PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

CARACTERIZACIÓN ESPACIO TEMPORAL DE LA ECOFISIOLOGÍA DE LA "APODANTHERA BIFLORA" UTILIZANDO MINERÍA DE PATRONES SECUENCIALES

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

JOSE LUIS BARTURÉN LARREA

Dirigido por

DR. HUGO ALATRISTA SALAS

Jurados

DR. HÉCTOR ANDRÉS MELGAR SASIETA

DR. HUGO ALATRISTA SALAS

DR. CÉSAR ARMANDO BELTRÁN CASTAÑÓN

Lima, 2016

RESUMEN

En los últimos años, los investigadores del Laboratorio de Ecología Evolutiva de la Universidad Peruana Cayetano Heredia (UPCH) han venido estudiando especies nativas del Bosque Seco Ecuatorial del norte del Perú. Este es el caso de la *Apodanthera Biflora,* raíz comestible de potencial uso alimentario e industrial. Con la finalidad de desarrollar planes de sostenibilidad y preservación de la especie, los expertos requieren realizar estudios más extensos sobre los factores que afectan las características nutricionales e industriales de la especie. Para determinar estos factores se deben descubrir correlaciones temporales a partir de fuentes de datos heterogéneas. Debido a la dificultad de explotar este tipo de datos no estandarizados ni agrupados, los métodos estadísticos tradicionales no son suficientes, por lo que se requiere herramientas permitan al experto identificar qué correlaciones temporales representan patrones frecuentes relevantes.

El presente trabajo evalúa el uso de las técnicas de minería de patrones secuenciales y visualización espacial, con el objetivo de determinar si su aplicación facilita la obtención de patrones frecuentes relevantes a partir de distintas fuentes de datos heterogéneos relacionados a la *Apodanthera Biflora.* Para lograr este objetivo, se utiliza una metodología basada en el *Descubrimiento de Conocimiento a partir de Bases de Datos* (KDD por sus siglas en inglés), el cuál define fases para la selección, pre procesamiento, transformación, minería y evaluación (visualización) de los datos.

Los resultados obtenidos demostraron que la técnica de minería de patrones secuenciales *PrefixSpan* y la visualización espacial, utilizando librerías de *Google Maps API* y *D3 Js*, permitieron a los expertos la obtención de patrones frecuentes relevantes. Así mismo, la técnica de transformación GIS para datos geográficos, y la técnica de discretización por entropía y frecuencia, han permitido el pre procesamiento de datos heterogéneos. A partir de las correlaciones descubiertas, los expertos identificaron patrones frecuentes relevantes, en las localidades de Chulucanas, Cerrato, El Morante, P. Mora y El Porvenir; principalmente relacionados a las características del suelo, precipitaciones y composición química de la raíz.

DEDICATORIA

A mis papás Rosa y Luis,
y a mi hermana Pierina
por su apoyo incondicional
durante toda mi vida.

AGRADECIMIENTOS

A mi familia por acompañar cada uno de mis pasos con amor y preocupación, especialmente a mi mamá Rosa, quien con su ejemplo y consejos me ayuda a ser una mejor persona.

A mi hermano Miguel Ángel y mi abuelo Juan, quienes desde el cielo me guían y protegen.

A los profesores y compañeros de la Escuela de Posgrado – Maestría en Ingeniería Informática por la formación y experiencia recibida. Un agradecimiento especial a mis amigos, Natalí y Dennis, con quienes pasé gratos momentos en estos dos años de estudio.

A mi asesor, el Dr. Hugo Alatrista, por su entusiasmo y motivación constante; y al Dr. Wilfredo Gonzales, Director del Laboratorio de Ecología Evolutiva en la Universidad Peruana Cayetano Heredia, por la oportunidad de realizar este proyecto.

A todas las personas que de una forma u otra me brindaron su apoyo para la realización del presente trabajo.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABLAS

CAPÍTULO 1

INTRODUCCIÓN

1.1. ANTECEDENTES

Actualmente la necesidad de analizar grandes cantidades de datos, de distintos tipos y fuentes (heterogeneidad), ha generado la evolución de campos de estudio enfocados a la optimización de técnicas para la extracción de información. Este es el caso de la minería de datos, la cual propone la utilización de técnicas de extracción de información útil para convertirla en una estructura comprensible que permita descubrir conocimiento y tomar decisiones.

Estas técnicas han sido exitosamente aplicadas en distintos campos de la ciencia, como la biología, física o química; permitiendo entender el comportamiento de las especies (Asher, L. et al. 2009), mejorar los tratamientos médicos (Baralis, E. et al. 2010), prevenir desastres naturales (Tadesse, T. et al. 2004), estudiar secuencias de ADN (Wang, K. et al. 2004), entre otros.

La información extraída gracias a los métodos de minería de datos, por sí misma, no representa nuevos conocimientos. Por ello se utilizan técnicas de visualización, las cuales permiten a los expertos validar los resultados obtenidos y entender mejor los fenómenos. Ambas técnicas vienen siendo aplicadas para resolver problemas en distintas partes del mundo; sin embargo, para problemas relacionados al Perú, se tienen pocos casos de aplicación. La mayoría de investigadores utilizan métodos estadísticos tradicionales aplicados a una sola fuente de información homogénea. Clark, D. et al. (2012) en sus estudios sobre clasificación taxonómica y la distribución geográfica de las especies de plantas dentro del Bosque Estacionalmente Seco de Perú (BES[1]), aplica el método estadístico “Análisis de

[1] El Bosque Estacionalmente Seco del Perú es una de las once ecorregiones que se extiende desde el golfo de Guayaquil (0º 30’ de latitud sur) hasta el departamento de La Libertad (7º 40’ de latitud sur), llegando hasta los 2.800 m.s.n.m. (Brack-Egg, E. 1986).

Varianza" (ANOVA) sobre una base de datos de composiciones químicas, recolectada por ellos mismos. Por otro lado, Rojas-Fox, J. (2012) en su estudio sobre la biología de la *Apodanthera Biflora* y su papel ecológico dentro del BES, utilizó el método ANOVA de una vía y anidado sobre una base de datos de muestras tomadas en distintas épocas. Ambas investigaciones, si bien han brindado aportes a la ciencia, aún hace falta estudios que relacionen otros tipos de fuentes de información que permitan a los expertos descubrir otro tipo de correlaciones.

1.2. PROBLEMÁTICA

Al entrevistar a expertos sobre las técnicas de almacenamiento de grandes cantidades de datos y del procesamiento para la obtención de patrones; se identificó la necesidad de analizar distintas fuentes de datos heterogéneas para extraer correlaciones temporales y/o espaciales. Este es el caso del laboratorio de Ecología Evolutiva de la Universidad Peruana Cayetano Heredia (UPCH), quienes realizaron un estudio sobre la *Apodanthera Biflora*, especie nativa que crece en el Bosque Seco Ecuatorial del norte del Perú.

La recolección de muestras de la especie se realizó entre los meses de marzo a junio del año 2009. Se exploraron 15 comunidades de los departamentos de Tumbes, Piura y Lambayeque. Los resultados confirman la importancia del valor nutricional de la *Apodanthera Biflora* como un recurso potencial en la industria (Clark, D. et al. 2012). Sin embargo, con el fin de desarrollar planes de sostenibilidad y preservación de la especie se requiere un estudio más extenso de los factores que afectan las características nutricionales e industriales, para identificar bajo qué condiciones se obtienen las mejores características ecofisiológicas de la *Apodanthera Biflora*. (Rojas-Fox, J. 2012)

En este contexto, actualmente existe la necesidad de contar con herramientas que utilicen técnicas de minería de datos y visualización, para descubrir correlaciones temporales (patrones secuenciales) relevantes a partir de fuentes de datos heterogéneos relacionadas a las características de la especie en las distintas zonas (Estudio realizado por el Laboratorio de Ecología Evolutiva - UPCH) y condiciones meteorológicas (Fuente: SENAMHI[2]) y geológicas (Fuente: INEI[3], MINAM[4]) cambiantes en el tiempo.

[2] Servicio Nacional de Meteorología e Hidrología del Perú
[3] Instituto Nacional de Estadística e Informática - Perú
[4] Ministerio del Ambiente - Perú

Las causas detectadas de por qué no se cuenta con esta herramienta, son:

- Dificultad para almacenar y pre procesar datos no estandarizados[5] ni agrupados.
- Dificultad para extraer patrones frecuentes a partir de las distintas características pre procesadas.
- Dificultad para identificar cuáles de los patrones frecuentes obtenidos son relevantes.

Por todo lo expuesto y con la finalidad de abordar esta problemática, se plantea la siguiente pregunta: ¿Qué impacto tienen las técnicas de minería de datos y visualización, sobre la obtención de patrones frecuentes relevantes a partir de distintas fuentes de datos heterogéneos? De esta interrogante se derivan las siguientes preguntas específicas:

- ¿Qué efecto tienen las técnicas de pre procesamiento, sobre los datos no estandarizados ni agrupados?
- ¿Qué efecto tienen las técnicas de minería de datos, sobre la extracción patrones frecuentes a partir de las características pre procesadas?
- ¿Qué efecto tienen las técnicas de visualización, sobre la identificación de patrones frecuentes relevantes?

[5] Los datos no estandarizados pueden ser temporales, geográficos, etc.

CAPÍTULO 2

REVISIÓN DE LA LITERATURA

En el presente capítulo se ha realizado una revisión de las técnicas de minería y visualización que permiten la extracción de patrones frecuentes. Así mismo se ha realizado una búsqueda de estudios previos donde se hayan aplicado técnicas de minería de patrones secuenciales para resolver problemas relacionados al ámbito biológico.

2.1. OBJETIVO DE LA REVISIÓN

El objetivo del estudio es identificar qué herramientas se utilizan para resolver problemas similares al planteado en el presente trabajo, así como identificar la relevancia de este tipo de investigaciones para el campo de las ciencias biológicas.

La búsqueda se realizó en la base de datos *Scopus* en mayo del 2015. Los términos de búsqueda fueron: *sequential, pattern, mining, visualization, biology, ecology.* Se filtraron los trabajos más recientes, es decir, documentos publicados entre enero del 2012 y diciembre del 2015. Adicionalmente se agregó a la revisión los artículos publicados por *Agrawal*, ya que fue uno de los primeros en introducir el tema de la minería de patrones secuenciales.

2.2. RESULTADOS DE LA REVISIÓN

La minería de datos nos permite extraer información relevante (útil) a partir de diferentes conjuntos de datos. Dentro de este campo, existen distintas técnicas que permiten extraer correlaciones temporales para descubrir patrones frecuentes. Además, existen técnicas de visualización que permiten a los expertos, de acuerdo a los tipos de datos, entender mejor los patrones obtenidos.

2.2.1. Técnicas de minería de datos aplicados a la extracción de patrones frecuentes

La minería de patrones secuenciales se utiliza para entender el comportamiento de fenómenos que cambian en el tiempo (Srikant, R., & Agrawal, R. 1996; Sunitha, G., & Mohan, R. 2014). Actualmente, muchas áreas de investigación como la bioinformática, la minería web o texto, entre otras, tienen que tratar con datos secuenciales temporales.

"La minería de secuencias frecuentes es un campo de la minería de datos relativamente nuevo. Rakesh Agrawal y Ramakrishnan Srikant dieron los primeros pasos al presentar los algoritmos *AprioriAll, AprioriSome y DinamicSome*. Muchos otros algoritmos con distinto enfoque se han desarrollado desde entonces." (Font Y. 2013).

2.2.1.1. Recolección y transformación de datos geográficos

Cuando se recolecta datos de tipo geográfico, muchas veces estos no se encuentran estandarizados. Por ejemplo, se puede tener, en una fuente de datos, coordenadas en formato decimal (DDM), mientras que en otro el formato es "grados minutos segundos". Para la etapa de aplicación de minería de datos, es necesario que las características espaciales estén previamente convertidas en "predicados espaciales".

Para ello se debe hacer uso de un sistema de coordenadas de referencia. El más utilizado es el WGS84 (4326) ya que es el estándar para ubicar cualquier punto en la Tierra (*Global Positioning System Interface Specification* 1984). Adicionalmente, la aparición de las tecnologías *GIS* (*Geographical Information Systems*) ha permitido el manejo de grandes volúmenes de datos espacio temporales (Alatrista, H. et al. 2013; Leong, K., & Chang, S., 2012).

Las bases de datos *PostGIS*[6] (*OS Geo Foundation* 2015) permiten almacenar atributos de tipo geográficos, utilizando el estándar WGS84, que luego pueden ser visualizados en programas como *QGIS* (*QGIS Project.* 2015) o *GeoServer* (*Open SourceGeospatialFoundation* 2015).

[6] PostgreSQL con soporte para extensiones GIS

Otros formatos en los que se podrían encontrar los datos geográficos, son los archivos vectoriales *shape* o *raster.* Para ello se requieren herramientas como *QGIS*, que permiten abrir un archivo de este tipo e importar los puntos encontrados a una base de datos con soporte *GIS*.

Si el objetivo es estudiar cambios en muestras recolectadas en distintos puntos geográficos, una forma de extraer patrones frecuentes es agregando información de cada punto en cado uno de los registros o filas de la tabla (Lee, A. J., Chen, Y. A., &Ip, W. C. 2009). Aplicando minería de datos, podemos representar un conjunto de características cambiantes en el tiempo y que son frecuentes en determinadas áreas geográficas.

2.2.1.2. Técnicas de discretización

La discretización es un proceso de transformación de datos cuantitativos en datos cualitativos (categoriales), con la finalidad de reducir los atributos en intervalos de acuerdo al rango de valores continuos. Los algoritmos para minería de patrones secuenciales requieren atributos categoriales para obtener resultados más eficientes.

Para discretizar un conjunto de datos, se debe tomar en cuenta el uso de etiquetas para los intervalos. Para ello, se debe considerar que los intervalos sean simétricamente distribuidos, es decir buscar la misma *frecuencia* de valores en los intervalos de un atributo (Fayyad U., & Irani K 1993). Esto evita desequilibrios en el balance de valores, siempre y cuando los puntos de corte tengan sentido al interpretarse. A este proceso se le conoce como discretización por frecuencia.

Otro enfoque de discretización es el basado en *entropía* (Fayyad U., & Irani K 1993). Para ello se deben definir el número de clases (K), el número de valores (m), el número de valores en el intervalo i-ésimo de una partición (m_i), el número de valores de la clase j en el intervalo i-ésimo de una partición (m_{ij}).

Se define la entropía del intervalo i-ésimo:

$e_i = -\sum_{i=1}^{k} p_{ij} log_2 p_{ij}$ donde p_{ij} es la probabilidad de que la clase *j* pertenezca al intervalo *i*, se calcula como $p_{ij} = \frac{m_{ij}}{m_i}$

2.2.1.3. Técnicas de minería de patrones secuenciales

El problema de minería de patrones secuenciales fue introducido por primera vez en *Agrawal* et al. (1995). Según los autores, una secuencia es una lista de transacciones ordenadas temporalmente, no necesariamente consecutivas, que a su vez pueden agrupar un conjunto de ítems. Estas secuencias corresponder a patrones de comportamiento o tendencia de un individuo.

Se han desarrollado distintas técnicas para extraer patrones frecuentes a partir de correlaciones temporales. Una de ellas es la minería de patrones secuenciales, donde una secuencia es una lista de transacciones ordenadas temporalmente, no necesariamente consecutivas, que a su vez agrupan un conjunto de ítems. Estas secuencias corresponden a patrones de comportamiento o tendencia de un individuo (Sunitha, G., & Mohan, R. 2014).

La minería de patrones secuenciales se aplica generalmente para el análisis del comportamiento de compras de un cliente en una tienda, tratamientos médicos, desastres naturales, patrones en llamadas telefónicas, flujo de *clicks* en una página web, secuencias de ADN o estructuras genéticas (Srikant, R., & Agrawal, R. 1996; Sunitha, G., & Mohan, R. 2014).

Los algoritmos utilizados para este tipo de minería son complejos y requieren un alto costo computacional, por ello la creación de algoritmos eficientes se hace difícil. Los principales aspectos que hay que tener en cuenta son el uso de estructuras de datos óptimas para la representación de secuencias, los mecanismos para reducir el conteo del soporte y minimizar el espacio de búsqueda del problema (Pei, J. et al. 2001; Font Y. 2013).

En la Tabla 1, se observan los atributos mínimos que debe contener una base de datos para que, a partir de ella, se pueda extraer secuencias. Además, se muestra una lista de transacciones ordenadas por ID de un individuo y la fecha.

Un ítem es un valor literal, mientras que un itemset es un conjunto no vacío de ítems. Por otro lado, se define una secuencia como una lista ordenada de itemsets. Por ejemplo, en la Tabla 1, el cliente con ID 1 compra el ítem 35 y posteriormente compra el ítem 95 (en dos estampillas temporales diferentes). Contrariamente, el cliente con ID 3 compra 3 artículos en la misma estampilla temporal (los artículos 35, 55 y 75).

Tabla 1. Ejemplo de base de datos transaccional

ID del individuo	Fecha de transacción	ID de Ítems
1 1	25/04/2015 30/04/2015	35 95
2 2 2	10/04/2015 15/04/2015 20/04/2015	15, 25 35 45, 65, 75
3	25/04/2015	35, 55, 75
4 4 4	25/04/2015 30/04/2015 25/05/2015	35 45, 75 95
5	12/04/2015	95

Agrupando la dimensión de análisis por fecha y por individuo, se obtiene secuencias por individuo, como se observa en la Tabla 2.

Tabla 2. Ejemplo de secuencias por individuo

ID del individuo	Secuencia por individuo
1	((35)(95))
2	((15 25)(35)(45 65 75))
3	((35 55 75))
4	((35)(45 75)(95))
5	((95))

Se dice que un individuo soporta (*soporte absoluto*) una subsecuencia “S”, si “S” está contenida en una secuencia, es decir, si una subsecuencia aparece por lo menos una vez en una secuencia que representa las compras de un cliente. El soporte relativo para una secuencia está determinado por una fracción del total de individuos que soportan esta secuencia (soporte relativo). El problema de la extracción de patrones secuenciales se define de la siguiente manera: sea σ un umbral llamado *soporte minimal*. Una sub-secuencia “S” es frecuente si y sólo si, tiene un soporte mayor o igual a σ. Por ejemplo, si se considera un *soporte minimal* de

25%, entonces la secuencia ((35)(45 75)) es frecuente ya que pertenece a los individuos 2 y 4 (40%). En el caso de la secuencia ((15 25)(35)), que solo está contenida en el individuo 2 (20%), se dice que no es frecuente ya que no supera el soporte mínimo relativo, como se puede observar en la Tabla 3. Todas las subsecuencias frecuentes son llamadas "patrones secuenciales".

Tabla 3. Ejemplo de soporte mínimo relativo

ID del individuo	Secuencia por individuo	¿Soporta ((35)(45 75))?	¿Soporta ((15 25)(35))?
1	((35)(95))	No	No
2	((15 25)(35)(45 65 75))	Si	Si
3	((35 55 75))	No	No
4	((35)(45 75)(95))	Si	No
5	((95))	No	No
		2/5 = 40%	1/5 = 20%

a) Definiciones

A continuación, se definen los conceptos más importantes relacionados al problema de la minería de patrones secuenciales, basados en el estudio comparativo realizado por Font Y. (2013).

Ítem: Un ítem es un valor literal.

Itemset: "Un *itemset* es un conjunto no vacío de ítems. Se denota un elemento *e* por (e_1, e_2, ..., e_m) donde cada e_j es un ítem. El *itemset*, también denominado transacción, es la base para la definición del concepto de secuencia." (Font Y. 2013).

Secuencia*: "*Una secuencia es una lista ordenada de *itemsets*. Se denota una secuencia S por $\langle S_1\ S_2...S_n\rangle$ donde cada S_j es un elemento de la secuencia [itemset]." (Font Y. 2013).

"Una secuencia $\langle a_1\ a_2...a_n\rangle$ está contenida en otra secuencia $\langle b_1\ b_2...b_n\rangle$ si se cumple que:

$i_1 < i_2 < ... < i_n$ tal que $a_1 \leq b_{i1}$, $a_2 \leq b_{i2}$, ... $a_n \leq b_{in}$. Por ejemplo, si se tienen $S_a = \langle (k)\ (j,\ i) \rangle$, $S_b = \langle (k,\ h)\ (j,\ i) \rangle$ se puede decir que S_a está contenida en la secuencia S_b."(Font Y. 2013).

Sub-secuencia: "Si una secuencia S_a está contenida en una secuencia S_b, S_a es llamada una sub-secuencia de S_b y S_b una súper-secuencia de S_a, denotada como $S_a \subseteq S_b$." (Font Y. 2013).

Tamaño de una secuencia: Cantidad de itemsets en la secuencia. Por ejemplo $S_a = \langle (a)\ (b\ c) \rangle$ es una secuencia de tamaño 2.(Font Y. 2013).

Longitud de una secuencia: Cantidad de ítems en la secuencia. "Una secuencia de longitud *k* es llamada *k*-secuencia" (Font Y. 2013). Por ejemplo la longitud de la secuencia $S_a = \langle (a)\ (b\ c) \rangle$ es 3 y se denomina 3-secuencia.(Font Y. 2013).

Soporte: "El soporte de una secuencia *S* es la cantidad de secuencias en la base de datos que contiene esa secuencia y se representa como, *min_sup(S)*" (Font Y. 2013). El soporte de una secuencia es definido por el usuario experto, y determina si una secuencia es frecuente o no. Por ejemplo según los datos de la Tabla 3, la secuencia $\langle (35)(45\ 75) \rangle$ tiene soporte *min_sup(S)* = 2, dado que la secuencia está contenida en las secuencias 2 y 4.

Secuencia Cerrada*: "*Una secuencia se dice que es cerrada si no existe una súper-secuencia S con el mismo soporte" (Font Y. 2013).

Secuencia Maximal: "Una secuencia es maximal si no es sub-secuencia de otra secuencia" (Font Y. 2013). De esta forma una secuencia "S" es maximal, si es que no existe una súper-secuencia que la contenga, suponiendo un soporte mínimo.

Prefijo: "Dadas dos secuencias $\alpha = \langle a_1\ a_2 ...\ a_n \rangle$, donde cada a_i corresponde a un elemento frecuente en [la secuencia] "S" y $\beta = \langle b_1\ b_2 ...\ b_m \rangle$ $(m \leq n)$. Se dice que β es un prefijo de α, si y sólo si:

- $b_i = a_i$; para $(i \leq m-1)$
- $b_m \subseteq a_n$
- Todos los ítems están ordenados alfabéticamente.

Por ejemplo, dado α=⟨(a) (a, b, c) (a, c) (d) (c, f)⟩, la secuencia β=⟨(a) (a, b, c)(a)⟩, es un prefijo con respecto a α."(Font Y. 2013).

Proyección de Secuencias: **"**Dadas dos secuencias α y β, tal que β es una subsecuencia de α. Una subsecuencia α' de la secuencia α, es llamada una proyección de α con respecto al prefijo β si y solo si:

- ➔ α' tiene prefijo β
- ➔ No existe una súper-secuencia α'' de α', tal que α'' es una subsecuencia de α y también tiene prefijo β.

Por ejemplo, dada la secuencia α=⟨(a) (a, b, c) (a, c) (d) (c, f)⟩ con prefijo β= ⟨(b, c) (a)⟩; entonces la subsecuencia α´= ⟨(b, c) (a, c) (d) (c, f)⟩, es la proyección de α con respecto al prefijo β."(Font Y. 2013).

Sufijo: "Dada una secuencia α= ⟨a_1 a_2... a_n⟩, donde cada a_i corresponde a un elemento frecuente en [la secuencia] "S". Sea β= ⟨b_1 b_2... b_m⟩ (m ≤ n) un prefijo de α.

Una secuencia γ= ⟨a_m'' a_m+1...a_n⟩ es llamada sufijo de α con respecto al prefijo β, y es denotada como γ= α/β, donde a_m''= (a_m – a_m').

Por ejemplo sea α' =⟨(a) (a, b, c) (a, c) (d) (c, f)⟩ una proyección de α con respecto al prefijoβ = ⟨(a) (a, b, c) (a)⟩, entonces γ = ⟨(_c) (d) (c, f)⟩ es el sufijo de α con respecto al prefijo β."(Font Y. 2013).

b) Taxonomía de los algoritmos de minería de patrones secuenciales

Dentro del área de estudio de la minería de patrones secuenciales existen diversos criterios para clasificarlos. Algunos autores han considerado agrupar a aquellos algoritmos que utilizan ciertas estructuras de datos para indexar en base de datos; otros los agrupan por la forma en que evalúan el soporte de las secuencias candidatas o por si las secuencias obtenidas incluyen restricciones o no. (Font Y. 2013).

Existen también algoritmos híbridos, pues utilizan varias técnicas o estrategias; así como algoritmos que han sido optimizados para solucionar un problema en particular. En la Figura 1 se muestra la taxonomía de algunos de los algoritmos, agrupados según las técnicas que emplean. (Font Y. 2013).

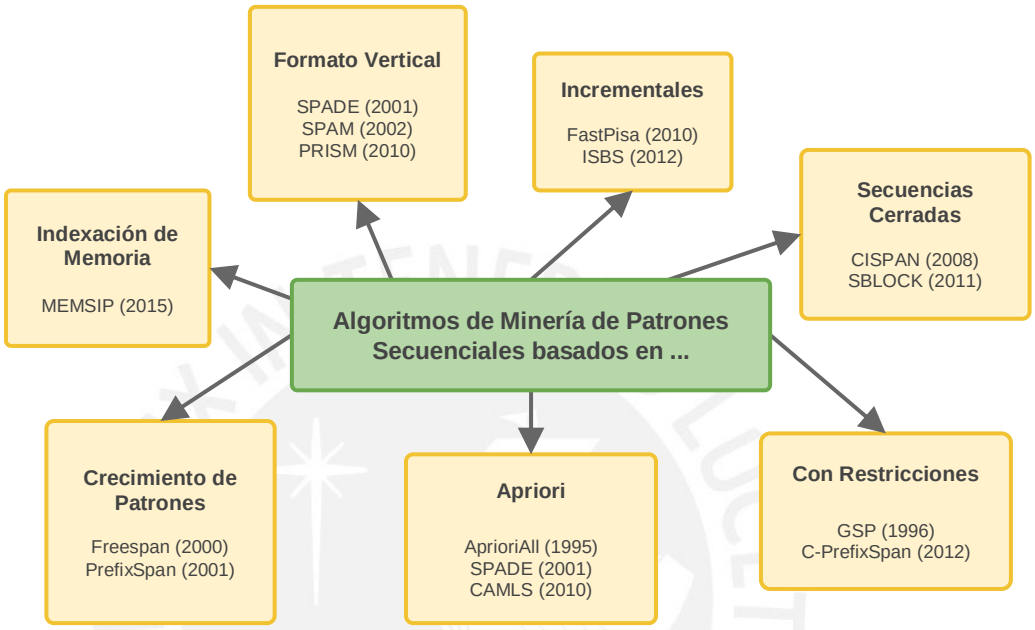

Figura 1. Taxonomía de Algoritmos de Minería de Patrones Secuenciales (Font Y. 2013).

La minería de patrones secuenciales es un problema difícil, ya que la extracción podría examinar un gran número de subsecuencias intermedias.

Entre los métodos basados en *Apriori* encontramos *GSP* y *SPADE*, los cuales recorren toda la data para encontrar "secuencias semilla". Estas semillas se utilizan para generar nuevas secuencias potenciales, las que se denominan "secuencias candidatas". De esta forma los patrones frecuentes se extraen en múltiples pasadas sobre la base de datos, encontrando cuáles de las secuencias candidatas son las más largas.

La mayoría de métodos desarrollados anteriormente, como *GSP* (*Generic Sequential Pattern*), exploran la generación de una gran cantidad de candidatos y pruebas. Sin embargo, este enfoque puede no ser eficiente en la minería de grandes bases de datos secuenciales donde se tienen numerosos patrones.

Sin embargo, este enfoque puede no ser eficiente en la minería de grandes bases de datos secuenciales donde se tienen numerosos patrones, ya que la extracción examina un gran número de subsecuencias intermedias, candidatos y pruebas. Por ello se ha desarrollado una nueva generación de métodos basados en patrones de crecimiento secuencial.

Estos nuevos métodos buscan una mayor eficiencia del algoritmo. Una base de datos secuencial es proyectada recursivamente en un conjunto de pequeñas bases de datos, y los patrones secuenciales son generados en cada una de estas, explorando sólo localmente fragmentos frecuentes. Entre estos algoritmos encontramos *PrefixSpan*, el cual desarrolla una técnica de pseudo-proyección que mejora el rendimiento, superando a *GSP*, *FreeSpan* y *SPADE* (Pei, J. et al., 2004). Además, esta metodología puede ser extendida a minería de patrones secuenciales con restricciones de tiempo específicas por individuo.

c) Algoritmos basados en crecimiento de patrones

"La idea principal de los algoritmos basados en crecimiento de patrones es tratar de reducir el tamaño del conjunto de datos explorados, mediante la realización de proyecciones de la base datos inicial y el crecimiento de patrones, sin generación de candidatos" (Font Y. 2013). Bajo la filosofía de "divide y vencerás", Pei et al. (2001) proponen el algoritmo *PrefixSpan*[7] (Pei, J., Han, J. et al., 2004).

El enfoque de patrones de crecimiento secuencial busca una mayor eficiencia del algoritmo. Una base de datos secuenciales es proyectada recursivamente en un conjunto de pequeñas bases de datos, y los patrones secuenciales son generados en cada una de estas, explorando sólo localmente fragmentos frecuentes (Pei, J., Han, J. et al., 2004).

d) PrefixSpan

PrefixSpan es un algoritmo para el descubrimiento de patrones secuenciales, propuesto por Pei et al. (2001). Está basado en el estudio inicial sobre patrones de crecimiento *FreeSpan*, y ofrece un crecimiento ordenado, así como reducción de las bases de datos proyectadas. *PrefixSpan* desarrolla una técnica de pseudo-proyección que mejora el rendimiento, superando a *GSP*, *FreeSpan* y *SPADE* (Pei, J. et al. 2004). Además, esta metodología puede ser

[7] Prefix Projected Sequential Pattern Mining

extendida a minería de patrones secuenciales con restricciones específicas por individuo (Pei, J. et al. 2001).

Una base de datos proyectada $S|_{\alpha}$ está formada por los sufijos de las secuencias en "S" con respecto al prefijo α. De forma general los pasos que sigue *PrefixSpan* para la detección de las secuencias frecuentes son los siguientes:

- ➔ Hallar el conjunto de ítems frecuentes.
- ➔ Dividir el espacio de búsqueda en subconjuntos donde cada uno representa un patrón secuencial. En la primera iteración los patrones secuenciales son el conjunto formado por los ítems frecuentes.
- ➔ Extraer los subconjuntos de patrones secuenciales construyendo las correspondientes proyecciones de las bases de datos y cada uno recursivamente, es decir, llamar nuevamente al mismo algoritmo, pero usando como base de entrada el patrón secuencial que se analiza.

"PrefixSpan es uno de los algoritmos más completos y estables para la minería de patrones secuenciales, aunque tiene una deficiencia cuando la base de datos contiene un gran número de patrones secuenciales y los ítems se repiten frecuentemente" (Font Y. 2013).

Tabla 4. Ejemplo de base de datos de secuencias

SID	**Secuencia**
1	<(a, b)(d, j, k)>
2	<(d, j, k)(h)>
3	<(d, j, k)>
4	<(b)(j, k)(d, i, m)>

Por ejemplo, dada la base de datos de la Tabla 4, en primer lugar, se debe determinar los ítems frecuentes. Considerando un soporte mínimo de 2, los ítems frecuentes de longitud 1 serían <b>: 2, <d>: 4, <j>: 4, <k>: 4. Este conjunto de patrones secuenciales se divide en tantos subconjuntos como patrones existan. Para el ejemplo, el conjunto se dividirá en 4 subconjuntos. Luego de hallar los patrones secuenciales, se proyecta una base de datos por

cada patrón. En las bases de datos proyectadas aparecerán todos aquellas sub-secuencias que tienen como prefijo al patrón secuencial que se quiere proyectar.

En la Tabla 5, se muestra la base de datos proyectada para la secuencia <(j)>. Posteriormente, partiendo de los nuevos patrones secuenciales formados y de forma recursiva, se debe llamar al algoritmo por cada uno de estos patrones (prefijos). El mayor problema que presenta el algoritmo es el costo en la proyección de las bases de datos. El tamaño requerido para el almacenamiento de las bases de datos que se van proyectando puede crecer considerablemente ya que las secuencias en estas bases pueden repetirse para varios prefijos.

Tabla 5. Ejemplo de base de datos proyectada con relación al prefijo <(j)>

Prefijo	**Sub-secuencias**
<(j)>	<(_k)>
	<(_k)(h)>
	<(_k)>
	<(_k)(d, i, m)>

En el peor de los casos el algoritmo puede llegar a multiplicar la cantidad de transacciones de la base de datos original por la cantidad de ítems frecuentes. Debido al problema planteado anteriormente *PrefixSpan* propone el uso de dos métodos de proyecciones, *bi-level projection* y *pseudo-projection*.

2.2.2. Técnicas de visualización de patrones frecuentes

La visualización es la acción de representar gráficamente (grafos, mapas, etc.) una información (relaciones, conceptos, etc.). También puede definirse como la interpretación visual de información.

El objetivo de la visualización depende del problema que se esté abordando. Los problemas pueden variar entre almacenar información, analizar datos, confirmar hipótesis, comunicar ideas, comprender fenómenos, entre otros.

La selección del método de visualización depende del contenido de los datos y del propósito de la visualización. Para garantizar el éxito del método, se debe procurar mostrar los datos sin distorsionarlos, así como, mostrarlos teniendo en cuenta la granularidad temporal y/o espacial (por niveles).

2.2.2.1. Técnicas de visualización espacial

Para visualización espacial, la técnica más común y que permite al usuario percibir los fenómenos, es utilizar un mapa geográfico. Sobre este mapa se dibujan los patrones frecuentes. Existe una amplia variedad de herramientas para este propósito. Las principales para visualización en navegadores web son:

- **Google Maps API**: Librería utilizada para la generación de mapas. Se utiliza la vista de Mapa y Satelital, así como las imágenes obtenidas por Street View en los alrededores de la zona.
- **D3 Js**: Librería utilizada para la generación de puntos sobre el mapa. Permite crear capas o *layers* de puntos, líneas o KML (Lenguaje de marcado basado en XML para representar datos geográficos).
- **Jquery**: Framework basado en JavaScript. Permite la manipulación de objetos y eventos dentro un contenido HTML.

2.2.3. Análisis de patrones frecuentes

La minería de patrones secuenciales permite la extracción de patrones frecuentes. Dependiendo el algoritmo utilizado y el soporte mínimo empleado, se puede llegar a obtener una gran cantidad de secuencias frecuentes. A partir de esta lista de secuencias se debe analizar bajo qué soporte mínimo se obtienen los patrones más interesantes. El soporte mínimo es generalmente dado por el experto, ya que es él quien conoce la frecuencia en la que se debe repetir un patrón para ser considerado importante.

Sin embargo, interpretar los patrones de una lista puede ser muy complejo. Es por ello la importancia de utilizar técnicas de visualización. Es a partir de esta representación gráfica, que se puede interpretar los fenómenos que permitan descubrir “patrones frecuente relevantes”. La

tarea de definir qué patrones son relevantes, está a cargo de los expertos, quienes son los propietarios de los datos.

2.2.4. Minería de patrones secuenciales aplicados a problemas biológicos

La minería de datos ha sido exitosamente aplicada en distintos campos de la ciencia para descubrir patrones y correlaciones. En las áreas de biología y ecología es posible extraer "patrones frecuentes" relacionados a fenotipos, genotipos o fisiología de una especie vegetal o animal. Todo ello se puede caracterizar en correlaciones con factores ambientales, espaciales y temporales.

Existen numerosos estudios biológicos que hacen uso de estos métodos para caracterizar el comportamiento de especies ante determinados factores (De Boer, F. K., & Hogeweg, P. 2012; Wang, K. et al. 2004). Por ejemplo, tenemos el caso presentado en Alatrista et al. (2012), donde los autores utilizan patrones secuenciales para analizar la calidad del agua en los ríos de Francia. En este trabajo se han explotado las propiedades espaciales inherentes a los datos para construir secuencias que denominan "secuencias espacialmente frecuentes". Las características se recolectaron en las estaciones de monitoreo ubicadas en distintos puntos de los ríos (Alatrista-Salas, H. et al. 2014).

Otro estudio es el realizado por Lei Wang et al. (2014), donde caracterizaron el comportamiento de los peces japoneses "*Medaka*" bajo condiciones químicas estresantes, como la acetona o el sodio metálico. Se observaron correlaciones temporales entre el ritmo circadiano, es decir oscilaciones de las variables biológicas en intervalos regulares de tiempo, y las concentraciones de diferentes sustancias tóxicas, utilizando patrones secuenciales (Wang, L. et al. 2014).

Un estudio más reciente fue el realizado por Mi-Jung Bae et al. (2015). En él se buscó caracterizar la reproducción del caracol de manzanas Golden, considerado como una de las plagas más serias en agricultura. Utilizando patrones secuenciales se encontró correlaciones entre la temperatura del agua y la supervivencia, la tasa de crecimiento, la reproducción y el comportamiento de los caracoles de agua dulce. El comportamiento de los caracoles de manzanas Golden fue examinado en diferentes temperaturas del agua (15 ° C, 20 ° C, 25 ° C y

30 ° C). Se clasificó los comportamientos de los caracoles en 12 categorías predefinidas cada minuto durante 2 días a cada temperatura (Bae, M. J., & Park, Y. S. 2015).

En el Perú contamos con distintas especies nativas que han sido investigadas ampliamente por laboratorios de biología en distintas universidades. Una de estas especies es la *Apodanthera Biflora* o comúnmente llamada “yuca del monte”. Esta especie ha sido estudiada a nivel estadístico para determinar sus valores nutricionales, lo cual ha permitido evaluar su potencial nutricional y su aplicación industrial en el bosque seco del norte del país.

2.3. CONCLUSIONES DE LA REVISIÓN

La minería de patrones secuenciales se ha venido aplicando al campo de la biología para resolver o entender problemas relacionados al comportamiento o tendencia de un fenómeno biológico.

En los últimos años ha existido una preocupación constante por optimizar las técnicas de minería de patrones secuenciales. En el caso de grandes bases de datos, el algoritmo *PrefixSpan* es el que ha demostrado un mayor rendimiento y rapidez en la generación de secuencias candidatas. Además, a partir de este algoritmo se han extendido variantes que permiten mejorar el rendimiento de acuerdo al problema a solucionar. Así por ejemplo existen técnicas que agregan restricciones de tiempo o individuo.

La revisión también ha revelado que no existen estudios que hayan caracterizado la ecofisiología de dicha especie utilizando técnicas de patrones secuenciales, para descubrir correlaciones temporales y poder visualizar espacialmente su aparición en las distintas zonas del Bosque Seco Ecuatorial.

CAPÍTULO 3

PLANTEAMIENTO DEL PROBLEMA

En el presente capítulo, se determinan los objetivos de la investigación, hipótesis y la utilidad del estudio para el campo de la ciencia.

3.1. OBJETIVO GENERAL

Determinar si las técnicas de minería de patrones secuenciales y visualización espacial, facilitan la obtención de patrones frecuentes relevantes a partir de fuentes de datos heterogéneos relacionados a la *Apodanthera Biflora.*

3.2. OBJETIVOS ESPECÍFICOS

[OE1] Analizar si las técnicas de transformación espacial GIS, y técnicas de discretización por entropía y frecuencia, permiten almacenar y pre procesar los datos no estandarizados ni agrupados relacionados a la *Apodanthera Biflora*.

[OE2] Evaluar si la técnica de minería de patrones secuenciales *PrefixSpan*, permite extraer patrones frecuentes a partir de las características pre procesadas relacionadas a la *Apodanthera Biflora*.

[OE3] Determinar si las técnicas de visualización espacial, facilitan la identificación de patrones frecuentes relevantes.

3.3. RESULTADOS ESPERADOS

a) Para el OE1 los resultados esperados son: la lista de secuencias transformadas y discretizadas.

b) Para el OE2, se espera la lista de patrones frecuentes extraídos por los algoritmos de minería de patrones secuenciales bajo distintos soportes mínimos; y los resultados de impacto del soporte mínimo en el tiempo de ejecución, consumo de memoria (MB), consumo de CPU (%) y número de secuencias frecuentes obtenidos del algoritmo.

c) Finalmente, para el OE3, se debe obtener un gráfico interactivo espacial que permita filtrar y seleccionar los patrones frecuentes relevantes encontrados.

3.4. VARIABLES

Tabla 6. Variables para la definición de la hipótesis

Variable	**Definición**
Técnicas de minería de patrones secuenciales y visualización espacial	Métodos que permiten extraer secuencias o patrones frecuentes a partir de un conjunto de datos. Como sub variables se encuentran: - Técnicas de transformación espacial - Técnicas de discretización
Patrones frecuentes relevantes	Patrones extraídos por la minería de patrones secuenciales, considerados “relevantes” gracias a la utilización de técnicas de visualización espacial.

3.5. HIPÓTESIS

Las técnicas de minería de patrones secuenciales y visualización espacial facilitan la obtención de patrones frecuentes relevantes a partir de distintas fuentes de datos heterogéneos relacionados a la *Apodanthera Biflora*.

Específicamente:

- Las técnicas de transformación espacial GIS, y técnicas de discretización por entropía y frecuencia, permiten almacenar y pre procesar los datos no estandarizados ni agrupados relacionados a la *Apodanthera Biflora*.

- La técnica de minería de patrones secuenciales *PrefixSpan*, permite extraer patrones frecuentes a partir de las características pre procesadas relacionadas a la *Apodanthera Biflora.*

- Las técnicas de visualización espacial, facilitan la identificación de patrones frecuentes relevantes

3.6. JUSTIFICACIÓN

La presente investigación se justifica por su implicancia práctica debido a que, la aplicación de técnicas de minería y visualización de datos, facilita la obtención de patrones frecuentes relevantes que les permitirá a los expertos tomar decisiones en relación a sus investigaciones.

3.7. ALCANCE

El presente trabajo tiene un alcance correlacional, ya que pretende responder a preguntas de investigación para evaluar el grado de asociación entre dos o más variables. Estas correlaciones se sustentan en la comprobación de hipótesis.

CAPÍTULO 4

MÉTODOS Y PROCEDIMIENTOS

Los métodos y procedimientos permiten describir cómo fue llevada a cabo la investigación. Para el presente trabajo se ha incluido detalles sobre el contexto, selección de muestras, método y procedimientos.

4.1. CONTEXTO DE LA INVESTIGACIÓN

La investigación fue realizada entre los meses de mayo a diciembre del 2015. Durante este período, el laboratorio de ecología evolutiva de la Universidad Privada Cayetano Heredia (UPCH), brindó acceso a los datos relacionados a la composición química de la *Apodanthera Biflora* para fines estrictamente académicos. Así mismo, se utilizó información pública encontrada en las páginas web de SENAMHI, INEI y MINAM.

4.2. SELECCIÓN DE LAS MUESTRAS

Se seleccionaron, 245 muestras en 3 departamentos del norte del país: Tumbes, Piura y Lambayeque, perteneciente al período comprendido entre enero y julio del 2009 (UPCH). Se agregaron características extraídas de fuentes externas como la temperatura y precipitaciones durante el año 2009 en el norte del Perú (SENAMHI), las poblaciones y localidades divididas por distritos en el año 2009 (Proyección Censo 2007 - INEI), los ríos y cuencas hidrográficas en el norte del Perú (MINAM).

4.3. MÉTODO: KNOWLEDGE DISCOVERY IN DATABASES (KDD)

Existe la necesidad de desarrollar herramientas de extracción de información útil (conocimiento) a partir de grandes y cambiantes volúmenes de datos. Esto ha originado el

surgimiento de un área de estudio denominado *Knowledge Discovery in Databases* (KDD) (Fayyad, U. et al.1996).

KDD se centra en todo el proceso de extracción de conocimiento a partir de los datos, incluyendo cómo los datos son almacenados y accedidos, cómo los algoritmos pueden escalar eficientemente a registros masivos de datos, cómo los resultados son interpretados y visualizados, y cómo toda la interacción humano computador puede ser útilmente modelada y soportada. Un patrón puede convertirse en conocimiento si tiene un grado de interés y utilidad potencial para un grupo de usuarios (Fayyad, U. et al.1996).

En Fayyad et al. (1996), se proponen los pasos que debe contener todo proceso KDD:

1. Desarrollar y entender el dominio de la aplicación, es decir entender el principal objetivo del proceso KDD desde el punto de vista del cliente. Existen dos tipos de objetivos: *Verificación* de la hipótesis de un usuario; y *Descubrimiento* autónomo de nuevos patrones, donde se aplican predicciones o descripción de los datos.
2. Seleccionar el conjunto de datos, variables, muestras en donde se extraerá el conocimiento.
3. Limpiar y pre procesar los datos. Remover registros que generen ruido a las muestras o completar la información faltante.
4. Reducción de la data y proyección. Encontrar características útiles que representen la data de acuerdo al objetivo de la tarea. En este paso es dónde se reduce eficientemente el número de variables o se encuentra una representación invariante de la misma.
5. Seleccionar el método de minería de datos que más se ajuste a los objetivos del proceso KDD definido en el paso 1. En este paso se incluye sumarización, clasificación, regresión, clusterización, etc.
6. Realizar un análisis y modelamiento exploratorio para generar una hipótesis. Decidir qué modelos y parámetros son los apropiados y asociarlos a un método de minería de datos en particular.
7. Minería de datos. Encontrar patrones de interés, reglas de clasificación, árboles, regresiones y clustering.
8. Interpretar los patrones extraídos. Este proceso también incluye visualización de dichos patrones.

9. Actuar sobre el conocimiento descubierto, por ejemplo, incorporándose en otros sistemas, documentarlo y reportarlo a las partes interesadas. Este paso también incluye revisar y resolver posibles conflictos con conocimientos descubiertos anteriormente.

El proceso KDD puede ser muy complejo y los pasos pueden cambiar significativamente dependiendo del origen de los datos. Por ejemplo, cuando los datos son geo referenciados requiere que sean transformados a un formato más compacto, abstracto y útil (Kwakkel, J. H. et al. 2014). Por ello sólo el uso de minería de datos puede conducir al descubrimiento de patrones que no tienen ningún significado para los expertos (Alatrista, H. et al. 2013).

En el presente trabajo se utiliza una metodología basada en el proceso KDD. Las fases del proceso KDD se pueden resumir en cinco, tal como se observa en la Figura 2.

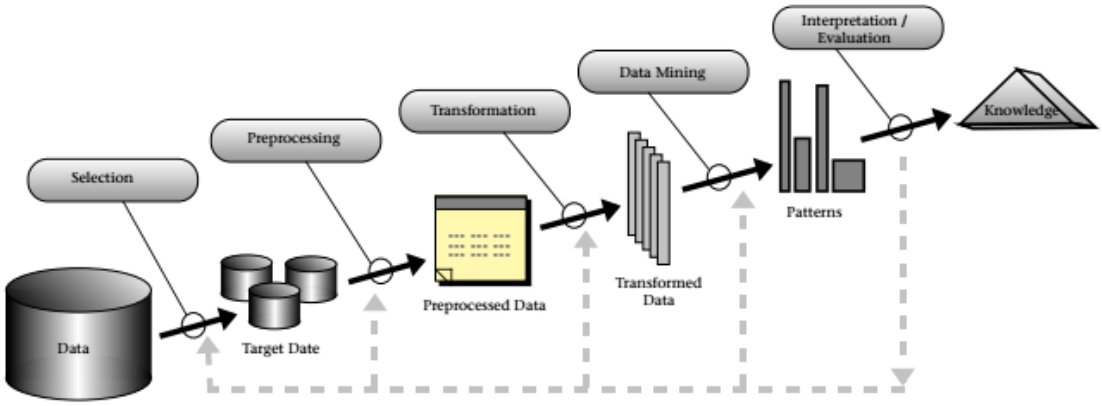

Figura 2. Fases del proceso KDD (Fayyad, U. et al.1996).

4.4. PROCEDIMIENTOS

A continuación, se describen los pasos seguidos de acuerdo a una metodología basada en las cinco fases del proceso KDD, definidos en el punto anterior.

4.4.1. Selección de Datos

El objetivo del presente trabajo es el descubrimiento de patrones secuenciales que permitan describir correlaciones temporales a partir de las principales características de la *Apodanthera Biflora,* los factores meteorológicos - geológicos, entre otros.

En este sentido, las principales características de la especie han sido recolectadas de la base de datos del Laboratorio de Ecología Evolutiva de la Universidad Peruana Cayetano Heredia. En total, se tomaron 245 muestras en 3 departamentos del norte del país: Tumbes, Piura y Lambayeque. Cada departamento, a su vez, se divide en comunidades de acuerdo a una determinada posición geográfica. La granularidad espacial se observa en la Tabla 7. La granularidad temporal es de 2 meses, es decir se recolectaron muestras en 3 estampillas temporales en un período de 6 meses.

Tabla 7. Comunidades donde se recolectaron las muestras de *Apodanthera Biflora*

Departamento	**Nro. de comunidades**	**Nro. de muestras**
Tumbes	2	30
Piura	9	130
Lambayeque	6	85

En total se tienen 17 comunidades, cuyo detalle se observa en la Tabla 8. En cada una de las muestras se recolectaron características asociadas al tallo y hojas de la *Apodanthera Biflora*. Estas características originalmente fueron registradas en fichas y posteriormente digitalizadas en formato Microsoft Excel.

Tabla 8. Comunidades donde se recolectaron las muestras de *Apodanthera Biflora*

	Comunidad	**Coordenada S**	**Coordenada W**	**Altitud (m)**
1	Tumbes 1	3° 32’ 57.0012’’	80° 18’ 27’’	0
2	Tumbes 2	3° 38' 21.9984"	80° 24' 14.0004"	10
3	Piura 1	5° 06' 12.999"	80° 34' 13.0008"	119
4	Piura 2	5° 08' 41.9994"	80° 32' 48.0006"	176
5	Piura 3	5° 12' 47.9988"	80° 23' 40.9992"	188

6	Piura 4	5° 13' 21"	80° 35' 57.9984"	154
7	Piura 5	5° 22' 58.9974"	80° 16' 50.9988"	243
8	Piura 6	5° 32' 39.0006"	80° 28' 18.0006"	246
9	Piura 7	5° 34' 51.999"	80° 14' 34.0008"	238
10	Piura 8	5° 44' 28.9998	80° 18' 38.9988"	199
11	Piura 9	5° 46' 22.998"	80° 12' 15.9978"	209
12	Lambayeque 1	6° 17' 17.9982"	79° 52' 58.9974"	68
13	Lambayeque 2	6° 22' 55.9986"	80° 1' 5.9982"	115
14	Lambayeque 3	6° 15' 35.9994"	80° 4' 19.9986"	139
15	Lambayeque 4	5° 47' 21.9978"	79° 59' 39.9984"	166
16	Lambayeque 5	6° 12' 39.3582"	79° 52' 32.7606"	115
17	Lambayeque 6	6° 14' 5.7582"	79° 55' 33.6"	68

Para el presente trabajo se seleccionaron 22 atributos, los cuales fueron convertidos a registros CSV y almacenados en una base de datos *PostgreSQL* con soporte para GIS (*Geographic Information System*). Para ello se utilizó la base de datos *PostGIS* (*OS Geo Foundation* 2015), la cual ofrece extensiones GIS que permiten almacenar atributos de tipo geográficos que luego pueden ser visualizados en programas como *QGIS* (QGIS Project. 2015) o *GeoServer* (*Open Source Geospatial Foundation* 2015). En la Tabla 9 se listan los atributos y tipos de datos almacenados en la base de datos.

Tabla 9. Atributos de la *Apodanthera Biflora* registrados por el Laboratorio de Ecología Evolutiva

	Atributo	**Tipo de Dato**	**Descripción**
1	time	date	Estampilla temporal de la toma de la muestra.
2	population_id	integer	Identificador de la población a la que pertenece la muestra.
3	population_name	string	Nombre de la población a la que pertenece la muestra.
4	population_department	string	Departamento a la que pertenece la muestra.
5	population_altitude	double	Altura (m.s.n.m.) de la zona.

6	population_temperature_max	double	Temperatura máxima (°C) de la zona.
7	population_temperature_avg	double	Temperatura media (°C) de la zona.
8	population_temperature_min	double	Temperatura mínima (°C) de la zona.
9	population_precipitation	double	Porcentaje de precipitaciones en la zona.
10	ground_ph	double	pH del suelo donde se recolectó la muestra.
11	ground_conductivity	double	Conductividad eléctrica del suelo.
12	ground_solid	double	Porcentaje de sólidos en el suelo.
13	ground_n	double	Porcentaje de nitrógeno en el suelo.
14	ground_p	double	Porcentaje de fósforo en el suelo.
15	ground_k	double	Porcentaje de potasio en el suelo.
16	stem_guide_num	integer	Número de guías en el tallo.
17	stem_guide_max_long	double	Longitud máxima (m) de las guías en el tallo.
18	stem_internodes_lenght	double	Largo (m) de los internodos del tallo.
19	stem_internodes_width	double	Ancho (m) de los internodos del tallo.
20	stem_leaves_num	integer	Número de hojas.
21	latitude	double	Latitud (minutos) de la toma de muestra
22	longitude	double	Longitud (minutos) de la toma de muestra

En la Figura 3, podemos visualizar la variabilidad de los atributos asociados a la dimensión de análisis. Las características relacionadas a conductividad de suelo (ground_conductivity), porcentaje de sólidos en el suelo (ground_solid) y porcentaje de potasio (ground_k) son las que presentan mayor variabilidad en el tiempo.

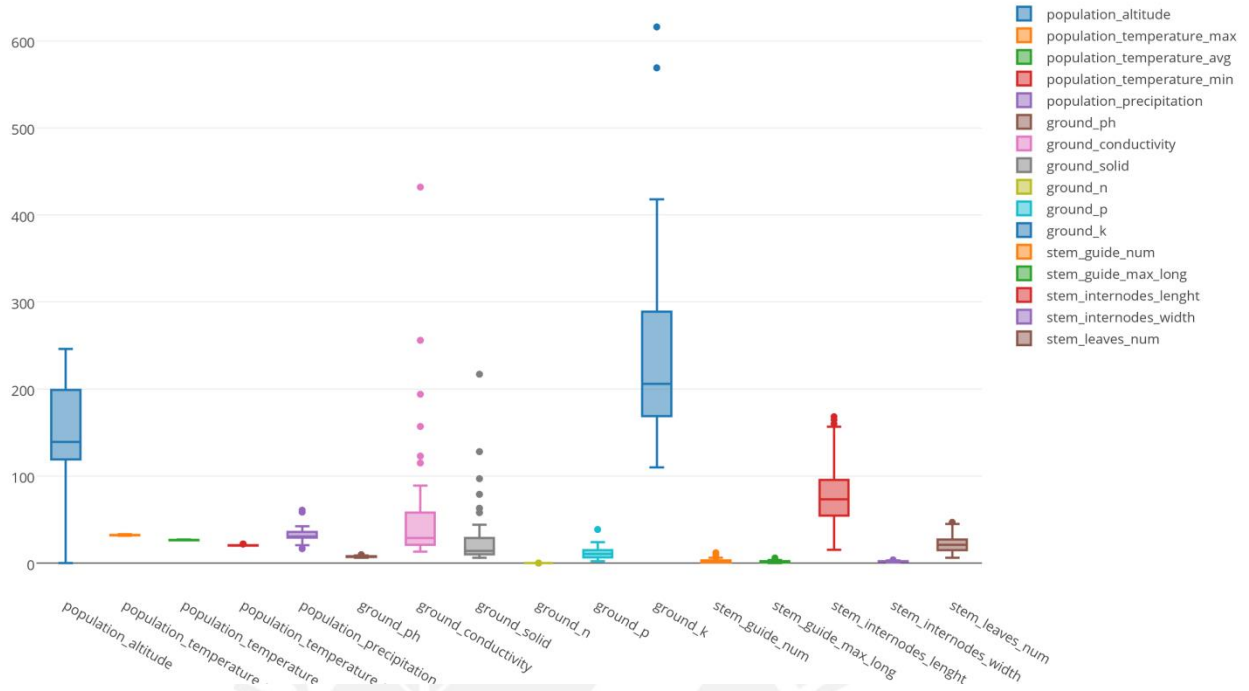

Figura 3. Atributos asociados a la dimensión de análisis VS valores numéricos

Adicionalmente a los datos recibidos por el Laboratorio de Ecología Evolutiva de la Universidad Peruana Cayetano Heredia, se han recolectado otros datos relacionados a partir de bases de datos abiertas. Estas fuentes adicionales se detallan en la Tabla 10.

Tabla 10. Fuentes de datos relacionadas a la *Apodanthera Biflora*

Datos	**Descripción**	**Formato**	**Fuente**
Hidrometeorológicos	Temperatura y precipitaciones durante el año 2009 en el norte del Perú.	Excel (CSV)	SENAMHI. Base de datos Hidrometeorógicos.
Demográficos	Poblaciones y Localidades divididas por distritos en el año 2009 (Proyección Censo 2007)	Shapefile (SHP)	INEI. Censo 2007
Hidrográficos	Ríos y cuencas hidrográficas en el norte del Perú.	Shapefile (SHP)	MINAM (Ministerio del Ambiente)

4.4.2. Pre procesamiento de Datos

Sobre la información proporcionada por el Laboratorio de Ecología Evolutiva, los datos almacenados son consistentes en valores y tipos de datos. No se detectan valores inválidos, sin embargo, fue necesario reemplazar la coma decimal por punto decimal en los atributos tipo *double*.

Para poder visualizar las muestras en un mapa, es necesario tener un atributo de tipo *Punto Geográfico* por cada uno de los registros. Las extensiones de *PostGis* ofrecen una función para crearlos a partir de la latitud y longitud registradas en coordenadas *Lambert*. Por ello, como parte del preprocesamiento, se transformaron dichos atributos al formato mencionado, ya que originalmente fueron almacenados en “grados minutos” (DDM).

Adicionalmente, para almacenar un punto se requiere un sistema de coordenadas de referencia, por lo que se seleccionó el WGS84 (4326) ya que es el estándar para ubicar cualquier punto en la Tierra (*Global Positioning System Interface Specification* 1984).

Para verificar si el *Punto Geográfico* generado es correcto, se utilizó el software *QGIS*, el cual permite ubicar en un mapa atributos de tipo geográficos. *QGIS* permite utilizar capas de tipo vectorial o *raster*, por lo que se descargó el archivo vectorial *shape* de los límites distritales de la página oficial del Instituto Geográfico Nacional (IGN).

Finalmente, se cargan los puntos geográficos mediante la extensión *PostGis* desde el *QGIS*. Para ello se configura la conexión a la base de datos y se selecciona el atributo que contiene el objeto geográfico a graficar. El resultado de la pre visualización se observa en la Figura 4. Los colores de los puntos han sido definidos por departamento; y las escalas de intensidad, por la altitud de los puntos (más oscuro, más alto).

Con respecto a las fuentes de información adicionales, los datos hidrometeorológicos se guardaron en base de datos como atributos: temperatura mínima, temperatura media, temperatura máxima y precipitación.

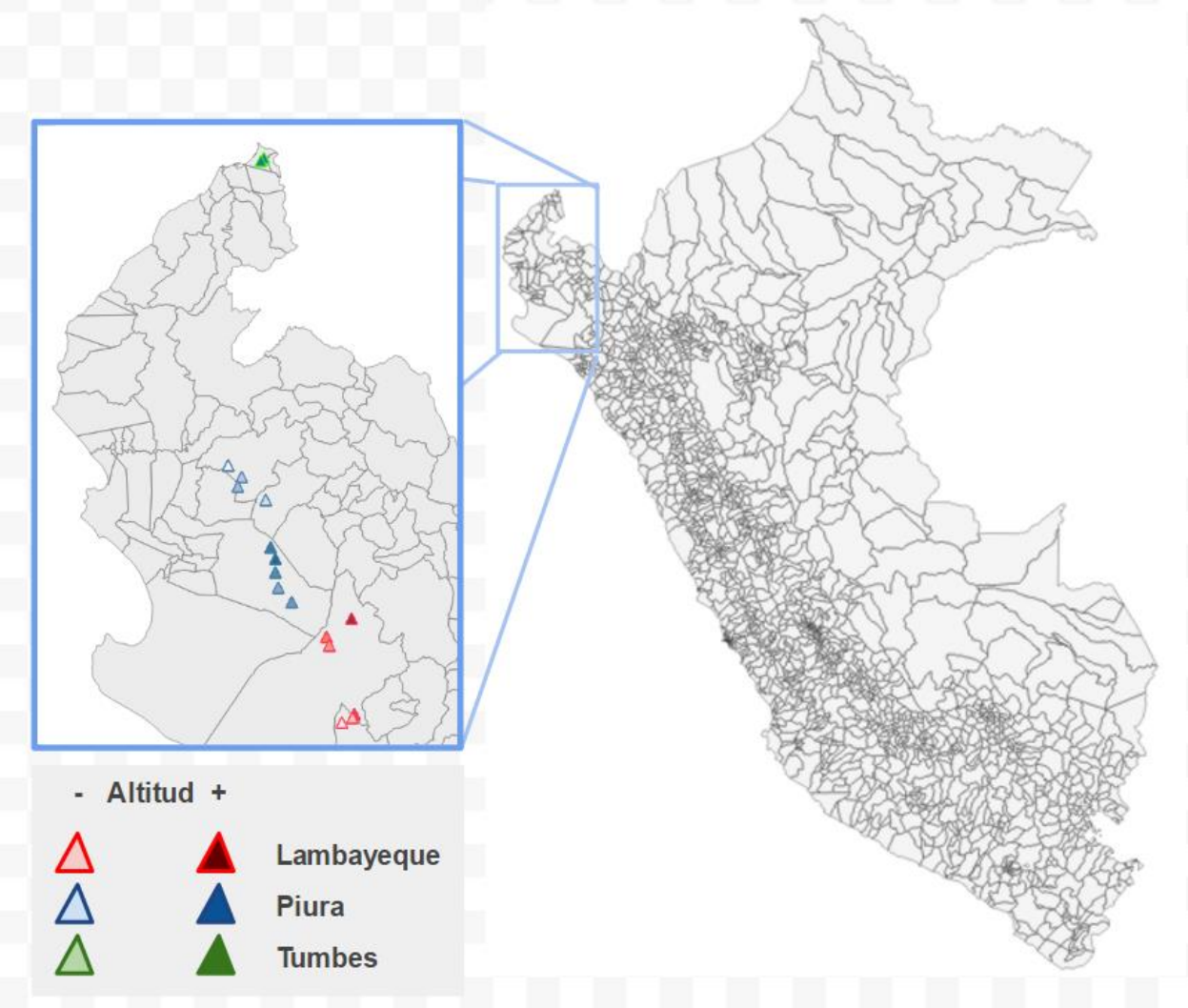

Figura 4. Pre visualización de los puntos geográficos utilizando QGIS.

Los datos demográficos e hidrográficos son de tipo *ShapeFile*, por lo que se guardaron en base de datos utilizando las herramientas proporcionadas por *QGIS*. Una vez almacenados se utilizaron funciones propias de *PostGis* para determinar, por cada Punto Geométrico que representa una muestra, cuántas Localidades (Punto Geométrico) y Ríos (Línea Geométrica) se encuentran en el radio de 5 Kilómetros. Se tomó el criterio de los 5 Kilómetros ya que corresponde a la distancia mínima entre las muestras. Finalmente, este valor numérico se guardó en base de datos como atributos: cantidad de localidades y cantidad de ríos.

4.4.3. Transformación de Datos

En el presente trabajo se utilizará dos métodos de discretización, por división no supervisado (basado en frecuencia) y supervisado (basado en entropía). En la Tabla 11 se describe el protocolo empleado para cada atributo.

Tabla 11. Protocolo de Discretización

ID	Atributo	Nro. de valores distintos	Método Discretización	Clases	
11	population_altitude	17	Frecuencia	1 2 3	**P < 120** Nro. valores: 80 **120 < P < 180** Nro. valores: 80 **P > 180** Nro. valores: 85
12	population_temperature_max	13	Entropía	1 2 3	**P = 31** Entropía: 0.8 Nro. valores: 75 **P = 32** Entropía: 0.6 Nro. valores: 125 **P = 33** Entropía: 0 Nro. valores: 45
13	population_temperature_min	11	Entropía	1 2	**P = 20** Entropía: 0.7 Nro. valores: 215 **P = 21** Entropía: 0 Nro. valores: 30
14	population_precipitation	15	Frecuencia	1 2	**P < 29** Nro. valores: 110 **P > 40** Nro. valores: 135
15	ground_ph	48	Entropía	1 2	**P < 5** Entropía: 0.9 Nro. valores: 68 **5 < P < 7**

				3	Entropía: 0.4 Nro. valores: 40 **P > 7** Entropía: 0.5 Nro. valores: 137
16	ground_conductivity	31	Entropía	1 2 3 4 5 6	**P < 30** Entropía: 0.4 Nro. valores: 127 **30 < P < 60** Entropía: 0.6 Nro. valores: 58 **60 < P < 90** Entropía: 0.5 Nro. valores: 34 **90 < P < 200** Entropía: 0.4 Nro. valores: 18 **200 < P < 300** Entropía: 0 Nro. valores: 5 **P > 300** Entropía: 0 Nro. valores: 3
17	ground_solid	27	Entropía	1 2 3 4 5	**P < 20** Entropía: 0.7 Nro. valores: 165 **20 < P < 40** Entropía: 0.6 Nro. valores: 45 **40 < P < 100** Entropía: 0.4 Nro. valores: 27 **100 < P < 200** Entropía: 0 Nro. valores: 5 **P > 200** Entropía: 0 Nro. valores: 3

18	ground_n	8	Entropía	1 2	**P < 0.03** Entropía: 0.7 Nro. valores: 186 **P > 0.03** Entropía: 0.5 Nro. valores: 59
19	ground_p	45	Entropía	1 2	**P < 7** Entropía: 0.7 Nro. valores: 67 **7 < P < 13** Entropía: 0.5 Nro. valores: 90 **P > 13** Entropía: 0.7 Nro. valores: 88
20	ground_k	45	Entropía	1 2 3	**P < 200** Entropía: 0.7 Nro. valores: 111 **200 < P < 300** Entropía: 0.4 Nro. valores: 89 **P > 300** Entropía: 0.5 Nro. valores: 45
21	stem_guide_num	9	Entropía	1 2	**P < 5** Entropía: 0.4 Nro. valores: 229 **P > 5** Entropía: 0.8 Nro. valores: 16
22	stem_guide_max_long	153	Entropía	1 2	**P < 3** Entropía: 0.8 Nro. valores: 229 **P > 3** Entropía: 0.2 Nro. valores: 16
23	stem_internodes_lenght	161	Entropía	1	**P < 50** Entropía: 0.3 Nro. valores: 46

				2 3	**50 < P < 100** Entropía: 0.5 Nro. valores: 149 **P > 100** Entropía: 0.2 Nro. valores: 50
24	stem_internodes_width	134	Frecuencia	1 2	**P < 2.2** Nro. valores:135 **P >= 2.2** Nro. valores: 110
25	stem_leaves_num	38	Entropía	1 2 3	**P < 17** Entropía: 0.5 Nro. valores: 79 **17 < P < 31** Entropía: 0.4 Nro. valores: 133 **P >= 31** Entropía: 0.5 Nro. valores: 33
26	river_count	3	Entropía	1 2	**P = 0** Entropía: 0.2 Nro. valores: 195 **P > 0** Entropía: 0.6 Nro. valores: 50
27	comunity_count	7	Frecuencia	1 2	**P < 5** Nro. valores: 155 **P > 5** Nro. valores: 90

4.4.4. Minería de patrones secuenciales

El objetivo de esta fase es encontrar patrones secuenciales frecuentes a partir de las secuencias categorizadas. Para este fin se ha utilizado la técnica de minería de patrones secuenciales *PrefixSpan,* descrita en el capítulo anterior.

La librería *SPMF* (Fournier-Viger, P. et al. 2014), creada por Philippe Fournier, ofrece soporte para minería de patrones secuenciales, entre ellas *PrefixSpan*. El algoritmo propuesto, requiere como datos de entrada la *base de datos de secuencias* y el *soporte mínimo* definido por el usuario.

Adicionalmente, el algoritmo requiere, para una gestión eficiente de consumo de memoria, que los ítems de una secuencia estén representados por un número entero positivo. Por ello los ítems se generaron con el ID del atributo y el ID de la clase a la que pertenece, como se muestra en la Tabla 10. Los ítems que correspondan a un mismo itemset, deben ser separados por un espacio. El separador de itemsets es el valor "-1" y el de secuencias es "-2".

El formato de salida es en un archivo de texto, donde cada línea representa un *patrón secuencial.* El patrón, de igual forma separa los itemset con el valor "-1". Al final de cada línea se muestra el *soporte mínimo* del patrón buscado, es decir en cuántas secuencias ha aparecido. Es necesario recalcar que, en el presente trabajo se tienen 17 secuencias de entrada, que representan las comunidades donde se tomaron las muestras.

4.4.5. Visualización espacial

Los patrones obtenidos pueden ser mejor analizados a través de técnicas de visualización. Para ello, se implementó una herramienta web accesible desde la mayoría de los navegadores modernos, de manera que facilite a los expertos la interpretación y validación de los patrones frecuentes.

El sistema cumplió con los siguientes requerimientos:

- Los patrones se podrán filtrar por fechas, departamentos, poblaciones y/o altitudes.
- Se podrán visualizar los patrones frecuentes más importantes. Al seleccionarlos se deben marcar las poblaciones donde se encontró dicho patrón.
- Al seleccionar un punto en el mapa, se debe ver la información detallada de la zona.

En la Figura 5, se observa un bosquejo de la aplicación. En la parte superior de la ventana, se visualizan los filtros de fechas, departamentos, poblaciones y/o altitudes. En el bloque de la izquierda, la lista de los principales patrones frecuentes. En el centro, se puede explorar el

mapa con los puntos donde se recolectaron las muestras. En el bloque de la derecha, se encuentra la descripción por cada comunidad de datos. Estos espacios o *frames* serán descritos más adelante.

Figura 5. Bosquejo de Visualización de Patrones Frecuentes

Para la visualización se ha utilizado las siguientes herramientas, que se ejecutan del lado del cliente:

- **Google Maps API**: Para delimitar el mapa correspondiente al Bosque Seco Ecuatorial, ubicado al norte del Perú. Departamentos de Tumbes, Piura y Lambayeque.
- **D3 Js**: Para graficar la aparición de patrones frecuentes.
- **Jquery**: Para agregar efectos interactivos: color, forma, movimiento.

Como se observa en la Figura 6, la aplicación web tiene 4 *frames* (cuadros) principales.

- **Frame 1:** Contiene listas desplegables para filtrar la visualización por fechas, departamentos, poblaciones y altitudes.
- **Frame 2:** Contiene la lista de patrones frecuentes más importantes. Al hacer *click* en alguno de ellos, se resaltan los puntos en el mapa (Frame 3) donde se ha encontrado la presencia de este patrón frecuente.

- **Frame 3:** Contiene el Mapa de los departamentos de Lambayeque, Piura, Tumbes. Los puntos son representados por colores, de acuerdo a la leyenda. Al hacer *click* en un punto se muestra el detalle del patrón frecuente. La opción de *zoom* está activada en el mapa presentado en este *frame*.
- **Frame 4:** Muestra el detalle del patrón frecuente o muestra seleccionada.

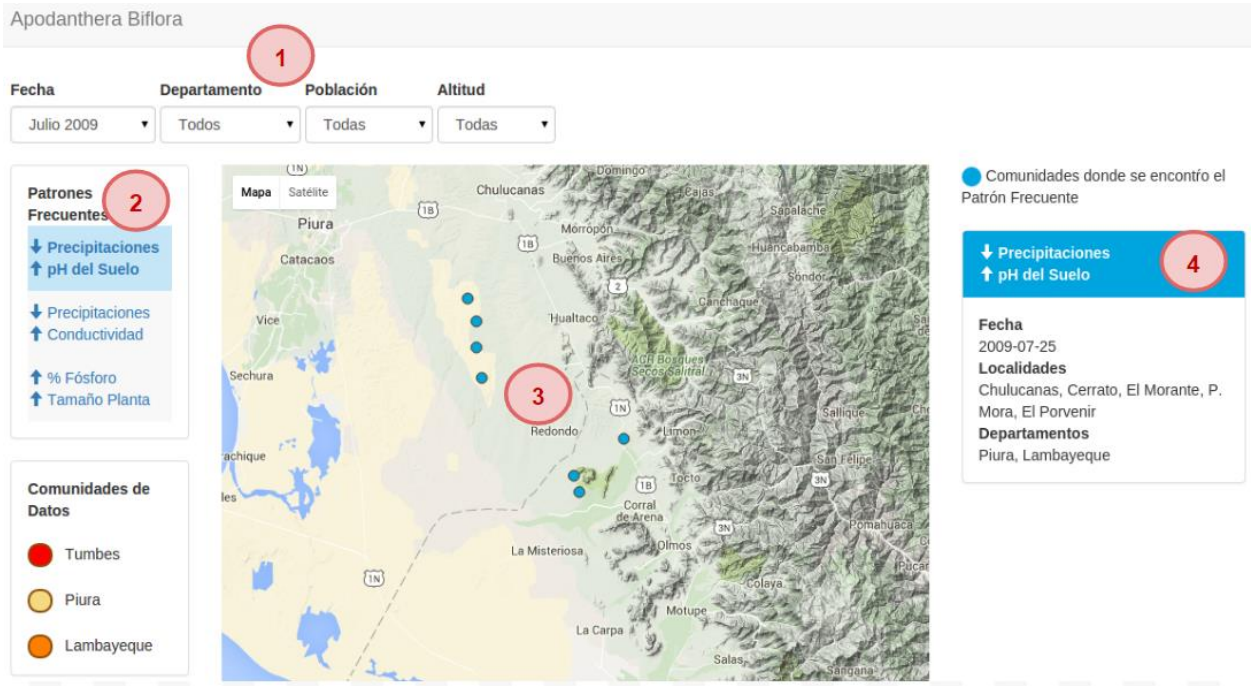

Figura 6. Aplicación Web de Visualización de Patrones Frecuentes

CAPÍTULO 5

RESULTADOS Y CONCLUSIONES

En el presente capítulo, se describen los resultados obtenidos al aplicar el proceso KDD sobre los datos descritos en el capítulo 4 del presente manuscrito.

5.1. INTERPRETACIÓN DE PATRONES FRECUENTES A PARTIR DE LA VISUALIZACIÓN

A partir de los resultados obtenidos, se encontraron patrones relevantes que valen la pena analizar. El formato de los resultados es:

itemset-**1** *itemset* **-1** *itemset* **-1** *itemset* **-1** #SUP: **10**, donde los *itemset* del patrón frecuente están separados por **"-1"** y el soporte mínimo (#SUP) es **"10"**, es decir que, de las 17 comunidades de datos, en más del 50% ocurren estos fenómenos.

A continuación, se muestra un ejemplo de patrón frecuente:
131 203 **-1** 203 **-1** 203 221 **-1** 242 **-1** 131 221 **-1** 131 161 191 221 242 271 **-1** #SUP: **10**

Debido a que algunos patrones frecuentes son muy largos, se utilizará el símbolo **[...]** para recortar la información no relevante. En lugar del separador **"-1"** se utilizará el símbolo →. Así mismo cada ítem es un número que representa un atributo y una clase. Por ejemplo, el número **131** representa el *Atributo ID = 13* y *Clase ID = 1.* La lista de atributos y clases se puede verificar en la Tabla 10.

a. Relación entre % precipitaciones y pH del suelo

142 -1 [...] →**142** 152*192* → [...] →**141** 163 [...], donde:

141: Precipitaciones menores a 29%; **142**: Precipitaciones mayores a 40%
152: pH estable entre 5 y 7; **153**: pH alcalino mayor a 7
192: Concentración de fósforo (P) mayor a 7%

Semántica: Se observa que en las zonas donde existe una mayor concentración de fósforo (P) y precipitaciones, el pH del suelo se encuentra entre los valores 5 y 7 (pH estable). En la misma secuencia se observa una disminución de las precipitaciones en el tiempo, ocasionando que el pH aumente. Esto posiblemente se debe a que el agua contribuye activamente en la neutralización del pH del suelo.

b. Relación entre ríos y conductividad del suelo

142 [...] **161 [...]** 262 →**162 [...]** 262 → [...] → [...], donde:

142: Precipitaciones mayores a 40%
161: Conductividad menor a 30; **162**: Conductividad mayor a 30 y menor a 60
262: Existen ríos cercanos alrededor de los 5 Km

Semántica: Se observa que las zonas con presencia de ríos cercanos tienen menor conductividad. Incluso cuando hay muchas precipitaciones la conductividad baja mucho más.

c. Relación entre características del tallo y concentración de fósforo (P)

[...] **191** → [...] →**191** [...] 211 241→ [...] →**192**212 →242, donde:

191: Concentración de fósforo (P) menor a 7%
192: Concentración de fósforo (P) mayor a 7%
211: Número de guías menor a 5; 212: Número de guías mayor a 5
241: Distancia entre nodos menor a 2.2 cm; 242: Distancia entre nodos mayor a 2.2 cm.

Semántica: Se observa que en las zonas con mayor concentración de fósforo (P), los tallos presentan una mayor distancia de entre nodos y número de guías. Esto quiere decir que la concentración de fósforo influye en el tamaño de la planta.

5.2. RENDIMIENTO Y EFICIENCIA DEL ALGORITMO PREFIXSPAN

Se ejecutó el algoritmo utilizando distintos valores de soporte mínimo. Los resultados en tiempo de ejecución, expresado en minutos, se muestran en la Figura 7. Como se puede observar, cuanto menor es el soporte mínimo, mayor es el tiempo de ejecución. Esto se debe a que, a menor valor de soporte mínimo, se hacen muchas más búsquedas de patrones en secuencias y subsecuencias. El tiempo máximo de espera fue de 225 minutos, utilizando un soporte mínimo de 10. Estos valores pueden optimizarse teniendo una computadora con mayor capacidad de procesamiento o ejecutando el algoritmo utilizando técnicas de *Cluster Computing*.

En cuanto al consumo de memoria y CPU, como se puede observar en las Figuras 8 y 9, es casi constante. El algoritmo requiere, aproximadamente, 312 MB de memoria y 25 % de CPU (Considerando un procesador Core i5 @ 1.7 GHz). Los resultados se van guardando en memoria y al final se almacenan en un archivo local en formato de texto.

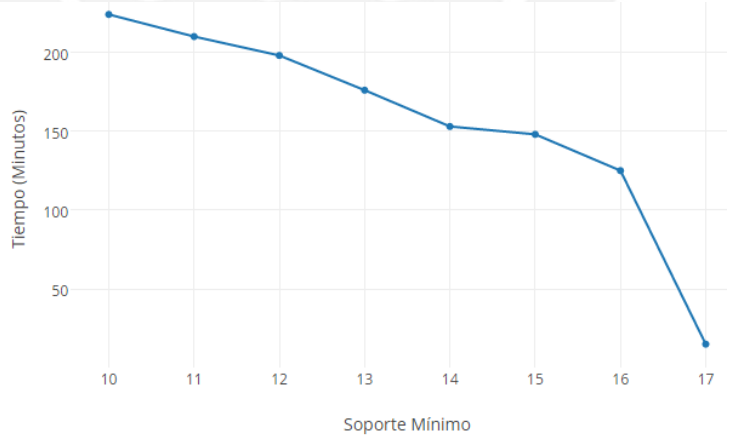

Figura 7. Comparación entre soporte mínimo y tiempo de ejecución del algoritmo

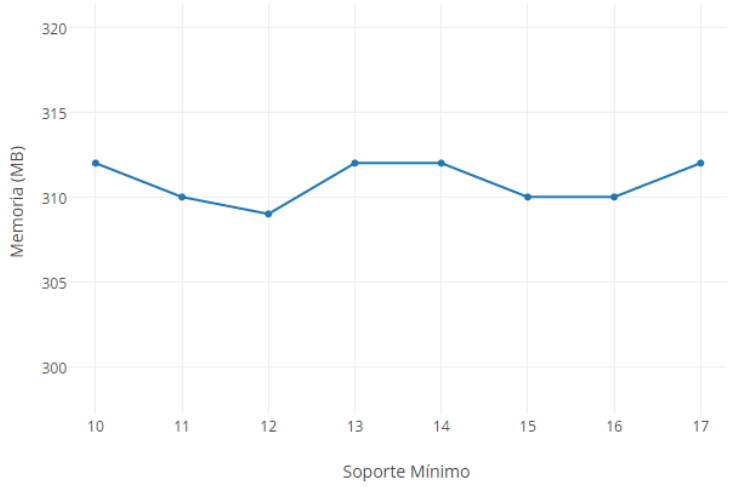

Figura 8. Comparación entre soporte mínimo y consumo de memoria (MB) del algoritmo

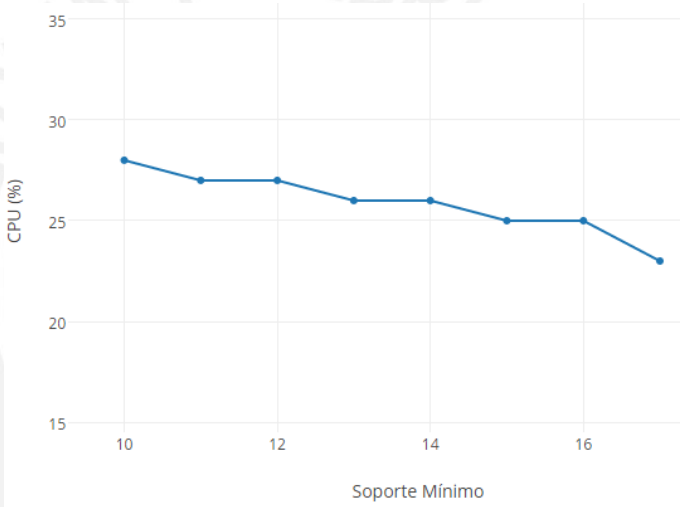

Figura 9. Comparación entre soporte mínimo y consumo de CPU (%) del algoritmo

Como se observa en la figura 10, se encontraron más de un millón de *patrones secuenciales frecuentes*. Esto se debe a que algoritmo por defecto busca los patrones frecuentes en subsecuencias, ocasionando que la cantidad de coincidencias aumente. Esto se puede optimizar utilizando algoritmos de minería de patrones secuenciales cerrados. Un patrón secuencial S es cerrado si no existe otro patrón S' con el mismo soporte tal que S⊂ S'. Identificar directamente patrones cerrados permite reducir el número de candidatos considerados.

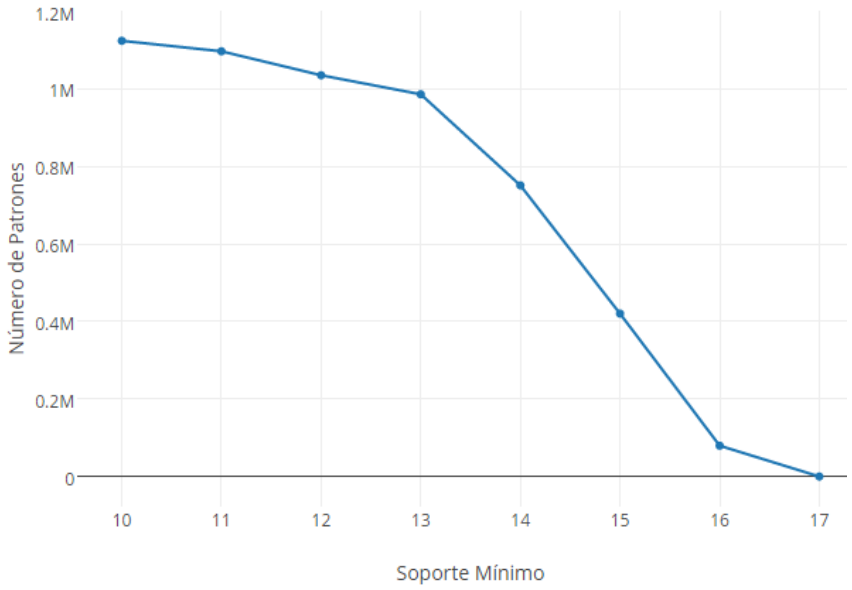

Figura 10. Comparación entre soporte mínimo y número de patrones obtenidos.

5.3. CONCLUSIONES

El presente trabajo surgió por la necesidad de contar con una herramienta que utilice técnicas de minería de datos y visualización, para descubrir patrones frecuentes relevantes a partir de fuentes de datos heterogéneos relacionadas a las características de la *Apodanthera Biflora* en las distintas zonas del Bosque Seco Ecuatorial (Estudio realizado por el Laboratorio de Ecología Evolutiva - UPCH) y condiciones meteorológicas (Fuente: SENAMHI) y geológicas (Fuente: INEI, MINAM) cambiantes en el tiempo.

Se puede afirmar, como conclusión general, que las técnicas de minería de patrones secuenciales y visualización espacial han permitido la obtención de patrones frecuentes relevantes a partir de distintas fuentes de datos heterogéneos relacionados a la especie. Entre los patrones frecuentes más relevantes encontramos la relación entre la disminución de precipitaciones y el aumento de pH del suelo; la relación entre mayor presencia de ríos cercanos y la baja conductividad del suelo; y la relación entre mayor distancia de entre nodos y alta concentración de fósforo.

De acuerdo a los resultados obtenidos en el capítulo anterior, específicamente:

- Las técnicas de transformación espacial GIS, han permitido almacenar los datos geográficos en un formato estándar (WGS84) para su posterior análisis y validación. Las técnicas de discretización por entropía y frecuencia, han permitido agrupar las 17 características más significativas en categorías.

- La técnica de minería de patrones secuenciales *PrefixSpan*, ha permitido extraer patrones frecuentes a partir de las características pre procesadas relacionadas a la *Apodanthera Biflora*. Se evaluaron distintos soportes mínimos, determinando que, a partir de 10, se obtienen los patrones frecuentes más interesantes. El tiempo de ejecución del algoritmo, utilizando este mismo soporte mínimo, es de aproximadamente 225 minutos. Así mismo se determinó que se requiere 312 MB de Memoria RAM y 25 % de CPU (Considerando un procesador Core i5 @ 1.7 GHz).

- Las técnicas de visualización espacial, utilizando librerías como *Google Maps API y D3 Js*, ha permitido la identificación de patrones frecuentes relevantes. Se obtuvieron más

de mil patrones frecuentes, sin embargo, los expertos identificaron 3 patrones relevantes en las localidades de Chulucanas, Cerrato, El Morante, P. Mora y El Porvenir, ubicadas en los departamentos de Piura y Lambayeque.

5.4. TRABAJOS FUTUROS Y RECOMENDACIONES

El presente trabajo representa un primer paso en el uso de técnicas de minería de patrones secuenciales y visualización espacial para descubrir correlaciones relacionadas a la caracterización de la *Apodanthera Biflora.* El resultado de la evaluación demuestra que es posible encontrar patrones frecuentes relevantes que permitan a los expertos complementar sus estudios sobre esta especie, es por ello que se propone, como trabajo a futuro, una investigación multidisciplinaria que permita a los biólogos utilizar estas herramientas para descubrir nuevas características de la especie.

Con el fin de enriquecer la minería de patrones secuenciales, se pueden agregar otro tipo de datos, como imágenes satelitales, de tal manera que se pueda enriquecer la información de los patrones obtenidos gracias a las características extraídas a partir de las imágenes, tales como, la temperatura, humedad, entre otras.

Los algoritmos de minería de patrones secuenciales requieren un mayor tiempo de ejecución cuando se tienen secuencias muy largas. Es por ello que se propone como trabajo futuro el uso de *Cluster Computing*, de tal manera que se pueda aprovechar la capacidad de cómputo de múltiples nodos en simultáneo. Se debe tener ciertas consideraciones como un sistema de archivos compartidos entre todos los nodos, así como el uso de un servidor principal con la capacidad para recibir y procesar los resultados de los nodos.

Así mismo se propone tener un repositorio centralizado de datos meteorológicos y geológicos para apoyar a futuras investigaciones que lo requieran. Los datos deben almacenarse en un formato estándar (Base de Datos) y de ser posibles normalizados.

Finalmente, se requiere que la aplicación de visualización esté disponible en internet y funcionando correctamente para que los expertos puedan tener acceso todo el tiempo a los resultados obtenidos.

BIBLIOGRAFÍA

Agrawal, R., &Srikant, R. (1995, March). Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on* (pp. 3-14). IEEE.

Alatrista, H., Bringay, S., Flouvat, F., Selmaoui-Folcher, N., &Teisseire, M. (2013). A Spatial-based KDD Process to Better Understand the Spatiotemporal Phenomena. In CAiSE'2013: 25th International Conference on Advanced Information Systems Engineering (Vol. 1001). CEUR-WS. org.

Alatrista-Salas, H., Azé, J., Bringay, S., Cernesson, F., Selmaoui-Folcher, N., &Teisseire, M. (2014). A knowledge discovery process for spatiotemporal data: Application to river water quality monitoring. *Ecological Informatics*.

AOAC. Official methods of Analysis of AOAC, 13AOAC (1980). Official methods of Analysis of AOAC, 13th edition. Washington D.C, USA. *Association of Official Analytical Chemists.*

Asher, L., Collins, L. M., Ortiz-Pelaez, A., Drewe, J. A., Nicol, C. J., & Pfeiffer, D. U. (2009). Recent advances in the analysis of behavioural organization and interpretation as indicators of animal welfare. Journal of the Royal Society Interface, 6(41), 1103-1119.

Bae, M. J., & Park, Y. S. (2015). Characterizing the effects of temperature on behavioral periodicity in golden apple snails (Pomaceacanaliculata). *EcologicalInformatics*.

Baralis, E., Bruno, G., Chiusano, S., Domenici, V. C., Mahoto, N. A., & Petrigni, C. (2010, September). Analysis of medical pathways by means of frequent closed sequences. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (pp. 418-425). Springer Berlin Heidelberg.

Brack-Egg, E. (1986). Las ecorregiones del Perú. *Boletín de Lima*, *44*, 57-70.

Clark, D., Tupa, M., Bazán, A., Chang, L., & Gonzáles, W. L. (2012). Chemical composition of Apodantherabiflora, a Cucurbit of the dry forest in northwestern Peru. Revista Peruana de Biología,19(2), 199-203.

De Boer, F. K., &Hogeweg, P. (2012). Co-evolution and ecosystem based problem solving. *Ecological Informatics*, *9*, 47-58.

Fabrègue, M., Braud, A., Bringay, S., Grac, C., Le Ber, F., Levet, D., &Teisseire, M. (2014). Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. *Ecological Informatics*,*24*, 210-221.

Fayyad U., &Irani K (1993). Multi-interval Discretization of Continuous-Valued Attributes for Classification Learning, In Proceedings of the 13th International JOint Conference on ARtificial Intelligence, Proceedings of the Fifth SIAM International Conference on Data Mining, Volume 119.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

Font Y. (2013). Análisis comparativo de algoritmos utilizados en la minería de secuencias frecuentes. Tesis para optar el grado de Máster en Cibernética Aplicada. ICIMAF. La Habana – Cuba.

Fournier-Viger, P., Gomariz, Gueniche, T., A., Soltani, A., Wu., C., Tseng, V. S. (2014). SPMF: a Java Open-Source Pattern Mining Library. Journal of Machine Learning Research (JMLR), 15: 3389-3393.

Global Positioning System Interface Specification. (1984). World Geodetic System 1984 (WGS84)

Gonzáles, W. (2009). Domesticación de especies silvestres de los bosques secos con potencial agroindustrial. *Innovación Agraria en el Norte del Perú: Integración de redes y cadenas productivas para la innovación.* 1-37.

Kwakkel, J. H., Carley, S., Chase, J., & Cunningham, S. W. (2014). Visualizing geo-spatial data in science, technology and innovation. *Technological Forecasting and Social Change*, *81*, 67-81.

Lee, A. J., Chen, Y. A., &Ip, W. C. (2009). Mining frequent trajectory patterns in spatial–temporal databases. *Information Sciences*, *179*(13), 2218-2231.

Leong, K., & Chang, S. (2012). STEM: A novel approach for Spatiotemporal Sequence Mining. *Asian Journal of Information Technology* 11(3): 94-99.

Liao, V. C. C., & Chen, M. S. (2014). DFSP: A Depth-First SPelling algorithm for sequential pattern mining of biological sequences. *Knowledge and information systems*, *38*(3), 623-639.

Open Source Geospatial Foundation. (2015). GeoServer. Recuperado en: http://docs.geoserver.org/

OSGeoFoundation. (2015). PostGISDocumentation. Recuperado en: http://postgis.net/documentation

Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., & Hsu, M. C. (2001, April). Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)* (pp. 0215-0215). IEEE Computer Society.

Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., ... & Hsu, M. C. (2004). Mining sequential patterns by pattern-growth: The prefixspan approach. *Knowledge and Data Engineering, IEEE Transactions on*, *16*(11), 1424-1440.

Poelen, J. H., Simons, J. D., &Mungall, C. J. (2014). Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *EcologicalInformatics*, *24*, 148-159.

QGIS Project. (2015). QGIS. Recuperado en: http://www.qgis.org/en/docs/

Rojas-Fox, J. (2012). "Patrones de variación fenotípica de la Apodanthera Biflora Gogn. (Cucurbitaceae) en el bosque estacionalmente seco del norte del Perú". Universidad Nacional Federico Villarreal. Lima -Perú.

Srikant, R., &Agrawal, R. (1996). *Mining sequential patterns: Generalizations and performance improvements* (pp. 1-17). Springer Berlin Heidelberg.

Sunitha, G., & Mohan, R. (2014). Mining frequent patterns from Spatiotemporal data sets: A Survey. *Journal of Theoretical and Applied Information Technology.* 265-274.

Tadesse, T., Wilhite, D. A., Harms, S. K., Hayes, M. J., & Goddard, S. (2004). Drought monitoring using data mining techniques: A case study for Nebraska, USA. Natural Hazards, 33(1), 137-159.

Wang, K., Xu, Y., & Yu, J. X. (2004, November). Scalable sequential pattern mining for biological sequences. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management* (pp. 178-187). ACM.

Wang, L., Ren, Z., Kim, H., Xia, C., Fu, R., & Chon, T. S. (2014). Characterizing response behavior of medaka (Oryziaslatipes) under chemical stress based on self-organizing map and filtering by integration. *Ecological Informatics*.