Análisis topológico del fenómeno *El Niño*
Uso de geometría y topología para ciencia de datos
Profesora Lilia Alanís López

A01197399 Diana Paola Cadena Nito
A01705747 Enrique García Varela
A01275180 Alexis Hernández Spinola
A00831314 Paola Sofía Reyes Mancheno
A01285041 María Fernanda Torres Alcubilla

Monterrey, Nuevo León

13 de junio de 2023

Resumen

El fenómeno meteorológico El Niño presenta una periodicidad complicada además de tener efectos negativos en diversos aspectos para el ser humano. Es por esto mismo que, a través de un análisis topológico y modelos de *machine learning*, se busca generar un modelo predictivo y de clasificación. Primero, se establece la periodicidad de los datos por medio del encaje de Takens y análisis de homologías persistentes. Posterior, se hace un suavizado de datos para poder realizar un modelo predictivo. Finalmente, se hace un modelo de clasificación para identificar la etapa del fenómeno predecida. Cabe mencionar que debido al periodo actual de El Niño, el modelo predictivo cuenta con áreas de mejora.

***Palabras clave**: Análisis topológico, Takens, MAPPER, predicciones, ML.*

1. Introducción

La Oscilación del Sur, ENOS, es un fenómeno meteorológico que tiene lugar en el océano Pacífico alrededor de la línea ecuatorial. (Guzman et al., 2020). Este proceso natural consta en el intercambio de aguas cálidas y frías a lo largo de esta franja en el pacífico tropical. A pesar de que se desconoce la causa inicial de este fenómeno, se han identificado dos componentes principales precedentes a su aparición: componente atmosférica y componente oceánica. (Maturana et al., 1997). La componente atmosférica se asocia con la fluctuación interanual del sistema de baja presión atmosférica superficial y el sistema de alta presión atmosférica superior. (STAFF, 2016).

Este fenómeno se define por sus dos etapas o fases, las cuales se identifican por un índice que indica la diferencia de la presión atmosférica media mensual entre la región de alta y baja presión para su componente atmosférica. En cuanto a la componente oceánica, esta se define por fuertes anomalías presentadas en la temperatura superficial del mar, ATSM, y en el cambio en el nivel del mar. Las anomalías se representan a través de la diferencia entre el valor observado menos la media climatológica de la zona en la que se realizó la medición. (Maturana et al., 1997). Estos cambios en la temperatura pueden ser positivos o negativos y prolongarse por meses de forma consecutiva.

La fase de El Niño se caracteriza un incremento en la temperatura superficial del mar, presentando anomalías positivas. Además, hay un incremento en el nivel del mar y decrecimiento de los vientos alisios. (Guzman et al., 2020). Este fenómeno regresa cada 2 y 7 años y tiene una duración entre 9 y 12 meses pero incluso se han registrado períodos prolongados de 3 a 4 años. (Maturana et al., 1997). Por otro lado, la fase de La Niña consta del enfriamiento de la temperatura superficial del mar, presentando anomalías negativas. Su aparición es menos frecuente a la de El Niño pero tiene una mayor duración, presentándose de 1 a 3 años. (Philander, 1998).

La importancia del estudio del ENOS recae en los efectos asociados en los patrones climáticos, incluso en regiones apartadas del Pacífico ecuatorial, central y oriental (Maturana et al., 1997). Una interferencia en la evolución puede significar un daño en el medio ambiente, el comercio e intereses sociales de la zona. Asimismo, El Niño Oscilación del Sur afecta a diversos fenómenos naturales presentes en las zonas afectadas por este. Como por ejemplo, inundaciones en zonas de centro América, incendios forestales a través de Australia del Este, oeste de África, en el norte y sur de América. De igual manera, hay un aumento de huracanes en el Pacífico del este y disminuyen en el Atlántico. (NOAA, 2023).

2. Metodología

Para la elaboración del proyecto, se utilizaron dos bases de datos y se implementaron conceptos topológicos para la identificación de periodicidad en los datos así como algoritmos de *machine learning* para predicción y clasificación. En cuanto a las bases de datos, la primera contiene información acerca de las temperaturas en todas las zonas de El Niño así como las anomalías percibidas. Las medidas en esta base de datos son mensuales, con un rango desde enero de 1982 hasta abril de 2023, dando un total de 496 registros; fue obtenida de `https://www.cpc.ncep.noaa.gov/data/indices/sstoi.indices`. La visualización de las anomalías para todas las zonas queda como se muestra a continuación.

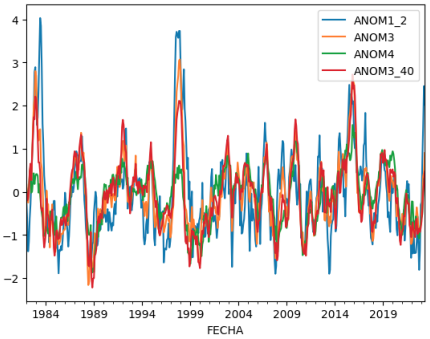


Figura 1: Anomalías registradas en todas las zonas.

La segunda base de datos contaba con registros trimestrales desde 1950 hasta 2023, abarcando únicamente la zona 3.40. De igual manera, los registros son de la temperatura y anomalías percibidas. Esta base de datos se obtuvo de `https://www.cpc.ncep.noaa.gov/data/indices/oni.ascii.txt` y para poder trabajar con ambas bases, esta se reduce a los registros desde 1982 a 2023. El comportamiento de la temperatura y anomalías de la zona 3.40 se puede observar en las siguientes figuras.

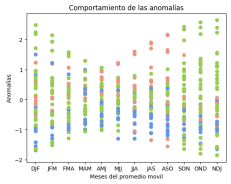


Figura 2: Anomalías El Niño 3.40

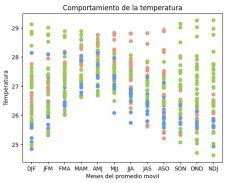


Figura 3: Temperatura El Niño 3.40

2.1. Takens

El encaje de Takens es una técnica de comprobación de periodicidad en datos determinísticos y no lineales; funciona a base de dos parámetros: dimensión y retraso de tiempo o *time delay*. (J. Stark, 1997). Con estos dos parámetros, se genera una imagen topológicamente equivalente a los datos originales, donde de mapean datos de una sola dimensión, al número de dimensiones que se definan. La información que se guarde dentro de la imagen de n-dimensiones, se define a través del retraso de tiempo, el cual define cada cuántas observaciones hacia el pasado se tomarán los n datos para cada dimensión. (Achigar Pereira, 2019). En sí, el retraso de tiempo permite encontrar relaciones entre observaciones que no se encuentran una después de la otra. Para la zona 3.40, se utilizó una dimensión de 3 y un retraso de 4.

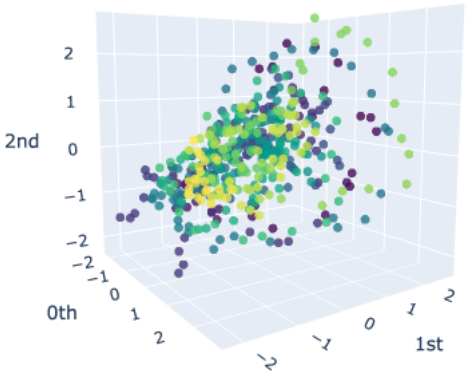


Figura 4: Nube de puntos El Niño 3.40

Asimismo, se utilizaron diagramas e imágenes de persistencia para la identificación de la cantidad de componentes principales y *1-hoyos* existentes en la topología. Los diagramas de persistencia dan una representación del nacimiento y muerte de las homologías. (García Castellanos, 2021). Sin embargo, para identificar cuáles de ellas se consideran ruido y cuáles son representativas de la topología, se realizan las imágenes de persistencia. En estas, las coordenadas ahora son nacimiento y persistencia y posteriormente se colorean como un mapa de color y permiten visualizar el número de *1-hoyos* representativos. (Vázquez Fernández, 2022). Además, se trabajó con distintas varianzas, $\sigma_1 = 0{,}01, \sigma_2 = 0{,}005$ y $\sigma_3 = 0{,}001$, y resoluciones.

2.2. Mapper

Con el objetivo de obtener características topológicas de nuestra base de datos, realizamos un análisis Mapper utilizando la librería de *kmapper*. Primero se realiza la inicialización del objeto de mapper al que se esten aplicando las funciones correspondientes. Entonces, a esta se le estará dando la función de proyección de las variables de fecha, año y de clasificación (Niño, Niña y Neutral) aplicadas a la temperatura. Además, se estaría asignando a una cubierta de n_cubos y con un porcentaje de traslape. Esto se estaría guardando finalmente dentro de un objeto con la información del complejo.

Más adelante, la aplicación de esta función permite obtener la visualización de dicha proyección a través del mapper. Con este, se puede identificar los elementos dentro de cada cluster, así como el color asociado a cierta variable. Para esta implementación se usaron los parámetros de 3 cubos con 4 clusters cada uno y un traslape del 0.2, con la variable de clasificación del fenómeno como función de color.

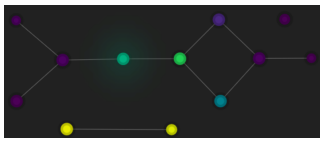


Figura 5: Topología de los datos

2.3. Modelación de datos

Para proceder con la modelación de datos y realizar predicciones acerca de El Niño y su clasificación, primero se requiere suavizar los datos. Esto con el propósito de lograr un funcionamiento correcto de los algoritmos y librerías a implementar.

2.3.1. Suavización de datos

La suavización de datos se realizó a través de dos métodos: Savitzky Golay y el súper-suavizador de Friedman. Savitzky Golay se basa en una regresión polinomial local, requiriendo distintos parámetros como la ventana y el grado del polinomio, principalmente. Se establecieron distintas combinaciones de los parámetros para encontrar un suavizado adecuado. Además, para medir el suavizado de datos, se utilizó la métrica del error medio cuadrado y se estableció un rango distinto por zona para desplegar los parámetros y el error asociado. Esto con la finalidad de no suavizar los datos de más, evitando perder su comportamiento. Para la zona de El Niño 3.40, se estableció un rango de error de (0.20, 0.25), obteniendo un modelo con una ventana de 40, polinomio de grado 12 y modo de *wrap*.

Por otro lado, el método del súper-suavizador de Friedman se basa de igual manera en una regresión lineal local. A diferencia del método anterior, las ventanas en Friedman se calculan de manera automática y suele ser utilizado cuando se trabaja con datos de estacionalidad complicada. En este método, la métrica a tomar en consideración fue la desviación estándar de los datos. Se optó por trabajar con dos métodos distintos para medir la eficacia de cada uno y poder seleccionar aquél que mostrara un mejor desempeño. En este caso, la desviación estándar de los datos en esta zona es de 0.8327 y cambia a 0.3985; con una frecuencia $f_3 = 0{,}50$, la desviación es de 0.4975.

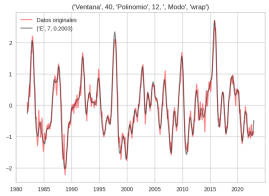


Figura 6: Savitzky-Golay El Niño 3.40

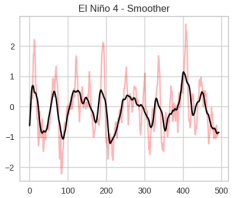


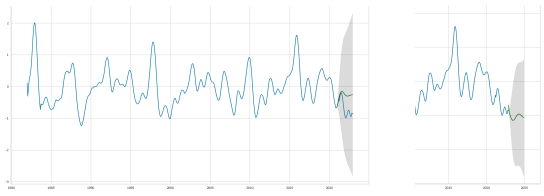
Figura 7: Suavizador Friedman El Niño 3.40 con frecuencia 0.5.

2.3.2. Series de tiempo

Para poder generar un modelo que prediga el estado de ENOS (Niño, Niña o Neutral) de manera mensual, se debe realizar un forecast o serie de tiempo que pronostique anomalías mensuales. Este forecast se hizo utilizando el Modelo Autorregresivo Integrado de Media Móvil, ARIMA por sus siglas en inglés, el cual se aplicó a los datos suavizados por el método de Savistky Golay y por Friedman. Previo a la aplicación dicho modelo, se utilizó la Prueba estacionaria de Dickey Fuller aumentada, para poder comprobar la hipótesis de que los dos sets de datos son estacionarios; misma que fue aceptada en ambos casos.

En el caso de los datos suavizados por Savitzky Golay, los parámetros utilizados fueron: $AR = 4$, lo cual muestra una correlación fuerte del n-mes con el mes n-4 (hecho que se evidencia también en el

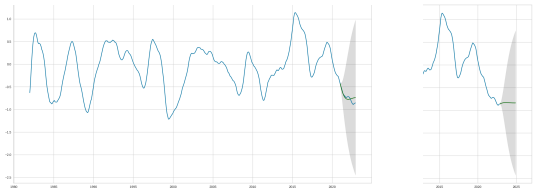
análisis de Takens); $I = 1$, la cual ayuda a tener un mejor ajuste en los datos; y finalmente $MA = 2$, que representa el promedio móvil que se utiliza para generar el modelo. Con estos parámetros, el Error Absoluto Medio Porcentual es del 0.51, valor que se da debido a que la serie de tiempo no alcanza los valores reales de las anomalías, sino que muestra la tendencia general de estas. Dicho comportamiento se puede observar en la figura 8, donde la 7a es la predicción de los datos de prueba (2021 y 2022) y 7b muestra el forecast de los años 2023 y 2024.



(a) Predicción 21-22 (b) Forecast 23-24

Figura 8: ARIMA (4,1,2) Savitzky Golay

Con respecto a los datos suavizados con Friedman, los parámetros escogidos fueron los siguientes: $AR = 4$, al igual que con el otro set de datos; $I = 1$, y $MA = 3$. En este caso, el valor de MAPE fue de 0.065, mostrando un modelo más ajustado a los datos. Esto se puede dar debido a que, como la desviación estándar de los datos se reduce a la mitad con Friedman, es más fácil que se pueda alcanzar los valores de las anomalías. Este modelo se puede observar en la figura 9.

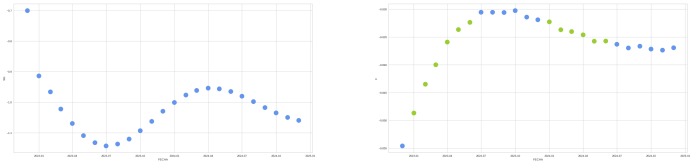


(a) Predicción 21-22 (b) Forecast 23-24

Figura 9: ARIMA (4,1,3) Friedman

2.3.3. Clasificación

Finalmente, se realizaron dos modelos de clasificación utilizando Extreme Gradient Boosting; el primero con los datos suavizados por Savitzky Golay, y el segundo con los datos de Friedman. En el primer caso, se obtuvo un accuracy del 0.82, y en el segundo del 0.8. La figura 10 muestra la clasificación del ENOS en el período comprendido entre 2023 y 2024. El color verde representa a un período *neutro*, y el azul significa que ese mes estaría ocurriendo el fenómeno de la Niña.



(a) Savitzky Golay (b) Friedman

Figura 10: Clasificación Forecast 2023-2024

3. Resultados

Al realizar el encaje de Takens como se visualiza en la figura 4 se aplicó un diagrama de persistencia para poder comprobar que exista un solo *1-hoyo* en la topología. Las imágenes de persistencia para la zona 3.40, mostradas a continuación, presentan un punto amarillo lo que indica la presencia de un hoyo; hecho que significa que existe periodicidad en los datos con un time-delay de 4 meses y una dimensión de 3. Es importante recalcar que dicho análisis también se condujo dividiendo en períodos de 12 años, donde dicha periodicidad se puede ver de forma más clara, pudiendo ser por los efectos del calentamiento global en la Oscilación del Sur.

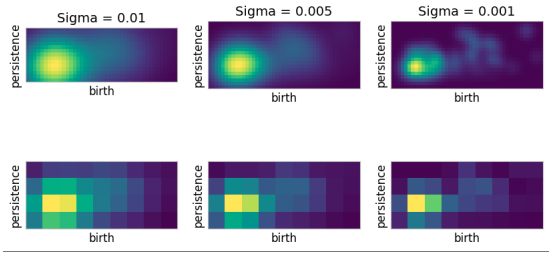


Figura 11: Imágenes de persistencia de El Niño 3.40.

Por otro lado, en el caso del análisis Mapper, las tonalidades moradas en la figura 5 representan una mayoría de datos neutros; las amarillas, Niños; y las azules, Niñas. Si se tiene verde, este representa una combinación entre Niña y Niño. Esto permite apreciar una separación por fenómenos, además de una separación por meses, donde el lado izquierdo y el centro comparten meses que no lo hacen con el lado derecho de la topología; esto se puede apreciar en la figura 12. En esta figura, también se puede observar como los fenómenos de El Niño y La Niña comparten casi los mismos nodos, los mismos que se componen por la mayoría de los meses. Esto se relaciona con los períodos de duración de los dos fenómenos, ya que ambos tienen períodos mayores

a 9 meses.

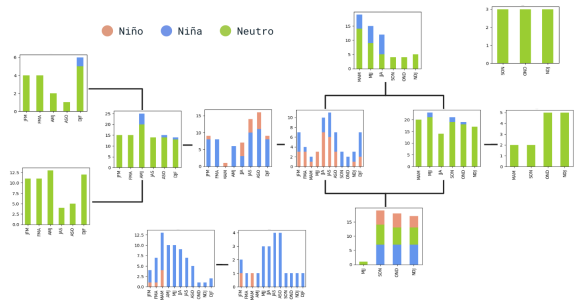


Figura 12: Gráfica de barras por meses de la topología

Así también, los *forecasts* obtenidos en las figuras 8 y 9, muestran en ambos casos un leve aumento en las anomalías para los años 2023 y 2024. Tomando en cuenta que durante el primer semestre de 2023 las anomalías han sido positivas (CLIMATICAS/NCEP/NWS, 2023), los *forecasts* no han podido aumentar lo suficiente como para mostrar valores positivos. Esto puede deberse a que, al final de 2022, se tiene un mínimo local y aumenta rápidamente para alcanzar el fenómeno El Niño desde mayo 2023 (CLIMATICAS/NCEP/NWS, 2023); hecho que dificulta que el modelo planteado simule con precisión dicho cambio tan brusco.

Por otro lado, como se puede observar en la figura 10, la mayoría de los meses están categorizados como fenómeno de La Niña. Al compararlo con la información real del primer semestre del 2023, siendo incorrecto. Esto sucede porque, a pesar de que la exactitud del modelo de clasificación es admisible, al agregar el error del modelo de series de tiempo, la exactitud de las predicciones se reduce considerablemente.

4. Conclusiones

De forma general, la complejidad de la Oscilación del Sur pudo ser observada a través de herramientas topológicas. Las mismas dieron la posibilidad de diseccionar ciertas propiedades de las anomalías que no son posibles de visualizar con estadística descriptiva. Dentro del análisis topológico de la Oscilación del Sur, se pudo evidenciar la existencia de una periodicidad de los datos. Aunque esta no puede revisarse a simple vista, el encaje de Takens brindó la posibilidad de encontrar esta correlación fuerte entre los datos de cada 4 meses, la misma que permitió generar una serie de tiempo con ARIMA. Así también, el análisis de mapper permitió encontrar que el suceso de los fenómenos de El Niño y La Niña no están restringidos a ciertos meses.

Con respecto a la predicción de fenómenos de El Niño o de La Niña, estos se vieron limitados al no ser capaces de predecir las anomalías mensuales de la zona 3.40. No obstante, muestran de forma general la tendencia de los datos dentro de un rango de 2 años, lo que permite saber a un corto plazo hacia dónde se puede estar dirigiendo el estado del ENOS.

4.1. Recomendaciones y trabajos futuros

Si bien el análisis topológico de los datos permitió conocer características generales del comportamiento de las anomalías, en futuros trabajos podría hacerse la exploración topológica separando los datos por cada fenómeno de El Niño. Este análisis permitiría encontrar otras propiedades o incluso preguntas, cómo ¿cuál es el impacto de la duración del fenómeno de El Niño en la duración del fenómeno de La Niña? Asimismo, en el ámbito de la predicción del comportamiento de ENOS, podría utilizarse bases de datos con más años de antelación, así como el aplicar algoritmos de redes neuronales que permitan tener *forecasts* más exactos. Por último, este es un fenómeno complejo en el que se involucran más variables, por lo que estas también podrían ser consideradas y generar mejores predicciones.

5. Anexos

Acceso al Github: https://github.com/dpcadenan/MA2007B

Referencias

Mauricio Achigar Pereira. Dinámica topológica expansiva: algunos aportes. 2019.

CENTRO DE PREDICCIONES CLIMATICAS/NCEP/NWS. El niÑo/oscilacion del sur(enso por sus siglas en inglés) discusion diagnostica, 2023. URL `https://www.cpc.ncep.noaa.gov/products/analysis_monitoring/enso_advisory/ensodisc_Sp.shtml`.

Alejandro García Castellanos. Robustez de la homología persistente: el teorema de estabilidad. 2021.

Emanuel Guzman, Carmela Ramos, and Ali Dastgheib. Influence of the el nino phenomenon on shoreline evolution. case study: Callao bay, peru. *Journal of Marine Science and Engineering*, 8(2):90, 2020.

M.E. Davies J. Huke J. Stark, D.S. Broomhead. Takens embedding theorems for forced and stochastic systems, nonlinear analysis: Theory, methods applications. 30:5303–5314, 1997.

Jenny Maturana, Mónica Bello, and Michelle Manley. Antecedentes históricos y descripción del fenómeno el niño, oscilación del sur. *El Niño-La Niña*, 2000:13–27, 1997.

NOAA. El niño southern oscillation (enso), 2023. URL `https://psl.noaa.gov/enso/`.

S George Philander. Who is el niño? *Eos, Transactions American Geophysical Union*, 79(13):170–170, 1998.

CLIMATE.GOV STAFF. El niño and la niña: Frequently asked questions, 2016. URL `https://www.climate.gov/news-features/understanding-climate/el-ni%C3%B1o-and-la-ni%C3%B1a-frequently-asked-questions`.

David Vázquez Fernández. Análisis topológico de datos y aprendizaje automático en la detección de ríos atmosféricos. 2022.