Instituto Tecnológico de Costa Rica
Escuela de Ingeniería en Computación
Programa de Maestría en Computación.

Cubic Spline Interpolation as a distance measure used in the discovery of meaningful rules from complex time series in presence of noise.

Cubic Spline Interpolation como medida de distancia utilizada en el descubrimiento de reglas significativas en series temporales complejas y en presencia de ruido

Propuesta de tesis sometida a consideración del Departamento de Computación, para optar por el grado de *Magíster Scientiae en Computación*, con énfasis en Ciencias de la Computación.

Autor:
David Elías Alfaro Barboza

Profesor Asesor:
Luis Alexánder Calvo Valverde

Julio 2016

Abstract

The ability to make short or long term predictions is at the heart of much of science. In the last decade, the data science community have been highly interested in foretelling about real life events by using data mining techniques to find out meaningful rules from different data types, including *Time Series*. Short-term predictions based on *the shape* of meaningful rules might lead to a vast number of applications. The discovery of *meaningful* rules can only be achieved as a result of algorithms equipped with a robust and accurate distance measure, capable to deal with noise in order to get the best possible similarity results between the elements of the time series. In this work, we believe that *Cubic Spline Interpolation* can be used as an efficient distance measure, to carry out the similarity computation in two specific algorithms: 1- *"Rule Bit Saves"* and 2- *"Find Antecedent Candidates"*, which were proposed by Mohammad Shokoohi-Yekta et al, to discover meaningful rules from complex time series, in presence of noise.

Resumen

La capacidad de hacer predicciones de largo o de corto plazo, ha estado siempre en el corazón de la ciencia. En la última década, la comunidad científica de datos ha mostrado un gran interés en vaticinar eventos de la vida real, mediante el uso de técnicas de minería de datos, para hallar reglas significaticas a partir de diversos tipos de datos, incluyendo el análisis de series temporales. Las predicciones basadas en *la forma* de la regla significativa puede dar lugar a un amplio número de aplicaciones. El descubrimiento de reglas significativas solo puede alcanzarse como resultado del uso de algoritmos equipados con una medida de distancia robusta y precisa, capaz de lidiar con el ruido, para obtener los mejores resultados de similitud posibles entre los elementos de las series temporales. En este trabajo, creemos que *Cubic Spline Interpolation* puede ser usado como una medida de distancia efficiente, para llevar a cabo el cómputo de la similitud en dos algoritmos específicos: 1- *"Rule Bit Saves"* Y 2- *"Find Antecedent Candidates"*, los cuales, fueron propuestos por Mohammad Shokoohi-Yekta et al, para el descubrimiento de reglas significativas en series temporales complejas, en presencia de ruido.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

1 Lista de Tablas

Tabla 1: Seudocódigo del algoritmo “Find Antecedent Candidates”.

Procedure *find_Antecedent_Candidates* (***T***, ***R***, ***sp***) Input: A time series subsequence, ***R***, extracted from a time series, ***T***; Split point for the antecedent/consequent, a number between zero and one, ***sp***; Output: locations of antecedents in ***T*** ordered by distances from ***R***'s antecedent, ***ac***;	
1	*antecedentLength* ← Length(*R*) × *sp*
2	*antecedent* ← *R* (1:*antecedentLength*)
3	*Distances* ← Euclidean (*antecedent*, each subsequence in *T*)
4	*AntecedentDistances* ← sort (localMinimums (*Distances*))
5	*AntecedentCandidates* ← Locations (*AntecedentDistances)*
6	*ac* ← *AntecedentCandidates*
7	**Return** *ac*

Tabla 2: Seudocódigo del algoritmo "Rule Bit Save".

Procedure *Rule_Bit_Saves* (***T***, ***R***, ***sp, n, ac, mlag***)
Input: A time series subsequence, ***R***, extracted from a time series, ***T***;
Split point for the antecedent/consequent, between zero and one, ***sp***;
The Number of instances of ***R*** to pick in the time series, ***n***; locations of antecedents in ***T*** ordered by distances from ***R***'s antecedent, ***ac;***
Maxlag allowed between the antecedent and consequent, ***mlag***;
Output: ***totalBitSave***;

```
1  antecedentLength ← Length(R) × sp
2  consequent ← R (antecedentLength:end)
3  Discretize and z-normalize (consequent)
4  AntecedentCandidates ← ac
5  totalBitSave ← 0
6  antecedentsSelected ← AntecedentCandidates (2 : n)
7    for i ← 1 to Length(antecedentsSelected) do
8      s1 ← AntecedentCandidates(i) + antecedentLength +1
9      s2 ← AntecedentCandidates(i) + Length(R)
10     subConsequent ← T (s1 : s1 + mlag)     // considering maxlag
11     consequentDist ← Euclidean (consequent, each subsequence
12                                  of subConsequent)
13     conseqLoc ← find (consequentDist == min(consequentDist))
14     subConsequent ← T (s1 + conseqLoc : s2 + conseqLoc)
15     Discretize and z-normalize (subConsequent)
16     subConsequentBits ← Huffman (subConsequent)
17     subConsequentMDLbits←MDL(subConsequent, consequent)
18     totalBitSave ← totalBitSave + subConsequentBits -
19                    subConsequentMDLbits
20   end for
21 Return totalBitSave – Huffman (consequent)  // Eq. 2
```

Tabla 3: Listado de entregables, objetivos relacionados y duración

Entregable	Objetivos	Duración (Dado en Semanas)
Modificación del software utilizado para la identificación de reglas motif y el hallazgo de reglas significativas	Incorporar todas medidas de distacia propuestas en ambos algoritmos.	4
Preprocesamiento de datos		1
Desarrollo e implementación de un ambiente de pruebas	Implementar un ambiente de pruebas que permita la ejecución de las diferentes versiones de ambos algoritmos.	3
Documento final con la recopilación y el análisis del diseño de experimentos	Ejecución del diseño de experimentos preparado, incluyendo la medición de las métricas definidas, para determinar si hay diferencias significativas entre los algoritmos.	2
Documento final del análisis de varianza no paramétrico y caracterización de los resultados obtenidos	Ejecutar y reportar el análisis de varianza no paramétrico, para aceptar o rechazar la hipótesis.	2
Desarrollo de un artículo científico	Preparar un artículo científico para ser presentado en una alguna revista afin al tema desarrollado en esta propuesta.	2
Documento final de tesis		2

Tabla 4: Cronograma

Entregable	Semanas															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Modificación del software utilizado para el hallazgo de reglas significativas																
Preprocesamiento de datos																
Desarrollo e implementación de un ambiente de pruebas																
Documento final con la recopilación y el análisis del diseño de experimentos																
Documento final del análisis de varianza no parametrico y caracterización de resultados																
Desarrollo de un artículo científico																
Documento final de tesis																

2 Lista de Figuras

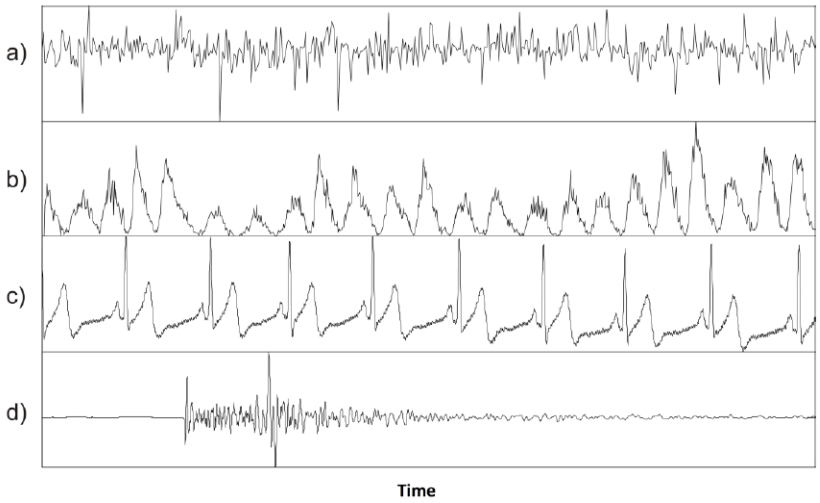

Figura 1: Ejemplos de Series Temporales

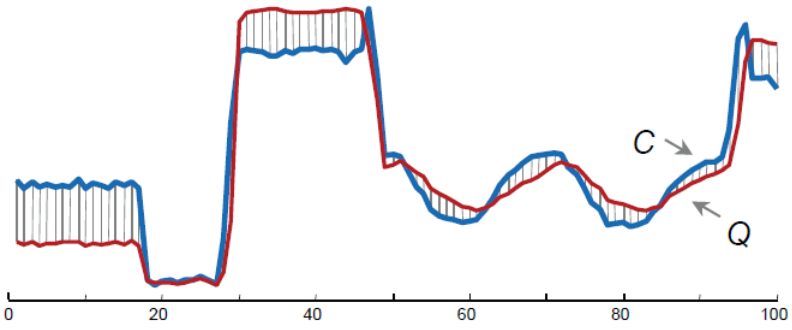

Figura 2: Visualización de la distancia Euclidiana de dos series temporales.

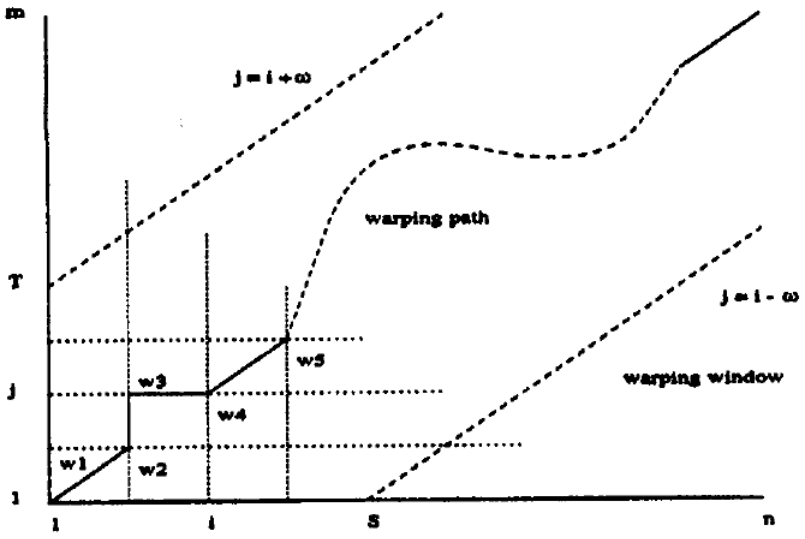

Figura 3: Visualización de la minimización de una ruta en DTW.

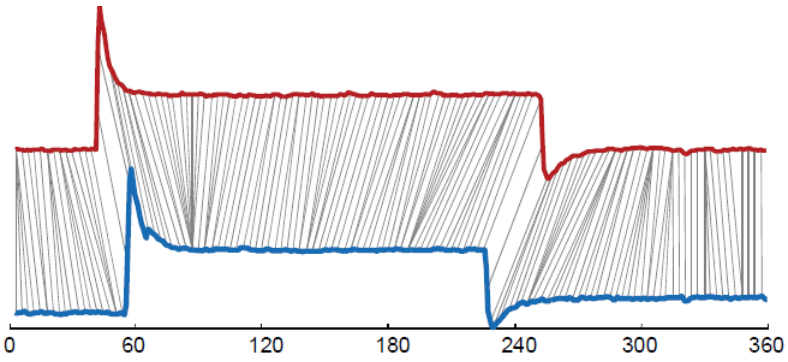

Figura 4: Visualización del calculo de DTW.

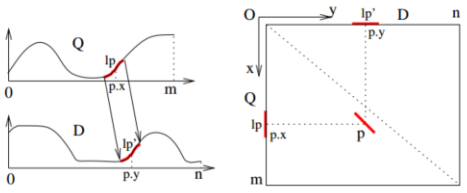

Figura 5: Ejemplo del cálculo de LPM.

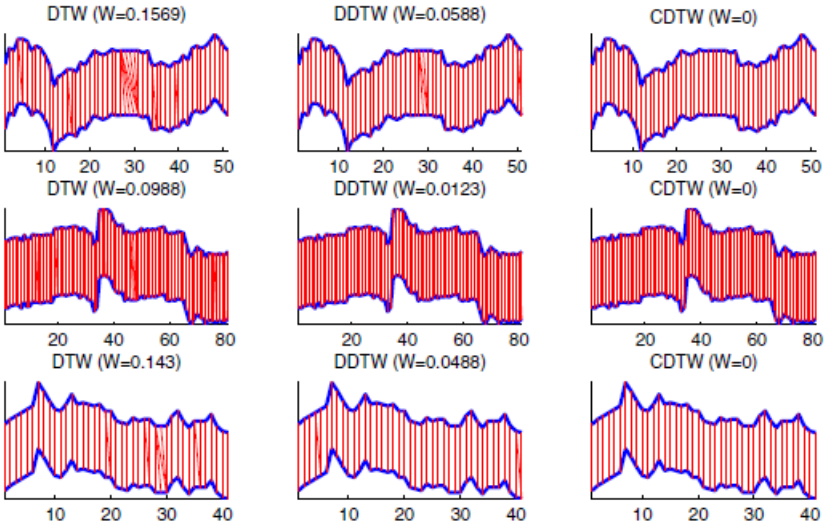

Figura 6: Visualización de deformaciones innecesarias en DTW.

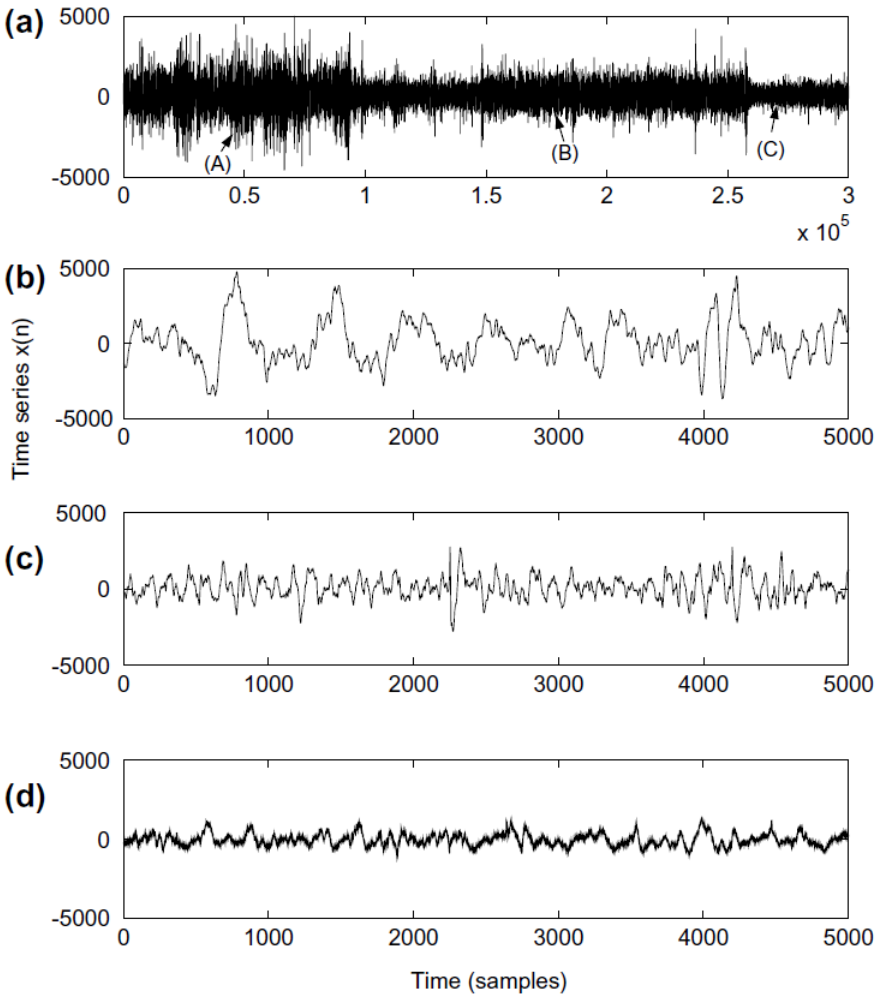

Figura 7: Ejemplo de ruido en las señales de un electoencefalograma.

3 Introducción

La habilidad de hacer predicciones acerca de acontecimientos o eventos de la vida real ha sido siempre un tema de gran interés para la ciencia.

En la última década, la comunidad de minería de datos se ha interesado vehementemente en el hallazgo de patrones o reglas que puedan ser útiles en la predicción de eventos de corto y de largo plazo [1].

La mayoría de trabajos de investigación recientes, orientados en la predicción de eventos de corto plazo mediante series temporales, se han enfocado principalmente en el análisis de los ***valores actuales*** del flujo de datos [13][18]. Sin embargo, en una basta cantidad de casos, el análisis de los valores actuales es irrelevante; en su lugar, la ***forma*** actual del patrón o la regla motivo y la detección oportuna en el flujo de datos, pueden ayudar a anticipar la ocurrencia de eventos futuros con mayor precisión [1].

Este trabajo de investigación tiene como objetivo principal, la implementación de la medida de distancia llamada *Cubic Spline Interpolation* en los algoritmos ***"Rule Bit Saves"*** y ***"Find Antecedent Candidates"***, utilizados respectivamente en el hallazgo y la detección de reglas significativas, para llevar a cabo predicciones de corto plazo, sobre series temporales complejas y en presencia de ruido.

Las predicciones de corto plazo sobre series temporales han tenido un auge importante, su aplicación y alcance se ha diversificado considerablemente. Las predicciones de corto plazo sobre texto durante las pulsaciones del teclado, predicciones sobre consultas de base de datos [22], predicciones sobre intervenciones médicas [23], son solo algunos ejemplos de predicciones sobre objectos discretos.

Recientemente, ha surgido una reto aún mayor; se requiere un mayor poder predictivo, lo que implica necesariamente la implementación de algoritmos de predicción mucho más precisos, más veloces y que puedan hallar patrones sobre conjuntos de datos mucho más grandes y complejos [19]. Por ejemplo, el radar Doppler utilizado en las últimas dos décadas para la detección de tornados, ha incrementado el tiempo promedio de alerta de 5.3 a 9.5 minutos, salvando un incontable número de vidas hu-

manas año con año. Sin embargo, aún se reportan alrededor de un 26% de tornados que no pueden predecirse mediante el uso de la tecnología existente [20]. McGovern et al. argumentan en [21], que las nuevas mejoras no vendrán necesariamente de sensores más sofisticados, sino, de algoritmos de predicción aún no inventados o algoritmos existentes aún no depurados, capaces de examinar series temporales complejas, para hallar reglas predictivas mucho más precisas y fiables.

La presente propuesta de investigación se encuentra distribuída de la siguiente manera: inicialmente, en la sección cuatro, se desarrollá el marco teórico como un grupo de ideas y conceptos fundamentales, que tienen como objetivo principal, guiar e involucrar al lector en el contexto de esta propuesta de investigación. En la sección cinco, se exponen los detalles más importantes de la propuesta de investigación, tales como el planteamiento del problema, la hipótesis, las métricas utilizadas y del por qué la importancia de llevar a cabo esta investigación. El objetivo general y los objetivos específicos de esta propuesta, se ofrecen en las secciones seis y siete respectivamente. El alcance y las limitaciones, serán puntualmente detalladas en la sección ocho, mientras que los entregables serán enlistados en la sección nueve. Por otra parte, en la sección diez, se describirá el diseño experimental y el ambiente de desarrollo que le darán forma a la metodología utilizada. Finalmente, en la sección once, se presenta el cronograma de actividades establecido, para la llevar a cabo la realización de este proyecto de investigación.

4 Marco Teórico

En la última década, ha surgido un enorme interés acerca de llevar a cabo tareas de mineria de datos sobre series temporales [34]. Literalmente, cientos de artículos han introducido nuevos algoritmos para indexar, clasificar, segmentar y agrupar series temporales [24]. Es por lo anterior, que iniciamos el desarrollo del marco teórico definiendo series temporales, como una pieza de conocimiento angular, para la comprensión y la defición del contexto de nuestra propuesta de tesis.

4.1 Series Temporales

Las series temporales, pueden definirse como: *"una secuencia de N observaciones (datos) ordenadas y equidistantes cronológicamente, sobre una característica (serie univariable o escalar) o sobre varias características (serie multivariable o vectorial), de una unidad observable, en diferentes momentos"* [25].

Una representación matemática común de una serie temporal *univariable* puede verse de la siguiente manera:

$y_1, y_2, ..., y_N; (y_t)^N t = 1; (y_t : t = 1, ..., N)$, donde y_t es la observación n^0 $t(1 \leq t \leq N)$ de la serie y N es el número de observaciones que componen la serie completa (el tamaño o la longitud de la serie) [27].

En el siguiente gráfico, se pueden observar algunos ejemplos de las diferentes formas que pueden adoptar las series temporales, con respecto a la naturaleza de la fuente de datos que se está representando. Algunas son más predecibles y constantes, mientras que otras fluctuan con mayor frecuencia en amplitud o simplemente presentan mayor cantidad de ruido.

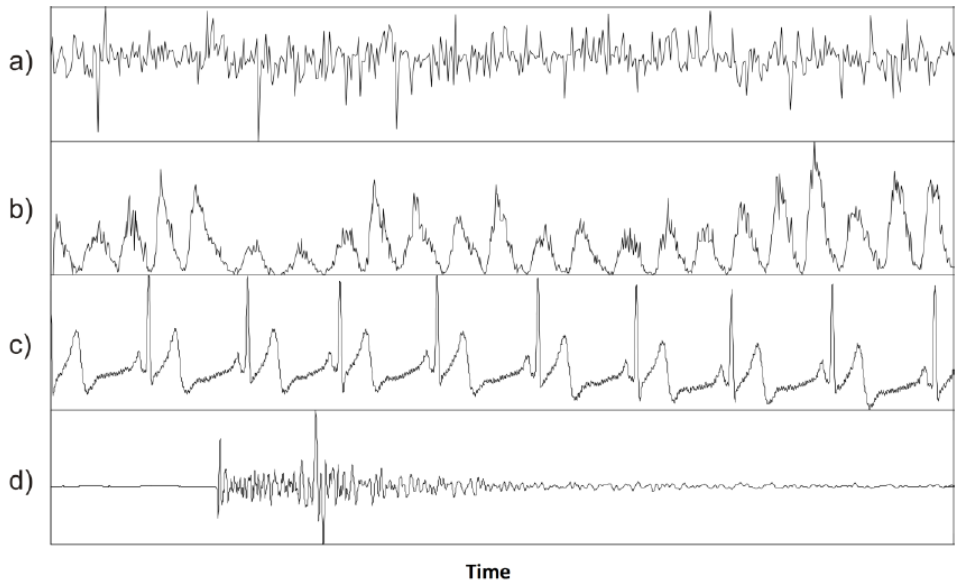

Figura 1. Ejemplo de la representación en series temporales de: **a)** Monzón, **b)** Manchas Solares, **c)** Electrocardiograma, **d)** Señales Sísmicas. **Fuente:** [27].

A diferencia de las series temporales no estarionarias ilustradas en el gráfico anterior, muchas series temporales tienen una tendencia creciente, por ejemplo, el número de automóviles en uso de un país, durante los últimos cincuenta años, o decreciente como el número de personas que trabajan en la agricultura. Existen muchas otras, sin embargo, que no tienen tendencia y son estacionarias, por ejemplo, la luminosidad a horas sucesivas, que varía cíclicamente a lo largo de las 24 horas del día [27].

La representación matemática más frecuente de una serie temporal multivariable, puede definirse de la siguiente manera:
$y_1, y_2, ..., y_N; (y_t)^N t = 1; (y_t : t = 1, ..., N)$, donde $y_t \equiv [y_t 1, y_t 2, ..., y_t M]'(m \geq 2)$ es la observación n^0 $t(1 \leq t \leq N)$ de la serie y N es el número de observaciones que conforman la serie completa.

Las observaciones pueden ser almacenadas en una matriz Y de orden NxM, como se muestra a continuación en la siguiente imagen:

$$Y \equiv \begin{bmatrix} y'_1 \\ y'_2 \\ . \\ . \\ . \\ y'_N \end{bmatrix} \equiv \begin{bmatrix} y'_{11} & y'_{12} & . & . & . & y_{1M} \\ y'_{21} & y'_{22} & . & . & . & y_{2M} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ y'_{N1} & y'_{N2} & . & . & . & y_{NM} \end{bmatrix} \quad (1)$$

donde y_{tj} es la observación n^0 $t(1 \leq t \leq N)$ sobre la característica o variable n^0 $j(1 \leq t \leq N)$, que es la misma en todo momento t [27].

Una vez definido el concepto de series temporales y como se explicará más adelante durante el desarrollo de este documento, la presente propuesta de investigación, se enfoca principalmente en el uso de "Cubic Spline Interpolation" como medida de distancia incorporada en dos algoritmos. Ambos algoritmos analizan los conjuntos de datos en forma de series temporales, para la identificación y la selección de "reglas significativas".

Una ***"regla significativa"*** debe entenderse como un patrón o "subsecuencia recurrente" de la serie de tiempo (usualmente llamado ***"motif"***), que describe un evento o comportamiento real del conjunto de datos que está siendo analizado. Una vez que el motif ha sido identificado, se debe analizar su capacidad predictiva en relación con el flujo de datos. Se debe tener presente entonces, el hecho de que no todas las reglas motif identificadas serán seleccionadas como significativas [1].

4.2 Análisis y Aplicaciones de Series Temporales

El análisis de las series temporales, en general, podría verse como la medición, el seguimiento y el estudio del comportamiento de algún fenómeno o actividad en el tiempo, con el objetivo final de encontrar conocimiento relevante [29]. El resultado de dicho análisis, es conocimiento de primera mano, que puede ser utilizado para

comprender mejor aquello que ha venido ocurriendo. Es posible también, a través del análisis, monitorear el estado actual de las series temporales, para explicar la relación causal o la estructura subyacente que producen los datos observados. Finalmente, es completamente viable además, realizar predicciones o pronósticos, que podrían proveer un control anticipado probable, sobre el fenómeno estudiado [1].

Por otra lado, existen innumerables aplicaciones e implementaciones existosas de la minería de datos sobre series temporales [27]. Por ejemplo, en el campo de la medicina, un epidemiólogo podría estar interesado en analizar, mediante reglas significativas, el número de casos de influenza observado durante un período de tiempo. De igual forma, a través de la trazabilidad y el análisis de las mediciones de la presión arterial de un individuo, se podrán anticipar complicaciones asociadas con mayor precisión, conduciendo a una mejor evaluación de los medicamentos utilizados en el tratamiento de ese mal [33]. De igual manera, los patrones resultantes de una resonancia magnética funcional de las ondas cerebrales (en forma de series temporales), pueden ser utilizados para vaticinar las reacciones cerebrales ante ciertos estímulos menores [33]. Finalmente, las aplicaciones más intensivas y sofisticadas en el análisis de series temporales, se enfocan en la resolución de problemas asociados a ciencias físicas y ambientales. Por ejemplo, el análisis de los niveles de contaminación de una región, el seguimiento de la temperatura global, el análisis de la actitividad sísmica, el reconocimiento de vos, entre muchos otros [33].

A continuación, se exponen algunas de las principales dificultades y retos que podrían presentarse durante las diferentes etapas del proceso de minería de datos sobre series temporales.

4.3 Grandes Retos Sobre la Minería de Series Temporales

El análisis y el descubrimiento de patrones sobre series temporales, por definición, presenta una serie de retos y complicaciones que deben abordarse, tales como: la super multidimensional, la presencia de grandes cantidades de datos que muchas veces resultan innecesarios o poco útiles durante el análisis, la sensibilidad al ruido o a la presencia de valores atípicos y el dinamismo durante la transmición de datos

o *"data streaming"*, debido a que se requiere en todo momento una gran capacidad de cómputo de datos, para analizar las fluctuaciones constantes sobre el conjunto de datos [1].

En adición a lo anterior, ***el cálculo de la similitud*** durante la comparación de dos o más series temporales es considerado un desafío aún mucho más crucial en la minería de series temporales, especialmente cuando se ha realizado previamente una reducción de la dimensionalidad, de la escala o la amplitud a través del tiempo. La carencia de un alineamiento oportuno sobre el eje tiempo o la amplitud, durante el cálculo de la similitud entre dos o más series temporales, es un problema serio debido a que ocasiona un impacto directo sobre el resultado de la comparación. [27].

El cálculo de la similitud es indispensable en la identificación de segmentos repetitivos, contenidos en la serie de tiempo. Dichos segmentos se denominan reglas ***"motifs"*** o *ocurrencias frecuentes de un subconjunto de la serie temporal* [1].

Los motifs pueden considerarse conocimiento significativo, cuando, como producto del análisis del flujo de datos de la serie temporal, se pueden utilizar para predecir un fenómeno o evento con una alta probabilidad [1].

La idenficación de reglas *motif* se logra fundamentalmente apartir del cálculo de las medidas de distancia entre los elementos de dos o más series temporales [1]; la precisión de dicho cálculo y el número de ocurrencias son vitales para determinar la calidad de dicha regla. Existen en la literatura, un número importante de medidas de distancia [9]. Las medidas de distancia más importantes para la comprensión de esta propuesta serán desarrolladas a continuación.

4.4 Medidas de Distancia para el Cálculo de la Similitud

Las medidas de similitud son de vital importancia cuando se ejecutan tareas de análisis y minería de datos sobre series temporales, tales como: descubrimiento de patrones, agrupamiento, clasificación, descubrimiento de reglas, analís de valores atípicos, entre otras [27].

Debido a la naturaleza númerica y continua de de los datos característicos de las series temporales, las medidas de similitud, típicamente se llevan a cabo en forma de

aproximaciones [9].

Los estudios existentes sobre series de tiempo se basan en cuatro categorías de funciones de distancia [16]. La primer categoría consiste en las *Lp-normas* o "*Lp-norms*". Son funciones de distancia métricas o lineales, que no soportar cambios en el desplazamiento local del eje tiempo, por ejemplo, la distancia Euclideana o la distancia Manhatthan. La segunda categoría, por otra parte, se conoce como *medidas elásticas*, y consisten en funciones de distancia capaces de lidiar con desplazamientos locales en el eje tiempo, pero no son métricas o lineales. La categoría tres, son medidas de distancia basadas en un umbral (por ejemplo TQuEST [9]). Por último, la categoría cuatro corresponde a todas aquellas medidas de distancia basadas en la *"forma"* o en un patrón de la serie temporal, tal es el caso de SpADe [35].

La complejidad inherente al cálculo de las medidas de similitud, imponen normalmente las principales limitaciones y restricciones de capacidad y tiempo de cómputo, sobre los algoritmos utilizados en el análisis y la minería de datos de series temporales [29]. Es decir, cuanto más rápido y preciso sea el cálculo de la medida de similitud definida en el algoritmo, menor será el tiempo de cómputo necesario para la ejecución del procedimiento completo de minería de datos sobre las series temporales.

Para el desarrollo de este proyecto, se utilizarán específicamente cinco medidas de distancia: 1- Distancia Euclidiana, 2- Swale, 3- Spade, 4- EPR y 5- Cubic Spline Interpolation. Cada una de las medidas de distancia anteriormente mencionadas se detallan a continuación, incluyendo además, Dynamic Time Warping (DTW) como un concepto previo fundamental, antes de exponer Cubic Spline Interpolation como medida de distancia.

4.4.1 Distancia Euclideana

La distancia Euclidiana es la *distancia "ordinaria" entre dos puntos de un espacio euclídeo* [30]. La distancia Euclideana tiene la característica de ser la medida de distancia más simple y la más utilizada para comparar series temporales [10]. Por ejemplo, si se require comparar dos series temporales Q y C, de largo n, donde $Q = q_1, q_2, ..., q_i, ..., q_n$ y $C = c_1, c_2, ..., c_i, ..., c_n$, se puede utilizar la distancia Euclidi-

ana obícua cuya fórmula matemática es la siguiente:

$$DE(Q,C) \equiv \sqrt[2]{\sum_{i=1}^{n} (q_i - c_i)^2} \tag{2}$$

Con se muestra en la **Figura 2**, la visualización del cálculo de la distancia Euclideana puede verse entonces como la raíz cuadrada de la suma de las diferencias al cuadrado; tal y como se observa en cada línea vertical para cada punto de datos desde C hasta Q y viceversa.

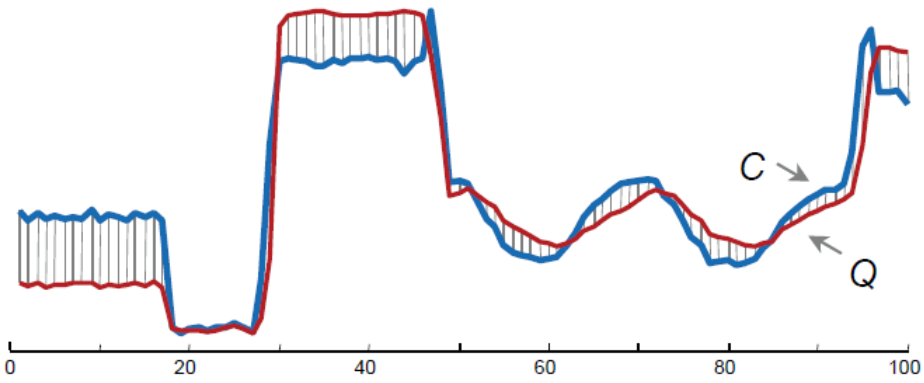

Figura 2. Visualización de la distancia Euclidiana entre dos series temporales C y Q.

Fuente: [30].

4.4.2 Distancia "Swale"

En [31], los autores proponen un modelo de similitud llamado ***"Swale"*** ("*Sequence Weighted ALignmEnt*", por sus siglas en Inglés); este modelo utiliza un sistema de puntuación para recompensar las similitudes y penalizar las disimilitudes durante la comparación de series temporales. El uso del marco de trabajo *Swale*, requiere necesariamente del ajuste de tres parámetros: 1- un umbral de similitud ϵ, 2- un valor o peso r utilizado para recompensar las similitudes y 3- un valor o peso p para penalizar los vacios o disimilitudes [31].

Más formalmente, la función de similitud de *Swale* puede definirse como:
Sean R y S dos series de tiempo de longitud m y n, respectivamente. Sea el costo de la diferencia o disimilitud gap_c y el costo de la similitud $reward_m$. Entonces dadas

dos series temporales R y S,

$Swale(R, S) =$

$$\begin{cases} n * gap_c, & \text{if m} = 0 \\ m * gap_c, & \text{if n} = 0 \\ reward_m+, & \text{if } \forall d, |r_d, 1 - S_d, 1| \leq \epsilon \\ Swale(Rest(R), Rest(S)), & \\ max\{gap_c + Swale(Rest(R), S), & \text{otherwise} \\ gap_c + Swale(R, Rest(S))\} & \end{cases}$$

El modelo Swale, posee una caraterística muy particular, ya que permite ajustar los pesos o los valores utilizados para castigar la disimilitud o bien, en sentido contrario, recompensar la similitud, con base en el criterio de un experto con conocimientos muy específicos de cierto dominio. Esto permite afinar la función de distancia explicada en la fórmula anterior, de manera tal, que pueda establecerse un rendimiento óptimo con base en la naturaleza de los datos, en lugar de tener una única técnica para cada dominio de datos [31].

4.4.3 Medida de distancia SpADe

En [35], se define Spatial Assembling Distance (SpADe por sus siglas en Inglés), como una medida de distancia capáz de manejar el desplazamiento y la amplitud en dimensiones de amplitud y tiempo. Adicionalmente, proponen una propuesta eficiente para la detección continua de patrones, utilizando SpADe, para evaluar la similitud de subsecuencias de series temporales sobre el flujo de datos.

Esta medida de distancia basa en la detección de la mejor combinación de un patrón de similitud local (Local Pattern Match LPM), mediante el cálculo de la ruta más corta en la matriz de similitud.

En la siguiente figura, se muestra un ejemplo del cómputo de LPM para dos segmentos de series de tiempo Q y D, en condiciones de desplazamiento y amplitud.

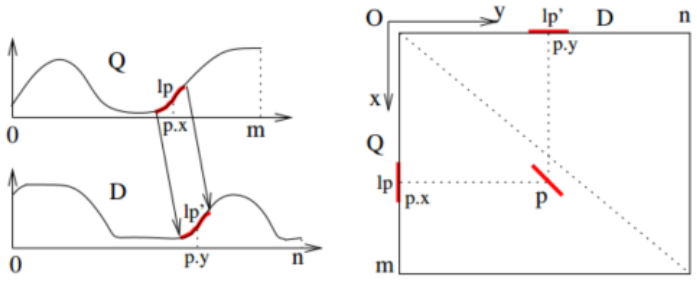

Figura 3. Ejemplo del cálculo de LPM.
Fuente: [31].

SpADe puede definirse matemáticamente como:

$$SDl(Dt, Q) = mini < teSD(D[t : i], Q) \tag{3}$$

Donde $SDl(Dt, Q)$ mide la distancia de la mejor similitud de la subsecuencia para (Q) iniciando en un punto t con respecto a D.

En resumen, SpaADe es una medida de distancia robusta basada en "la forma" de la subsecuencia de la serie temporal. En [35] se argumentan además, que no es sensible al desplazamiento o a la amplitud de las dimensiones durante el análisis del flujo de datos de la serie temporal y que por ende, alcanza mejores resultados con respecto a medidas de distancia más comunes tales como distancia Euclidiana o DTW.

4.4.4 Medida de distancia ERP

En [36], los autores proponen la distancia ERP ("**E**dit distance with **R**eal **P**enalty" por sus siglas en Inglés). ERP podría representarse de la siguiente manera, dadas dos series temporales R y S,

$Swale(R, S) =$

$$\begin{cases} \sum_{1}^{n} |s_i - g|, & \text{if m = 0} \\ \sum_{1}^{m} |r_i - g|, & \text{if n = 0} \\ min\{ERP(Rest(R), Rest(S)) + dist_{erp}(r_1, s_1), & \text{otherwise} \\ ERP(Rest(R), S) + dist_{erp}(r_1, gap), RRP(R, Rest(S) + dist_{erp}(S_1, gap)\} & \end{cases}$$

Donde g es un valor constante, normalmente igual a 0 según la recomendación de los autores en [36]. Los autores, definen ERP como una variante de *L1-norm*, con la excepción de que esta medida de distancia si es capaz de soportar el desplazamiento local en el eje tiempo, en otras palabras, lo autores lo consideran "el matrimonio perfecto" entre una medidas de distancia "L1-norm" y una medida de distancia de categoría elástica, ya que se asemeja a "L1-norm" en ser una función de distancia métrica, pero también, se asemeja a distancias elásticas como DTW, en su capacidad misma de poder soportar los desplazamientos o transiciones en el eje tiempo [36].

4.4.5 DTW (Dynamic Time Warping)

En el año 1994, Berndt y Clifford [32] proponen ***"Dynamic Time Warping"***. En el análisis de series de tiempo, DTW es un algoritmo para medir elásticamente la similitud entre dos secuencias temporales que pueden variar en velocidad y por ende en alineamiento, en relación con el eje tiempo [27].

A diferencia de la distancia Euclidiana, el cálculo de la similitud no se hace en forma lineal, por el contrario, las secuencias son "deformadas" de manera no lineal, para determinar la medida de similitud independiente con respecto a ciertas variaciones en el eje tiempo [32]. Por ejemplo, la tarea de detección de patrones implica la búsqueda de una serie de tiempo S, para instancias de una plantilla T, donde $S = s_1, s_2, ..., s_i, ..., s_n$ y $T = t_1, t_2, ..., t_i, ..., t_n$.

Las secuencias S y T, pueden ser acomodadas para conformar un plano de m por n o un cuadrante, en donde cada punto del cuadrante, (i, j), corresponde a un

alineamiento entre elementos s_i y t_i.

Un camino W, alinea los elementos que pertenecen a S y a T, tal que la distancia entre ellos es minimizada.

$$W = w_1, w_2, ..., w_k \tag{4}$$

Es decir, W es una secuencia o un camino específico de puntos en el cuadrante, en donde cada w_k corresponde a un punto $(i, j)_k$, como se muestra en la figura 3.

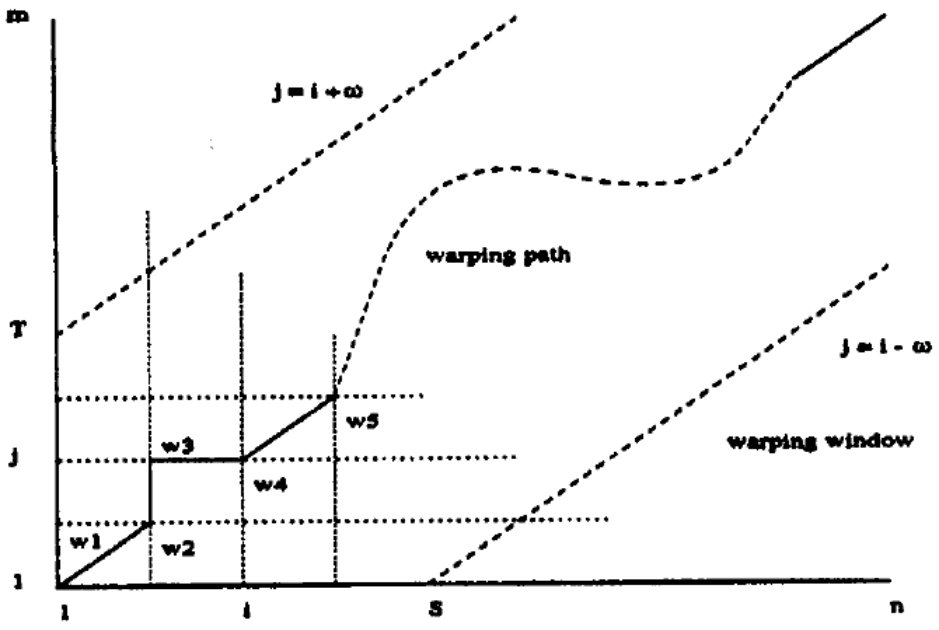

Figura 4. Visualización de un camino w en un cuandrante de m por n.
Fuente: [32].

Posteriormente, por definición, se debe definir una medida de distancia. En [32], los autores proponen $\delta(i, j) = (s_i - t_j)^2$. Una vez definida la medida de distancia, se puede definir formalmente DTW como la minimización sobre los caminos o rutas potenciales ***no lineales***, basados en la distancia acumulada para cada ruta, en donde δ es una medida de distancia entre dos puntos de datos o elementos de las series temporales.

$$DTW(S, T) = min_W[\sum_{k=1}^{p} \delta(w_k)] \tag{5}$$

En la Figura 4, se muestra observa un ejemplo claro de la forma que adoptan las

deformaciones no lineales calculadas mediante el algoritmo DTW, para adaptar la diferencia de los puntos de datos con respecto al eje tiempo.

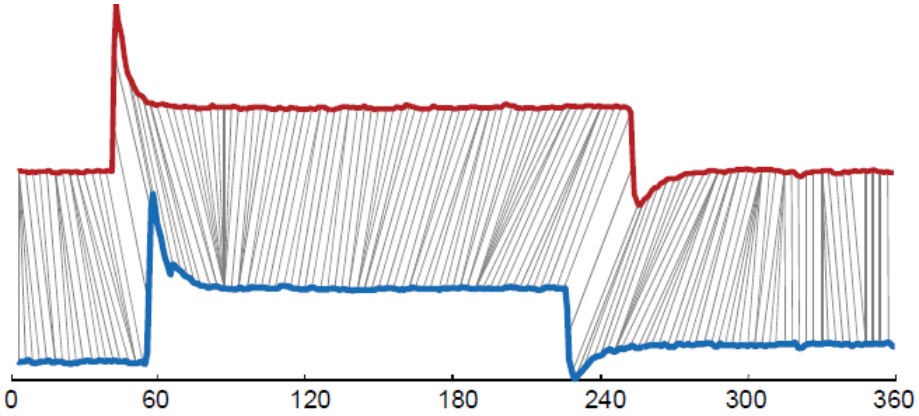

Figura 5. Visualización de dos segmentos de series temporales acerca del comportamiento de insectos se alinean con invarianza a la deformación computada mediante el uso de DTW. **Fuente:** [32].

4.4.6 DTW basado en Cubic Spline Interpolation

Dynamic **T**ime **W**arping basado en Cubic **S**pline **I**nterpolation (SIDTW), también conocido como "Cubic Dynamic Time Warping" (CDTW), fue propuesto recientemente como una novedosa y mejorada extensión de DTW [6].

Acerca de la función SIDTW:

Segun los autores en [6], inicialmente, se debe calcular la derivada de cada punto de la series de temporal mediante "Cubic Spline Interpolation", a este proceso se le conoce como "**interpolación**". Lo anterior es necesario para reemplazar las derivadas que fueron estimadas por Derivative Dymanic Time Warping (DDTW). Después de la interpolación, se utilizan las secuencias basadas en derivadas para representar las series de tiempo originales, para lograr una mejor descripción la tendencia original de las series temporales, haciéndolas más fácil de deformar ("warping") [6].

En [6] se argumenta que mediante el uso de esta medida de distancia, se pueden producir menores singularidades (menos "warping") y obtener en cambio, la mejor ruta de deformación posible, con el menor largo. Además, se considera una versión de medida de distancia alternativa de DTW cuando los conjuntos de datos de las series

temporales no son adecuados para ser medidos a través de DTW [6].

La principal motivación de SIDTW como propuesta es en esencia obtener derivadas mucho más precisas para reflejar mejor la tendencia de las series de tiempo que estan siendo comparadas y así, mejorar también la efectividad de la medida de similitud para las distancias acumuladas de tres elementos adyacentes [6].

Por ejemplo,

$$r(i,j) = d(i,j) + min \begin{cases} r(i,j-1) \\ r(i-1,j-1) \\ r(i-1,j) \end{cases}$$

Sin embargo, en DDTW, los pasos son los mismos que en DTW, además de la diferencia de los cuadrados entre q_i y c_j.

La estimación de la derivada $d'(i,j)$ en DDTW es utizada para reemplazar la distancia $d(i,j)$ en DTW. De esta forma, tenemos la siguiente derivada:

$$d'(i,j) = (d_i(q) - d_j(c))^2 \tag{6}$$

donde,

$$d_l(x) = \frac{((x_l - x_{l-1}) + (x_{l-1} - x_{l-1})/2)}{2} \tag{7}$$

Como se muestra en la siguiente imagen, una de las principales ventajas de CDTW es el hecho de que se pueden evitar la creación de deformaciones innecesarias.

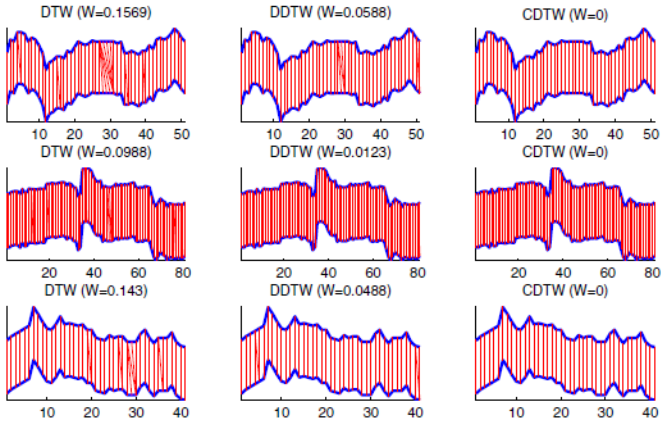

Figura 6. Visualización de deformaciones innecesarias en DTW. **Fuente:** [6].

En la imagen anterior, W corresponde a la cantidad de deformaciones producidas, en donde para DTW y DDTW, se observan valores mayores zero, mientras que en CDTW no produce deformaciones.

Los experimentos sobre conjuntos de datos de series temporales desarrollados en [6], reflejan que SIDTW puede reducir significativamente el número de singularidades y alinear más adecuadamente los puntos entre las dos series de tiempo (facilitando notablemente el cómputo de la similitud). En resumen, mientras más precisa sea la pendiente, más efectivo será también el cómputo de la deformación sobre las series de tiempo. Finalmente, cuando existen muchos puntos adyacentes con valores iguales en la serie de tiempo, los autores proponen el uso de **CDTW**, como una alternativa viable y superior (por ejemplo en comparación con DTW y DDTW) para evitar la producción de deformaciones innecesarias [6].

4.4.7 El Ruido y la Interpolación en Series Temporales

La recolección de los datos de una serie de tiempo se encuentra siempre degradada por ruido en cierto grado [12]. Incluso una estimación aproximada del nivel de ruido contenido en la serie temporal es crucial, previo al análisis de los datos.

En la práctica, las series de tiempo no estacionarias son muy comunes en campos diversos como la geofísica, las finanzas y las ciencias biológicas [27].

En el siguiente gráfico se pueden observar, por ejemplo, segmentos *no estacionarios* de señales de un electroencefalograma a partir de la medición clinica intracraneal de un individuo [12].

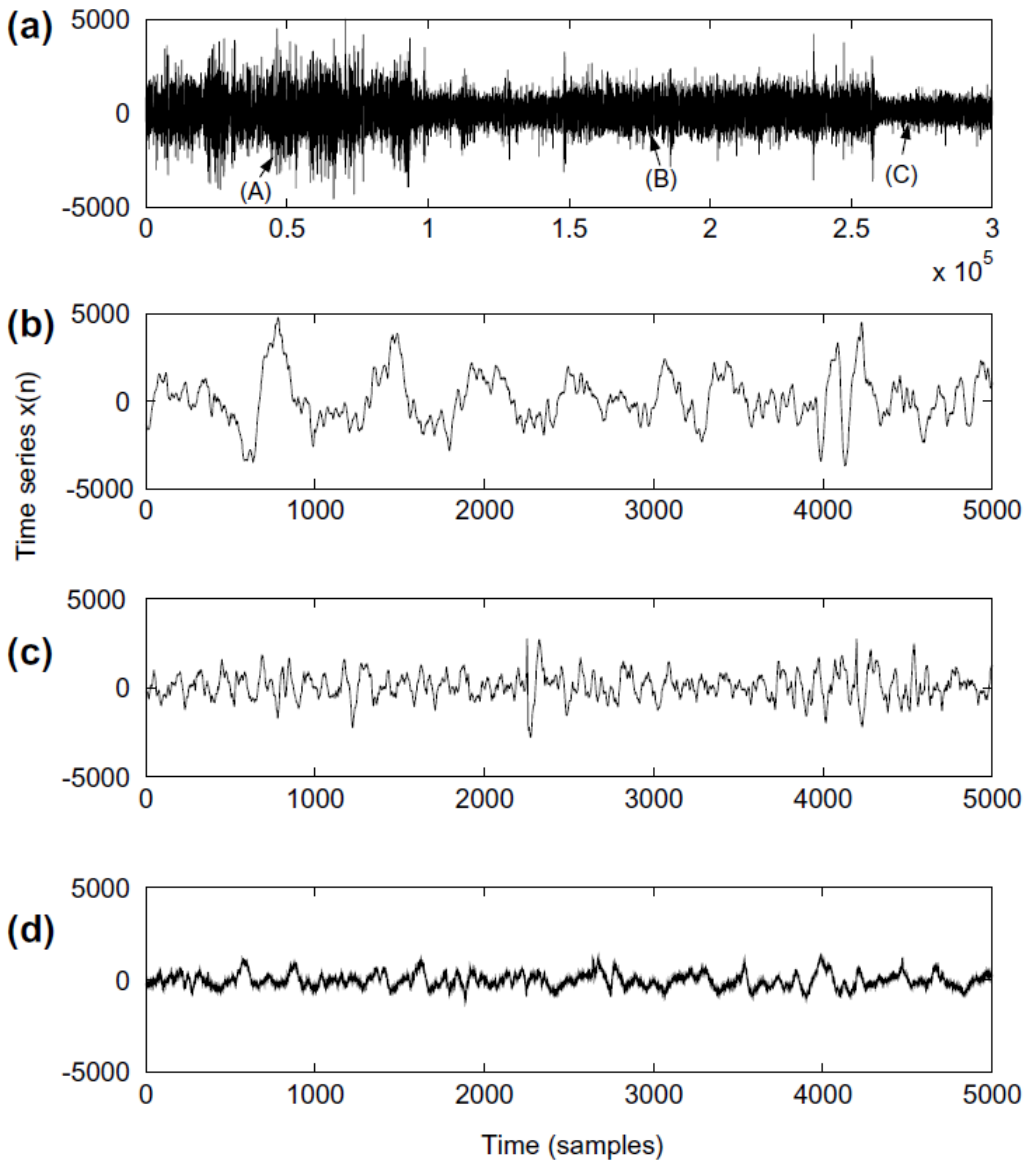

Figura 7. Medición Clínica Intracraneal. Visualizacíon de ruido.

Fuente: [12].

En la figura 7, se puede apreciar la presencia de ruido en la serie temporal **(a)**. Es posible también, observar los segmentos subyacentes **(b)**, **(c)** y **(d)**, superposicionados como parte de **a** y luego evaluados como segmentos individuales, con aún presencia de "***ruido blanco***".

El ruido blanco es una señal aleatoria (conocida también como proceso estocástico) que se caracteriza por el hecho de que sus valores en el eje tiempo, no guardan correlación estadística. En resumen, el ruido blanco es no correlativo, es decir, en el

eje del tiempo la señal toma valores sin ninguna relación unos con otros [27].

Matemáticamente, podría definirse entonces como una sucesión de variables aleatorias (procesos estocástico) con una esperanza (media) cero y una varianza constante e independientes de cualquier valor de t (covarianza nula) [12].

Los valores faltantes, al igual que el ruido, representan igualmente un problema serio que debe abordarse previo al análisis de series temporales.

La interpolación por su parte, es un concepto que puede ser aplicado a los datos de series temporales para determinar valores faltantes. En primera instancia, se debe comprender el modelo de datos apropiado con el que se esta trabajando (autorregresivo, autorregresivo de media móvil, autorregresivo integrado de media móvil o lineal). Una vez definido el modelo y de acuerdo con la posición de la observación faltante, se deben calcular las ponderaciones apropiadas para hacer un pronóstico en ambas direcciones, hacia adelante o hacia atrás. Finalmente, utilizando una combinación lineal del pronóstico anterior, se podrán predecir entonces los valores faltantes bajo las condiciones dadas [27].

5 Propuesta de Proyecto

5.1 Planteamiento del Problema

El cálculo de la similitud en series temporales ha sido un tema muy estudiado en la última década [13]. La precisión, la velocidad de cómputo y la tolerancia al ruido, son factores claves (particularmente en conjuntos de datos grandes y complejos) a la hora de elegir una medida de distancia robusta para comparar dos series de tiempo de largo n [4].

En la literatura, la medida de distancia más utilizada para comparar series temporarles es sin duda la distancia *Euclidiana* (o alguna de sus variaciones). Dicha medida es muy utilizada en el descubrimiento y la comparación de reglas motif o patrones en series temporales [10][11].

Existen pruebas empíricas fiables, que demuestran que la distancia Euclidiana es muy competitiva e incluso superior a medidas mucho más complejas en una amplia variedad de dominios, particularmente cuando el conjunto de datos se vuelve cada vez más grande [9][15].

Sin embargo, algunos aportes más recientes al estado del arte indican que *Dynamic Time Warping*, en una o varias dimensiones, puede incluso comportarse de forma más robusta y estable que la distancia *Euclidiana* [5]. Este argumento se soporta principalmente en la sensibilidad conocida que presenta la distancia *Euclidiana* ante la presencia de ruido, pero fundamentalmente y por su naturaleza lineal [30], intolerante ante pequeñas distorciones observadas al comparar dos series temporales desfasadas con respecto al eje tiempo [6].

Por último, la presencia de ruido y de valores faltantes en los datos de una serie de tiempo, son problemas difíciles de tratar; lo que sí es seguro, es que son retos inevitables, prácticamente inherentes, que al igual que las limitaciones de las medidas de distancia, deben abordarse previo al análisis y que, finalmente definen el planteamiento del problema [12].

El problema anteriormente planteado, se atacará mediante la siguiente propuesta de investigación.

5.2 Propuesta del Proyecto

Apoyados en la premisa anterior, el proyecto pretende estudiar el nivel de acierto obtenido como resultado del descubrimiento de *"reglas significativas"* en series de tiempo, utilizando *Cubic Spline Interpolation* como una medida alternativia de distancia aparentemente superior a la distancia *Euclidiana*, principalmente ante la presencia de distorciones en el conjunto de datos y otras limitaciones ya planteadas.

El proyecto se enfoca en remplazar la distancia Euclidiana y probar que la utilización de otras medidas de distancia (particularmente el uso de *Cubic Spline Interpolation*) pueden ser mucho más tolerantes al ruido y aún así, garantizar al menos el mismo nivel de acierto en el descubrimiento de reglas significativas en series temporales.

5.3 Trabajos Relacionados

La presente propuesta de investigación se basa en trabajo de Mohammad Shokoohi-Yekta et al [1], quienes proponen una serie de nuevos algoritmos que permiten el descubrimiento veloz de reglas significativas de alta calidad a partir de grandes conjuntos de datos, las cuales, pueden predecir la ocurrencia de eventos futuros.

Es importante destacar el hecho de que a pesar de que otras medidas de distancia fueron evaluadas (se menciona por ejemplo: DTW, Swale, Spade y ERP), ***los algoritmos propuestos en [1] utilizan la distancia Euclidiana*** para la identificación de reglas motif y su selección posterior como reglas significativas.

A diferencia de otras propuestas mencionadas a continuación, el trabajo propuesto en [1], se basa el descubrimiento de reglas motif basadas en *la forma* para vaticinar los eventos futuros. En contraposición, los trabajos anteriores intentan lograr predicciones basadas en los *valores actuales* del flujo de datos [37].

En una secuencia de trabajos que culminaron en [38], Park y Chu, investigaron un mecanismo para hallar reglas sobre series de tiempo. Sin embargo, el algoritmo únicamente es valorado con respecto a la velocidad y datos aleatorios (*random walk data*). No se presentaron pruebas de que el algoritmo pudiera en realidad encontrar

las reglas generalizables en series de tiempo.

Los trabajos de Wu y colegas en [39], también utilizan representaciones lineales de segmentos para apoyar el descubrimiento de reglas en series de tiempo. Ellos probaron su algoritmo sobre datos financieros reales, reportando aproximadamente un 68% de "corrección de la predicción de la tendencia de datos" sobre la serie temporal. Curiosamente, sin embargo, los autores corrieron su algoritmo en datos proporcionados por otros y obtuvieron exactamente los mismos resultados.

Acerca de los algoritmos propuestos en [1]

Una vez más, es importante subrayar que en nuestra propuesta, se busca la modificación de dos algoritmos propuestos en [1], particularmente el reemplazo de la medida de distancia Euclidiana. Ambos algoritmos serán explicados a continuación en las tablas 1 y 2.

En la siguiente tabla, se explica el seudocódigo del primer algoritmo llamado "***Find Antecedent Candidates***". Este algoritmo, sirve para la búsqueda de reglas motif (potenciales reglas significativas).

Como se detalla en la tabla 1, recibe los siguientes argumentos: un segmento subsecuente R extracído de la serie temporal T y un punto de división sp, que es simplemente un número entre cero y uno, utilizado para dividir los segmentos antecedentes y consecuentes de la regla [1].

El algoritmo devolverá las ubicaciones de los segmentos antecedentes encontrados en T, ordenados por las distancias a partir de los R segmentos antecedentes almacenados y retornados como ac [1].

Procedure *find_Antecedent_Candidates* (***T***, ***R***, ***sp***) Input: A time series subsequence, ***R***, extracted from a time series, ***T***; Split point for the antecedent/consequent, a number between zero and one, ***sp***; Output: locations of antecedents in ***T*** ordered by distances from ***R***'s antecedent, ***ac***;	
1	*antecedentLength* ← Length(*R*) × *sp*
2	*antecedent* ← *R* (1:*antecedentLength*)
3	*Distances* ← Euclidean (*antecedent*, each subsequence in *T*)
4	*AntecedentDistances* ← sort (localMinimums (*Distances*))
5	*AntecedentCandidates* ← Locations (*AntecedentDistances*)
6	*ac* ← *AntecedentCandidates*
7	**Return** *ac*

Tabla 1. Seudocódigo del algoritmo "Find Antecedent Candidates".

Como se observa en la linea 3 del algoritmo, la distancia Euclideana es implementada para el cómputo de las medidas de distancia, para encontrar "segmentos antecedentes" en cada subsecuencia de la serie temporal T [1].

En la tabla 2, por su parte, se muestra el seudocódigo del segundo algoritmo llamado "***Rule Bit Saves***".

Este algoritmo, recibe como argumento, un segmento o subsecuencia R extraído de la serie temporal T, un punto de división sp, que es un número entre cero y uno, el número de instancias de R que serán seleccionadas se almacenarán en n, la ubicación del segmento antecedente en la serie temporal T ordenada por las distancias a partir de los segmentos antecedentes R y finalmente el $maxlag$ o tiempo máximo de rezago.

El algoritmo devolverá como resultado el número total de bit ahorrados, es decir, la calificación acumulada de la regla motif (segmento R) con respecto a capacidad predictiva sobre la serie temporal T [1].

```
Procedure Rule_Bit_Saves (T, R, sp, n, ac, mlag)
Input: A time series subsequence, R, extracted from a time series, T;
Split point for the antecedent/consequent, between zero and one, sp;
The Number of instances of R to pick in the time series, n;  locations
of antecedents in T ordered by distances from R's antecedent, ac;
Maxlag allowed between the antecedent and consequent, mlag;
Output: totalBitSave;
 1  antecedentLength ← Length(R) × sp
 2  consequent ← R (antecedentLength:end)
 3  Discretize and z-normalize (consequent)
 4  AntecedentCandidates ← ac
 5  totalBitSave ← 0
 6  antecedentsSelected ← AntecedentCandidates (2 : n)
 7    for i ← 1 to Length(antecedentsSelected) do
 8      s1 ← AntecedentCandidates(i) + antecedentLength +1
 9      s2 ← AntecedentCandidates(i) + Length(R)
10      subConsequent ← T (s1 : s1 + mlag)     // considering maxlag
11      consequentDist ← Euclidean (consequent, each subsequence
12                                          of subConsequent)
13      conseqLoc ← find (consequentDist == min(consequentDist))
14      subConsequent ← T (s1 + conseqLoc : s2 + conseqLoc)
15      Discretize and z-normalize (subConsequent)
16      subConsequentBits ← Huffman (subConsequent)
17      subConsequentMDLbits←MDL(subConsequent, consequent)
18      totalBitSave ← totalBitSave + subConsequentBits -
19                          subConsequentMDLbits
20    end for
21  Return totalBitSave – Huffman (consequent)  // Eq. 2
```

Tabla 2. Seudocódigo del algoritmo "Rule Bit Save".

En la línea 11 de la tabla 2, se observa claramente el uso de la distancia Euclideana para calcular la distancia entre entre el segmento consecuente, para cada occurrencia del segmento subsecuente R que sea encontrado en T [1].

Es importante resaltar el uso de MDL (Minimun Description Length, por sus siglas en Inglés), en la línea 17, inspirada en la estrategia de búsqueda más simple posible, que evalúa las potenciales reglas significativas, con base en cuan bien pueden comprimir los datos [1].

5.4 Hipótesis

Con base en la definición del problema y en la propuesta de proyecto, se define la siguiente hipótesis:

El uso de la medida de distancia Cubic Spline Interpolation, mejora el nivel de exactitud en los algoritmos "Rule Bit Saves" y "Find Antecedent Candidates" propuestos por Mohammad Shokoohi-Yekta y co-

laboradores, en el hallazgo de reglas significativas en series de tiempo complejas y en presencia de ruido.

5.5 Métricas

El análisis comparativo de los niveles de exactitud obtenidos a partir de la ejecución de los algoritmos según la distancia utilizada, requerirá de las siguientes métricas:

- **Exactitud (Q):**

$$\frac{Total\_Aciertos}{Total\_Predicciones} \tag{8}$$

En el caso más general, se utilizará inicialmente la distancia Euclidiana entre la parte consecuente predicha y las $\boldsymbol{F}$ ubicaciones halladas desde donde la regla fue disparada, un valor denotado como ***"Ferror"***, también conocido como la *media cuadrática*.

Sobre el mismo conjunto de prueba, y mediante el uso del mismo segmento consecuente de la regla, se disparará aleatoriamente $\boldsymbol{F}$ veces y se medirá la distancia *Euclidiana* (Cubic Spline Interpolation y otras ya mencionadas en el marco teórico), entre el segmento consecuente predicho y la ubicaciones aleatorias F.

Ese valor será denotado como ***Rerror*** (el cual, se promediará entre aproximadamente 1000 ejecuciones aleatorias).

En resumen, la medida de calidad reportada puede definirse como:

$$Q = \frac{Ferror}{Rerror} \tag{9}$$

Los valores cercanos a uno, sugieren que las reglas a prueba, no se consideran mejores que las encontradas en la estimación aleatoria. Los valores significativamente menores a uno, indican que la regla en efecto encuentra una estructura verdadera en los datos. En la mayoría de los experimientos se utilizará un retraso máximo (***maxlag***) [1] igual a cero.

5.6 Justificación del Proyecto

La realización de este proyecto es importante, porque a través de sus resultados se podrían llegar a obtener los siguientes beneficios:

- **Un aporte al estado del arte**. El estudio del estado del arte realizado indica que "Cubic Spline Interpolation", como medida de distancia, no ha sido utilizado aún en algoritmos para el descubrimiento de reglas significativas sobre series temporales. La idea es, por ende, prometedora e innovadora y podría tener un impacto significativo sobre el rendimiento de algoritmos existentes que actualmente utilizan otras medidas de distancia más utilizadas como por ejemplo, la distancia *Euclideana*.

- **Ataca un problema de investigación real**. Muchas de las aplicaciones producto del análisis de series temporales, son posibles a partir del descubrimiento de reglas *motif* y su uso posterior sobre el conjunto de datos [1]. El descubrimiento de dichos patrones en series temporales, se puede resumir a un problema de *similitud*. El cómputo de la similitud (o disimilitud) se obtiene a partir de la medida de distancia entre los puntos de datos de las series temporales [1]. Mientras más eficiente y robusta sea la medida de distancia, mejor será el cálculo de la similitud y por ende, mejores reglas *motif* se obtendrán. Como consecuencia de lo anterior, la precisión, la velocidad del cómputo y la tolerancia al ruido serán superior y por consiguiente, la calidad predictiva de la regla sobre el conjunto de datos será mucho más fiable [16].

- **Las potenciales aplicaciones y su impacto en la sociedad**. Como se desarrolló en el marco teórico, las aplicaciones de la minería sobre series temporales son amplias y diversas [1]. Las predicciones de corto plazo mediante el uso de reglas significativas en campos triviales como el análisis del mercado de acciones, el estudio de las condiciones meteorológicas (por ejemplo, el incremento del tiempo promedio de alerta de tornados [20]) son cada día más utilizadas. En robótica, por ejemplo, se han realizado avances significativos en la exploración

de la anticipación (predicciones de corto plazo) de las fuerzas futuras percibidas por un robot con base en las intenciones motoras de otro agente y así adaptar su movimiento, ejecutando un nuevo plan de acción [19].

6 Objetivo General

El objetivo general de esta propuesta de investigación es:

Estudiar el nivel de exactitud obtenido a partir del uso de *Cubic Spline Interpolation* como medida de distancia utilizada para el descubrimiento de reglas significativas, en series temporales complejas y en presencia de ruido.

7 Objetivos Específicos

Los objetivos específicos de este proyecto son los siguientes:

1. Proponer el uso de *Cubic Spline Interpolation* como medida de distancia utilizada en los algoritmos ***"Rule Bit Saves"*** y ***"Find Antecedent Candidates"***, creados por Mohammad Shokoohi-Yekta y colaboradores, para el descubrimiento de reglas significativas sobre series temporales.

2. Realizar un análisis comparativo del nivel de exactitud obtenido al utilizar diferentes medidas de distancia en los algoritmos mencionados en el objetivo anterior.

3. Explicar los resultados obtenidos en el objetivo específico anterior, con el propósito de aprobar o rechazar la hipótesis que ha sido planteada.

8 Alcance de la Investigación

El alcance de esta propuesta de investigación se enfoca específicamente en estudiar el efecto de utilizar *"Cubic Spline Interpolation"* como medida de distancia en los algoritmos ***"Rule Bit Saves"*** y ***"Find Antecedent Candidates"*** propuestos por Mohammad Shokoohi-Yekta y colaboradores, utilizados para el hallazgo de reglas significativas en series de tiempo complejas y en presencia de ruido.

El diseño experimental incluye también la comparación y el análisis de cinco versiones de ambos algoritmos. Cada versión incorpora la implementación de cinco medidas de distancia distintas: 1- Euclideana, 2- Swale, 3- Spade, 4- EPR y 5- Cubic Spline Interpolation.

Como resultado de esta investigación, se entregarán los siguientes productos:

- La implementación de las cinco versiones de ambos algoritmos para cada una de las medidas de distancia.
- Programas auxiliares para la ejecución y el control del entregable anterior.
- Análisis estadístico que contraste los resultados de los experimentos para aceptar o rechazar la hipótesis.
- Un artículo científico que se entregará al comite editorial de alguna revista o conferencia, con miras a su publicación.

Es necesario delimitar esta investigación por motivos de tiempo y extensión. Es por ello que a continuación se detallan las siguientes excepciones:

- No se tomarán en cuenta otras medidas de distancia.
- No se utilizará ningún otro algoritmo para la identificación de reglas *motif*.
- No se utilizará ningún otro algoritmo para la detección de segmentos antecedentes de una potencial regla significativa.
- Cualquier otro resultado, documento, software o producción que no se encuentren contemplados en los entregables, no será considerado como parte del alcance de este proyecto.

En resumen, debe considerarse el hecho de que el objetivo principal de esta investigación, se enfoca principalmente en la implemetación de las medidas de distancia sobre ambos algoritmos, para su posterior análisis comparativo, aplicado a cada unos de los cinco diferentes conjuntos de datos; utilizando *"Cubic Spline Interpolation"* como el elemento más importante de la hipótesis planteada.

9 Entregables

Los entregables son los siguientes:

- **Modificación del software utilizado para el hallazgo de reglas significativas:** incorporar las medidas de distancia propuestas para ambos algoritmos.
- **Desarrollo e implementación de un ambiente de pruebas:** este ambiente permitirá ejecutar y medir la exactitud de ambos algoritmos, en el hallazgo y la selección de reglas significativas.
- **Documento final con la recopilación y el análisis del diseño de experimentos:** este entregable implica la ejecución del diseño de experimentos y la construcción de una tabla resumen de los resultados obtenidos.
- **Documento final del análisis de varianza no paramétrico y la caracterización de los resultados obtenidos:** corresponde al análisis de varianza no paramétrico, para aceptar o rechazar la hipótesis mediante la caracterización de los resultados.
- **Documento de tesis:** recopila los entregables anteriores, las conclusiones y el resultado de la revisión de la hipótesis.
- **Desarrollo de un artículo científico:** en borrador en español de un artículo científico.

10 Metodología

10.1 Diseño de Experimentos

Para describir el planeamiento pre-experimental para el diseño de experimentos de este trabajo (con la información disponible hasta el momento) se usan los *lineamientos* desarrollados en el libro de *Douglas C. Montgomer*y [2]. El esquema del procedimiento recomendado en los lineamientos para el desarrollo de esta etapa, incluye lo siguiente:

1. **Reconocimiento y definición del problema:** consiste en desarrollar una declaración clara y sencilla del problema. Una clara definición del problema, normalmente contribuye substancialmente a una mejor comprensión del fenómeno que está siendo estudiado y a la solución final de dicho problema.

2. **Selección de factores, niveles y rangos:** consiste en enumerar todos los posibles factores que pueden influenciar el experimento. Incluye tanto los factores de diseño potencial (los que potencialmente se podrían querer modificar en los experimentos) y los factores perturbadores (los que no se quieren estudiar en el contexto del experimento). También, se deben seleccionar los rangos sobre los que varían los distintos factores y los niveles específicos sobre los que se aplicarán las iteraciones del experimento.

3. **Selección de la variable de respuesta:** debe proveer información útil sobre el fenómeno que está siendo estudiado.

4 **Selección del diseño de experimental:** se refiere a aspectos claves del experimento tales como el tamaño de la muestra, la selección del orden adecuado para la ejecución de los intentos experimentales y la decisión de bloquear o no algunas de las restriciones de aleatoriedad en la pruebas.

5 **Llevar a cabo el experimiento:** en esta etapa, es de vital importancia, monitorear el proceso cuidadosamente para asegurar la correcta ejecución del experimento con respecto a lo planeado.

10.1.1 Declaración del Problema

Estudiar el comportamiento de *Cubic Spline Interpolation* como medida de distancia utilizada en el descubrimiento de reglas significativas en series temporales complejas y en presencia de ruido.

Cubic Spline Interpolation y otras medidas medidas de distancia serán incorporadas en dos los algoritmos específicos llamados ***"Rule Bit Saves"*** y ***"Find Antecedent Candidates"*** propuestos por Mohammad Shokoohi-Yekta y colaboradores en [1]. Es importante señalar, que por cada medida de distancia utilizada, se creará una nueva versión de ambos algoritmos.

La precisión en el hallazgo de reglas significativas será medido a través de la ejecución, la comparación y el análisis de cada versión de los algoritmos, mediante la utilización de al menos cinco fuentes de datos temporales de complejidad variada y en presencia de diferentes grados de ruido.

10.1.2 Factores

En el diseño de experimentos, un factor es aquel componente que tiene cierta influencia en las variables de respuesta [2]. El objetivo de un experimento es determinar esta influencia. A su vez, cada factor cuenta con varios niveles posibles con los cuales experimentar.

Usando la información recolectada en esta etapa de la investigación, así como la experiencia adquirida por el estudiante y expuesta en los capítulos anteriores, se han seleccionado inicialmente los siguientes dos factores para su estudio:

1. **Las métricas o medidas de distancia**

   Se utilizarán dos algoritmos para la ejecución del diseño de experimental:

   1- ***"Rule Bit Saves"*** utilizado en la identificación de potenciales reglas o patrones significativos (*detección de reglas Motif en [1]*) y 2- ***"Find Antecedent Candidates"***, creado para probar capacidad predictiva de las reglas motif identificadas en el algoritmo anterior, sobre cada uno de los conjuntos de datos. Ambos algoritmos serán moficados y adaptados a cada medida de distancia y

su ejecución sobre cada conjunto de datos, será realizada en forma controlada e independiente.

Las medidas de distancia utilizadas, son las siguientes:

- ***Distancia Euclidiana***
- ***Swale***
- ***Spade***
- ***EPR***
- ***Cubic Spline Interpolation***

2. **El conjunto de datos, tamaño y complejidad**

   Los niveles de ruido, la valoración de la complejidad y el tamaño de cada conjunto de datos, se encuentran aún pendientes de determinar. Dicha clasificacíon estará dada en al menos tres niveles que permitan medir la complejidad de cada conjunto de datos.

   En el experimento, se utilizarán los siguientes conjuntos de datos:

   - 1. **"Energy disaggregation dataset":** contiene el amperaje del consumo diaro de una casa promedio durante un año.
   - 2. **"Zebra finch vocalizations":** contiene las grabaciones del canto de un pájaro *"Zebra"* durante sus primeros 100 dás de vida.
   - 3. **"Daily basis activity data set":** este conjunto de datos contiene información telemétrica de actividades cotidianas de una persona.
   - 4. **"NASA telemetry data":** contiene medidas de voltajes erróneas producidas por las válvulas utilizadas en los transbordadores espaciales de la NASA, utilizadas para el estudio y la detección de anomalías .

Los conjuntos de datos anteriores, fueron seleccionados pensando primordialmente en facilitar la ejecución del diseño de experimentos con base en los resultados obtenidos en [1]. Lo anterior permitirá generar resultados mucho más confiables durante la ejecución de las pruebas.

En la tabla 1, presentada a continuación, se resumen los factores que se utilizarán para evaluar los resultados de ambos algoritmos, para cada conjunto de datos.

	Factores	
Algoritmos	**Conjunto de Datos**	**Medida de Distancia**
Rule Bit Saves Find Antecedent Candidates	Energy Disaggregation Zebra Finch Vocalizations Daily Basis Activity NASA Telemetry Data	Distancia Euclidiana Swale Spade EPR Cubic Spline Interpolation

Tabla 1. Resumen de los factores que se utilizarán en el diseño de experimentos.

10.1.3 Variables de Respuesta

Dado que la hipótesis afirma maximizar el nivel de exactitud en la identificación de reglas significativas y el hallazgo de segmentos antecedentes sobre los diferentes conjuntos de datos, se ha seleccionado ***la exactitud*** como la única variable de respuesta para ambos algoritmos:

1. ***Medición de la exactitud sobre el algoritmo "Rule Bit Saves"***
   Cálculo del total de aciertos en la identificación de reglas *motif* (potenciales reglas significativas) sobre cada conjunto de prueba, mediante la ejecución de las cinco diferentes versiones del algoritmo.

2. ***Medición de la exactitud sobre "Find Antecedent Candidates"***
   Una vez que las reglas motif han sido identificadas, se requiere calcular para cada una de ellas, el total de aciertos en el hallazgo de segmentos antecedentes (*predicciones*), sobre cada conjunto de datos, para cada una de las medidas de distancia implementadas en las cinco versiones del algoritmo.

10.1.4 Recolección de Datos

Las variables de respuesta serán recolectadas de forma automática, una vez concluida la ejecución de cada una de las versiones de ambos algoritmos.

La automatización de la recolección de las variables de respuesta será posible mediante la implementación del ambiente de pruebas.

10.1.5 Análisis Estadístico

Una vez concluída la ejecución del experimento, se deben analizar los resultados para comparar ambos algoritmos en sus diferentes versiones. Para esta comparación, se debe analizar si existe una diferencia significativa en los promedios obtenidos de la variable de respuesta para cada uno de los distintos grupos. Un análisis de varianza permite saber si la diferencia en la media de varias poblaciones es significativa debido a la influencia de alguno de los factores [3].

En primera instancia, se harán las pruebas de normalidad utilizando el valor p para mostrar si los resultados cumplen o no los requisitos para una prueba paramétrica. Posteriormente, una vez realizada la prueba de normalidad, se utilizará el método estadístico, no paramétrico llamado ***"Kruskal-Wallis-Test"***. Mediante el uso de este método, no se tiene que asumir que las poblaciones siguen una distribución normal, únicamente se utilizarán muestras independientes provenientes de distintas fuentes de datos; poblaciones no relacionadas cuyas muestras no se afectan unas con otras [26].

10.2 Ambiente de Desarrollo

Para el desarrollo de la propuesta de investigación actual, se implementará una plataforma que cumpla con dos características principales:

1. Que la plataforma se pueda ejecutar sobre los sistemas operativos Windows y Linux: esto implica que el código fuente debe ser escrito en un lenguaje de programación capaz de correr en los dos sistemas operativos, con un número mínimo de cambios y que las bibliotecas utilizadas estén disponibles para los dos sistemas operativos.

2. Que las herramientas y bibliotecas a utilizar sean gratuitas al menos para uso académico.

Basado en estas dos características, se ha elegido una lista de posibles soluciones de software a utilizar. Como se indica, esta lista es preliminar, por lo que se preveen posibles cambios durante el desarrollo de la tesis.

- **Sistema Operativo:** Windows 8.1 y Linux Ubuntu 16.04.1 LTS.
- **Lenguaje de programación:** Matrix Laboratory (MATLAB).

11 Plan de Trabajo

El plan de trabajo de esta propuesta de investigación debe verse como la secuencia de actividades que se deberán ejecutar para producir los entregables mencionados en la sección nueve de este documento.

En la tabla 3, se presenta el detalle de cada uno de los entregables, sus respectivos objetivos y su respectiva duración.

Entregable	Objetivos	Duración (Dado en Semanas)
Modificación del software utilizado para la identificación de reglas motif y el hallazgo de reglas significativas	Incorporar todas medidas de distacia propuestas en ambos algoritmos.	4
Preprocesamiento de datos		1
Desarrollo e implementación de un ambiente de pruebas	Implementar un ambiente de pruebas que permita la ejecución de las diferentes versiones de ambos algoritmos.	3
Documento final con la recopilación y el análisis del diseño de experimentos	Ejecución del diseño de experimentos preparado, incluyendo la medición de las métricas definidas, para determinar si hay diferencias significativas entre los algoritmos.	2
Documento final del análisis de varianza no paramétrico y caracterización de los resultados obtenidos	Ejecutar y reportar el análisis de varianza no paramétrico, para aceptar o rechazar la hipótesis.	2
Desarrollo de un artículo científico	Preparar un artículo científico para ser presentado en una alguna revista afin al tema desarrollado en esta propuesta.	2
Documento final de tesis		2

Tabla 3. Listado de entregables, objetivos relacionados y duración.

Finalmente, como se observa en la tabla 4, el cronograma consta de un plazo de dieciséis semanas para la completitud del proyecto.

Entregable	Semanas															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Modificación del software utilizado para el hallazgo de reglas significativas																
Preprocesamiento de datos																
Desarrollo e implementación de un ambiente de pruebas																
Documento final con la recopilación y el análisis del diseño de experimentos																
Documento final del análisis de varianza no parametrico y caracterización de resultados																
Desarrollo de un artículo científico																
Documento final de tesis																

Tabla 4. Cronograma.

Referencias

[1] Shokoohi-Yekta, Chen, Bilson, Bing Hu, Zakaria and Eamonn Keogh. University of California, Riverside. *"Discovery of Meaningful Rules in Time Series".* KDD 2015, Proceedings of the 21th ACM SIGKDD. International Conference on Knowledge Discovery and Data Mining. Pages 1085-1094.

[2] D. C. Montgomeryx. *"Guidelines for designing experiments, design and analysis of experiments."* 5th Edition, 2000, pp. 13-17".

[3] R. L. Mason. *"Statistical Design and Analysis of Experiments With Applications to Engineering and Science."* Second Edition. John Wiley & Sons. 2003.

[4] M. Vlachos, G. Kollios, and D. Gunopulos. *"Discovering similar multidimensional trajectories."* Proc 18th Int. Conf. Data Eng. pp. 673684, 2002.

[5] S. Chu, E. Keogh, and D. Hart, *"Iterative Deepening Dynamic Time Warping for Time Series"*. pp. 195212, 2002.

[6] H. Li, X. Wan, Y. Liang, and S. Gao. *"Dynamic Time Warping Based on Cubic Spline Interpolation for Time Series Data Mining."* 2014 IEEE Int. Conf. Data Min. Work., pp. 1926, 2014.

[7] C. Ratanamahatana and E. Keogh, *"Everything you know about dynamic time warping is wrong."* Third Work. Min. Temporal Seq. Data, pp. 2225, 2004.

[8] G. Al-Naymat, S. Chawla, and J. Taheri. *"SparseDTW: A novel approach to speed up dynamic time warping."* Conf. Res. Pract. Inf. Technol. Ser., vol. 101, no. December 2003, pp. 117127, 2009.

[9] H. Ding, G. Trajcevski, and P. Scheuermann, *"Querying and mining of time series data: experimental comparison of representations and distance measures."* Proc. VLDB Endow., vol. 1, no. 2, pp. 15421552, 2008.

[10] A. Mueen, E. Keogh, Q. Zhu, S. Cash, and B. Westover, *"Exact Discovery of Time Series Motifs"*. Proc. 2009 SIAM Int. Conf. Data Min., pp. 473484, 2009.

[11] A. M. Denton, C. A. Besemann, and D. H. Dorr, *"Pattern-based time-series subsequence clustering using radial distribution functions."* Knowl. Inf. Syst. Vol. 18, no. 1, pp. 127, 2009.

[12] J. Hu, J. B. Gao, and K. D. White, *"Estimating measurement noise in a time series by exploiting nonstationarity."* vol. 22, pp. 807819, 2004.

[13] G. Das, K. Lin, H. Mannila, G. Renganathan, and P. Smyth. *"Rule discovery from time series."* Knowl. Discov. Data Min, pp. 1622, 1998.

[14] G. Al-Naymat, S. Chawla, and J. Taheri. *"SparseDTW: A novel approach to speed up dynamic time warping."* Conf. Res. Pract. Inf. Technol. Ser., vol. 101, no. December 2003, pp. 117127, 2009.

[15] E. Keogh, Exact indexing of dynamic time warping, in: Processings of the 28th VLDB Conference, 2005, pp.358-380.

[16] Michael Morse, Jignesh M. Patel, *"An Efficient and Accurate Method for Evaluating Time Series Similarity". University of Michigan*

[17] Park, S and Chu, S.W. Discovering and Matching Elastic Rules from Sequence Databases. Fundam. Inform. 47, 2001.

[18] Wu, H., Salzberg, B., and Zhang, D., Online Event-driven Subsequence Matching over Financial Data Streams, SIGMOD Conference, 2004: 23-34.

[19] Gribovskaya, E., Kheddar, A., and Billard, A. Motion Learning and Adaptive Impedance for Robot Control during Physical Interaction with Humans. ICRA 2011.

[20] Brotzge, J. and Erickson, S., Tornadoes without NWS warning. Weather Forecasting, 25, 159-172. 2010.

[21] McGovern, et al. Identifying Predictive Multi-Dimensional Time Series Motifs: An application to severe weather prediction. Data Mining and Knowledge Discovery. 2010.

[22] Li, G, Ji, S., Li, C, and Feng, J. Efficient type-ahead search on relational data: a TASTIER approach. SIGMOD Conference 2009: 695-706.

[23] Weiss, S, Indurkhya, N, and Apte, C. Predictive Rule Discovery from Electronic Health Records. ACM IHI, 2010.

[24] Eamonn Keogh, Shruti Kasetty, University of California, Riverside. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration.

[25] Tak-chung Fu, A review on time series data mining, Department of Computing, Hong Kong Polytechnic University, Hunghom, Kowloon, Hong Kong, 2010.

[26] Yvonne Chan, Roy P Walmsley. Learning and Understanding the Kruskal-Wallis One-Way Analysis-of-Variance-by-Ranks Test for Differences Among Three or More Independent Groups. Published on December 1997.

[27] Carmelo Cassisi, Placido Montalto, Marco Aliotta, Andrea Cannata and Alfredo Pulvirenti. Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining. 2012 Cassisi et al.

[28] Han and Kamber (2005). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, CA.

[29] Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., and Keogh, E. (2012). Searching and mining trillions of time series subsequences under 54 dynamic time warping. In Proceedings of the 18th ACM

SIG KDD international conferenceon Knowledge discovery and data mining, pages 262-270. ACM.

[30] Gustavo Batista, Xiaoyue Wang, Eamonn J. Keogh, A Complexity-Invariant Distance Measure for Time Series, University of California, Riverside, University of Sao Paulo - USP.

[31] Michael Morse, Jignesh Patel. An Efficient and Accurate Method for Evaluating Time Series Similarity. University of Michigan. SIGMOD07, June 1114, 2007, Beijing, China.

[32] D. Berndt, J. Clifford. Using Dynamic Time Warping to Find Patterns in Time Series. In AAAI-94 Workshop on Knowledge Discovery in Databases, pages 359370, 1994.

[33] Robert Shumway, David Stoffer. Time Series Analysis and Its Applications - With R Examples. Third edition. Pages 1-7. 2011.

[34] Chotirat Ann Ratanamahatana, Jessica Lin, Dimitrios Gunopulos, Eamonn Keogh. Mining Time Series Data. University of California, Riverside. 2014.

[35] Yueguo Chen Mario A, Nascimento Beng Chin Ooi, Anthony K. H. Tung. National University of Singapore. University of Alberta. On Shape-based Pattern Detection in Streaming Time Series.

[36] Lei Chen (University of Waterloo), Raymond Ng (University of British Columbia). On The Marriage of Lp-norms and Edit Distance,

[37] Makridakis, S., Wheelwright, S., and Hyndman, R. J., Forecasting: methods and applications. New York: John Wiley & Sons. ISBN 0-471-53233-9. 1998.

[38] Park, S., and Chu, S.W. Discovering and Matching Elastic Rules from Sequence Databases. Fundam. Inform. 47, 2001.

[39] Wu, Salzberg and Zhang. Online Event-driven Subsequence Matching over Financial Data Streams. SIGMOD Conference, 2004: pages 23-34.