Pontificia Universidad Católica del Perú
Escuela de Postgrado

DISEÑO DE UN PROCESO COMPUTACIONAL BASADO EN TÉCNICAS DE MINERÍA DE DATOS PARA EL ANÁLISIS DEL FENÓMENO DE "EL NIÑO"

Tesis para optar el grado de Magíster en Informática con mención en Ciencias de la Computación, que presenta

Oscar Antonio Díaz Barriga

Dirigido por

Dr. HUGO ALATRISTA SALAS

Jurado

Dr. HÉCTOR ANDRÉS MELGAR SASIETA
Dr. HUGO ALATRISTA SALAS
Dr. CÉSAR ARMANDO BELTRÁN CASTAÑÓN

San Miguel, 2017

Resumen

El Perú es afectado recurrentemente por el fenómeno El Niño, el cual es un fenómeno climático que consiste en el aumento de la temperatura del mar en el Pacífico Ecuatorial. Este 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. En la actualidad mediante un juicio experto se analizan las diversas fuentes de datos heterogéneas para poder encontrar posibles correlaciones útiles entre ellos. En el presente trabajo se propone un proceso computacional basado en técnicas de minería de datos que permita determinar la existencia de correlaciones espacio-temporales en relación a la temperatura superficial del mar y las variables meteorológicas pertenecientes a las regiones de la costa norte del Perú, en el periodo 2015 al 2016, último intervalo de tiempo en el que se presentó El Niño. Para esto se utiliza una metodología basada en KDD (Knowledge Discovery in Database), la cual está conformada por una serie de pasos como: la recolección de diferentes fuentes de datos, la integración en una base de datos explotable, limpieza y pretratamiento de los datos, creación de escenarios que permitan validar las posibles correlaciones, extracción de patrones mediante la librería SPMF y finalmente una propuesta de visualización, de los patrones encontrados, que permita comprender mejor el fenómeno. Los resultados obtenidos muestran la existencia de correlaciones espacio-temporales en las regiones del norte del Perú principalmente entre la temperatura de la superficie del mar y el caudal de los ríos de la costa, siendo estas correlaciones validadas por un experto miembro del IGP.

Abstract

Peru is recurrently affected by the El Niño phenomenon, which is a climatic phenomenon that consists in the increase of sea temperature in the Equatorial Pacific. This in turn forms part of the ENSO (El Niño - Southern Oscillation) that has a fluctuation period of 2 to 7 years, with a warm phase known as El Niño and a cold phase, La Niña. At present, through an expert judgment, the various sources of heterogeneous data are analyzed in order to find possible useful correlations between them. The present work proposes a computational process based on data mining techniques to determine the existence of temporal space correlations in relation to sea surface temperature and meteorological variables pertaining to the regions of the north coast of Peru, in the Period 2015 to 2016, the last time interval in which El Niño was presented. For this, a methodology based on KDD (Knowledge Discovery in Database) is used, which is conformed by a series of steps like: the collection of different sources of data, the integration in a database explotable, cleaning and pretreatment of the data, creation of scenarios that allow to validate the possible correlations, extraction of patterns through the SPMF library and finally a proposal of visualization, of the patterns found, that allows a better understanding of the phenomenon. The results obtained show the existence of spatiotemporal correlations in the northern regions of Peru, mainly between sea surface temperature and coastal river flow, these correlations being validated by an expert member of the IGP.

Dedicado a mis padres.

Agradecimientos

A mis padres y mi hermana Marita que siempre me apoyaron y que desde el cielo me protegen.

A los profesores y compañeros de la Escuela de Posgrado, gracias por la formación y la experiencia recibida.

A mi asesor, el Dr. Hugo Alatrista, por su apoyo y motivación constante.

A todas las personas que de una forma u otra colaboraron para la realización del presente trabajo.

Índice general

Índice de tablas

Índice de figuras

CAPÍTULO 1

Generalidades y Motivación

1.1. Introducción

El Niño, es un fenómeno climático que consiste en el aumento de la temperatura del mar en el Pacífico Ecuatorial. Este forma parte del ENSO (El Niño - Oscilación del Sur) el cual tiene un periodo de fluctuación de 2 a 7 años (Dewitte y cols., 2014), con una fase cálida conocida como El Niño y una fase fría, La Niña.[1]

Uno de los principales indicadores de la presencia de El Niño es la variación de la temperatura de la superficie del mar (TSM).

En la Figura 1.1 se observa que durante la presencia de El Niño, baja la fuerza de los vientos alisios que van de este a oeste lo que reduce el afloramiento de la corriente fría submarina en dirección a la costa (de este a oeste).

El monitoreo del ENSO se realiza principalmente en 4 regiones del Pacífico Ecuatorial conocidas como: Niño 4; Niño 3.4, de longitud 120° W hasta 170° W; Niño 3; y Niño 1+2, de longitud 80° W hasta 90° W y latitud° a 10° S.

En la Figura 1.2 se muestran las regiones anteriormente mencionadas, donde se observa que la región Niño 1+2 se encuentra frente a la costa norte del Perú.

Se considera que, las variaciones de TSM que son iguales o superiores a 0,5 °C (0,9 °F) en la región Niño 3.4, son indicadores de la presencia del fenómeno El Niño.

En el Perú, existe un Comité multisectorial Encargado del Estudio Nacional del Fenómeno El Niño (ENFEN)[2] conformado por:

- Autoridad Nacional del Agua (ANA)
- Dirección de Hidrografía y Navegación de la Marina de Guerra del Perú (DHN)
- Instituto de Defensa Civil (INDECI)
- Instituto del Mar del Perú (IMARPE)
- Instituto Geofísico del Perú (IGP)
- Servicio Nacional de Meteorología e Hidrología (SENAMHI)

[1] http://www.imarpe.pe/imarpe/index.php?id_seccion=I01780100000000000000000

[2] http://www.met.igp.gob.pe/variabclim/enfen/

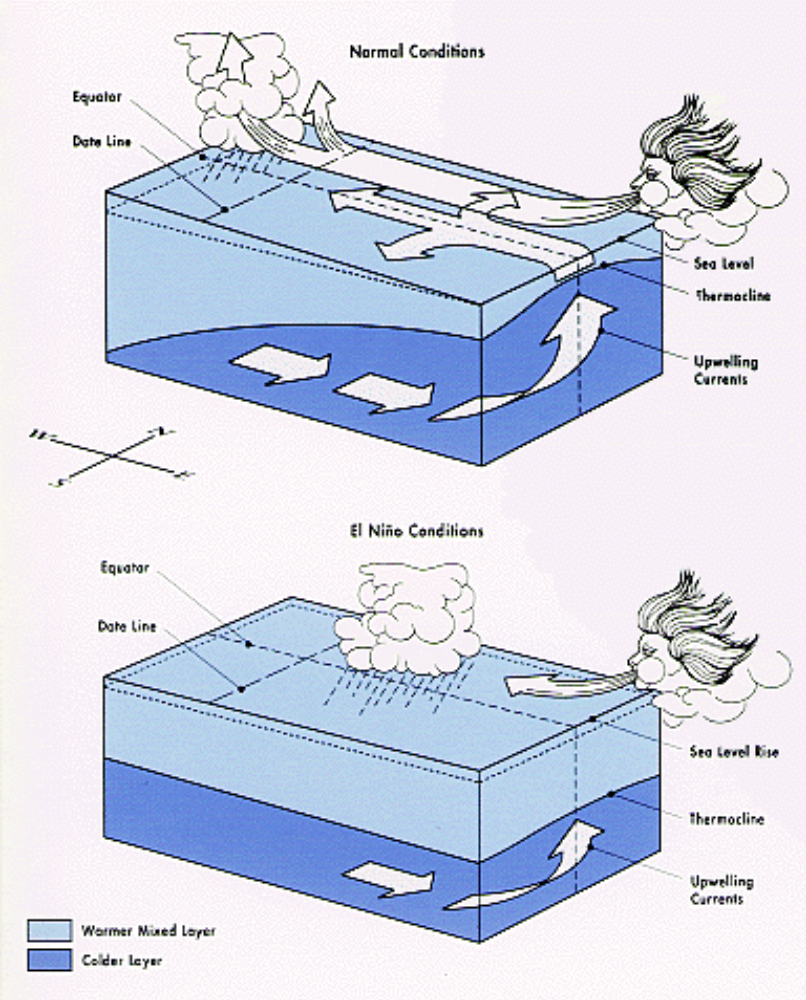

Figura 1.1: *El Niño - Oscilación del Sur (ENSO).*
Fuente: National Oceanic and Atmospheric Administration (NOAA).[3]

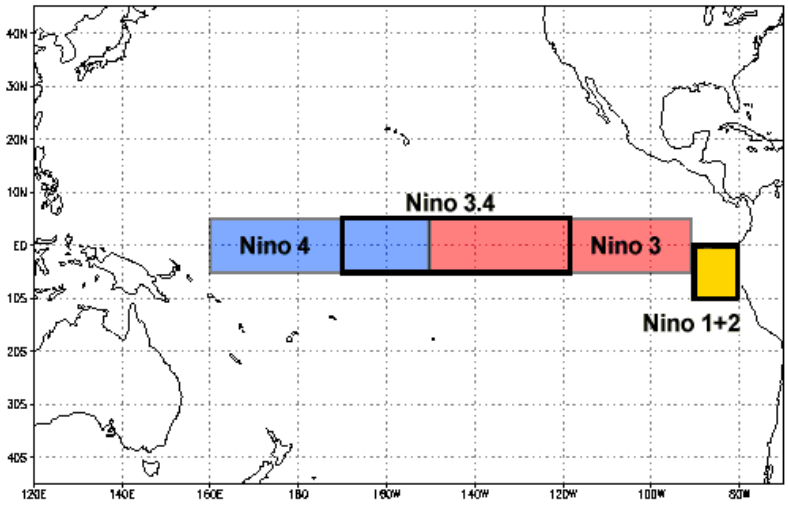

Figura 1.2: *Temperaturas de la superficie del mar en el Pacífico ecuatorial (traducción propia).*
Fuente: National Oceanic and Atmospheric Administration (NOAA).[4]

[3] https://www.ncdc.noaa.gov/teleconnections/enso/enso-tech.php.
[4] https://www.ncdc.noaa.gov/teleconnections/enso/indicators/sst.php.

Estos representantes se reúnen de manera regular para analizar mediante juicio experto la información, cada uno en su área de conocimiento, y actualizar las diversas fuentes de datos con información de: las condiciones meteorológicas, oceanográficas, biológico - pesqueras, hidrológicas y modelos climáticos.

El tener que analizar diversas fuentes de datos heterogéneos, pudiendo ser estos archivos: csv, txt, imágenes, pdf, etc; mediante un juicio experto, representa un problema que exige a cada una de las personas entender los diferentes tipos de datos y así poder encontrar posibles correlaciones útiles entre ellos.

En la actualidad, existen técnicas como las proporcionadas por la minería de datos que permiten procesar grandes fuentes de datos heterogéneos, con el objetivo de encontrar patrones que representen correlaciones entre los datos.(Tan, Steinbach, y Kumar, 2006)

La presente tesis busca desarrollar un proceso computacional para el análisis y determinación de la existencia de correlaciones espacio-temporales. Para ello, se hará uso de fuentes de datos de diferentes tipo como: txt, CSV (*Comma-Separated Values*), PDF (*Portable Document Format*); con información de variables meteorológicas y oceanográficas de la costa norte del Perú, durante el periodo 2015 al 2016, teniendo en cuenta que el último intervalo de tiempo en el que se presentó el fenómeno El Niño fue del 18 Marzo 2015 al 21 Abril 2016 (ENFEN, 2015, 2016). Las correlaciones que se descubran serán validadas por un miembro del IGP y así poder ayudar a los investigadores del fenómeno El Niño en su análisis y mejor comprensión de éste.

1.2. Problemática

El Niño es un fenómeno que se presenta de forma recurrente con una periodicidad de 2 a 7 años (Dewitte y cols., 2014) , con el aumento de la temperatura de la superficie del mar, ocasionando cambios en el clima en el mundo. En la costa norte del Perú frente a la zona Niño 1+2 los efectos del fenómeno causan pérdidas materiales como humanas. Por ello, uno de los principales problemas, es la dificultad de poder medir el impacto que produce la temperatura en la superficie del mar (TSM) en las variables meteorológicas registradas en el litoral peruano en el contexto del fenómeno de El Niño. Teniendo en cuenta lo anterior, se tiene que algunas de las principales causas de este problema son:

- La mayoría de estudios se realizan en la zona 3.4 y pocos en la Zona Niño 1+2 la cual se encuentra frente a la costa norte del litoral peruano.

- La falta de datos explotables que representen el fenómeno. Los datos tales como: la temperatura de la superficie del mar, las lluvias, el caudal de los ríos, etc; se encuentran y representan de forma variada e independiente lo que hace difícil el poder encontrar correlaciones espacio-temporales entre los datos.

- Falta de modelos que tomen en cuenta la dimensión espacial de grano fino, no se tiene una forma de poder representar el efecto de TSM en determinada región de la costa norte del Perú.

- No se cuenta con indicadores que permitan representar la dinámica espacio-temporal del fenómeno El Niño.

Como efectos de las principales causas mencionadas se tienen:

- El poco conocimiento del fenómeno en la Zona El Niño 1+2 en comparación a las otras zonas (4, 3.4 y 3) por ejemplo, no permite determinar de los efectos en una determinada ubicación de la costa norte del Perú.
- No se cuentan con datos estructurados que representen del fenómeno.
- Como se mencionó anteriormente, no se puede prever el efecto o alcance de El niño en una determinada zona del litoral peruano.
- Dificultad en la comprensión del fenómeno de El Niño dado que el juicio experto no permite analizar los indicadores de manera conjunta. .

En la presente tesis se busca implementar un proceso computacional que permita la construcción de indicadores con dinámicas espacio-temporales, a partir de fuentes de datos heterogéneas, que permitan medir el impacto de la temperatura del mar en las variables meteorológicas registradas en el litoral peruano, principalmente en las regiones de la costa norte del Perú las cuales se encuentran frente a la zona 1+2, en el contexto del fenómeno de El Niño, mediante el uso de técnicas de minería de datos.

1.3. Objetivos

1.3.1. Objetivo General

Implementar un proceso computacional basado en técnicas de minería de datos que permita medir el impacto de la temperatura superficial del mar en las variables meteorológicas registradas en el litoral norte del Perú en el contexto del Fenómeno de El Niño.

1.3.2. Objetivos Específicos

Con el fin de lograr el objetivo general se plantean los siguientes objetivos específicos:

- Construir una base de datos explotables que representen el fenómeno estudiado.
- Definir escenarios que representen la dinámica espacio-temporal del fenómeno estudiado a fin de medir las correlaciones entre los datos que representan los eventos suscitados en el mar y sus posibles efectos en tierra.
- Identificar correlaciones espacio-temporales para cada uno de los escenarios antes identificados, a fin de facilitar la comprensión del fenómeno.
- Restituir y Visualizar los patrones obtenidos que permitan la mejor comprensión del fenómeno.

En la Tabla 1.1 se tienen los Objetivos y los resultados finales.

Tabla 1.1: *Objetivos y Resultados*

OBJETIVO GENERAL	**RESULTADOS FINALES**
Implementar un proceso computacional basado en técnicas de minería de datos que permita medir el impacto de TSM en las variables meteorológicas registradas en el litoral norte del Perú en el contexto del fenómeno de El Niño	En qué variables meteorológicas impactan la temperatura de la superficie del mar
OBJETIVOS ESPECÍFICOS (COMPONENTES)	**RESULTADOS INTERMEDIOS**
Construir una base de datos explotables que representen el fenómeno estudiado.	Una base de datos espacio-temporal
Definir escenarios que representen la dinámica espacio-temporal del fenómeno estudiado.	Dos escenarios posibles
Identificar correlaciones espacio temporales entre los datos asociados y zonas homogéneas.	Una lista de patrones secuenciales
Restituir y Visualizar los patrones obtenidos que permitan la mejor comprensión del fenómeno.	Prototipo de visualización de patrones secuenciales

En el siguiente capítulo se abordarán temas relacionados con el estado del arte y el marco conceptual.

CAPÍTULO 2

Estado del Arte y Marco Conceptual

2.1. Estado del Arte

Se realizó una búsqueda en diferentes bases de datos bibliográficas (Scopus y IEEExplore) sobre investigaciones relacionadas a "El Niño" donde se utilizaron técnicas de minería de datos y búsqueda de patrones. Se encontró que la mayoría de las investigaciones están enfocadas en la relación del El Niño-ENSO con otros fenómenos, como el aumento o falta de lluvias; o en la validación de técnicas de machine learning. Considerando lo anteriormente mencionado, las investigaciones relacionadas a El Niño se pueden clasificar en 2 tipos:

- Las que estudian una relación de El Niño con otros fenómenos naturales.
- Las que usan los datos de El Niño para validar nuevos modelos.

2.1.1. Relación de El Niño con otros fenómenos naturales

En este tipo de trabajos, se busca una relación de El Niño con otros fenómenos naturales: lluvias, sequías, incendios, etc., utilizando diferentes técnicas de minería de datos. A continuación, se tienen 3 de los artículos más citados:

- En un estudio, se utiliza el método de máquina de soporte vectorial (SVM) para la búsqueda de patrones 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 (PDO), la Oscilación del Atlántico Norte (NAO), la Oscilación Multidecadal del Atlántico (AMO), 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. (Kalra, Miller, Lamb, Ahmad, y Piechota, 2013)
- La extracción de 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. (Dhanya y Kumar, 2009)

- Para el estudio de las teleconexiones climáticas con las sequías meteorológicas, se desarrollaron modelos de predicción utilizando máquinas de vectores de soporte (SVM) y copulas[5] enfocada sobre la región Rajasthan Occidental (India). Donde se estudió como un análisis del clima a gran escala se relaciona con diferentes índices climáticos como El Niño - Oscilación del Sur, etc. (Ganguli y Reddy, 2014)

2.1.2. Validar modelos

También se encontró que en otros casos donde, se utilizan los datos de El Niño validar modelos (nuevos o ya existentes). Algunos de los artículos más citados son los siguientes:

- Se desea comparar tres métodos de aprendizaje automático: Red neuronal bayesiana (BNN), apoyo vector de regresión (RVS) y el proceso de Gauss (GP) con la regresión lineal múltiple (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 (AO) y la Oscilación del Atlántico Norte (NAO). (Rasouli, Hsieh, y Cannon, 2012)

- Se utiliza 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.(Janicke, Bottinger, Mikolajewicz, y Scheuermann, 2009)

- Uno de los artículos, se enfoca 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. (Kawale y cols., 2013; Kawale, Steinbach, y Kumar, 2011)

A continuación, se presenta la Tabla 2.1 donde se tiene una lista de los documentos encontrados. Adicionalmente, se tienen columnas que indican el tipo de relación con el fenómeno de El Niño y los métodos utilizados.

[5]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.(**?**)

Tabla 2.1: *Documentos relacionados al fenómeno El Niño*

Buscador	Título	Autor	Año	Citaciones	Tipo	Métodos utilizados
Scopus	Using large-scale climatic patterns for improving long lead time streamflow forecasts for Gunnison and San Juan River Basins	Kalra, A., Miller, W.P., Lamb, K.W., Ahmad, S., Piechota, T	2013	24	Relación de El Niño con otros fenómenos Naturales	Máquina de soporte vectorial (SVM)
Scopus	Data mining for evolving fuzzy association rules for predicting monsoon rainfall of India	Dhanya, C.T., Kumar, D.N.	2009	5	Relación de El Niño con otros fenómenos Naturales	FP-Growth
Scopus	Ensemble prediction of regional droughts using climate inputs and the SVM-copula approach	Ganguli, P., Reddy, M. J.	2014	3	Relación de El Niño con otros fenómenos Naturales	Máquina de soporte vectorial - Copula
IEEEXplore	Spatio-Temporal Analysis of the Relationship between South American Precipitation Extremes and the El Niño Southern Oscillation	Wu, E., Chawla, S.	2007	0	Relación de El Niño con otros fenómenos Naturales	Teoria de valor extremo y Estadístico Moran's I
Scopus	Short lead-time streamflow forecasting by machine learning methods, with climate variability incorporated	Rasouli, K., Hsieh, W.W., Cannon, A.J.	2010	0	Relación de El Niño con otros fenómenos Naturales	Support vector regression (SVR), Gaussian process (GP), Bayesian neural network (BNN), Multiple linear regression (MLR)

Buscador	Título	Autor	Año	Citaciones	Tipo	Métodos utilizados
Scopus	Long range forecast of streamflow using support vector machine	She, N., Basketfield, D.	2005	0	Relación de El Niño con otros fenómenos Naturales	Máquina de soporte vectorial (SVM)
IEEExplore	Multisensor analysis of teleconnection signals in relation to terrestrial precipitation and forest greenness in North and Central America	Chang, N. B. , Imen, S., Mullen, L. Chen, C. F.; Valdez, M., Yang, J.	2014	0	Relación de El Niño con otros fenómenos Naturales	Coeficiente de correlación de Pearson
Scopus	Daily streamflow forecasting by machine learning methods with weather and climate inputs	Rasouli, K., Hsieh, W.W., Cannon, A.J.	2012	17	Validar modelos	Support vector regression (SVR), Gaussian process (GP), Bayesian neural network (BNN), Multiple linear regression (MLR)
IEEExplore	Visual Exploration of Climate Variability Changes Using Wavelet Analysis	Janicke, H., Bottinger, M., Mikolajewicz, U. ; Scheuermann, G.	2009	8	Validar modelos	Análisis Wavelet
Scopus	Discovering dynamic dipoles in climate data	Kawale, J., Steinbach, M., Kumar, V.	2011	6	Validar modelos	Vecino más cercano
Scopus	A graph-based approach to find teleconnections in climate data	Kawale, J., Liess, S., Kumar A., Steinbach, M., Snyde, P, Kumar, V., Ganguly, A. Samatova, N., Semazzi, F.	2013	3	Validar modelos	Vecino más cercano recíproco compartido (SRNN)

Buscador	Título	Autor	Año	Citaciones	Tipo	Métodos utilizados
Scopus	An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland	Deo, R.C., Şahin, M.	2016	0	Validar modelos	Aprendizaje de máquina extremo
IEEExplore	Weather forecasting using deep learning techniques	Salman, A. G., Kanigoro, B., Heryadi, Y.	2015	0	Validar modelos	Aprendizaje profundo

De la tabla anterior, se observa que en los casos en los que se busca alguna relación de El Niño con otro fenómeno natural, se utilizaron métodos de clasificación y en la mayoría de estos Máquina de Soporte Vectorial (SVM); y para el caso que se requiere realizar alguna validación de modelos, el método utilizado es muy variable.

2.2. Marco Conceptual

En la presente sección se mencionarán algunos conceptos que se tendrán en cuenta durante el desarrollo de la tesis.

2.2.1. Datos espaciales

Son datos que contienen información espacial, como su ubicación geográfica. Por ejemplo, la información que proporciona una boya en el mar, está conformada por la temperatura de la superficie y las coordenadas Lambert que muestran su posición durante la captura de datos. Los datos espaciales o geográficos a menudo exhiben propiedades de dependencia espacial y heterogeneidad espacial. La dependencia espacial es la tendencia en que las observaciones que son más próximas en el espacio geográfico tienden a exhibir mayores grados de similitud o disimilitud (en función de los fenómenos). La proximidad se puede definir en términos muy generales, incluyendo la distancia, la dirección y/o de la topología. (Miller, 2008)

2.2.2. Datos espacio-temporales

Son datos espaciales que también contienen la información temporal. Por ejemplo: la temperatura o velocidad del viento capturada en una estación meteorológica en un lugar del litoral peruano en una fecha determinada.

2.2.3. Patrones secuenciales

Es una técnica de minería de datos, que ayuda a encontrar secuencias que se repiten frecuentemente y a su vez descubrir posibles correlaciones entre los datos. En los patrones secuenciales se tiene en cuenta una serie de conceptos como:

- Item, un valor literal
- Itemset, conjunto de valores literales
- Secuencia, lista ordenada de itemset
- Soporte, valor numérico que se obtiene al dividir el número de secuencias en la que se encuentran los patrones secuenciales frecuentes entre la cantidad de secuencias totales.

Por ejemplo: en el caso de El Niño se puede tener una serie de secuencias como las mostradas en la Tabla 2.2.

Tabla 2.2: *Patrones Secuenciales*

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)

Donde se tienen los items:

Ta: Temperatura alta.
Tb: Temperatura baja.
Pa: Presión alta.
Pb: Presión baja.
Ca: Caudal alto.
Cb: Caudal bajo.
Va: Velocidad del viento alta.
Vb: Velocidad del viento baja.

Donde:

- Un Itemset es (Ta, Pb, Cb, Va).
- La Secuencia, formada por 3 itemsets que representando 3 tiempos diferentes:
  - Tiempo 1 : (Ta, Pb, Cb, Va)
  - Tiempo 2 : (Ta, Pa, Cb)
  - Tiempo 3 : (Ta, Pb, Ca)
- La zona, representa la región geográfica para un determinado escenario de estudio.
- Una secuencia frecuente con soporte igual 2/3 es: (Ta, Pb, Cb)(Tb)(Ta), dado que esta secuencia está presente en la Zona 1 y 2 de un total de 3 Zonas.

En la minería de patrones secuenciales existen muchas clases de algoritmos, los basados en: el esquema Apriori, el crecimiento de patrones, base de datos en formato vertical, etc. Cada uno de estos algoritmos busca obtener las mayor cantidad de secuencias en

el menor tiempo y con el menor uso de recursos para esto los algoritmos basados en el Crecimiento de patrones, intentan reducir el número de secuencias candidatos en una larga secuencia de base de datos, entre ellos se tienen: PSP (Parameter Space Partition), FreeSPAN (Frequent Pattern-Projected Sequential Pattern Mining), PrefixSpan (Prefix-projected Sequential pattern mining), LAPIN (LAst Position INduction), PRISM (PRIme-Encoding Based Sequence Mining), etc.(Yadira, 2013)

2.2.4. Escenarios

Los escenarios, son la forma en que se agrupan los datos con el objetivo de crear secuencias, ver la Tabla 2.2, permitiendo definir hipótesis que ayuden a determinar posibles correlaciones entre ellos. Por ejemplo, un posible escenario puede estar definido por los datos agrupados en función de las regiones de la costa norte del país y su proyección en el mar en la Zona 1+2, ver Figura 2.1; mientras que otro escenario puede estar definido por los datos agrupados en función de las regiones que forman parte de la proyección de la región Zona 1+2 en la costa del país, ver Figura 2.2. En ambas figuras la región de color corresponde la Zona 1+2

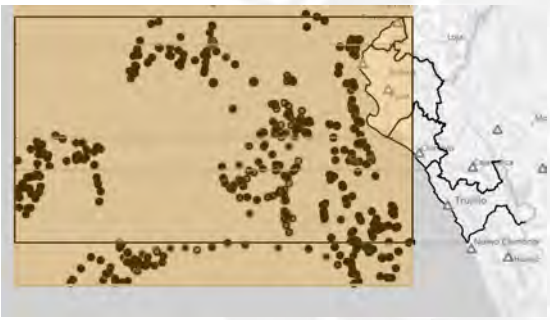

Figura 2.1: *Escenario A*

Figura 2.2: *Escenario B*

En el siguiente capítulo se verá la metodología propuesta que permite alcanzar los objetivos propuestos.

CAPÍTULO 3

Metodología

La metodología utilizada en el presente trabajo está basada en el KDD (*Knowledge Discovery in Database*)(Fayyad, Piatetsky-Shapiro, y Smyth, 1996a, 1996b), la cual permite el descubrimiento de conocimiento a partir de bases de datos, está compuesto por una serie de pasos, tal como se muestra en la Figura 3.1.

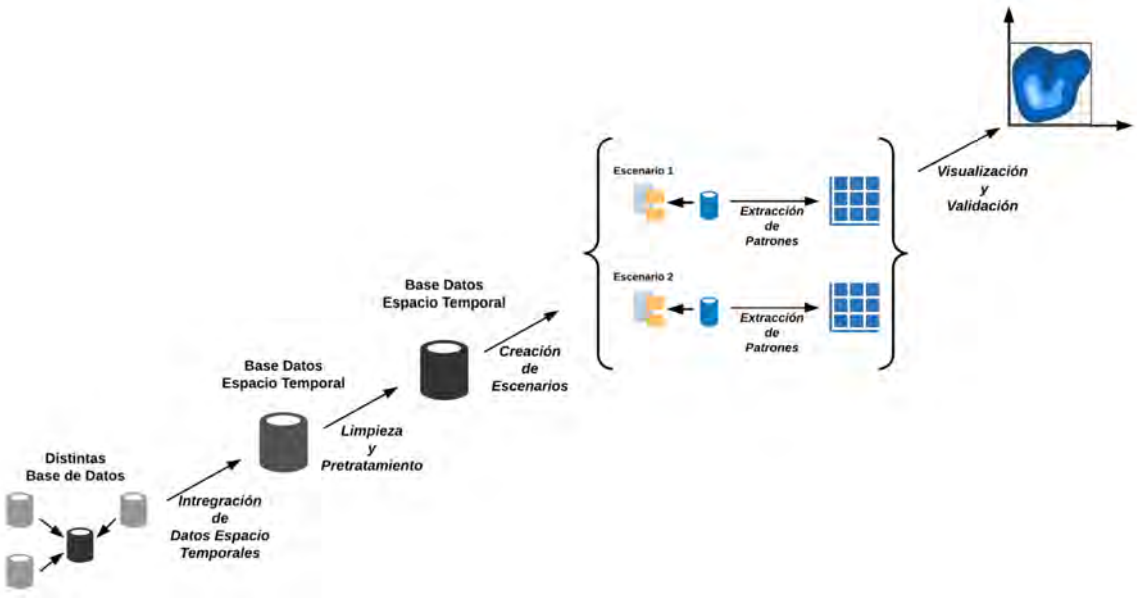

Figura 3.1: *Metodología*

Entre los pasos a realizar se tienen:

1. Ubicar diferentes fuentes de datos heterogéneas, ya sean en archivos pdf, archivos de texto CSV (los datos están separados por coma), archivos en formato netCDF (*Network Common Data Form*) estándar internacional del *Open Geospatial Consortium*, etc.

2. Los datos recolectados se integran en una base de datos espacio-temporal, donde se deben seleccionar las variables que serán utilizados en el proceso de descubrimiento de conocimiento planteado en esta tesis.

3. En la etapa de limpieza o pre-tratamiento de datos, se deben realizar operaciones básicas como: la eliminación de ruido, determinación de estrategias para el manejo

de los campos de datos que faltan o en blanco, inconsistentes o fuera de rango, reducción de datos en caso tengan un comportamiento invariante en el tiempo.

4. Escenarios: Se crean los escenarios donde se definen las hipótesis que permitirán validar las posibles correlaciones.

5. Minería de datos: la búsqueda de patrones secuenciales espacio-temporales mediante el algoritmo de minería de patrones secuenciales llamado PrefixSpan.

6. Visualización: los patrones extraídos se deben visualizar en función de los escenarios seleccionados de tal forma que permitan una mejor comprensión de estos.

7. Los resultados obtenidos son evaluados y validados por un experto, con el objetivo de determinar si estos se consideran como conocimiento novedoso.

Un resumen de lo anterior y las acciones a realizar se muestran en la Tabla 3.1:

Tabla 3.1: *Resumen de Tareas a realizar*

Pasos	Tareas
Recolección de datos	Creación de programas para la extracción de datos de fuentes heterogéneas.
Integración	Creación de un Base de Datos Espacio Temporal en Postgresql con soporte de objetos espaciales
Limpieza y Pretratamientos	Normalizar los datos, agrupar los datos en rangos, por ejemplo: mediante el uso de cuantiles.
Creación de escenarios	Definir hipótesis a partir de los datos espacio temporales
Extracción de Patrones	Uso de algoritmos para el descubrimiento de patrones secuenciales, como: PrefixSpan
Visualización	Mediante una aplicación web
Validación	Experto

3.1. Alcance

Teniendo en cuenta lo mencionado en el capítulo 2, el alcance de la presente tesis comprende:

- En la presente tesis se implementará un proceso computacional basado en técnicas de minería de datos que permitan estudiar el impacto de la temperatura de la superficie del mar en las variables meteorológicas del norte y centro del litoral peruano, con relación a la zona Niño 1+2.

- La creación una base datos espacio-temporal con los datos recolectados de diferentes fuentes de datos.

- La creación escenarios donde los datos estarán agrupados de forma secuencial permitiendo definir hipótesis de posibles correlaciones.

- La búsqueda de posibles patrones espacio-temporales mediante el uso del algoritmo PrefixSpan para minería de patrones secuenciales.
- La implementación del prototipo de una aplicación web la cual permita la visualización de los patrones encontrados.

3.2. Límites

- Los datos a utilizar corresponden a información de libre acceso en internet.
- Los datos respecto a la temperatura de la superficie del mar se obtendrán de la información recolectada por las boyas ubicadas frente al litoral norte y centro, con relación a la zona Niño 1+2.
- Los datos recolectados corresponderán al último evento del fenómeno de El Niño, periodo 2015-2016.

Los datos serán recolectados de diferentes fuentes que son de acceso público vía web.

CAPÍTULO 4

Experimentación y Resultados

Como se indicó en el capítulo anterior la metodología a utilizar está basada en el KDD (*Knowledge Discovery in Database*), esto permite el descubrimiento de conocimiento a partir de bases de datos, Figura 3.1. El código utilizado esta viable en GitHub bajo licencia GPL v3.[6]

4.1. Bases de datos

Con el objetivo de medir el impacto de la temperatura de la superficie del mar en las variables meteorológicas registradas en el litoral peruano en el contexto del fenómeno de El Niño, se utilizaron diversas fuentes de datos. A continuación, se describen las diversas fuentes de datos y la información que proporcionan.

4.1.1. 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[7], 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, luego estos datos son enviados por satélite desde la superficie, como se observa en la Figura 4.1. (Argo-España, s.f.)

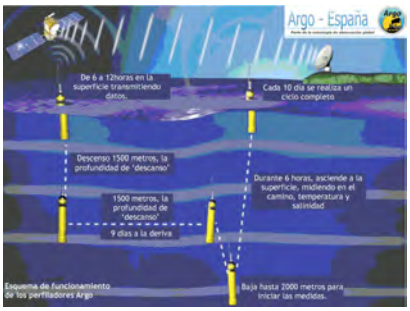

Figura 4.1: *Comportamiento de una Boya.*
Fuente: Argo-España[7]

[6] https://github.com/oscardbpucp/Comp-Process-STPatterns
[7] http://www.oceanografia.es/argo

Los datos de las boyas son almacenados por el sistema oceanografía operativa Coriolis[8] quien a su vez pone en acceso público la información recolectada, en archivos en formato NetCD [9] o CSV (*Comma-Separated Values*).
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' y '30/06/2016', obteniéndose 30 archivos de texto CSV, donde cada archivo corresponde a los datos recolectados por una determinada boya. Las principales características que se obtienen por cada boya se muestran en la Tabla 4.1

Tabla 4.1: *Principales características*

Nombre del valor	Tipo de dato	Descripción
ARGOS_ID	Nominal	Identificador de la boya
DATE (YYYY-MM-DDTHH:MI:SSZ)	Discreta	Fecha de registro
Latitud	Continua	
Longitud	Continua	
Presión	Continua	Decibar
Temperatura	Continua	Grados celcius
Salinidad	Continua	PSU

4.1.2. Variables Meteorológicas en Litoral peruano

La información de meteorológica del litoral peruano se obtuvo de la NOAA (*National Oceanic and Atmospheric Administration*) la cual es una agencia científica del Departamento de Comercio de los Estados Unidos. Una de las varias tareas que realiza la NOAA es registrar información de las estaciones meteorológicas localizadas en todo el mundo y a su vez poner en acceso público vía web esta información.[10]
Los datos que se obtienen son 2 archivos:

- Un archivo con la información de las estaciones meteorológicas en el mundo, desde febrero del 2015 hasta agosto del 2016, con los siguientes parámetros, Tabla 4.2
- Un segundo archivo en formato CSV conteniendo la información registrada por las estaciones meteorológicas, en la Tabla 4.3 se muestran algunas de las características registradas.

4.1.3. Variables hidrometeorológicas

La información hidrometeorológicas, que está referida a el caudal de los ríos, se obtiene del sitio web del Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI)[11] extrayéndose la información para el periodo 2015-03-12 al 2016-07-10, la cual se encuentra en (432) archivos en formato PDF (*Portable Document Format*) uno por día. La Tabla 4.4 muestra algunas de las características que se tienen en los archivos.

[8] http://www.coriolis.eu.org/Data-Products/Data-Delivery/Data-selection
[9] http://www.unidata.ucar.edu/software/netcdf/examples/files.html
[10] https://www7.ncdc.noaa.gov/CDO/country
[11] http://www.senamhi.gob.pe/?p=0320

Tabla 4.2: *Parámetros de lista de estaciones meteorológicas*

Nombre	Descripción
USAF	ID de la estación
WBAN	Número NCDC WBAN (Weather Bureau Air Force Navy)
CTRY	Identificador de código de país FIPS, ejemplo Perú = PE
ST	Estado de Estados Unidos de Norteamérica
ICAO	Señal de llamada ICAO (International Civil Aviation Organization)
LAT	Latitud en milésimas de grados decimales
LON	Longitud en milésimas de grados decimales
ELEV	Elevación en décimas de metro
BEGIN	Fecha de inicio de registro
END	Fecha de fin de registro

4.2. Integración

En esta etapa se constituye una sola base de datos a partir de las diferentes fuentes de datos.

Primero, la información del caudal de los ríos se pasa a un archivo CSV y así se logra normalizar el tipo de archivo de fuente de datos, para esto se implementó una aplicación en Python que utiliza el programa *pdftohtml*[12] el cual convierte los 432 archivos PDF en archivos HTML. Luego se realiza un *Web Scraping*[13] de los archivos HTML con el objetivo de extraer la información de los ríos localizados en la costa norte del Perú. El archivo CSV resultante contiene la siguiente información:

- Fecha
- Departamento, de forma manual se busca la ubicación de las cuencas.
- Cuenca,

[12]https://linux.die.net/man/1/pdftohtml
[13]Procedimiento por el cual se extrae información de un sitio en web

Tabla 4.3: *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 4.4: *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

- Hidroeléctrica,
- Caudal,
- Anomalía Hídrica

Finalmente, se implementan una serie de programas en Python que pasan la información de los archivos CSV (temperatura de la superficie del mar, estaciones meteorológicas y caudal de los ríos) a una base de datos PostgreSQL, cuyo esquema es el que se muestra en la Figura 4.2

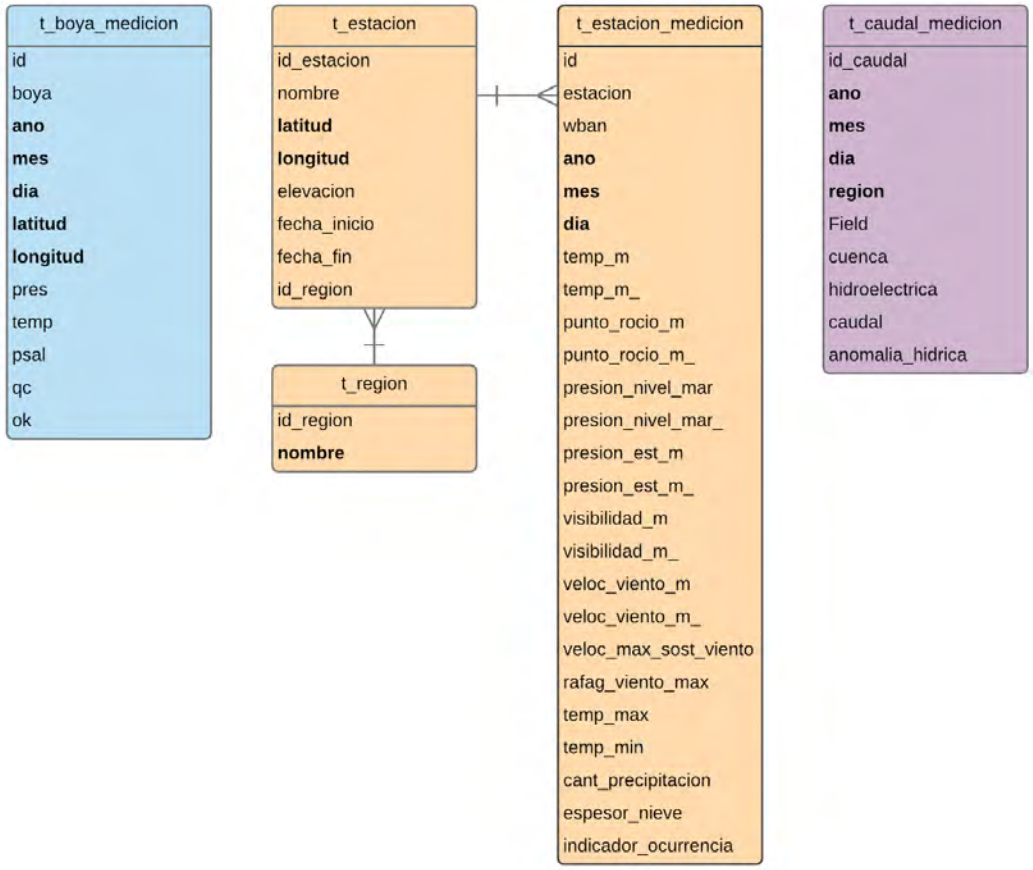

Figura 4.2: *Base de Datos Integrada*

4.3. Limpieza y Pre-tratamiento

En esta etapa, los datos son normalizados y se extraen 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 la Tabla 4.5 .

Tabla 4.5: *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

Para normalizar los datos, se procede a discretizar la información ya sea en:

- Terciles (3)
- Quintiles (5)

Los campos con valores no válidos, o no se tiene valor, se le reemplazan por un valor fuera de rango: 999999.
Dado que la información de las boyas se registran aproximadamente cada 10 días, se decide agrupar los valores en un periodo en el que se tenga por lo menos una boya en un determinada región, para esto se calcula el porcentaje de datos faltantes, relacionado a las boyas, agrupados para diferentes rangos de días: El resultado se muestra en la Figura 4.3.

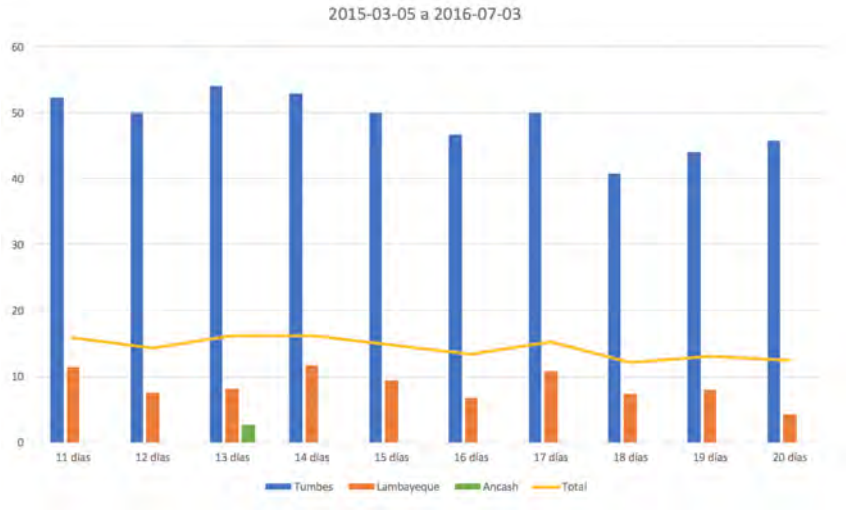

Figura 4.3: *Porcentaje de datos faltantes - Boyas*

En la Figura 4.3, 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. De la Figura 4.3 se excluyen las regiones Piura y La Libertad dado que no se presentaron perdidas de datos en dichas regiones.

Luego se determina la cantidad de datos para un periodo de 18 días como se observa en la Figura 4.4, obteniéndose 14 grupos de datos validos es decir, se tiene información para todas las regiones en el mismo rango de tiempo.

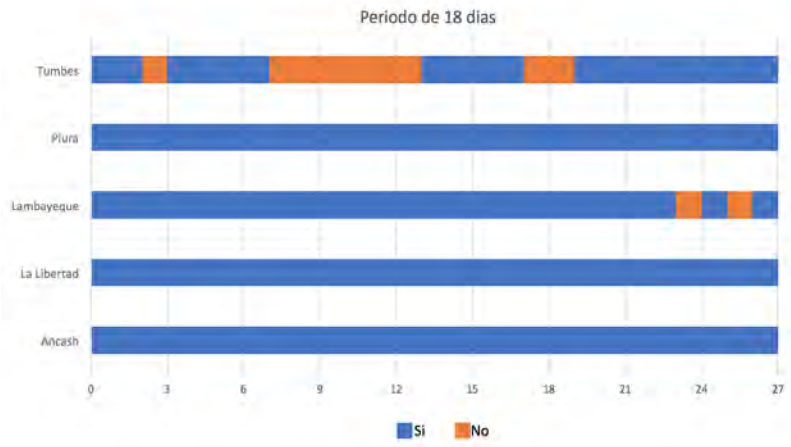

Figura 4.4: *Datos agrupados en periodo de 18 días - Boyas*

4.4. Escenarios

Luego que se tienen los datos agrupados, calculados a partir del promedio de los valores por región y en periodos de 18 días, se procede a definir los escenarios. Estos están conformados por las regiones y su correspondiente proyección en el mar dentro de la Zona 1+2 de El Niño. En esta etapa se desarrolla una aplicación web que permite visualizar los escenarios y el comportamiento en el tiempo de las boyas. Los escenarios que se analizarán son:

- Escenario 1: 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. Ver Figura 4.5.

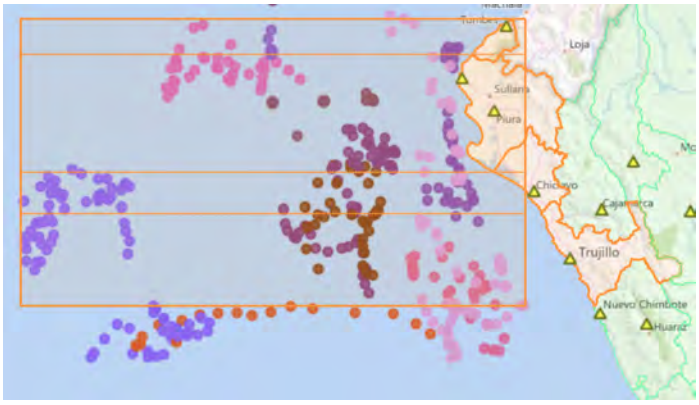

Figura 4.5: *Escenario 1*

- Escenario 2: Conformada por todas las regiones que se encuentran frente a la Zona 1+2: Tumbes, Piura, Lambayeque, La Libertad y Ancash, esta última, es una región que 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. Ver Figura 4.6.

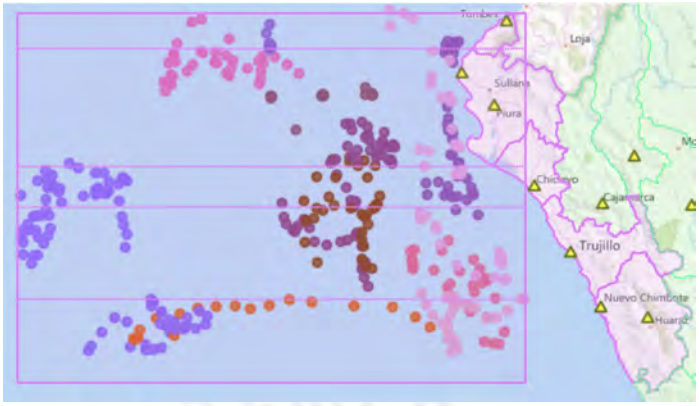

Figura 4.6: *Escenario 2*

Cabe mencionar que, en las Figuras 4.5 y 4.6, los puntos ubicados en el mar representan las boyas en un determinado rango de tiempo.

4.5. Minería de datos

Luego de definirse los escenarios, se construyen las secuencias por medio de una serie de programas en Python y el uso de la aplicación Orange[14]. Cada secuencia corresponde a una región y esta a su vez está conformada por Itemsets, los cuales están formados por items que son los valores discretizados de las características seleccionas mencionadas en la Subsección 4.3. Un ejemplo de una secuencia para el escenario Escenario 1 se observa en la Tabla 4.6:

Tabla 4.6: *Ejemplo de secuencia para la región Tumbes en el Escenario 1*

Id	Secuencia
Tumbes	(boya-temp_<24.578, boya-salinidad_<35.003, estac-temp_>=81.139, estac-pto-rocio_>=69.952, est_presion_1011.4:1483.608, est_presion_est_m_1008.168:1480.439, est_veloc_viento_m_<7.182, est_temp_max_82.482:88.287, est_temp_min_>=73.379, caudal_>=111.528), (boya-temp_>=26.23, boya-salinidad_<35.003, . . . , est_temp_min_69.668:73.379, caudal_>=111.528), ... (boya-temp_<24.578, boya-salinidad_35.003:35.129, . . . , est_temp_min_69.668:73.379, caudal_43.145:111.528)

Luego de la implementación de los escenarios se procede a la extracción de patrones secuenciales, como se menciona en la Sección 2.2.3 existen muchas clases de algoritmos

[14]Open source machine learning and data visualization for novice and expert. Interactive data analysis workflows with a large toolbox. http://orange.biolab.si/

para el descubrimiento de patrones secuenciales siendo uno de los más importantes PrefixSpan (Prefix-projected Sequential pattern mining) el cual es un algoritmo que permite el descubrimiento de patrones secuenciales frecuentes dentro de una base de datos de secuencias, fue propuesto por Pei et al. (2001). La principal idea de PrefixSpan es que en lugar de proyectar las bases de datos de secuencias considerando todas las posibles ocurrencias de subsecuencias frecuentes, la proyección sólo se basa en prefijos frecuentes debido a que cualquier subsecuencia frecuente siempre se puede encontrar creciendo un prefijo frecuente.

Para la extracción de patrones mediante el algoritmo PrefixSpan (*Prefix-projected Sequential pattern mining*) se utiliza la librería de minería de datos SPMF[15] la cual es de código abierto, esta librería requiere que la base de datos de secuencias a analizar cumpla con un determinado formato, donde el valor de **-1** se utiliza para indicar el fin de un itemsets y el valor de **-2** el fin de una secuencia. En la Tabla 4.7 se muestra un ejemplo de cómo quedarían las secuencias luego del cambio de formato para la base de datos de secuencias en el caso del Escenario 1.

Tabla 4.7: *Ejemplo de base de datos de secuencias para el Escenario 1 cumpliendo el formato de SPMF*

Id	Secuencia
Tumbes	boya-temp_<24.578 boya-salinidad_<35.003 estac-temp_>=81.139 estac-pto-rocio_>=69.952 est_presion_1011.4:1483.608 est_presion_est_m_1008.168:1480.439 est_veloc_viento_m_<7.182 est_temp_max_82.482:88.287 est_temp_min_>=73.379 caudal_>=111.528 **-1** boya-temp_>=26.23, boya-salinidad_<35.003 ... est_temp_min_69.668:73.379 caudal_>=111.528 **-1** ... boya-temp_<24.578 boya-salinidad_35.003:35.129 ... est_temp_min_69.668:73.379 caudal_43.145:111.528 **-1 -2**
Piura	boya-temp_<24.578 ... **-1** ... **-1** ... **-1** ... **-1** ... **-1** ... **-1 -2**
Lambayeque	boya-temp_<24.578 ... **-1** ... **-1** ... **-1** ... **-1** ... **-1 -2**
La Libertad	boya-temp_<24.578 ... **-1** ... **-1** ... **-1** ... **-1** ... **-1 -2**

La extracción de patrones se realizó para diferentes valores de soporte mínimo con el objetivo que los patrones encontrados estén presentes en la mayoría de las regiones y dado que se diferentes niveles de discretización con el propósito de observar si se obtienen más patrones al aumentar el nivel de discretización. Los resultados obtenidos se muestran en la Tabla 4.8, cabe mencionar que el Escenario 1 está conformado por 4 secuencias y el Escenario 2 por 5 secuencias.

De la Tabla 4.8, se descartan primero aquellos resultados que tienen muchas secuencias

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

Tabla 4.8: *Pruebas y Resultados*

Escenario	Niveles Discretización	Soporte Mínimo	Nro Secuencias	Max. Memoria (MBytes)	Tiempo Total
1	3	3	>20000	740	>336 min 00 seg
1	**3**	**4**	**7502**	**262.31**	**19 min 47 seg**
1	5	3	>20000	666	217 min 52 seg
1	5	4	552	32.56	1 min 48 seg
2	3	4	>20000	666.68	551 min 42 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

(>20000) ya que estas corresponden a todas las combinaciones posibles de los items y segundo, a los resultados donde los patrones secuenciales están conformados por ítems individuales que no agregan valor, obteniendo se mejores resultados para:

- Escenario 1, con 3 niveles de discretización y 4 de soporte mínimo.
- Escenario 2, con 5 niveles de discretización y 4 de soporte mínimo.

Tabla 4.9: *Escenario 1 - Patrones Secuenciales más relevantes*

	Patrones Secuenciales
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
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
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	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
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

En la Tabla 4.9 se muestran los patrones secuenciales más relevantes correspondientes a el Escenario 1, donde se observa que:

- En la secuencia 1, se observa que cuando se tiene un caudal alto, mayor ó igual a 111.528 m^3/s, y luego la temperatura de la superficie del mar disminuye pasando

de un valor alto, 26.23 °C a un valor entre 24.578 °C a 26.23 °C, el caudal también disminuye, a un valor entre 43.145 m^3/s y 111.528 m^3/s.

- En la secuencia 2 se tiene, que luego de producirse una disminución del caudal de un valor alto, mayor ó igual a 111.528 m^3/s, a un valor intermedio entre 43.145 m^3/s y 111.528 m^3/s, la temperatura de la superficie del mar disminuye drásticamente de un valor alto, mayor a 26.23 °C a uno menor de 24.578 °C.
- Para la secuencia 3, se observa que la temperatura de la superficie del mar desciende de un valor mayor a 26.23 °C a uno menor de 24.578 °C, luego que en un mismo periodo de tiempo se presenta un caudal medio con un valor entre 43.145 m^3/s y 111.528 m^3/s, después la temperatura de la superficie del mar baja a un valor menor a 24.578 °C.
- En la secuencia 4, se tiene que el caudal inicialmente se encuentra en un valor alto, mayor ó igual a 111.528 m^3/s, luego se produce una disminución de la temperatura de la superficie del mar de un valor mayor o igual a 26.23 °C a uno intermedio entre 24.578 °C a 26.23 °C, y finalmente el caudal disminuye a un valor entre 43.145 m^3/s y 111.528 m^3/s.
- Finalmente en la secuencia 5, se tiene que el caudal inicialmente se encuentra en un valor alto, mayor ó igual a 111.528 m^3/s y luego de producirse una disminución de la temperatura de la superficie del mar de un valor mayor o igual a 26.23 °C a uno intermedio entre 24.578 °C a 26.23 °C, el caudal disminuye a un valor entre 43.145 m^3/s y 111.528 m^3/s y después la temperatura de la superficie del mar baja a un valor menor de 24.578 °C.

En la Tabla 4.10 se muestran los patrones secuenciales más relevantes correspondientes a el Escenario 2, donde se observar que:

- En la secuencia 1, se observa que cuando se tiene 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.
- En la secuencia 2 se tiene, que el caudal disminuye de 161.761 m^3/s a 500346.84 m^3/s a un valor entre 28.612 m^3/s y 68.884 m^3/s, cuando la temperatura de la superficie del mar disminuye en este caso de un valor entre 26.105 °C y 26.769 °C a un valor entre 24.089 °C y 25.121 °C.
- Para la secuencia 3 y secuencia 4, se tiene una 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, la temperatura de la superficie del mar baja a un valor menor a 24.089 °C.
- Finalmente en la secuencia 5, se tiene que cuando el caudal se encuentra en un valor alto, entre 161.761 m^3/s a 500346.84 m^3/s y luego este disminuye a un valor entre 68.884 m^3/s y 161.76 m^3/s la temperatura de la superficie del mar baja de un valor entre 24.089 °C y 25.121°C a menos de 24.089 °C .

En el caso del Escenario 2, compuesto de 5 regiones, el soporte mínimo es 4, es decir solo en 4 regiones se encuentra los patrones obtenidos, para determinar dichas regiones se aplica

Tabla 4.10: *Escenario 2 - Patrones Secuenciales más relevantes*

	Patrones Secuenciales
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
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
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	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
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

una restitución donde se obtienen las siguientes regiones: Tumbes, Piura, Lambayeque y La Libertad. No es necesario ejecutar una restitución para el caso de los patrones obtenidos en el Escenario 1 ya que el soporte mínimo de 4 es decir se encuentra en todas las regiones del respectivo escenario.

4.6. Visualización

El último paso es la visualización de los patrones secuenciales encontrados, Tablas 4.9 y 4.10, para ello se implementa la aplicación web llamada ViSTPatterns Soft la cual hace uso de la librería de javascript D3[16] para la generación de gráfico como se muestra en la Figura 4.7.

La aplicación desarrollada, genera gráficos basado en grafos donde se tienen segmentos de recta que representan una determinada característica, cuenta adicionalmente con una leyenda representado por un mapa de color, donde una mayor intensidad de color hace referencia a un mayor de valor de una determinada característica, la cantidad de variación del nivel de intensidad de color corresponde al nivel de discretización utilizado. En la aplicación los patrones secuenciales se ingresan mediante una interface de configuración donde se selecciona el nivel de discretización y las secuencias que van hacer graficadas, esta interface se observa en la Figura 4.8. Adicionalmente, la aplicación permite una mayor información de las características al colocar el puntero sobre estas.

[16]https://d3js.org/

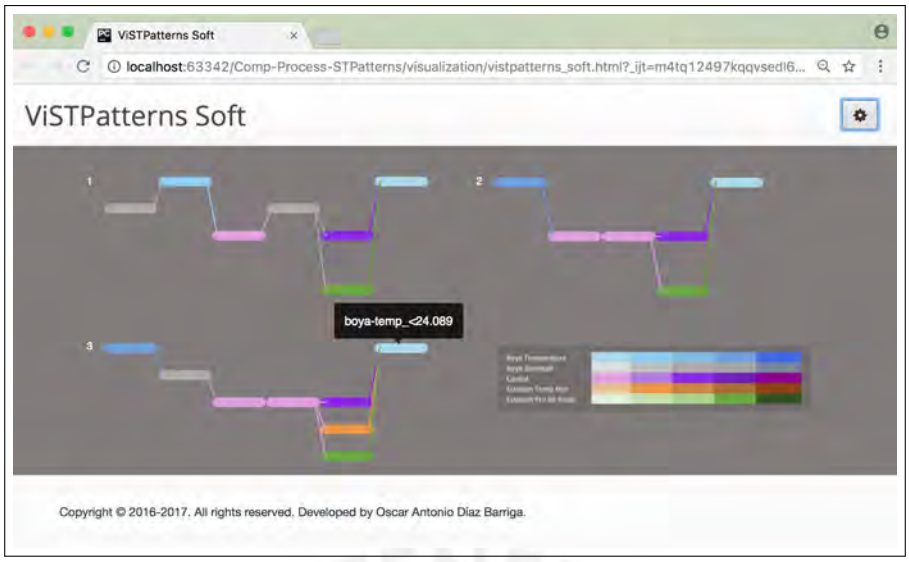

Figura 4.7: *Aplicación ViSTPatterns Soft*

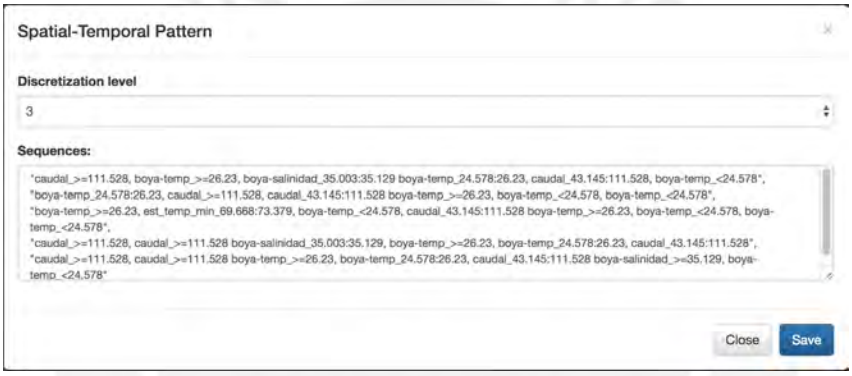

Figura 4.8: *Configuración de ViSTPatterns Soft*

La lectura de los gráficos se debe realizar de derecha izquierda, donde cada segmento corresponde a un periodo de tiempo, si se tienen dos o más segmentos conectados a un mismo punto de origen, como se observa en los gráficos de la Figura 4.7, implica que las características se presentaron en un mismo periodo de tiempo.

En la Figura 4.9 se muestran los gráficos correspondientes a los patrones secuenciales obtenidos de la Tabla 4.9 pertenecientes al Escenario 1. A partir de los gráficos se puede interpretar:

- En el gráfico 1 correspondiente a la primera secuencia, se observa por la variación de color correspondiente a la temperatura de la superficie del mar (boya temperatura) que mientras su valor va disminuyendo el nivel de caudal también disminuye.
- De los gráficos 2, 3 y 5 correspondiente a las secuencias 2, 3 y 5 respectivamente se tiene, que una disminución del caudal viene acompañada de una disminución de la temperatura del mar.
- En el gráfico 4 correspondiente a la secuencia 4, se observa que una disminución del caudal es precedida por una disminución de la temperatura de la superficie del mar.
- En la mayoría de los gráficos se observa que una disminución de temperatura de la superficie del mar viene precedida por un valor medio de caudal.

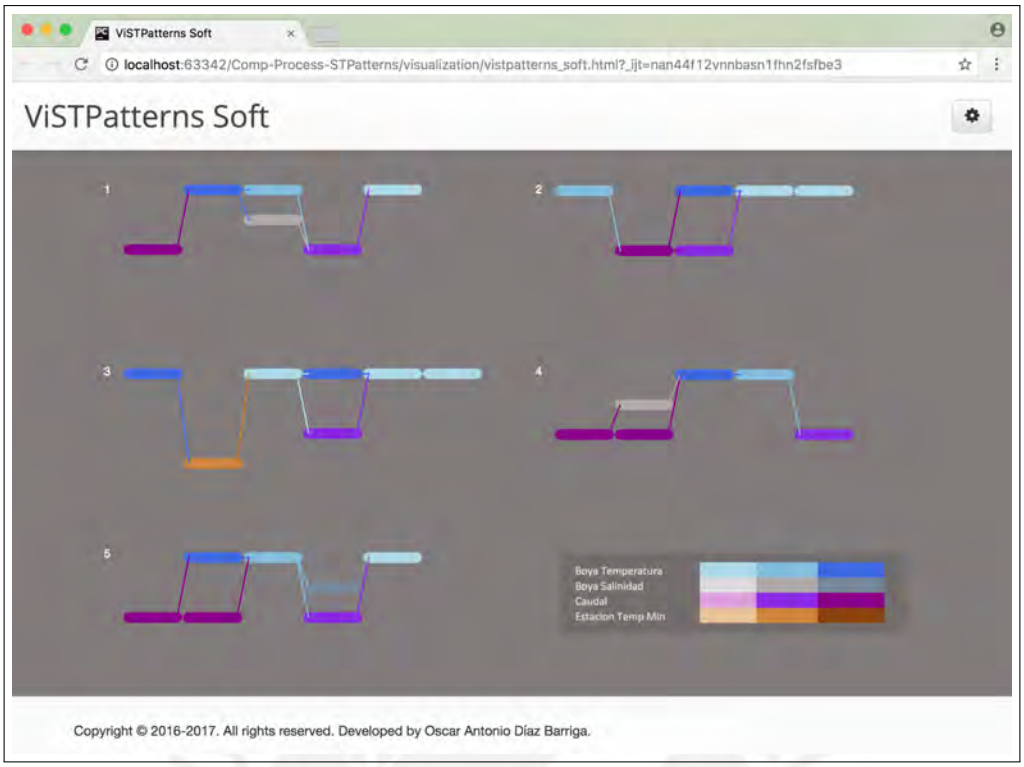

Figura 4.9: *Gráficos correspondientes al Escenario 1*

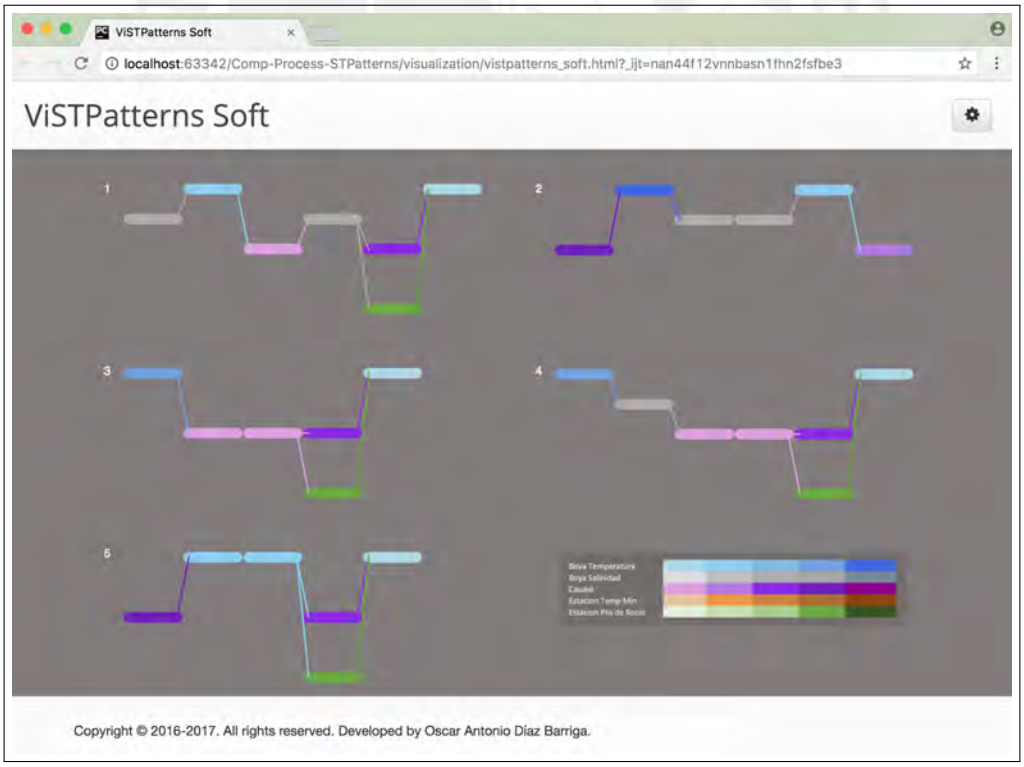

Figura 4.10: *Gráficos correspondientes al Escenario 2*

Mientras que, en la Figura 4.10 se tienen los gráficos correspondientes a los patrones secuenciales obtenidos de la Tabla 4.10 pertenecientes al Escenario 2. A partir de los gráficos se puede interpretar:

- En los gráficos 1, 3 y 4 correspondiente a las secuencias 1, 3 y 4, se observa una disminución de la temperatura de la superficie del mar (boya temperatura) mientras el valor del nivel de caudal aumenta a un valor medio.
- En los gráficos 2 y 5 correspondiente a las secuencias correspondientes, se tiene que durante la disminución de la temperatura de la superficie del mar (boya temperatura) el valor del nivel de caudal también disminuye.
- De los gráficos 1, 3, 4 y 5 se observa que durante la disminución de la temperatura de la superficie del mar a sus valores más bajos esta va precedido de un caudal medio y una temperatura media de rocío alto.

En general de los gráficos mostrados para la Figura 4.9, correspondientes al Escenario 1, los resultados resultan ser más consistentes al presentar un comportamiento similar, mientras que en el caso de los gráficos mostrados para la Figura 4.10 correspondientes al Escenario 2 se observan resultados más variables.

CAPÍTULO 5

Discusión, Conclusiones y Trabajos Futuros

5.1. Discusión

En el presente documento se desarrolla un proceso computacional basado en técnicas de minería de datos para el análisis del fenómeno El Niño, para ello se propone una metodología basada en el KDD (*Knowledge Discovery in Database*)(Fayyad y cols., 1996a, 1996b) compuesta de varias etapas: recolección de datos, integración, limpieza y pre-tratamientos, creación de escenarios, extracción de patrones, visualización y validación; de donde se obtuvieron una serie de resultados, patrones secuenciales espacio-temporales.

Se construyó una base de datos a partir de fuentes heterogéneas y no estructuradas tales como archivos de tipo: CSV y PDF. En el caso de los archivos PDF correspondiente a la información hidrometeorológica de un determinado día, se encontró que algunos de los documentos no siempre tienen el mismo formato por lo que se tuvo que realizar una corrección manual del formato previa extracción de la información en la etapa de Integración, en dicha etapa se definió una base datos compuesta por una serie de tablas correspondientes a cada una de las fuentes de datos y donde se tiene información en común como los es: lugar y fecha, lo cual es muy importante dado que se buscan patrones secuenciales espacio-temporales.

En el proceso de definición de escenarios se desarrolló una herramienta visual (aplicación web) que permitió mostrar las regiones de la costa norte y las boyas ubicadas en la zona 1+2 en diferentes periodos de tiempo, esto ayudo a comprender mejor el comportamiento espacio temporal de las boyas, el cual afecta a la recolección de información de la temperatura de la superficie del mar, dado que se observó que existen días en la que en una o más regiones no se tiene como mínimo una boya en frente del mar. Por lo anterior, en la etapa de pre-tratamiento se tuvo que agrupar la información en periodos de 18 días y luego se procedió definir los Escenarios 1 y 2, diferenciándose en que el ultimo escenario toma en cuenta todas las regiones que están frente a la zona 1+2 a pesar de que la región Ancash no se encuentra completamente frente a la zona 1+2 a comparación de las otras regiones: Tumbes, Piura, Lambayeque y La Libertad.

En la etapa de minería de datos teniendo en cuenta que la información se agrupa en 18 días se formaron secuencias de 14 itemsets, cada secuencia representando una región, de los cuales mediante el uso del algoritmo de PrefixSpan se lograron obtener múltiples secuencias espacio-temporales, 7502 y 2201 para el Escenario 1 y Escenario 2 respectivamente. De

entre ellas se escogieron las más relevantes, siendo aquellas que pudieran proporcionar mayor información, para esto debían cumplir con: estar presente la temperatura superficial del mar, incluir la mayor cantidad de itemsets y tener varias características en un mismo periodo de tiempo, reduciéndose la cantidad de resultados.

Los patrones secuenciales relevantes obtenidos fueron analizados por un experto, dando como válidos aquellos patrones en los que se observa, para el caso del fenómeno El Niño, que luego de un valor máximo de temperatura de la superficie del mar se tiene una disminución abrupta de esta como lo indica Dewitte (Dewitte y cols., 2014) y adicionalmente, cuando se tiene una reducción de este valor máximo de temperatura de la superficie del mar las lluvias bajan por ende el caudal se reduce como lo indica Ramos (Ramos, 2015), lo anterior se presenta en los patrones 1, 2, 5 del Escenario 1, correspondientes a las regiones: Tumbes, Piura, Lambayeque y La Libertad.

En la visualización de los patrones secuenciales espacio-temporales obtenidos mediante la aplicación desarrollada, ViSTPatterns Soft, en la que se representan diferentes tipos características simultáneamente y como es la relación de ellos en el tiempo, no es un tipo gráfico usual para los oceanógrafos y meteorólogos por lo que se dificulta su comprensión por parte de ellos, tal vez mayores elementos de ayuda o documentación puedan reducir la dificultad en su entendimiento.

5.2. Conclusiones

De forma general se puede concluir, que es posible implementar un proceso computacional basado en técnicas de minería de datos que permiten medir el impacto de la temperatura de la superficie del mar en las variables meteorológicas registradas en la costa norte del litoral peruano, dentro del contexto del estudio del fenómeno El Niño, como se muestra en el capítulo anterior. Adicionalmente, se concluye que:

- Se construyó una base de datos transaccional en función de diversas fuentes de información lo cual permitió obtener patrones secuenciales.
- Se definieron 2 escenarios que han permitido representar la dinámica espacio-temporal del fenómeno El Niño y ubicar posibles correlaciones espacio-temporales.
- Se identificaron algunos patrones espacio-temporales, ver Tabla 4.9 y 4.10, de estos resultados el experto dio como válidos aquellos patrones donde se presenta una correlación entre la temperatura de la superficie del mar y el nivel del caudal, lo cual se registró principalmente en las regiones: Tumbes, Piura, Lambayeque y la Libertad.
- Se implementó un prototipo de aplicación web el cual se observa en la sección 4.6 del capítulo anterior, este permite visualizar los patrones obtenidos con el fin de facilitar la comprensión del fenómeno de El Niño, por ejemplo, se puede observar que un descenso de la temperatura del mar va acompañado de un descenso del nivel de caudal de los ríos de la costa norte del Perú.

5.3. Trabajos Futuros

La presente investigación se realiza dentro del contexto del fenómeno El Niño, como una extensión de este trabajo se recomienda añadir el caso de La Niña ya que ambos son partes del mismo fenómeno cíclico, El ENSO (El Niño - Oscilación del Sur) y así tener un mejor entendimiento de dicho fenómeno.

Debido a la gran cantidad de patrones secuenciales obtenidos se recomienda revisar algunas técnicas de selección de patrones automáticas que permitan la reducción de cantidad de patrones.

Otro trabajo futuro, es el añadir otras características como la información de migración de las aves, la cantidad de peces, etc., teniendo en cuenta que algunas de estas características ya son tomadas en cuenta por el Comité multisectorial Encargado del Estudio Nacional del Fenómeno El Niño (ENFEN) en su análisis de El Niño.

Referencias

Argo-España. (s.f.). *Esquema de funcionamiento de los perfiladores argo.* `http://argo.oceanografia.es/node/99`.

Dewitte, B., Takahashi, K., Goubanova, K., Montecinos, A., Mosquera, K., Illig, S., ... others (2014). Las diversas facetas de el niño y sus efectos en la costa del perú. *Montes*, *1*, 3.

Dhanya, C., y Kumar, D. N. (2009). Data mining for evolving fuzzy association rules for predicting monsoon rainfall of india. *Journal of Intelligent Systems*, *18*(3), 193–210.

ENFEN. (2015). *Estudio nacional del fenómeno del niño.* `https://www.dhn.mil.pe/Archivos/oceanografia/enfen/comunicado-oficial/04-2015.pdf`.

ENFEN. (2016). *Estudio nacional del fenómeno del niño.* `https://www.dhn.mil.pe/Archivos/oceanografia/enfen/comunicado-oficial/08-2016.pdf`.

Fayyad, U., Piatetsky-Shapiro, G., y Smyth, P. (1996a). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37. doi: 10.1609/aimag.v17i3.1230

Fayyad, U., Piatetsky-Shapiro, G., y Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34. doi: 10.1145/240455.240464

Ganguli, P., y Reddy, M. J. (2014). Ensemble prediction of regional droughts using climate inputs and the svm–copula approach. *Hydrological Processes*, *28*(19), 4989–5009.

Janicke, H., Bottinger, M., Mikolajewicz, U., y Scheuermann, G. (2009). Visual Exploration of Climate Variability Changes Using Wavelet Analysis. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 1375–1382. doi: 10.1109/TVCG.2009.197

Kalra, A., Miller, W. P., Lamb, K. W., Ahmad, S., y Piechota, T. (2013). 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. doi: 10.1002/hyp.9236

Kawale, J., Liess, S., Kumar, A., Steinbach, M., Snyder, P., Kumar, V., ... Semazzi, F. (2013). A graph-based approach to find teleconnections in climate data. *Statistical Analysis and Data Mining*, *6*(3), 158–179. doi: 10.1002/sam.11181

Kawale, J., Steinbach, M., y Kumar, V. (2011). Discovering Dynamic Dipoles in Climate Data. *Society for Industrial and Applied Mathematics. Proceedings of the SIAM International Conference on Data Mining*(Dmi), 107–118.

Miller, H. J. (2008). Geographic Data Mining and Knowledge Discovery. En *The handbook of geographic information science* (pp. 352–366). doi: 10.1002/9780470690819.ch19

Ramos, Y. (2015, Agosto). Generación de modelos climáticos para el pronóstico de la

ocurrencia del Fenómeno El Niño. Corrigiendo los escenarios climáticos para la costa norte del Perú. *Instituto Geofísico del Perú*, *2*, 4–8. http://www.met.igp.gob.pe/publicaciones/Divulgacion_PPR_El_Nino_IGP_201508.pdf.

Rasouli, K., Hsieh, W. W., y Cannon, A. J. (2012). Daily streamflow forecasting by machine learning methods with weather and climate inputs. *Journal of Hydrology*, *414-415*, 284–293. doi: 10.1016/j.jhydrol.2011.10.039

Tan, P.-N., Steinbach, M., y Kumar, V. (2006). Introduction to data mining.

Yadira, F. R. (2013). Análisis comparativo de algoritmos utilizados en la minería de secuencias frecuentes.