DIVULGACIÓN CIENTÍFICA

MACHINE LEARNING Y CAUSALIDAD

Ken Takahashi[1]

[1] Instituto Geofísico del Perú (IGP), Lima, Perú

Ken Takahashi es investigador científico principal del Instituto Geofísico del Perú, especializado en El Niño y aplicaciones de la inteligencia artificial. Es director de la Subdirección de Ciencias de la Atmósfera e Hidrósfera y responsable del Observatorio del Conocimiento Científico sobre Cambio Climático. Doctor en Ciencias Atmosféricas por la Universidad de Washington e investigador distinguido en RENACYT.

Palabras clave: *Machine learning*, causalidad, IA

Citar como Takahashi, K. (2023). Machine learning y causalidad. *Boletín científico El Niño*, Instituto Geofísico del Perú, vol. 10 n.° 11, págs. 4-9.

Resumen

La inteligencia artificial, particularmente el *machine learning*, está mostrando grandes avances en muchos campos de la ciencia. Sin embargo, los patrones que estas técnicas identifican para hacer sus predicciones no consideran las relaciones causa y efecto entre las variables consideradas, lo cual puede resultar en modelos que no sean robustos ante cambios en la naturaleza de los datos, como podría ocurrir con El Niño ante el cambio climático. Es necesario potenciar el entendimiento humano experto en los dominios del conocimiento, como la variabilidad y cambio climático, para que sea la guía para la evaluación y mejora de los modelos de *machine learning*.

1. Introducción

El año 2023 se caracterizó por un crecimiento acelerado en el uso directo y explícito de la inteligencia artificial (IA) por parte del público en general, particularmente de las herramientas de IA generativa como el ChatGPT. Estas herramientas han demostrado una extraordinaria capacidad para producir textos, imágenes, música, video, entre otros, con habilidad comparable o incluso superior a la de las personas promedio en algunas tareas.

Sin embargo, la IA ya tiene varios años de rápidos avances gracias al uso de las diversas técnicas de *machine learning* (aprendizaje automático), que van desde la regresión lineal más sencilla hasta el *deep learning* (aprendizaje profundo) basado en redes neuronales artificiales. El *machine learning* permite que los algoritmos descubran y aprendan relaciones y patrones a partir de grandes volúmenes de datos, lo que posibilita realizar predicciones con un alto nivel de exactitud. Esto se logra sin necesidad de que expertos en el tema específico programen cada regla o fórmula necesaria para decidir o calcular el resultado, como se hace con los denominados "sistemas expertos", labor muy difícil para problemas complejos con múltiples aspectos. En cambio, con el aprendizaje automático, se preparan conjuntos de datos de entrada y sus correspondientes resultados esperados. A través de un proceso automático, el modelo se ajusta por sí mismo hasta alcanzar la capacidad de producir dichos resultados, sin mayor intervención humana.

Como ejemplo, la definición operacional de El Niño en la costa del Perú se basa únicamente en el Índice Costero El Niño (ICEN). Este índice es una medida de la temperatura de la superficie del mar en la región Niño 1+2, ubicada frente a nuestra costa norte. La magnitud de los eventos se determina según los valores umbral que sean excedidos por el ICEN por al menos tres meses (ENFEN, 2014; Takahashi et al., 2012). La especificación del índice, los umbrales y el criterio de los tres meses se realizó en consenso por el grupo

de especialistas de la comisión ENFEN. Por tanto, este procedimiento operacional, empleado para determinar si estamos en presencia de El Niño, se consideraría un "sistema experto". Ahora, si bien el uso de algoritmos es relativamente complejo, lo sería mucho más si adicionalmente se quisiera considerar variables como las lluvias extremas, inundaciones en la costa, aumento en el nivel del mar, ocurrencia de golpes de calor, enfermedades infecciosas como el dengue, impactos en la pesquería o agricultura, ya que sería necesario incorporar reglas para cada una de estas variables adicionales, lo cual sería un ejercicio bastante subjetivo en tanto que, por ejemplo, un físico y un biólogo tendrían una diferente noción de lo que es importante. Otra opción sería que el grupo de especialistas se ponga de acuerdo con el listado de eventos El Niño históricos y sus magnitudes, basado en criterios no especificados, con el cual luego se podría "entrenar" un modelo de aprendizaje automático para que desarrolle un algoritmo que considere todas las diferentes variables para establecer la ocurrencia de El Niño y su magnitud.

A grandes rasgos, si bien este segundo enfoque tiene mucho potencial, se requiere un gran volumen de datos para entrenar un modelo confiable, lo cual no se cumple en el caso de la variabilidad climática. Se puede pensar que por tener 50 años de registros climáticos tenemos una gran base de datos, pero esto dependerá mucho de la aplicación, particularmente de la escala temporal. Si queremos entrenar un modelo para el pronóstico de lluvias intensas en el Perú, podríamos considerar datos de precipitación horaria. Como primera aproximación, podríamos tener 24 horas por día multiplicado por 90 días por pico de temporada de lluvia anual, multiplicado por 50 años de registro, lo que daría 108 000 datos efectivos. Por otro lado, para el caso del pronóstico de El Niño, dado que este ocurre —a lo más— una vez por año, tenemos solo 50 años efectivos de datos con la misma longitud de registro. Esto sería insuficiente, ya que un modelo de aprendizaje profundo básico para el pronóstico de El Niño podría tener cientos de parámetros o más, los cuales tendrían que ser ajustados durante el entrenamiento.

Una forma de superar la limitación de lo corto del registro de El Niño es el uso de datos sintéticos para entrenar estos modelos, técnica que se está aplicando con bastante éxito en muchos campos, incluyendo el pronóstico de El Niño (Ham et al., 2019). En el caso del reciente modelo para el pronóstico de El Niño que desarrollamos en el Instituto Geofísico del Perú (IGP), este fue entrenado inicialmente usando miles de años de simulaciones de modelos climáticos globales con una representación adecuada de El Niño y, posteriormente, fue afinado con datos observacionales (Rivera-Tello et al., 2023). Sin embargo, una limitación en este ejemplo es que no se tiene certeza de que dichos modelos globales ni el registro observacional representen la diversidad de eventos El Niño que puedan presentarse en la naturaleza, por lo que siempre existirá la posibilidad de que un próximo evento no pueda ser pronosticado adecuadamente, sobre todo considerando que el contexto del cambio climático implica un cambio constante en la estadística de los datos, algo que se conoce como "deriva de datos" (*data drift*) en el ámbito del aprendizaje automático. Esta falta de representatividad es una limitante fundamental de todos los modelos empíricos generados solo mediante datos, a diferencia de los modelos climáticos globales y otros basados en las leyes de la naturaleza, aún cuando son aproximados, ya que, en principio, podrían tener validez fuera del rango de situaciones para los cuales fueron desarrollados.

Una diferencia fundamental en la generalizabilidad de los modelos empíricos típicos y los modelos climáticos basados en la física es que los primeros identifican correlaciones, pero no la causalidad. El *machine learning* está orientado a identificar patrones en las bases de datos que corresponden a determinados resultados. En la medida en que las bases de datos sean suficientemente amplias, no solo en cantidad de datos sino en la diversidad de posibilidades y situaciones que pudieran darse en la realidad, los modelos podrían desempeñarse bien con nuevos datos, es decir, que podrán "generalizar" adecuadamente sin necesidad de identificar las causas y efectos. Lamentablemente, como se indicó arriba, el mundo está en constante cambio, lo que implica una deriva, de manera que los datos del pasado no reflejarán necesariamente lo que se presentará en el futuro. Por lo tanto, estos patrones podrían dejar de ser válidos.

Por ejemplo, sabemos que para que ocurran fuertes lluvias en la costa norte del Perú es necesario tener una alta temperatura del mar en la costa, por encima de un umbral de aproximadamente 26 °C (Woodman

y Takahashi, 2014); sin embargo, sabemos que este modelo empírico se volverá inadecuado en el futuro dado que, con el cambio climático, la temperatura umbral también aumentaría en el futuro debido al calentamiento de la atmósfera (Johnson y Xie, 2012). Conociendo los mecanismos físicos, se puede incorporar esta corrección a los modelos empíricos (Jáuregui y Takahashi, 2016), pero esto requiere participación de expertos en clima tropical y El Niño en el diseño y ajuste de los modelos. Un modelo puro de *machine learning* identifica patrones, pero no distingue causa y efecto, es decir, no considera los mecanismos. Las correlaciones entre los datos pueden ser fuertes, pero pueden sugerir relaciones de causalidad erradas y los modelos basados en estas relaciones pueden fallar ante nuevos datos.

En el caso de los modelos climáticos globales, sus ecuaciones matemáticas se basan en las leyes de la física, las cuales, en teoría, son aplicables en nuevas situaciones, incluyendo el cambio climático (excepto por ciertas relaciones empíricas que puedan ser incorporadas al modelo por necesidad). Las leyes dinámicas incorporan la causalidad explícitamente y codifican los mecanismos que determinan cómo deben cambiar las variables del clima dada la situación actual.

2. Análisis causal

En las últimas décadas, el análisis de la causalidad ha pasado a ser algo que se puede evaluar empíricamente. Como explica Judea Pearl en "El Libro del Por Qué" (Pearl & Mackenzie, 2018), se puede aplicar la estadística utilizando datos preexistentes, incluso para responder preguntas como ¿qué pasaría con B si modificamos A?, aun cuando dichos datos no provengan de experimentos controlados diseñados para este fin. No obstante, esto es posible siempre y cuando se tenga una noción clara de las relaciones causa-efecto, lo cual debe provenir del conocimiento y entendimiento experto de la situación específica bajo estudio.

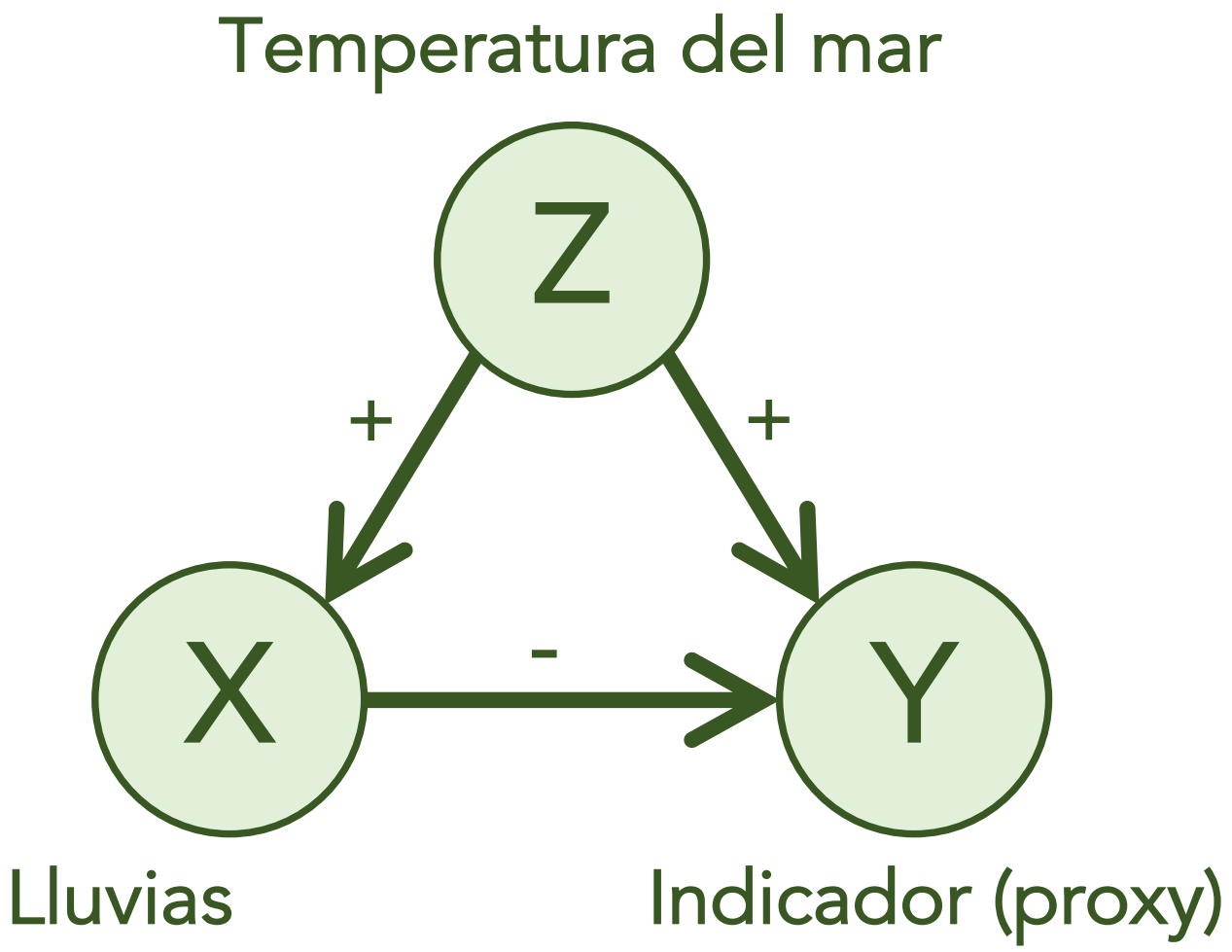

Figura 1. Diagrama de causalidad para las variables, X, Y y Z. Las flechas apuntan de causa a efecto. Los símbolos + y - indican si la causa genera un efecto en la misma dirección o en el opuesto, respectivamente.

La importancia de las relaciones causales en los resultados estadísticos se puede ilustrar con el siguiente ejemplo. Consideremos las variables Z, X e Y, las cuales representan, respectivamente, la variabilidad de la temperatura superficial del mar asociada a El Niño y La Niña, las lluvias en la costa norte que son afectadas por la temperatura del mar, y una variable hipotética que podría ser un indicador paleoclimático de El Niño que es influenciado tanto por la temperatura del mar como por las lluvias, es decir, Z es causa de X, mientras que X y Z son causas de Y. Estas relaciones de causalidad se ilustran en la Figura 1, en la que las flechas apuntan de las causas a los efectos. Además, en la figura, las flechas tienen un signo más (+) o menos (-) según si, un incremento en la variable causa, produce directamente un incremento o una reducción, respectivamente, en la variable de efecto. Implícito en esta figura es que tanto X, Y y Z pueden estar influenciadas por otros factores externos que son causalmente independientes entre sí y que no son afectadas por las tres variables, generando "ruido" en su variabilidad.

Estas relaciones de causalidad las hacemos cuantitativas con las siguientes ecuaciones:

- Z = (causas externas de Z) (1)
- X = Z + (causas externas de X) (2)
- Y = 2.5 Z - X + (causas externas de Y) (3)

En las que las causas externas se representan mediante números aleatorios con distribución normal. En estas ecuaciones, las causas están en el lado derecho y los efectos al lado izquierdo. Con estas ecuaciones podemos generar datos como los mostrados en la Figura 2.

La Figura 2 es una gráfica de dispersión de datos de Y contra X generados por las ecuaciones 1-3. Sabemos que X debería producir un efecto negativo en Y según la ecuación 3, ¡pero la figura parece mostrar lo contrario!, es decir, a mayores valores de X tenemos mayores valores de Y en promedio, lo cual se verifica con una línea de tendencia correspondiente a un ajuste de regresión lineal, cuya ecuación:

- Y = 0.63 X + 0.02 (4)

Tiene un coeficiente positivo para X de 0.63, mientras que la ecuación 3 indica que el coeficiente debería ser -1.

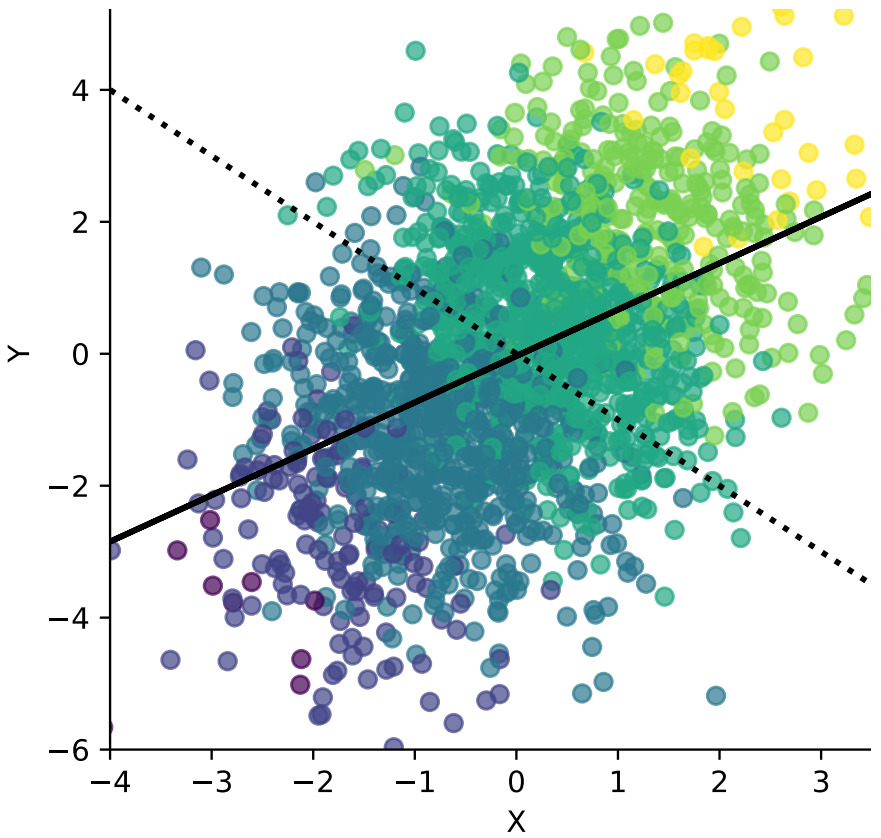

Figura 2. Diagrama de dispersión entre los valores de X e Y generados por las ecuaciones 2-4; los colores indican valores de Z (valores menores son más oscuros). El efecto causal de X sobre Y es negativo, pero el diagrama de dispersión y el ajuste de regresión lineal (línea negra, ecuación 1) sugieren un efecto positivo debido a la acción del confusor Z. La línea punteada indica el efecto real de X sobre Y (Y = -X).

La explicación de esta situación se encuentra en la variable Z, la cual afecta tanto a X como a Y (ecuaciones 2 y 3) y no es influenciada ni por X ni Y (ecuación 1). Dado que genera confusión en la relación entre X e Y, a Z se le denomina "variable de confusión". La regresión lineal en la Figura 2 en realidad está mostrando el efecto combinado de X y de Z sobre Y. Si anuláramos el efecto de la variabilidad de Z sobre Y, por ejemplo, fijándonos en los puntos de un solo color en la Figura 2, los que aproximadamente corresponden a un valor de Z específico, podemos verificar que la relación entre Y y X es negativa, ilustrada con la línea punteada. Así que, si sabemos *a priori* que Z es un confusor, sería incorrecto hacer simplemente la regresión lineal de la ecuación 4 para estimar el efecto directo de X sobre Y, aun cuando el ajuste sea cuantitativamente bueno. Más bien, si usamos tanto X como Z como predictores, obtendremos un resultado más real, con una ecuación de regresión lineal múltiple cercana a la ecuación 3.

Esto nos podría sugerir que, en el caso de que no conozcamos bien las relaciones causales indicadas, como se muestra en la Figura 1, en general deberíamos incluir todas las posibles variables relacionadas a

Y como predictores para corregir los efectos de los potenciales confusores; sin embargo, esto tampoco es una solución confiable. Por ejemplo, si nuestro interés fuera ahora estimar X y usáramos Y y Z como predictores, obtendremos X = 1.50 Z - 0.32 Y, lo cual es causalmente erróneo, ya que, al compararlo con la ecuación 2, no existe una influencia causal de Y sobre X. Además, la magnitud estimada del efecto de Z sobre X es 50 % mayor que el real (coeficiente de 1.5 vs. 1.0). En este caso, el incorporar Y como predictor distorsiona el resultado, ya que Y es efecto de X, no causa, así como también es efecto de Z. En esta situación, Y es conocido como un "colisionador" debido a que las flechas colisionan en Y en la Figura 1. Si más bien, sabiendo que Y no debe ser predictor, solo consideramos a Z, obtenemos un resultado consistente con la ecuación 2 ($X \approx Z$). Sin embargo, un punto crucial es que la regresión lineal causalmente errónea presenta un mejor ajuste que la regresión lineal causalmente correcta (menor error cuadrático medio de 0.57 vs. 0.70). Esto significa que, si no conociéramos las relaciones causales, es probable que elijamos el modelo errado como una mejor representación de la realidad. Con la tendencia actual de desarrollar grandes modelos de IA construidos ciega y exclusivamente sobre la base de datos (*data driven*), con la mínima intervención humana en su diseño y ajuste, prevalecerán las relaciones o patrones que resultan en mejor ajuste, aun cuando representen erróneamente la causalidad.

Pero ¿qué importa si la causalidad en el modelo es errada, si lo que nos interesa es el mejor ajuste para la predicción? Desde una perspectiva práctica, ¿no es esto lo único que nos interesa? Más allá de conocer y entender el funcionamiento de la realidad, un modelo causalmente correcto, como los modelos climáticos globales, es más robusto si algo cambia en las condiciones en las que es aplicado. El caso más obvio es el cambio climático, el cual genera gran incertidumbre acerca de si los modelos empíricos puramente basados en datos mantengan su validez en un clima futuro. En nuestro ejemplo, si el *proxy* paleoclimático Y incorporara una tendencia que no tiene relación directa con la temperatura Z o la lluvia X (con una magnitud aproximadamente igual a la del efecto de Z sobre Y), por ejemplo, debido a cambios en el uso del suelo producidos por la acción humana, el desempeño del modelo desarrollado con X y Z como predictores ahora presentaría un mayor error cuadrático medio (0.73) que el modelo que solo considera X. Este último es más robusto, ya que representa más fielmente al proceso real que genera los datos.

3. Perspectivas

Ante el rápido avance en el *machine learning*, no podemos decir "jamás" en relación con las capacidades de la inteligencia artificial, particularmente en comparación con las humanas. Sin embargo, en este momento, los modelos no respetan necesariamente la causalidad de las relaciones entre las variables consideradas y sigue estando en el dominio de los humanos el asegurar que el diseño de estos modelos lo hagan. Si bien hay intentos de automatizar la identificación de relaciones causales (por ej., Runge et al., 2019), por lo pronto, las estructuras de causalidad, como en la Figura 1, deben ser construidas por expertos humanos sobre el balance de los resultados de investigaciones previas, así como su propia experiencia e intuición. Este último punto de la intuición implica que las estructuras causales en muchos casos serán hipótesis que deben ser validadas.

Con la revolución que está trayendo la IA, los científicos humanos no pueden desaprovechar las ventajas que traen las nuevas herramientas. Sin embargo, una reciente encuesta de la revista Nature a más de 1600 investigadores (Van Noorden y Perkel, 2023) indicó que casi el 70 % de estos consideran que uno de los impactos negativos de la IA a la investigación científica sería generar una mayor dependencia en la identificación de patrones (es decir, correlaciones) en lugar de entendimiento (causalidad), sobre todo en aquellos casos en los que más adelante podría producir la degradación de la base de la ciencia. Es esencial asegurar que los investigadores que incursionan en el uso de la IA mantengan una perspectiva crítica hacia esta y prioricen el entendimiento como un objetivo, al menos, tan importante como la predicción. De esta manera, el expertise humano en los dominios del conocimiento, como la variabilidad y cambio climático, sirva de guía para la evaluación y mejora de los modelos de *machine learning*.

Referencias

Ham, YG., Kim, JH. & Luo, JJ., 2019: Deep learning for multi-year ENSO forecasts. *Nature* 573, 568–572. https://doi.org/10.1038/s41586-019-1559-7

Jauregui, Y.R., and K. Takahashi, 2017: Simple physical-empirical model of the precipitation distribution based on a tropical sea surface temperature threshold and the effects of climate change. *Clim. Dyn* 50, no. 2217. https://doi.org/10.1007/s00382-017-3745-3.

Johnson, Nathaniel C, and Shang-Ping Xie, 2010: Changes in the sea surface temperature threshold for tropical convection." *Nature Geoscience* 3, no. 12, 842–45. https://doi.org/10.1038/ngeo1008.

Pearl. J., y Mackenzie, D., 2018: *The Book of Why. The New Science of Cause and Effect*, Basic Books. 418 pp.

Rivera Tello, G.A., Takahashi, K. & Karamperidou, C., 2023: Explained predictions of strong eastern Pacific El Niño events using deep learning. *Scientific Reports* 13, 21150. https://doi.org/10.1038/s41598-023-45739-3

Runge, J., Bathiany, S., Bollt, E. et al., 2019: Inferring Causation from Time Series in Earth System Sciences. *Nature Communications* 10, no. 1: 2553. https://doi.org/10.1038/s41467-019-10105-3.

Takahashi, K., Mosquera, K. y Reupo, J. (2014). El Índice Costero El Niño (ICEN): historia y actualización. *Boletín técnico: Generación de modelos climáticos para el pronóstico de la ocurrencia del Fenómeno El Niño*, Instituto Geofísico del Perú, 1 (2), 8-9. http://hdl.handle.net/20.500.12816/4639

Van Noorden, Richard, and Jeffrey M. Perkel, 2023: AI and Science: What 1,600 Researchers Think. *Nature* 621, no. 7980: 672–75. https://doi.org/10.1038/d41586-023-02980-0.

Woodman, R. y Takahashi, K. (2014). ¿Por qué no llueve en la costa del Perú (salvo durante El Niño)? *Boletín técnico: Generación de modelos climáticos para el pronóstico de la ocurrencia del Fenómeno El Niño*, Instituto Geofísico del Perú, 1 (6), 4-7. http://hdl.handle.net/20.500.12816/5046