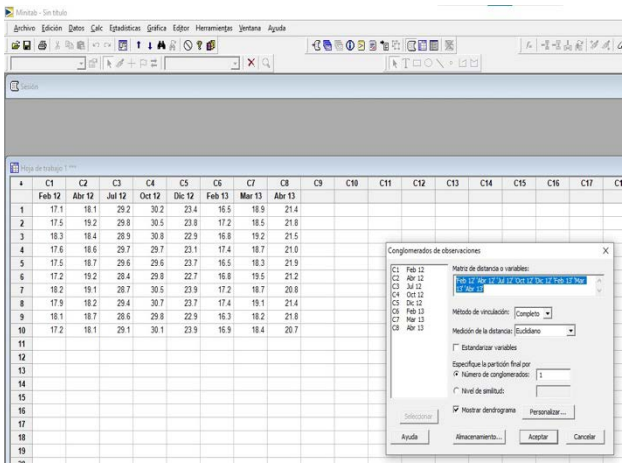


Figura 6. Seleccionar el método de medición de distancia entre los conglomerados.

Por último, ya que se tiene definido el método de enlace, es necesario seleccionar el método de la distancia, para el caso práctico de la presente investigación se eligió el método euclidiano, el cual es la medida de distancia más común utilizado para el análisis clúster o de conglomerados, mismo que calcula la raíz cuadrada de la suma de las diferencias al cuadrado. Una vez definido el método de enlace y la medición de la distancia, seleccionamos mostrar dendograma (Fig. 6).

Análisis de imágenes satelitales

Con la finalidad de identificar las temporalidades con una mayor resolución, así como fenómenos oceanográficos que debido al diseño de los muestreos podrían no identificarse, como es el caso de los eventos de surgencias, se obtuvieron composiciones mensuales globales de la temperatura superficial del mar (tsm) del sensor ModisAqua disponibles en el sitio web http://oceandata.sci.gsfc.nasa.gov/MODISA/Mapped/8Day/4km/SST/, con una resolución espacial de 4x4 km por pixel (utilizando los compuestos de noche para evitar la contaminación por reflectancia). La serie de los compuestos obtenidos de tsm comprende el periodo de octubre de 2011 a enero de 2013. Para recortar el área de estudio se utilizó el paquete computacional MatLab (Matrix Laboratory).

Análisis estadístico

En el presente trabajo los análisis estadísticos fueron realizados mediante el manejo del paquete Statistica v8.0. Se aplicó una prueba de normalidad y debido a que no mostraron una distribución normal, se realizó la transformación logarítmica de la base de datos para normalizar los datos y aplicarles el estadístico correspondiente. Con la finalidad de identificar las temporalidades del área de estudio, se aplicó un análisis de grupos (Clúster) utilizando el método de unión completa y distancias euclidianas a la tsm.

Con los datos obtenidos *in situ* de TSM, se realizaron mapas de distribución espacial por fecha de muestreo, así como por temporada fría y cálida en la zona marino-costera perteneciente al área de estudio con ayuda del programa de interpolación bidimensional Surfer® Ver. 9.0, por último, la representación temporal de variables en estudio se graficó con el programa SigmaPlot 10.0.

Resultados y discusión

Identificación de temporalidades respecto a la temperatura superficial del mar (TSM)

Durante el período de estudio, los valores de tsm fluctuaron entre 15.8 a 30.9 °C, el valor promedio mensual mínimo en febrero de 2013 (17.21°C) y el máximo en octubre de 2012 (29.76°C) (Tabla 1; Fig. 7).

		Temporada fría	Temporada cálida
Temperatura ºC	Máxima	23.20	30.90
	Mínima	15.80	27.60
	Promedio	19.20	29.40

Tabla 1. Valores promedio, máximos y mínimos de temperatura superficial del mar.

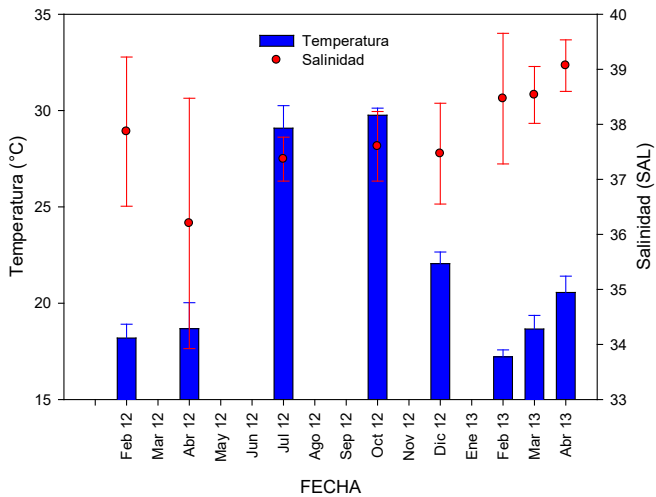


Figura 7, *Distribución temporal de tsm (ºC) y Salinidad (Sal), en la zona marino-costera del municipio de Guasave, Sinaloa, México, durante un periodo de febrero de 2012 a abril de 2013.*

Del análisis de grupos (Clúster) aplicado a la tsm se identificaron 2 grupos a un nivel de corte de 10 unidades de distancia aplicado a la temperatura (Fig. 8), y se eligió el criterio de la temperatura para realizar el análisis de los resultados de las variables analizadas, debido a que separa claramente los resultados en dos temporadas climáticas: fría y cálida, correspondiendo a la primera el periodo de diciembre a abril con temperaturas de 19.20° C en promedio, y de junio a octubre a la cálida con un promedio de 29.40° C.

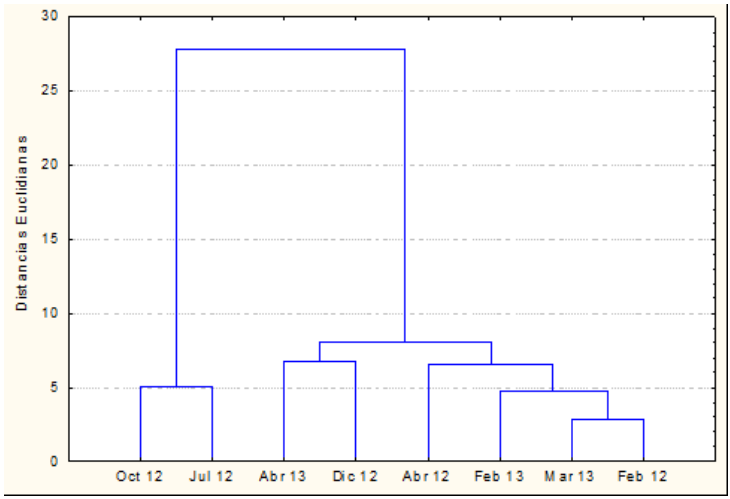


Figura. 8. Análisis de grupos de tsm en la zona marino-costera del municipio de Guasave, Sinaloa, México.

Respecto a la distribución espacial, los mayores valores de temperatura se asociaron a las estaciones más cercanas a la línea de costa, específicamente en las bocas del sistema lagunar y frente al Río Sinaloa (Fig. 9).

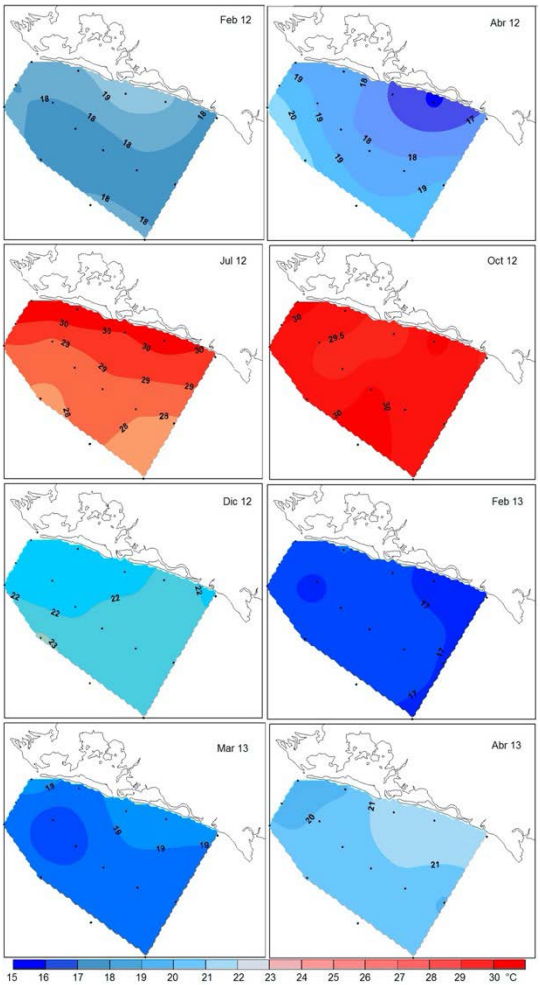


Figura 9. Distribución espacial de temperatura superficial del mar (°C) en la zona marino-costera del municipio de Guasave, Sinaloa, México, durante un periodo de febrero de 2012 a abril de 2013.

Al realizar el análisis por épocas climáticas, durante la temporada fría se obtuvieron temperaturas promedio de 15.80 a 23.20 °C. Espacialmente se observó una zona de menor temperatura hacia el norte del área de estudio en el transecto central respecto a la línea de costa (Fig. 10a). En la temporada cálida el rango promedio fue de 27.60 a 30.90 °C, con los valores menores en la zona oceánica y máximos en las estaciones más cercanas a la línea de costa (Fig. 10b).

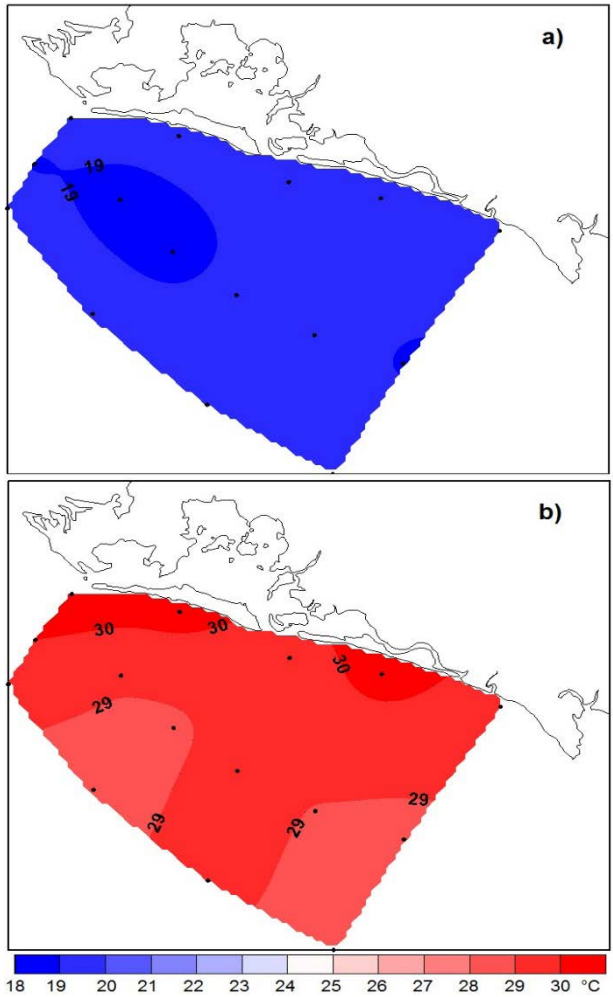


Figura 10. Distribución espacial promedio de tsm (°C) durante: a) temporada fría y b) temporada cálida en la zona marino-costera del municipio de Guasave, Sinaloa, México.

El patrón general de las temperaturas en el área de estudio presenta una clara influencia de la señal estacional en la que pueden agruparse en temporada fría y cálida, dicha agrupación fue posible con el análisis clúster o de conglomerados aplicado en el presente estudio, además, dicha estacionalidad se corrobora con las imágenes satelitales.

Durante la temporada fría, el promedio de temperaturas mostró un claro proceso de enfriamiento en la zona de estudio típico de los eventos de surgen-

cias costeras, con aguas más cálidas hacia la región sur del golfo de California (Fig. 11).

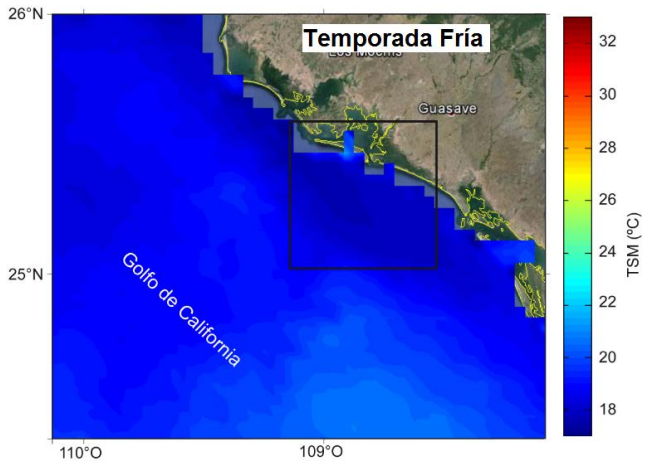


Figura 11. Distribución espacial de Temperatura Superficial del Mar (TSM) en temporada fría en la zona marino-costera del municipio de Guasave, Sinaloa, México.

Para la temporada cálida por otro lado, las mayores temperaturas se ubicaron en la zona costera, con un claro gradiente de mayores a menores temperaturas hacia la zona oceánica (Fig. 12).

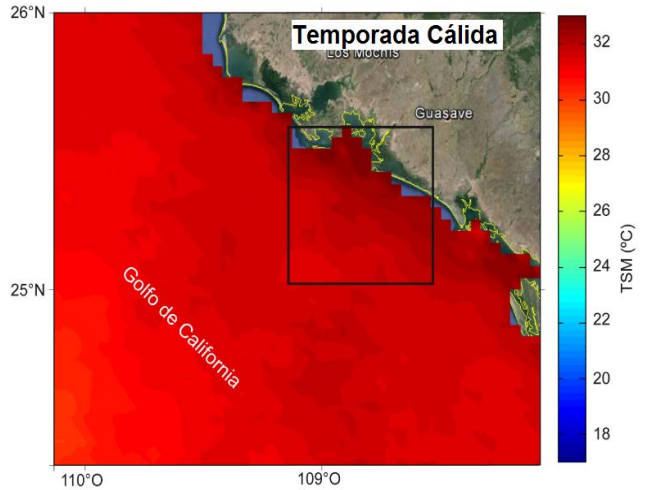


Figura 12. Distribución espacial de Temperatura Superficial del Mar (TSM) en temporada cálida en la zona marino-costera del municipio de Guasave, Sinaloa, México.

Durante la temporada fría, se registraron las menores temperaturas, con los valores mínimos ubicados en la zona norte del área de estudio, lo cual concuerda con lo reportado por Ulloa-Pérez (2005) y Valencia-Martínez (2012), quienes describen este mismo patrón de temperaturas. Esta distribución de temperaturas puede estar relacionado con la presencia de aguas oceánicas de menor temperatura que las costeras y que se internan en el Sistema Lagunar Navachiste durante el intercambio por mareas debido a fenómeno descrito para la zona por Escobedo-Urías, (2010).

Hay numerosos trabajos que analizan distintas variables del agua aplicando métodos de conglomerados, como es el caso de (Díaz y Mormeneos, 2002) quienes determinan que por la estructura de los datos en su investigación este método superó otros siete métodos jerárquicos o difusos por la capacidad de estudio aun con un grupo reducido de objetos. Robles (2018) aplico el método para el análisis de la Temperatura Superficial del Mar (TSM) en la zona costera de sonora para la caracterización de procesos físicos y biológicos, y puede ser afectada por procesos en escalas espacio temporales muy diversas.

Por otro lado, al contrastar los valores promedios de temperatura para el área con los trabajos anteriores, se observó que el promedio obtenido en este estudio (19.20 °C) es menor al reportado por Ulloa-Pérez (2005) quien obtuvo un promedio de 20°C, y que se debió a que en los trabajos mencionados tuvieron una cobertura espacial más reducida con estaciones costeras, debido a lo cual registraron aguas más cálidas que en este estudio. Sin embargo, al comparar este valor promedio con el obtenido por Valencia-Martínez (2012), (TSM: 18.45 °C), el promedio en este trabajo fue mayor, lo cual pudo deberse a que este monitoreo tuvo una cobertura temporal más amplia (febrero de 2012 a abril 2013) a la del trabajo citado (febrero-octubre 2012), además que el monitoreo de Valencia-Martínez (2012) se realizó en condiciones La Niña (Fig. 13), fenómeno oceanográfico caracterizado por presentar valores de temperatura más bajos que en condiciones normales (Hayward et al, 1999), por lo cual aunque también se obtuvieron datos en la misma época, al promediarlos con los valores medidos en 2013 cuando el efecto del evento había finalizado, se obtuvieron temperaturas mayores.

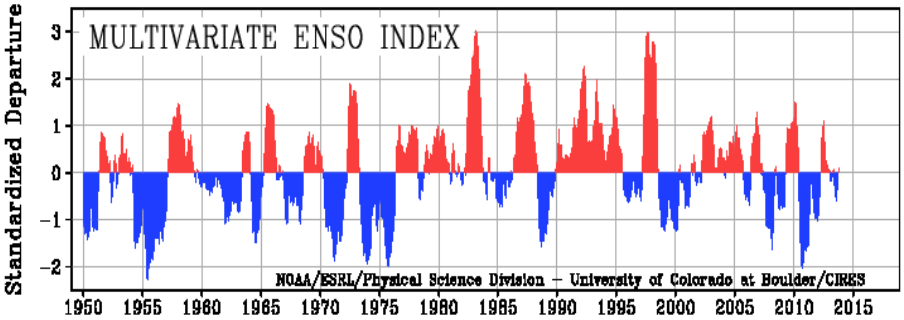


Figura 13. Índice multivariado del ENOS (El Niño Oscilación del Sur) para el periodo de 1950 a 2015.

Conclusiones

- El análisis clúster o análisis de conglomerados es una técnica estadística multivariante que agrupa individuos con el fin de lograr la máxima homogeneidad en cada grupo, en la presente investigación fue posible identificar dos temporadas en el área de estudio de acuerdo a los resultados que arrojó el análisis clúster.
- Los diferentes patrones espaciales y temporales de la temperatura superficial del mar en el área de estudio, permiten identificar claramente dos temporadas climáticas: una fría durante los meses de diciembre a abril y una cálida de julio a octubre.
- Las imágenes de satélite de temperatura superficial del mar comprobaron la temporalidad de los eventos de surgencias costeras durante la temporada fría con inicio en diciembre y terminó en abril, la mayor intensidad fue en enero, febrero y marzo de 2012.

Referencias

Blanco, A., Muñoz, I., E., M., de Miguel, C., & Casas, E. (2017). Aplicaciones de la segmentación jerárquica en medición y evaluación de programas educativos. Ejemplos con un programa de educación financiera., 235-257.

Burbano, V., Edy, L., & Moreno, E. (2018). Análisis de Conglomerados del Norte del valle del cauca. Caso estudio Cartago, Zarzal y la Unión. Ingenieria Industrial, 1, 78-91.

Castellarin, A., Burn, D., & Brath, A. (2001). Assessing the effectiveness of hydrological similarity measures for flood frequency analysis. Journal of Hydrology, 241, 270-285.

Escobedo_Urías, D. (2010). Diagnóstico y descripción del proceso de eutrofización en lagunas costeras del Norte de Sinaloa. Tesis de Doctorado. CICIMAR-IPN, 298. La Paz, Baja California Sur, México.

Gutiérrez, A., & Ciancio, A. (2023). Análisis del destino de la producción en la economía argentina: una aplicación de análisis de clúster. Revista de la Facultad de Ciencias Económicas, 20-40.

Hair, J., Anderson, R., Tatham, R., & Black, W. (1999). Análisis multivariante. Madrid: Prentice Hall.

Hayward, T., Baumgartne, T., C. D., R., D., G., G.-C., Hyrenbach, K., M., T. (1999). The State of the California Current, 1998-1999: transition to cool-water conditions. California Cooperative Oceanic Fisheries Investigations. 29-62: Report 40.

Lin, G., & Chen, L. (2006). Identification of homogeneous regions for regional frequency analysis using the self-organizing map. Journal of Hydrology, 324, 1-9.

Meneses, J. (2019). Introducción al análisis multivariante. Primera Edición. Universidad Oberta de Catalunya,. Barcelona, 39-43.

Palacio, F., Apodaca, M., & Crisci, J. (2020). Análisis multivariado para datos biológicos: Teoría y su aplicación utilizando el lenguaje R. Fundación Natural de Féliz de Azara, 268.

Peña, D. (2002). Análisis de Datos Multivariantes. España: McGraw-Hill.

Pérez, S., Montes, J., & Vázquez, J. (2004). Managing knowledge: the link between culture and organizational learning. Journal of knowledge management, 8, 93-104.

Peterson, L. (2002). CLUSFAVOR 5.0: hierarchical cluster and principal-component analysis of microarray based transcriptional profiles. (B. C. Medicine, Ed.) Texas, One Baylor Plaza, ST-924, USA. Obtenido de http://genomebiology.com/2002/3/7/software/0002.

Price, A., Patterson, N., Plenge, R., Meinblatt, M., Shadick, N., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, 38, 904.

Rao, A., & Srinivas, V. (2006). Regionalization of waters heds by hybrid-cluster analysis. Journal of Hydrology, 318, 37-56.

Robles, C. (2018). Variabilidad oceanográfica de la zona costera del estado de Sonora. (Tesis de Postgrado). Universidad de Sonora.

Rodriguez, M., Comin, C., Casanoca, D., Bruno, O., Amancio, D., Costa, L., & Rodriguez, F. (2019). Clustering algorithms: A comparative approach. PLoS ONE. e0210236, 14(1).

Starstedt, M., & Mooi, E. (2014). Cluster Analysis. In a Concise Guide to Market Research. Berlin, Heidelberg: Springer.

Tussel, C. (2023). Calidad de la democracia en América Latina, 2013-2018: una clasificación con observaciones de conglomerado y dendrograma. Estado & comunes, revista de politicas y problemas públicos., 2, 39-56.

Ulloa-Pérez, A. (2005). Influencia de la disponibilidad de nutrientes sobre los cambios espacio-temporales de la comunidad de fitoplancton en el litoral del municipio de Guasave, Sinaloa. Tesis de Maestría, CIIDIR-IPN, 100. Unidad Sinaloa.

Valencia-Martínez, S. (2012). Caracterización del área de alimentación de tortugas marinas en la zona marino-costera del complejo insular San Ignacio-Navachiste-Macapule, Sinaloa, Golfo de California. Tesis de Mestría, CIIDIR-IPN, 86. Unidad Sinaloa.