What do Sequential Patterns Say About the "El Niño" Phenomenon?

O. Díaz-Barriga, M. Nunez-del-Prado, *Member, IEEE,* and H. Alatrista-Salas, *Member, IEEE*

***Abstract*—El Niño phenomenon starts with an increase in the temperature of the sea surface in the equatorial zone of the Pacific Ocean. This increase is characterized by the arrival of a superficial mass of warm waters into the sea, which generates anomalous climate changes on land. These unusual events can be floods, droughts, intense rains, which endanger the urban population and infrastructure of cities. To be able to launch early warnings of possible catastrophic events in populated areas, it is necessary to know how and within how long the change in sea temperature impacts on continental characteristics. The present work describes a computational process based on techniques of extraction and visualization of sequential patterns to capture temporal variations of the variables describing the El Niño phenomenon. Results show the existence of correlations between the sea surface temperature and the flow of the rivers of the coast. These correlations can be used as monitoring tools for early warning releases.**

***Index Terms*—El Niño phenomenon, Data mining, Sequential pattern mining, Pattern Visualization.**

I. Introducción

El Niño es un fenómeno climático que consiste en el aumento de la temperatura del mar en el Pacífico Ecuatorial. Éste a su vez forma parte del ENSO (El Niño - Oscilación del Sur) que tiene un periodo de fluctuación de 2 a 7 años, con una fase cálida conocida como El Niño y una fase fría, La Niña [1]. El monitoreo del ENSO se realiza principalmente en 4 regiones del Pacífico Ecuatorial conocidas como: Niño 4, Niño 3.4, Niño 3; y Niño 1+2. Esta última zona se encuentra frente a la costa sur de Ecuador y a la costa norte del Perú. En este último país, El Niño ha golpeado muy fuertemente en estos últimos años. En el 2016, el incremento de la temperatura superficial del mar favoreció el aumento en la frecuencia de lluvias de magnitud muy fuerte, principalmente en la costa norte del Perú. Esto originó inundaciones con posteriores consecuencias en pérdidas materiales y humanas.

Específicamente, el fenómeno ocurre porque la Temperatura de la Superficie del Mar (TSM), que tradicionalmente es fría en el otoño e invierno, se calienta y ello trae también un aumento de la temperatura general del aire. Si la variación del TSM es igual o superior a $0,5°C$ $(0,9°F)$ en la región Niño 3.4, se dice que el fenómeno empezó.

Si bien es cierto que la temperatura del mar nos indica la presencia del fenómeno El Niño, es difícil cuantificar el impacto de este incremento en la tierra. Es decir, si la temperatura del mar aumenta en $0,5°C$, ¿Cómo será el impacto en la temperatura de las áreas costeras o en el caudal de los ríos en tierra?, ¿Si un incremento en la temperatura del mar es percibido, en cuánto tiempo se notarán los efectos en tierra?, ¿Cómo podemos utilizar esta información para lanzar alertas tempranas a las poblaciones instaladas en el litoral?

Para responder a estas preguntas, se deben utilizar técnicas que permitan analizar datos con características temporales (cuándo apareció el evento) y espaciales (dónde apareció el evento). Además, las técnicas seleccionadas deberían permitir capturar la evolución temporal de un conjunto de características que describen el fenómeno (*i.e.*, El Niño). En este sentido, las técnicas de minería de datos, específicamente la técnica de extracción de patrones secuenciales, permiten procesar grandes cantidades de datos heterogéneos con el objetivo de encontrar patrones que representan correlaciones temporales entre las características que describen el fenómeno en estudio [2].

En este sentido, el presente trabajo tiene como contribución principal el uso de técnicas de minería de patrones para capturar el impacto de la temperatura del agua del mar en las características continentales durante un evento anómalo. Para este fin, proponemos un proceso computacional para la extracción de patrones secuenciales a partir de datos heterogéneos asociados a variables meteorológicas, hidrográficas y oceanográficas de la costa norte del Perú, lugar donde el fenómeno El Niño tuvo lugar durante el período 2015-2016.

El resto del documento está organizado de la siguiente manera: la Sección II muestra el estado del arte. La Sección III describe la técnica utilizada para la extracción de patrones. Posteriormente, la Sección IV detalla las experimentaciones y los resultados obtenidos. El artículo finaliza con las conclusiones y los trabajos futuros descritos en la Sección V.

II. Estado del Arte

Al construir el estado de la literatura en temas relacionados al estudio del fenómeno El Niño utilizando técnicas de minería de datos, se encuentran investigaciones que buscan construir relaciones entre el fenómeno El Niño y otros fenómenos naturales, tales como lluvias, sequías, incendios, etc. Por ejemplo, Dhanya y Kumar [3] proponen una metodología para extraer patrones del tipo de reglas de asociación difusas entre los índices atmosféricos y la lluvia del monzón de verano de toda la India y dos regiones homogéneas. En este caso, los datos de El Niño - Oscilación del Sur (ENSO) y el índice de viento zonal de la oscilación Ecuatorial del Océano Índico se utilizaron como variables causales.

Por otro lado, existen otros trabajos basados en técnicas de aprendizaje de máquinas. Por ejemplo, en Rasouli *et al.* [4], los

Pontificia Universidad Católica del Perú, Av. Universitaria 1801, San Miguel, Lima, Peru.

Universidad del Pacífico, Av. Salaverry 2020, Jesús María, Lima, Peru.

autores compararon tres métodos de aprendizaje automático: *Bayesian Neural Network* (BNN), *Support Vector Regression* (SVR), *Gaussian Process* (GP) con la *Multi Linear Regression* (MLR) para el pronóstico de los caudales diarios de una pequeña cuenca en la Columbia Británica, Canadá, durante un periodo de 1 a 7 días. Para esto, se seleccionaron diferentes índices climáticos como son: la temperatura de la superficie del mar en la región El Niño 3.4, el Pacífico-Norteamérica (PNA), la Oscilación del Ártico (OA) y la Oscilación del Atlántico Norte (OAN).

Posteriormente, Kalra *et al.* [5] utilizaron *Support Vector Machines* (SVM) con el objetivo de mejorar los pronósticos de caudales en las cuencas de los ríos Gunnison y San Juan. Para ello se usa la información de los índices oceánicos-atmosféricos promedio anuales que consisten en: la Oscilación Decadal del Pacífico (ODP), la Oscilación del Atlántico Norte (OAN), la Oscilación Multidecadal del Atlántico (OMA), El Niño - Oscilación del Sur (ENSO) y la temperatura de la superficie del mar (TSM) para la región de Hondo en el período de 1906-2006.

Recientemente, Ganguli y Reddy [6] estudiaron las teleconexiones climáticas con las sequías meteorológicas. En este trabajo, los autores desarrollaron modelos de predicción utilizando *Support Vector Machines* (SVM) y copulas[1] enfocada sobre la región Rajasthan Occidental (India), donde se estudió cómo un análisis del clima a gran escala se relaciona con diferentes índices climáticos como El Niño - Oscilación del Sur.

Por otro lado, J. Kawale *et al.* [7], [8], se enfocan en el descubrimiento de dipolos de presión, fenómenos climáticos de larga distancia, con el objetivo de construir una red de anomalías climáticas utilizando la correlación de series de tiempo de las variables climáticas de todos los lugares de la Tierra, entre ellas El Niño.

Además, existen trabajos que miden el impacto del fenómeno El Niño en patrones de consumo. Por ejemplo, en [9], los autores usan la metodología *Box-Cox* para medir la influencia del fenómeno El Niño en el mercado eléctrico en Colombia. En vista de que el 80% de la energía proviene de fuentes hídricas, una inundación provocada por el fenómeno puede alterar los patrones de consumo. En este contexto, los autores cuantifican este impacto a partir del *Índice Oceánico El Niño*. De la misma manera, Caicedo *et al.* [10] estudian aumento de la incertidumbre hidrológica en Colombia producto del efecto El Niño y/o por la propuesta de cambio regulatorio del mecanismo de contratación de los servicios energéticos en ese país.

Finalmente, existen otros trabajos más exploratorios. Por ejemplo, en el trabajo propuesto por Janicke *et al.* [11] se utilizó la exploración visual de la variabilidad del clima mediante análisis *Wavelet*, usando como información la variación de la Temperatura Superficial del Mar (TSM) en la zona Niño 3.

La revisión del estado del arte en temas relacionados a la explotación de datos asociados al fenómeno El Niño para la obtención de conocimientos nuevos muestra que no se han realizado esfuerzos en términos del uso de técnicas de minería de datos para el estudio de dicho fenómeno y las aplicaciones de estas técnicas para comprender mejor el impacto del fenómeno en la sociedad. Específicamente, este trabajo focaliza en capturar el impacto de la temperatura del mar en las características meteorológicas e hidrológicas continentales mediante el uso de la técnica de extracción de patrones secuenciales, la cual es descrita en la sección siguiente.

[1]Las copulas, son objetos matemáticos que capturan completamente la estructura de dependencia entre las variables aleatorias, ofreciendo una gran flexibilidad en la construcción de modelos estocásticos multivariantes.

III. Patrones Secuenciales

La técnica de extracción de patrones secuenciales [12] permite descubrir correlaciones temporales a partir de un conjunto de características que describen un fenómeno y que cambian temporal y espacialmente. Básicamente la idea es, a partir de una base de datos de secuencias, extraer las sub-secuencias frecuentes. La técnica de extracción de patrones frecuentes reposa sobre ciertas nociones. Un *item* I es un valor literal y puede ser, por ejemplo, la humedad entre 60% y 80%. Un *itemset* IS es un conjunto de *items* que describen un área geográfica durante una estampilla temporal (*e.g.*, el conjunto de características meteorológicas de Tumbes, Perú el 18/02/2016). Además, una *secuencia* S es una lista ordenada de *itemsets*. Una secuencia representa la evolución temporal de un conjunto de características que describe un fenómeno (*e.g.*, la variación de características meteorológicas de una ciudad en los últimos 3 meses).

Por otro lado, una secuencia S' es una sub-secuencia de S si y solo si $S' \in S$. Finalmente, el soporte $supp(S')$ de una secuencia S' es el número de secuencias que contienen la sub-secuencia S'. El problema de extracción de patrones secuenciales se define como: dado un conjunto de secuencias y un soporte minimal σ (umbral), extraer todas las sub-secuencias tal que su soporte sea superior o igual al soporte minimal σ.

Por ejemplo, imaginemos que estudiamos tres zonas representadas por ciertas características. Las secuencias para estas zonas (base de secuencias) se muestran en el Cuadro I, donde P, C, V y T representan la presión, el caudal, la velocidad del viento y la temperatura, respectivamente. Los sub-índices a y b representan la intensidad (alta y baja). Por ejemplo, Pb es un *item* (presión baja). (Ta, Pb, Ca) es un *itemset* y representa las características meteorológicas de una zona (zona 1) en un instante dado (tiempo 3). Finalmente $(Ta, Pb, Cb, Va)(Tb, Pa, Cb)(Ta, Pb, Ca)$ es una secuencia y representa la evolución temporal de un conjunto de características que describen la zona 1. En la misma tabla podemos ver que la sub-secuencia $(Ta, Pb, Cb)(Tb)(Ta)$ aparece en dos zonas (zonas 1 y 3). Si el soporte minimal $\sigma = 2$ entonces la sub-secuencia antes mostrada sería un patrón secuencial.

TABLA I
Base de Secuencias

Zona	Secuencias
1	(Ta, Pb, Cb, Va)(Tb, Pa, Cb)(Ta, Pb, Ca)
2	(Ta, Pb, Cb)(Tb, Pb, Cb, Va)(Ta, Pa, Cb)
3	(Ta, Pb, Cb, Vb)(Ta, Pb, Ca)(Tb, Pb, Cb)

En la literatura existen muchos algoritmos que permiten la extracción de patrones secuenciales, algunos de ellos son PSP, FreeSPAN, PrefixSpan, LAPIN, PRISM y COPPER [13]. En este trabajo, se utilizará el algoritmo PrefixSpan, que es uno de los algoritmos más robustos cuando es aplicado en bases de datos densas [14] gracias a que PrefixSpan utiliza la estrategia de dividir y conquistar, al realizar una exploración en profundidad del espacio de búsqueda con proyecciones sucesivas de la base de datos.

Los siguientes párrafos muestran algunas definiciones para comprender cómo funciona el algoritmo PrefixSpan [15].

1) Prefijo de una secuencia: se define la función de prefijo como $S \times N \rightarrow S$, donde S es un conjunto de secuencias, N es un conjunto de enteros positivos y $prefijo(s,k) = s[1 : k]$. En otras palabras, el $prefijo(s,k)$ devuelve los primeros k elementos de la secuencia s.
2) Sufijo de una secuencia: se define la función de sufijo como $S \times S \rightarrow S$ tal que el $sufijo(s,s') = s[m+1 : n]$, si y solo si s' es un prefijo de s con m elementos y s es una secuencia que contiene n elementos o *items*.
3) Proyección de una base de secuencias: sea s una secuencia de la base de secuencias $seqBD$. La base de datos proyectada de s, denotada $seqBD|_s$, es un conjunto de sufijos de secuencias $seqBD$ prefijadas por s.

El algoritmo PrefixSpan toma como entrada una secuencia s y un soporte minimal σ y devuelve como resultado todas las secuencias frecuentes con el prefijo s, es decir, aquellas que aparecen al menos en σ secuencias de la base de secuencias $seqBD$. El Algoritmo 1 muestra un pseudo-código minimalista de PrefixSpan.

Algoritmo 1: PrefixSpan(α, $seqBD|_\alpha$)

Datos: un prefijo α y la base de secuencias proyectada $seqDB|_\alpha$
Resultado: una lista de patrones con prefijo α

1 Encontrar el elemento x tal que $soporte(seqDB|_\alpha(x)) \geq \sigma$;
2 Agregar x a α para extender el patrón secuencial αx y guardarlo ;
3 Construir la base αx-proyectada $(seqDB|_\alpha)|_x$ para cada αx y llamar al procedimiento a PrefixSpan($\alpha x, (seqDB|_\alpha)|_x$) ;

Los patrones secuenciales se construyen progresivamente durante una exploración en profundidad del espacio de búsqueda. Primero, todos los elementos frecuentes x se extraen de la base de datos proyectada $seqBD|_\alpha$. Cabe señalar que, en la primera llamada, $seqBD|_\alpha$ corresponde a la base de secuencias inicial $seqBD$, debido a que $\alpha = \{\}$.

Luego, para cada elemento x, el algoritmo extiende el patrón secuencial α con x. Dos extensiones son posibles en esta etapa: 1) agregando x al último conjunto de elementos de α, es decir (αx); o 2) insertando x después del último conjunto de elementos de α, es decir $(\alpha)(x)$ (la siguiente estampilla temporal). El soporte para estos dos patrones se calcula y solo se conservan aquellos que son frecuentes. Finalmente, para cada patrón secuencial, el algoritmo realiza otra proyección de la base de datos usando $seqBD|_\alpha|_x$ y extiende el patrón recursivamente invocando la función PrefixSpan nuevamente. El algoritmo se detiene cuando ya no se puede generar más proyecciones.

IV. Experimentación y Resultados

Este estudio se realizó en cinco ciudades del litoral Peruano: Tumbes, Piura, Lambayeque, La Libertad y Ancash. Para ello, se utilizaron diversas fuentes de datos. En esta sección se describen dichas fuentes y cómo ellas fueron tratadas para luego ser explotadas por el algoritmo PrefixSpan.

A. Bases de Datos

La base de datos utilizada en este estudio contiene tres juegos de datos, los cuales se describen a continuación.

1) Temperatura de la superficie del mar: La información de la temperatura de la superficie es obtenida por boyas perfiladoras ubicadas frente al litoral costero en la Zona Niño $1+2$. Estas boyas forman parte del programa internacional Argo[2]. Cada diez días las boyas descienden hasta los 2000 metros de profundidad para luego iniciar el ascenso a la superficie, midiendo en su camino principalmente la temperatura, la salinidad y la presión. Posteriormente, estos datos son enviados por satélite desde la superficie.

Para el presente estudio se seleccionó la información correspondiente a la superficie del mar recolectada por las boyas ubicadas en la Zona Niño $1+2$, en el periodo $01/02/2015$ al $30/06/2016$, obteniéndose los datos de 30 boyas. Las principales características que se obtienen por cada boya se muestran en el Cuadro II.

TABLA II
Principales características de los datos oceanográficos

Característica	Descripción
ARGOS_ID	Identificador de la boya
DATE (YYYY-MM-DDTHH:MI:SSZ)	Fecha de registro
Latitud	Coordenada Lambert
Longitud	Coordenada Lambert
Presión	Medida en Decibares
Temperatura	Medida en Grados Celcius
Salinidad	Medida en USP (Unidades Prácticas de Salinidad)

2) Variables meteorológicas en litoral Peruano: La información meteorológica del litoral Peruano se obtuvo de la NOAA (*National Oceanic and Atmospheric Administration*)[3] la cual es una agencia científica del Departamento de Comercio de los Estados Unidos. Los datos obtenidos son almacenados en 2 archivos: 1) uno con la información de las estaciones meteorológicas en el mundo, por ejemplo, la coordenadas Lambert de cada estación; y 2) otro con la información registrada por las estaciones meteorológicas, recuperadas en el período comprendido entre el $01/02/2015$ hasta $31/08/2016$ (ver Cuadro III).

[2]http://www.oceanografia.es/argo
[3]https://www7.ncdc.noaa.gov/CDO/country

3) Variables hidrológicas: La información hidrológica, que está asociada al caudal de los ríos, se obtiene de la web del Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI)[4]. Los datos fueron extraídos para el periodo $12/03/2015$ al $10/07/2016$. El Cuadro IV muestra las características de los datos hidrológicos.

Finalmente, estos tres juegos de datos son integrados, teniendo en cuenta que pertenezcan a la misma estampilla temporal (que hayan sucedido en el mismo tiempo) y a la misma zona (localidad) para su posterior pre-procesamiento.

B. Limpieza y Pre-tratamiento

En esta etapa, los datos son normalizados y se seleccionan aquellos que corresponden a las regiones de la costa norte del Perú: Tumbes, Piura, Lambayeque, La Libertad y Ancash. La información extraída está conformada por las características mostradas en el Cuadro V.

Posteriormente, los datos fueron discretizados en Terciles y en Quintiles utilizando la técnica de frecuencias iguales [16], el cual agrupa los datos en 3 o 5 intervalos conteniendo - aproximadamente - el mismo número de muestras. Los campos con valores no válidos, o que no tienen valor, se reemplazan por un valor fuera de rango: 999999. Posteriormente, dado que la información de las boyas se registran aproximadamente cada 10 días, ésta es agrupada por periodos en los cuales se tenga por lo menos la información de una boya en un determinada región. Para ello, se calcula el porcentaje de datos faltantes, agrupándolas para diferentes rangos de días.

En la Figura 1 se muestra el porcentaje de datos faltantes respecto a las boyas por cada región para diferentes periodos de tiempo, observándose que en el periodo de 18 días se tiene un menor porcentaje de datos faltantes.

Finalmente, se determina la cantidad de datos para un periodo de 18 días, obteniéndose 14 grupos de datos válidos es decir, se tiene información para todas las regiones en el mismo rango de tiempo.

Una vez que los datos fueron agrupados utilizando el promedio de los valores por región y en periodos de 18 días, se procede a definir dos escenarios.

[4]http://www.senamhi.gob.pe/?p=0320

TABLA III
ALGUNAS CARACTERÍSTICAS REGISTRADAS POR LAS ESTACIONES METEOROLÓGICAS

Nombre	Descripción
STN—	ID de la estación de la Fuerza Aérea
WBAN	Número NCDC WBAN (Weather Bureau Air Force Navy)
YEARMODA	Año Mes Día
TEMP	La temperatura media de rocío medida en décimas de grados Fahrenheit
DEWP	La media del punto de rocío medida en décimas de grados Fahrenheit
SLP	Presión media nivel del mar, medida en décimas de mili bares
STP	Presión media de la estación, medida en décimas de mili bares
VISIB	Visibilidad media del día, medida en décimas de milla
WDSP	Velocidad media del viento, medida en décimas de nudos
MXSDP	Velocidad máxima reportada del viento sostenida, medida en décimas de nudos
GUST	Ráfaga de viento máxima reportada, medida en décimas de nudos
MAX	Temperatura máxima registrada décimas de grados en Fahrenheit
MIN	Temperatura mínima registrada décimas de grados en Fahrenheit

TABLA IV
INFORMACIÓN HIDROLÓGICA DIARIA

Nombre	Descripción
Cuencas	Nombre de la cuenca
Estación Hidrométrica	Nombre de estación hidrométrica
Caudal	En metros cúbicos por segundo
Anomalía Hídrica	Variación de los caudales frente a valores históricos (en porcentaje)
Tendencia respecto al anterior	Ascendente, Leve Ascendente, Estable, Leve Descendiente, Descendente

TABLA V
CARACTERÍSTICAS SELECCIONADAS

	Característica
Mar	Temperatura
	Salinidad
Estación Meteorológica	Temperatura
	Punto de rocío
	Presión a nivel del mar
	Presión en estación
	Velocidad del viento
	Temperatura máxima
	Temperatura mínima
Estación Hidrológica	Caudal

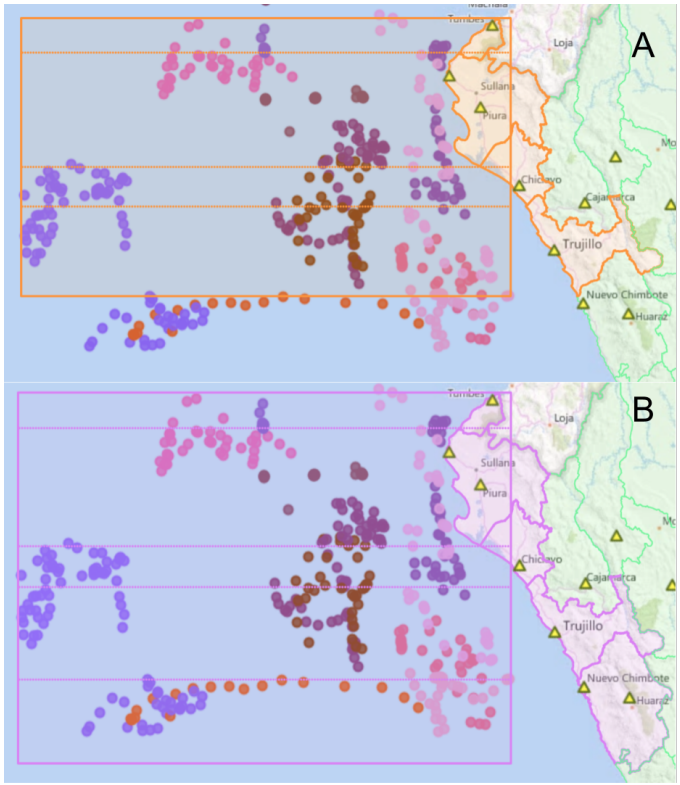

Fig. 1. Porcentaje de datos faltantes - Boyas.

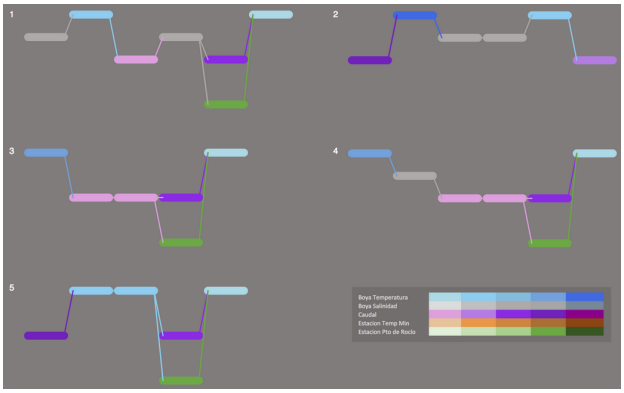

Fig. 2. Ilustración de escenarios. A) Escenario 1 B) Escenario 2.

C. Escenarios

Los escenarios están conformados por las regiones y su correspondiente proyección en el mar dentro de la Zona $1+2$ de El Niño.

El Escenario 1 (Figura 2A) toma en cuenta únicamente las regiones de la costa norte del Perú: Tumbes, Piura, Lambayeque y La Libertad las cuales se encuentran frente a la Zona $1+2$ de El Niño. El Escenario 2 (Figura 2B) incluye todas las regiones que se encuentran frente a la Zona $1+2$: Tumbes, Piura, Lambayeque, La Libertad y Ancash, siendo esta última una región que se puede considerar que no está localizada en la costa norte del Perú, pero sigue estando dentro de los límites de la Zona $1+2$ de El Niño.

Cabe mencionar que, en las Figuras 2A y 2B, los puntos ubicados en el mar representan las boyas en un determinado rango de tiempo y las líneas representan los límites de las zonas construidas.

D. Minería de Datos

Luego de definirse los escenarios, se construyen las secuencias. Cada secuencia corresponde a una región y ésta a su vez está conformada por *itemsets*, los cuales están formados por *items* que son los valores discretos de las características mencionadas en la Subsección IV-B.

Luego de la implementación de los escenarios se procede a la extracción de patrones secuenciales utilizando el algoritmo PrefixSpan implementado en la librería SPMF[5]. La extracción de patrones se realizó para diferentes valores de soporte minimal con el objetivo de que los patrones encontrados estén presentes en la mayoría de las regiones. Además se utilizó el algoritmo sobre datos con dos niveles de discretización para ver el efecto de ella en el proceso de extracción (ver Sección IV-B). El Cuadro VI muestra la cantidad de patrones obtenidos, la cantidad de memoria y el tiempo de procesamiento utilizados para el proceso de extracción, para cada escenario y para diferentes soportes minimales. Como se puede apreciar en este cuadro, si el soporte minimal decrece entonces el número de patrones extraídos aumenta, al igual que los recursos del sistema (*i.e.*, memoria y tiempo de proceso). Este comportamiento es característico cuando se utilizan algoritmos de extracción de patrones. En este mismo contexto, es importante recordar que el Escenario 1 está conformado por cuatro secuencias y el Escenario 2 por cinco secuencias. Si el soporte minimal $\sigma = 4$ para ambos escenarios, el algoritmo debe extraer aquellas sub-secuencias que aparecen en al menos en 4 zonas.

Una vez que los patrones fueron extraídos, se realizó un filtro para descartar las secuencias frecuentes con una gran cantidad de *itemsets* de tamaño 1, es decir, que contiene solo un *itemset* ya que no ofrecen información temporal semánticamente rica.

Los Cuadros VII y VIII muestran algunos ejemplos de patrones secuenciales, para el Escenario 1 y para el Escenario 2 respectivamente. Cada cuadro muestra 5 patrones secuenciales, que desde el punto de vista del experto del Instituto Geofísico del Perú, son los que aportan información interesante. En cada uno de los patrones secuenciales, los *itemsets* están separados por comas y los *items* por espacios. Por ejemplo, la primera secuencia frecuente del Cuadro VII se puede interpretar como: al principio se tiene un caudal alto (mayor ó igual a 111.528 m^3/s), luego la temperatura de la superficie del mar disminuye pasando de un valor alto (mayor a 26.23 °C) a un valor medio (24.578 °C a 26.23 °C), el caudal también disminuye, a un valor medio (43.145 m^3/s y 111.528 m^3/s). El resto de los patrones secuenciales para el Escenario 1 pueden ser interpretados de la misma manera.

De igual manera, en el Cuadro VIII se muestran algunos ejemplos de patrones secuenciales correspondientes al Escenario 2, que a juicio de expertos son relevantes. La secuencia 1 puede ser interpretada como: la secuencia inicia con una temperatura de la superficie del mar entre 24.089 °C y 25.121°C, y luego se produce un aumento moderado de caudal de 28.612 m^3/s a un valor entre 68.884 m^3/s y 161.76 m^3/s, la temperatura de la superficie del mar baja a un valor menor a 24.089 °C. La secuencia 3 empieza con la temperatura de la superficie del mar entre 26.105 °C y 26.769 °C, y luego se produce un aumento moderado de caudal de 28.612 m^3/s a un valor entre 68.884 m^3/s y 161.76 m^3/s junto con la aparición de lluvias moderadas (entre 69.809 y 74.111). Finalmente, la temperatura de la superficie del mar baja a un valor menor a 24.089 °C.

E. Visualización de Patrones Secuenciales

Los patrones secuenciales obtenidos son difíciles de interpretar por los expertos. Es por ello que es importante buscar una forma de mostrar los patrones de manera gráfica. El objetivo de la visualización es resumir los patrones y facilitar el aprendizaje a partir de ellos. Para esto, se implementó un prototipo llamado *ViSTPatterns Soft* el cual hace uso de la librería *javascript D3*[6] para la generación de las imágenes.

La idea detrás del prototipo de visualización es representar cada característica (por ejemplo, la temperatura) por un segmento de recta. Para representar el valor discreto que corresponde a cada característica utilizamos diferentes tonalidades de color, donde una mayor intensidad de color hace referencia

[5]http://www.philippe-fournier-viger.com/spmf/

[6]https://d3js.org/

TABLA VI
PRUEBAS Y RESULTADOS

Escenario	Niveles de discretización	Soporte mínimo	Número de secuencias	Memoria utilizada (MBytes)	Tiempo Total
1	**3**	**4**	**7502**	**262.31**	**19 min 47 seg**
1	5	4	552	32.56	1 min 48 seg
2	3	5	124	20.02	1 min 02 seg
2	**5**	**4**	**2201**	**54.67**	**4 min 45 seg**
2	5	5	29	10.91	0.26 seg

TABLA VII
ESCENARIO 1 - EJEMPLO DE PATRONES SECUENCIALES

Nro.	Patrón Secuencial	Soporte
1	caudal_>=111.528, boya-temp_>=26.23, boya-salinidad_35.003:35.129 boya-temp_24.578:26.23, caudal_43.145:111.528, boya-temp_<24.578	4
2	boya-temp_24.578:26.23, caudal_>=111.528, caudal_43.145:111.528 boya-temp_>=26.23, boya-temp_<24.578, boya-temp_<24.578	4
3	boya-temp_>=26.23, est_temp_min_69.668:73.379, boya-temp_<24.578, caudal_43.145:111.528 boya-temp_>=26.23, boya-temp_<24.578, boya-temp_<24.578	4
4	caudal_>=111.528, caudal_>=111.528 boya-salinidad_35.003:35.129, boya-temp_>=26.23, boya-temp_24.578:26.23, caudal_43.145:111.528	4
5	caudal_>=111.528, caudal_>=111.528 boya-temp_>=26.23, boya-temp_24.578:26.23, caudal_43.145:111.528 boya-salinidad_>=35.129, boya-temp_<24.578	4

TABLA VIII
ESCENARIO 2 - EJEMPLO DE PATRONES SECUENCIALES

Nro.	Patrón Secuencial	Soporte
1	boya-salinidad_34.928:35.05, boya-temp_24.089:25.121, caudal_<28.612, boya-salinidad_35.05:35.142, caudal_68.884:161.761 estac-pto-rocio_69.809:74.111, boya-temp_<24.089	4
2	caudal_161.761:500346.845, boya-temp_26.105:26.769, boya-salinidad_34.928:35.05, boya-salinidad_34.928:35.05, boya-temp_24.089:25.121, caudal_28.612:68.884	4
3	boya-temp_26.105:26.769, caudal_<28.612, caudal_<28.612, caudal_68.884:161.761 estac-pto-rocio_69.809:74.111, boya-temp_<24.089	4
4	boya-temp_26.105:26.769, boya-salinidad_34.928:35.05, caudal_<28.612, caudal_<28.612, caudal_68.884:161.761 estac-pto-rocio_69.809:74.111, boya-temp_<24.089	4
5	caudal_161.761:500346.845, boya-temp_24.089:25.121, boya-temp_24.089:25.121, caudal_68.884:161.761 estac-pto-rocio_69.809:74.111, boya-temp_<24.089	4

a un mayor valor de una determinada característica. Estas variaciones de la intensidad de color serán mostradas en una leyenda para mejorar su comprensión. Adicionalmente, la aplicación permite una mayor información de las características al colocar el puntero del ratón sobre éstas.

La lectura de los gráficos se debe realizar de izquierda a derecha, donde cada segmento corresponde a un periodo de tiempo. Si se tienen dos o más segmentos en paralelo conectados a un mismo punto de origen implica que las características se presentaron en un mismo periodo de tiempo. Cabe recalcar que los gráficos muestran la evolución temporal de un conjunto de características que están asociadas al fenómeno El Niño. Esta evolución representa un patrón secuencial, donde el espacio entre cada *itemsets* es de 18 días por lo menos.

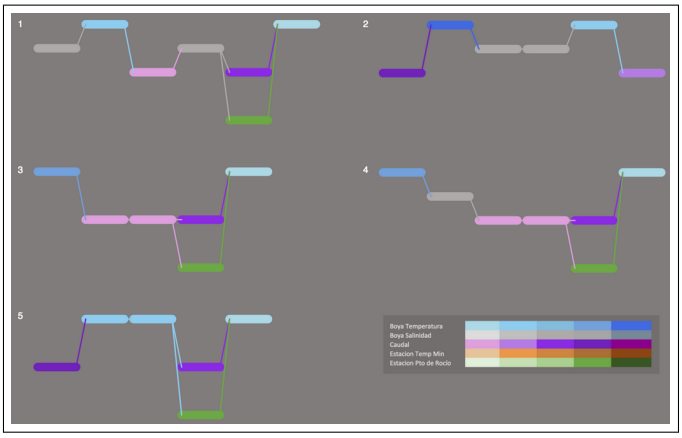

Fig. 3. Patrones secuenciales correspondientes al Escenario 2.

En la Figura 3 se muestran los gráficos correspondientes a los patrones secuenciales obtenidos a partir del Cuadro VIII pertenecientes al Escenario 2. A diferencia del Escenario 1, los resultados encontrados para este escenario tiene patrones más ricos semánticamente. En los gráficos 1, 3 y 4, correspondientes a las secuencias 1, 3 y 4, se observa una disminución de la temperatura de la superficie del mar mientras el valor del nivel de caudal aumenta de manera considerable, junto con la aparición de lluvias en el mismo período de tiempo (quinto *itemset*). A diferencia, en los gráficos 2 y 5, se tiene que durante la disminución de la temperatura de la superficie del mar el valor del nivel de caudal también disminuye, pero de manera menos perceptible que en los tres casos anteriores.

A partir de los gráficos que representan los patrones secuenciales 1 y 3 (*c.f.*, Figura 3), se puede crear la alerta siguiente: cuando la temperatura del mar se encuentre entre 26.1 y 26.7 °C, el caudal de los ríos aumentará en cuatro veces, acompañado de lluvias moderadas, todo ello en al menos los siguientes 54 días (cada itemset representa 18 días). Dicho de otro modo, la representación visual de los patrones puede servir para predecir eventos atípicos. En el ejemplo anterior, la aparición de un segmento celeste (temperatura del mar moderada) seguida de la aparición de un segmento rosa (caudal bajo de los ríos) producirá un aumento en el caudal de los ríos (segmento morado) y lluvias moderadas (segmento verde).

V. CONCLUSIONES Y TRABAJOS FUTUROS

Este artículo describe un proceso que permite la extracción y visualización de patrones secuenciales a partir de un conjunto de datos heterogéneos asociados al fenómeno El Niño. Una

secuencia se convierte en patrón si su aparición es frecuente. Entonces, un patrón representa un comportamiento típico, el cual aparece recurrentemente en las zonas de estudio. Los resultados de este estudio muestran que es posible medir el impacto de la temperatura de la superficie del mar en las variables meteorológicas registradas en la costa norte del litoral peruano. Además, permite estimar el tiempo en que un evento puede impactar en otro.

Con respecto a los trabajos futuros, se puede estudiar también el caso de La Niña ya que ambos son partes del mismo fenómeno cíclico y así tener una visión global del problema. Por otro lado, debido a la gran cantidad de patrones secuenciales obtenidos se revisarán algunas técnicas de selección automática de patrones. Finalmente, otro trabajo futuro, es el añadir otras características como la información de migración de las aves, la cantidad de peces, entre otros.

REFERENCIAS

[1] B Dewitte, K Takahashi, K Goubanova, A Montecinos, K Mosquera, S Illig, I Montes, A Paulmier, V Garçon, S Purca, et al. Las diversas facetas de el niño y sus efectos en la costa del perú. *Montes*, 1:3, 2014.

[2] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Introduction to data mining. 2006.

[3] CT Dhanya and D Nagesh Kumar. Data mining for evolving fuzzy association rules for predicting monsoon rainfall of india. *Journal of Intelligent Systems*, 18(3):193–210, 2009.

[4] Kabir Rasouli, William W. Hsieh, and Alex J. Cannon. Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology*, 414-415:284–293, 2012.

[5] Ajay Kalra, William P. Miller, Kenneth W. Lamb, Sajjad Ahmad, and Thomas Piechota. Using large-scale climatic patterns for improving long lead time streamflow forecasts for Gunnison and San Juan River Basins. *Hydrological Processes*, 27(11):1543–1559, 2013.

[6] Poulomi Ganguli and M Janga Reddy. Ensemble prediction of regional droughts using climate inputs and the svm–copula approach. *Hydrological Processes*, 28(19):4989–5009, 2014.

[7] Jaya Kawale, Michael Steinbach, and Vipin Kumar. Discovering Dynamic Dipoles in Climate Data. *Society for Industrial and Applied Mathematics. Proceedings of the SIAM International Conference on Data Mining*, (Dmi):107–118, 2011.

[8] Jaya Kawale, Stefan Liess, Arjun Kumar, Michael Steinbach, Peter Snyder, Vipin Kumar, Auroop R. Ganguly, Nagiza F. Samatova, and Fredrick Semazzi. A graph-based approach to find teleconnections in climate data. *Statistical Analysis and Data Mining*, 6(3):158–179, 2013.

[9] J. D. Velasquez, I. Dyner, and C. J. Franco. Modeling the effect of macroclimatic events on river inflows in the Colombian electricity market. *IEEE Latin America Transactions*, 14(10):4287–4292, October 2016.

[10] G. Caicedo, H. Rudnick, and E. Sauma. Auction mechanisms for long-term electricity contracts: Application to the colombian market. *IEEE Latin America Transactions*, 12(4):609–617, June 2014.

[11] H. Janicke, M. Bottinger, U. Mikolajewicz, and G. Scheuermann. Visual Exploration of Climate Variability Changes Using Wavelet Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1375–1382, 2009.

[12] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, ICDE '95, pages 3–14, Washington, DC, USA, 1995. IEEE Computer Society.

[13] Agustin Guevara-Cogorno, Claude Flamand, and Hugo Alatrista-Salas. Copper - constraint optimized prefixspan for epidemiological research. *Procedia Computer Science*, 63:433 – 438, 2015. EUSPN 2015/ ICTH-2015.

[14] Sunita Mahajan, Prajakta Pawar, and Alpa Reshamwala. Performance analysis of sequential pattern mining algorithms on large dense datasets. *International Journal of Application or Innovation in Engineering & Management*, 3(2):345–351, February 2014.

[15] J Pei, J Han, Mortazavi B Asl, H Pinto, Q Chen, U Dayal, and M C Hsu. PrefixSpan Mining Sequential Patterns Efficiently by Prefix Projected Pattern Growth. In *Proc.17th Int'l Conf. on Data Eng.*, pages 215–226, 2001.

[16] Marc Boulle. Optimal bin number for equal frequency discretizations in supervized learning. *Intell. Data Anal.*, 9(2):175–188, March 2005.

Oscar Días Barriga es Magíster en Informática con mención en Ciencias de la Computación de la Pontificia Universidad Católica del Perú, donde también culminó sus estudios de Ingeniería Electrónica. Participo en múltiples proyectos como el desarrollo de una aplicación para la computación en grid llamada "Legión Framework" la cual aprovecha la capacidad de procesamiento de las computadoras de los laboratorios de un Campus Universitario. Sus áreas de interés incluyen Inteligencia Artificial, Patrones Secuenciales, y Reconocimiento de patrones.

Miguel Nunez-del-Prado doctor en informática por la Universidad de Toulouse. Obtuvo este título por su trabajo sobre ataques de inferencia en datos geolocalizados y su impacto en la privacidad de los usuarios en el LAAS-CNRS Francia. Es Ingeniero en Computación, Redes y Telecomunicaciones. Tiene dos maestrías, una en Informática y Telecomunicaciones y otra en Gestión estratégica de la Innovación. Trabajó como científico de datos en el Grupo INTERSEC (París, Francia).

Hugo Alatrista-Salas es doctor en Ciencias de la Computación de la Universidad de Montpellier en Francia. El tema de investigación con el cual obtuvo su título está relacionado con la Minería de Datos espacio-temporal con aplicación en el medio ambiente y la salud pública. Además tiene una maestría en Calculabilidad, Algorítmica, Seguridad y Administración de Redes de la misma Universidad. Actualmente, es profesor investigador en la Universidad del Pacífico y vice decano del programa de Ingeniería de la Información.