UNIVERSIDAD DE SONORA

DIVISIÓN DE INGENIERÍA

POSGRADO EN INGENIERÍA INDUSTRIAL

EXTRACCIÓN DE REGLAS ASOCIATIVAS DE PATRONES DE ELEMENTOS METEOROLÓGICOS QUE IMPACTAN EL PROCESO DE PRODUCCIÓN DEL CULTIVO DE LA UVA DE MESA

T E S I S

PRESENTADA POR

DANIEL ALBERTO ZÚÑIGA VÁZQUEZ

Desarrollada para cumplir con uno de los
requerimientos parciales para obtener
el grado de Maestro en Ingeniería

DIRECTOR DE TESIS DR. LUIS FELIPE ROMERO DESSENS
CODIRECTOR DE TESIS DR. JUAN MARTÍN PRECIADO RODRÍGUEZ

HERMOSILLO, SONORA. SEPTIEMBRE 2012

Universidad de Sonora

Repositorio Institucional UNISON

RESUMEN

Este proyecto representa el trabajo conjunto de investigadores de la Universidad de Sonora y del Centro de Investigación en Alimentación y Desarrollo, en donde se cumple con el objetivo de identificar reglas asociativas de patrones de elementos meteorológicos que aceleren o retarden el desarrollo vegetativo en el proceso de producción del cultivo de la uva de mesa. Se comienza con una descripción de las generalidades de la vid, métodos para medir su desarrollo vegetativo y una descripción del impacto de los elementos meteorológicos en este. Posteriormente, se presenta la historia de las predicciones de elementos meteorológicos, basadas en el movimiento de las estrellas y los planetas, en las nubes, las fases lunares y el movimiento del viento, o con el uso de instrumentos como el termómetro, el anemómetro y el pluviómetro que permitieron hacer predicciones más precisas. Hasta el día de hoy, estas predicciones son realizadas por modelos matemáticos aplicando herramientas de minería de datos. En este trabajo se extraen reglas asociativas de patrones de elementos meteorológicos que impactan el desarrollo vegetativo del proceso de producción de la uva de mesa. En la metodología implementada, se comienza por identificar la ubicación del viñedo, así como los datos disponibles de la estación agroclimática adjunta a éste. Los datos fueron seleccionados, convertidos a series de tiempo y filtrados con algoritmos desarrollados en el software Matlab. Para la extracción de las reglas asociativas en series de tiempo, se aplicaron los algoritmos de segmentación Haar, Wavelet y la suma acumulada de razones de menor coeficiente. Se extrajeron cuatro reglas asociativas que identifican el cambio de una fase fenológica a la siguiente. Adicionalmente, se desarrolló un modelo de predicción de elementos meteorológicos utilizando el algoritmo del k vecino más cercano. Este modelo permite predecir los elementos meteorológicos de temperatura, radiación solar y humedad relativa con una anticipación de veinticuatro horas. Ambos modelos se validaron utilizando la técnica de uno contra todos.

ABSTRACT

This project is a combined effort of researchers from the University of Sonora and the Food and Development Research Center (CONACYT), in order to identify association rules patterns of meteorological elements that accelerate or retard vegetative growth in the grape production process. It begins with a description of the vine generalities, methods to measure the vegetative growth and a description of the meteorological elements impact on the vine. Subsequently, the history of meteorological elements predictions is presented; based on the movement of stars and planets, clouds, moon phases and the movement of the wind, or the use of instruments like the thermometer, anemometer, and the rain gauge that allowed more accurate predictions. Until today, these predictions are made by applying data mining mathematical modeling tools. In this work, association rules patterns of meteorological elements that accelerate or retard vegetative growth in the grape production process are extracted. The methodology implemented begins by identifying the vineyard location and the available data from the nearest agro climatic station. Data was selected, converted into time series and filtered through Matlab algorithms. For the time series association rules extraction, the segmentation algorithm Haar Wavelet and a cumulative sum of ratios with the lowest coefficients were applied. Four association rules were extracted that identify the phenological phase change. Additionally, a meteorological elements prediction model was develop using the k nearest neighbor algorithm. This model can predict meteorological elements of temperature, solar radiation and relative humidity with a twenty-four hour advance. Both models were validated using the one versus all technique.

AGRADECIMIENTOS

Agradezco a Dios por la oportunidad que me ha dado para vivir esta gran experiencia.

A Juan José Zúñiga, Norma Lucy Vázquez y Alejandra Fonseca por su aliento y ánimo que me permitió seguir adelante.

Al Dr. Juan Martín Preciado, mi más sincero agradecimiento por su paciencia, conocimiento, dedicación y apoyo incondicional.

Al Dr. Luis Felipe Romero por sus consejos, apoyo, observaciones y colaboración para realizar éste proyecto.

A las empresas que estuvieron en el anonimato dentro de esta tesis y sin embargo me brindaron su apoyo de información veraz y precisa.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) y al Programa Integral de Fortalecimiento Institucional (PIFI 2011) por su apoyo económico.

Y por último, agradezco a la Institución que me abrió la puerta y confió en mí potencial, a la Universidad de Sonora, especialmente al Departamento de Ingeniería Industrial y de Sistemas.

¡GRACIAS!

ÍNDICE GENERAL

ÍNDICE DE TABLAS

Pág.

ÍNDICE DE FIGURAS

Pág.

ÍNDICE DE ECUACIONES

Pág.

1. INTRODUCCIÓN

En este capítulo se presentan las condiciones bajo las cuales se va a realizar este proyecto. Se comienza planteando la ambientación a la problemática en el apartado de antecedentes. A partir de los antecedentes se plantea el problema, se definen un objetivo general, dos objetivos específicos, se formula una hipótesis, se establecen alcances del proyecto y se presenta la justificación de la realización del proyecto en sí.

1.1. Antecedentes

El Centro de Investigación en Alimentación y Desarrollo (CIAD), A.C. responde a la problemática del sector alimentario de México realizando estudios, asesorías, consultorías y servicios en los sectores agroalimentario, pesquero, industrial y comercial, considerando su impacto en tres ámbitos básicos:

1) La producción, conservación, calidad y comercialización de los alimentos.
2) La salud y el desarrollo biológico del ser humano.
3) La repercusión social y económica de los procesos de desarrollo económico e integración internacional.

En el año 2000, el área de Desarrollo Regional del CIAD comenzó un proyecto para realizar un sistema de administración de un viñedo productor de uva de mesa para exportación localizado al norte de la ciudad de Hermosillo, Sonora. El sistema administrará la estructura de costos, la parte técnica del viñedo, los micronutrientes y macro nutrientes, el riego y la parte fenológica de la uva.

La investigación realizada por Preciado (2011) identificó agrupamientos de elementos meteorológicos en las etapas tempranas de la uva de mesa. También identificó si los agrupamientos de elementos meteorológicos aceleran o retardan el desarrollo vegetativo en ciertas etapas fenológicas. A partir de esta investigación, se sugiere el

uso de modelos predictivos para detectar con anticipación patrones de elementos meteorológicos que aceleren o retarden el desarrollo vegetativo de la uva de mesa.

1.2. Planteamiento del Problema

Se requiere detectar con anticipación patrones de elementos meteorológicos que aceleren o retarden el desarrollo vegetativo de la uva de mesa.

1.3. Objetivo General

Identificar reglas asociativas de patrones de elementos meteorológicos que aceleren o retarden el desarrollo vegetativo en el proceso de producción del cultivo de la uva de mesa.

1.4. Objetivos Específicos

- Identificar patrones en los elementos meteorológicos que faciliten la extracción de reglas asociativas de impacto significativo en el proceso de producción del cultivo de la uva de mesa.
- Identificar detonadores de las reglas asociativas que permitan detectar los patrones que activan dichas reglas y sustentar acciones de apoyo al proceso de cultivo.

1.5. Hipótesis

"Es posible identificar patrones de elementos meteorológicos que aceleren o retarden el desarrollo vegetativo en el proceso de producción del cultivo de la uva de mesa mediante la identificación de reglas asociativas de patrones de elementos meteorológicos."

1.6. Alcances y Delimitaciones

Los datos utilizados corresponden a observaciones de las etapas fenológicas comprendidas entre brotación, inflorescencia y florescencia de la uva de mesa *Flame Seedless* en un viñedo situado al norte de Hermosillo, para los ciclos productivos

comprendidos entre 2001 al 2005. Las observaciones de elementos meteorológicos (Temperatura, Humedad Relativa, Radiación Solar y Presión de Vapor) corresponden a las del Sistema de Información Agroclimática (SIA) (Fundación Produce Sonora, 2004).

1.7. Justificación

El modelo permitirá anticipar oportunamente condiciones de tiempo atmosférico que aceleren o retarden el desarrollo vegetativo, facilitando un mejor apoyo al proceso de toma de decisiones antes de que el desarrollo vegetativo de la uva se vea afectado por los elementos meteorológicos. Además, el modelo presentará la propiedad de ser replicable a otras fases fenológicas del cultivo.

En el capítulo dos se presentan generalidades del cultivo de la vid, así como una descripción histórica de la predicción de los elementos meteorológicos. Después se aborda una metodología de minería de datos para la extracción de reglas asociativas en series de tiempo a través de diversos algoritmos de agrupamiento. El capítulo tres desarrolla una metodología para la identificación de patrones en el cultivo de la vid en forma de reglas asociativas. En el capítulo cuatro se presentan los resultados obtenidos, así como las reglas asociativas extraídas y un modelo de predicción de tres EM (Elementos Meteorológicos).

2. PRODUCCIÓN DE UVA DE MESA Y CLIMA

En este capítulo se presentan conceptos esenciales y su interrelación que conforman la base de esta investigación. Se hace la aclaración de que algunos conceptos son tratados más a fondo que otros debido a su importancia en la estructura de este proyecto. Se comienza con una descripción de las generalidades del cultivo de la vid, enfocándose en daños causados a vid por variaciones de determinados elementos meteorológicos. Después se presenta cómo se ha tratado de anticipar el tiempo atmosférico a través de la historia y se hace énfasis en porqué esta tarea resulta tan compleja. Posteriormente, se presenta un conjunto de técnicas que pudieran permitir el desarrollo de modelos de predicción de elementos meteorológicos más precisos y exactos, así como ejemplos de dichos modelos utilizando estas técnicas.

2.1. Generalidades del Cultivo de la Vid

La vid es una planta leñosa perenne, que florece y produce frutos durante varios años. Los elementos de la producción vitícola (figura 2.1) se clasifican en dos grupos, permanentes y culturales. Dentro de los factores permanentes, el clima es posiblemente el que con mayor intensidad determina las posibilidades de producción (Hidalgo, 2002). Isbister (1918) define clima como un tiempo atmosférico generalizado, relacionada con un área grande y por un periodo de tiempo largo, generalmente mayor a 30 años, y define tiempo atmosférico (TA) como los elementos meteorológicos de la atmósfera en cualquier momento y tiempo en particular. Así mismo, también define elemento meteorológico (EM) como las condiciones particulares de la atmósfera como: temperatura, humedad, presión, viento, precipitación, radiación solar, nubosidad y polvo, en un lugar y tiempo determinado.

El ciclo anual de producción de uva puede ser alineado con la variación estacional del TA (Preciado, 2011). La relación entre los elementos meteorológicos y los

fenómenos periódicos en la vida vegetal es estudiada por la fenología (Castellví y Elías, 2001).

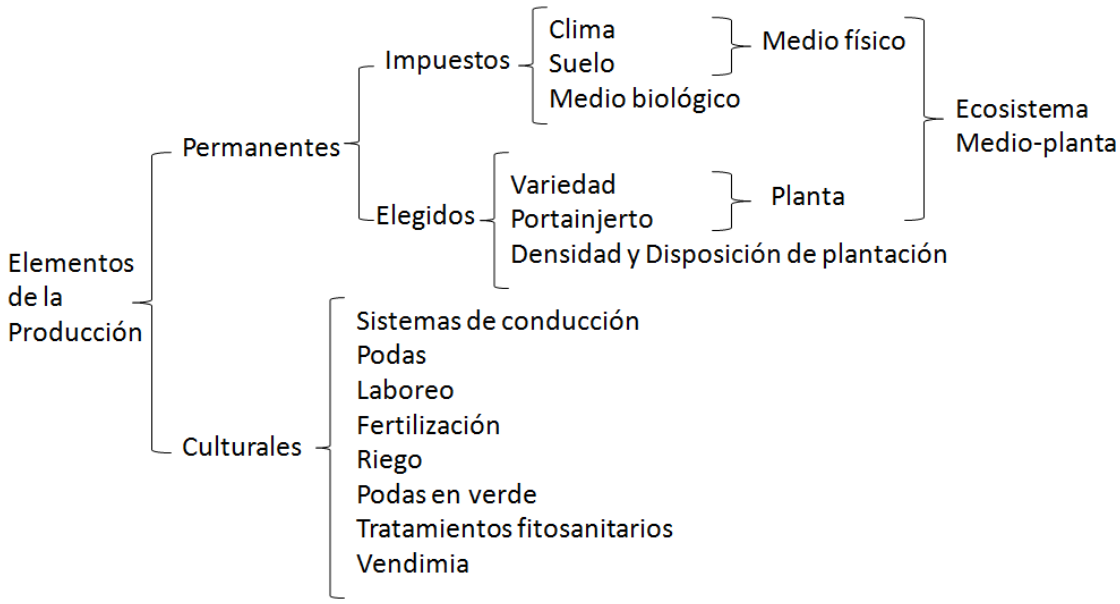


Figura 2.1. *Elementos de la Producción Vitícola (Hidalgo, 2002)*

Eichhorn y Lorenz desarrollaron un sistema de identificación de las etapas de fenológicas de la vid, el cual fue modificado por Coombe (1995) (figura 2.2). En este sistema, el desarrollo de la vid se divide en 47 etapas fenológicas, comenzando con los retoños de invierno y finalizando con la caída de las hojas en invierno, agrupadas a su vez en 5 etapas mayores (brotación e inflorescencia, florescencia, desarrollo de baya, madurez y senectud).

Las exigencias de TA de la vid, así como los daños causados por determinados elementos meteorológicos, varían dependiendo de la etapa fenológica y de la variedad de vid (Reynier, 2002). Dentro de los elementos meteorológicos que dañan la vid se encuentran las heladas, y pueden ser provocadas por diferentes fenómenos como (Hidalgo, 2002):

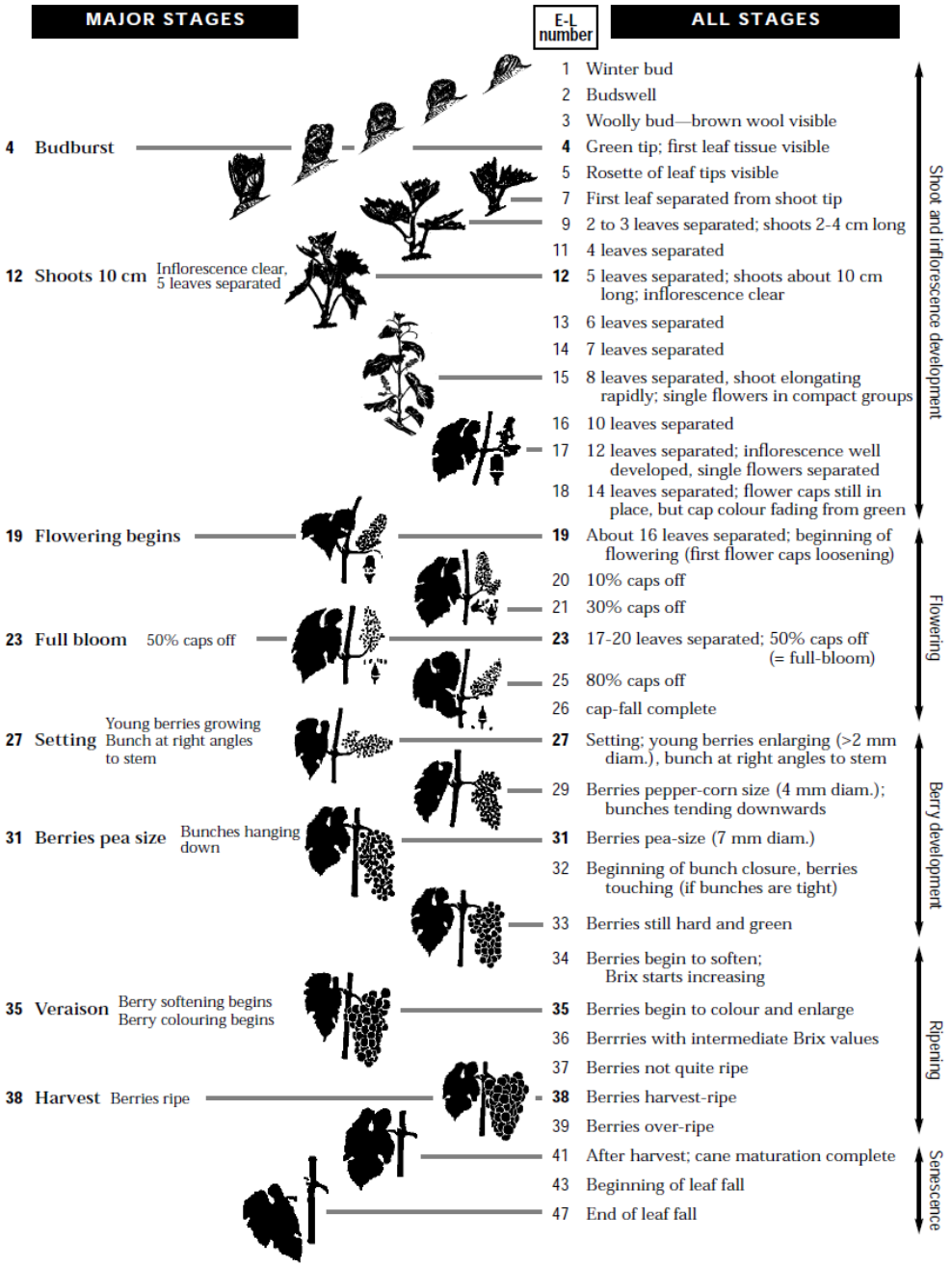


Figura 2.2. *Etapas de Desarrollo de la Vid – Sistema E-L Modificado (Coombe, 1995)*

Heladas por convección.- Se producen por la invasión de una masa de aire frío, generalmente acompañada de vientos y precipitaciones.

Heladas por irradiación.- Cuando la irradiación (emisión constante de energía calorífica) terrestre es menor a la irradiación atmosférica y solar.

Heladas por evaporación.- Cuando la humedad relativa desciende después de una lluvia, ocasiona la evaporación del agua, causando la reducción de la temperatura de la planta.

En las heladas de invierno, las yemas de la vid sufren daños con temperaturas menores de -10 °C (Hidalgo, 2002). Después de la brotación, las heladas por debajo de -2 °C destruyen el cultivo (Marro, 2000). En primavera, las plantas son muy sensibles a las heladas y basta con temperaturas menores de -0.5 °C para causar pérdida de gran parte del cultivo (Hidalgo, 2002). La sensibilidad de los órganos de la vid aumenta de 1 a 2 °C cuando están mojados por la lluvia. En otoño, la vid sufre daño cuando la temperatura desciende por debajo de -2.5 °C (Reynier, 2002).

Las temperaturas de 38-40°C son bien soportadas, aunque es posible que se presente asurado o escaldado (Hidalgo, 2002). Las temperaturas mayores a 30°C pueden quemar las hojas y los racimos si van acompañadas de sequedad, viento caliente y seco (Marro, 2000). Cuando la temperatura elevada y la humedad relativa baja son concomitantes, los golpes de sol pueden provocar el escaldado de los racimos y quemaduras en la parte de las hojas expuestas (Reynier, 2002). Cuando la temperatura es superior a 42°C, se producen desecaciones y pardeamientos en las hojas y los racimos. A temperaturas mayores de 55°C la planta muere (Hidalgo, 2002).

Las temperaturas bajas detienen la fotosíntesis, mientras que las temperaturas elevadas causan el cierre de los estomas por déficit higrométrico (diferencia entre presión de vapor saturado y presión de vapor), afectando la acumulación de azucares en la uva (Marro, 2000).

El viento puede desgarrar el limbo de las hojas y arrancar pámpanos jóvenes ocasionando una pérdida de cosecha. Además, si transporta aire muy caliente, puede provocar pardeamineto general. El viento también favorece el transporte de esporas de los hongos causantes de enfermedades del follaje a distancias de hasta 60 km (Reynier, 2002).

Debido a las exigencias de TA de los cultivos, es importante predecir los elementos meteorológicos con anticipación, para reducir así, los impactos negativos en la agricultura debido a sus variaciones (Sivakumar, 2004). De acuerdo con Sivakumar, Hansen (2007), los elementos meteorológicos clave en un modelo de predicción de TA son la lluvia, temperatura, radiación solar, humedad relativa y velocidad del viento.

2.2. Predicción de Elementos Meteorológicos

El uso de predicciones de EM para la toma de decisiones en la agricultura se remonta varios milenios atrás. En el año 700 A.C. los griegos asociaban los EM con el movimiento de las estrellas y los planetas. En el año 650 A.C. los Babilonios realizaban predicciones de EM en base a patrones en las nubes (Alter, 1994). En el año 904 D.C. se observaban las fases lunares y el movimiento del viento para realizar predicciones de EM (Rashed, Morelon, 1996).

A través del tiempo, se desarrollaron herramientas para realizar mediciones de EM que permitieron hacer predicciones más precisas. Alrededor de 1593 Galileo inventó el termómetro; en 1643 Toricelli inventó el barómetro; el anemómetro y el pluviómetro también se inventaron alrededor de dichas fechas (Alter, 1994).

En 1903, el noruego Vilhelm Bjerknes propuso tratar la evolución de la atmósfera según las leyes de la termodinámica y de la mecánica de fluidos, defendiendo que la predicción de los EM es un problema determinista (Lezaun, 2002). Sin embargo, los EM tiene un comportamiento caótico (Sneyers, 1998), aunque sea en rigor un

sistema determinista, es decir, que su comportamiento futuro puede ser determinado completamente por sus condiciones iniciales, la naturaleza determinista de este sistema no la hace completamente predecible, debido a que pequeñas variaciones en condiciones iniciales pueden implicar grandes diferencias en el comportamiento futuro, haciendo más compleja la predicción a largo plazo (Kellert, 1993).

Después de 1950, las predicciones de EM a través de modelos matemáticos produjeron resultados más cercanos a la realidad con la llegada de la simulación por computadora (Lynch, 2007). Con el desarrollo tecnológico fue posible la elaboración de radares, estaciones climáticas y satélites, que permitieron desarrollar modelos más complejos e incrementar la precisión y exactitud de los predicciones (Alter, 1994).

Según Nimbus Weather Service (2005) actualmente existen diversos métodos para predecir los EM como el método de la persistente, donde hoy es igual a mañana y asume que las condiciones atmosféricas no cambiarán en el tiempo. El método climatológico involucra el uso de datos estadísticos de los elementos atmosféricos de años anteriores. El método de predicción numérica de EM utiliza complejos programas de cómputo conocidos como modelos numéricos, que procesan datos de elementos atmosféricos como la temperatura, presión atmosférica, viento, humedad y precipitación en supercomputadoras.

Dentro de los métodos de predicción numérica de EM se encuentran:

El modelo ETA-SENAMHI: utiliza las salidas de los modelos americanos de aviación (AVN) y WAFS como condiciones iniciales. En la figura 2.3 se presenta una imagen de este modelo.

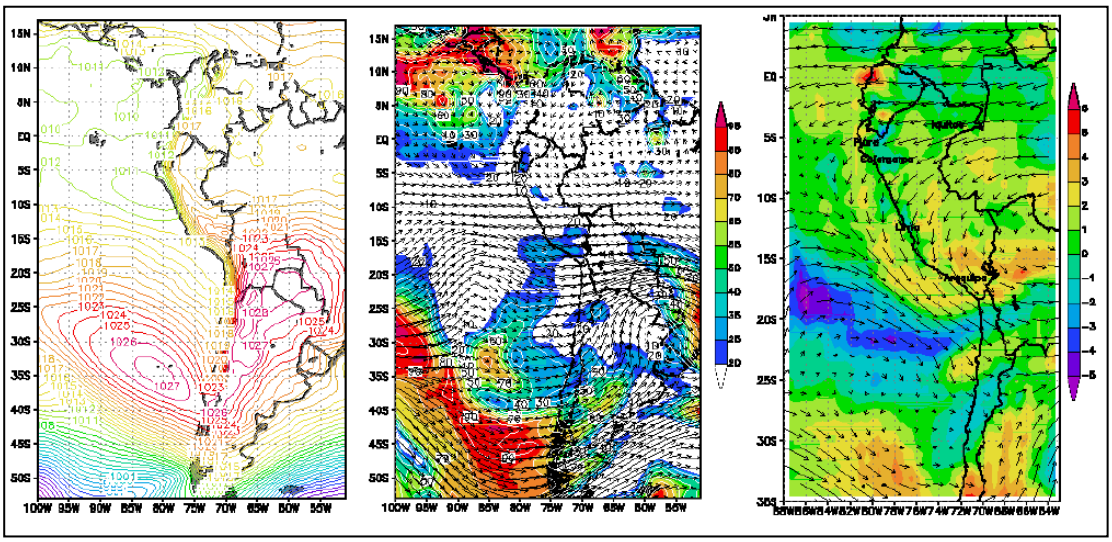


Figura 2.3. *Imagen de Modelo ETA-SENAMHI (Nimbus Weather Service, 2005).*

Modelo RAMS: modelo muy versátil que permite bajar la resolución a menos de un kilómetro. En la figura 2.4 se presenta una imagen de este modelo.

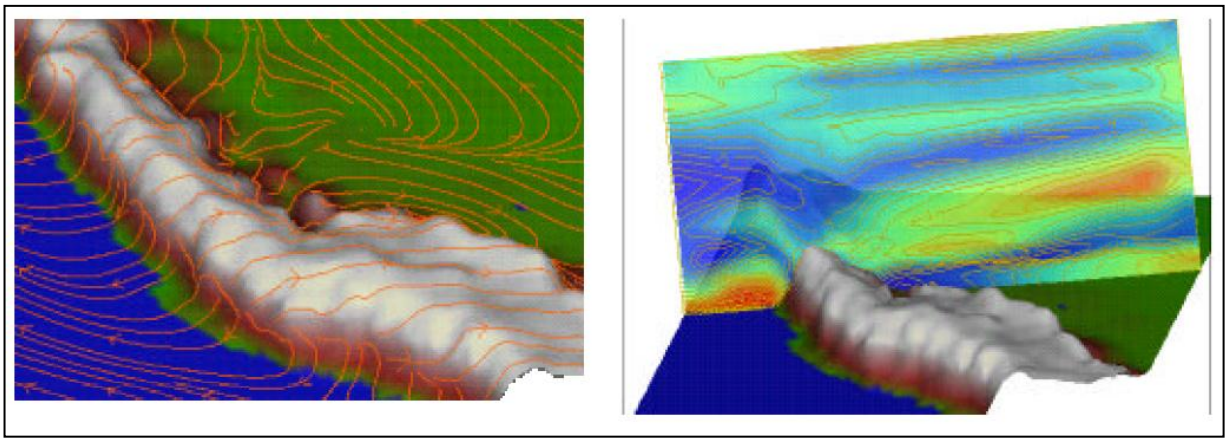


Figura 2.4. *Imagen de Modelo RAMS (Nimbus Weather Service, 2005).*

Modelo CCM3: modelo climático global acoplado océano-atmosférico-tierra. En la figura 2.5 se presenta una imagen de este modelo.

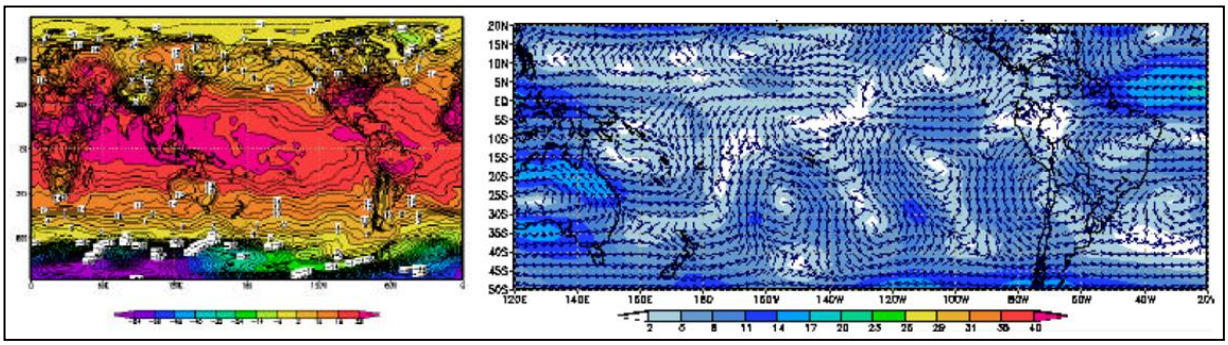


Figura 2.5. *Imagen de Modelo CCM3 (Nimbus Weather Service, 2005).*

Los métodos de predicción numérica se pueden aplicar para un modelo regional o un modelo global. En el modelo global, se puede tener predicciones confiables hasta con una semana de anticipación y la confiabilidad aumenta si el periodo es menor a cinco días. Un modelo regional puede ser útil para el pronóstico del tiempo con alta resolución (desde 500-600 metros hasta 60-100 kilómetros) con antecedencia de 48 horas (Nimbus Weather Service, 2005).

En la actualidad, los modelos predictivos de EM utilizados para la toma de decisiones en la agricultura, están basados en la premisa que se debe tomar ventaja de las bases de datos de elementos meteorológicos (Sivakumar, 2004). Dichos modelos predictivos pueden ser desarrollados a través de modelos matemáticos para la identificación de patrones o Minería de Datos (MD) (McGovern et al., 2011) sin embargo, la aplicación de herramientas de MD para el pronóstico en la agricultura es un proceso relativamente reciente (Veenadhari et al., 2011).

2.3. Minería de Datos

El incremento de capacidad y accesibilidad económica de las tecnologías de generación y almacenamiento automático de datos, ha provocado un desfase marcado en la generación de datos y su transformación en información. “Conforme el volumen de datos aumenta, la cantidad de personas que los comprenden disminuye en proporción alarmante. Esto ocasiona que información potencialmente útil no sea

transformada en explícita y permanezca oculta en los datos" (Witten, Frank, 2005), causando así que las organizaciones se vuelvan ricas en datos y pobres en conocimiento (Tang, McLennan, 2005).

Fue la disponibilidad de grandes volúmenes de datos lo que transformó los métodos, técnicas y formas de analizar los datos, orientándolo hacia determinadas técnicas especializadas englobadas bajo el nombre de MD (Pérez y Santín, 2007). Witten, Frank (2005) definen la MD como el proceso de descubrimiento de patrones en los datos. Fayyad et al. (1996) la definen como la aplicación del análisis de datos y algoritmos de descubrimiento que producen un conjunto particular de patrones (o modelos) ocultos en los datos, definición con la cual concuerdan Tang, McLennan (2005), pero agregan que el descubrimiento de patrones debe ser por medios automáticos. De acuerdo con Bow (2002), "un patrón puede ser definido como una descripción cuantitativa o estructural de un objeto o algunas otras entidades de interés".

Algunos de los métodos y técnicas actuales incluidas dentro de la MD datan de la década de 1950, pero el término "Minería de Datos" no fue acuñado hasta una década después. Sin embargo, en ese entonces, el término MD era utilizado para describir la detección de patrones sin una significancia estadística (Bowen, 2006). Desde entonces, se le han dado diversos nombres además de MD incluyendo extracción de conocimiento, descubrimiento de información, arqueología de datos, identificación de patrones y procesamiento de patrones de datos (Fayyad et al., 1996).

No fue sino hasta 1989 cuando surgió el término Descubrimiento de Conocimiento en Bases de Datos (KDD, por sus siglas en inglés, *Knowledge Discovery in Databases*) en el primer taller de KDD (Piatetsky-Shapiro, 1989), para enfatizar que el conocimiento es el producto final de los descubrimientos basados en bases datos.

El proceso de KDD (figura 2.6) inicia con el planteamiento del problema (objetivo a alcanzar) y establece cinco actividades antes de descubrir el conocimiento buscado. Conviene destacar que la propiedad interactiva representada por la retroalimentación de cada etapa, puede implicar el regreso a cualquiera de las etapas anteriores. Las etapas del KDD se describen a continuación (Fayyad et al., 1996):

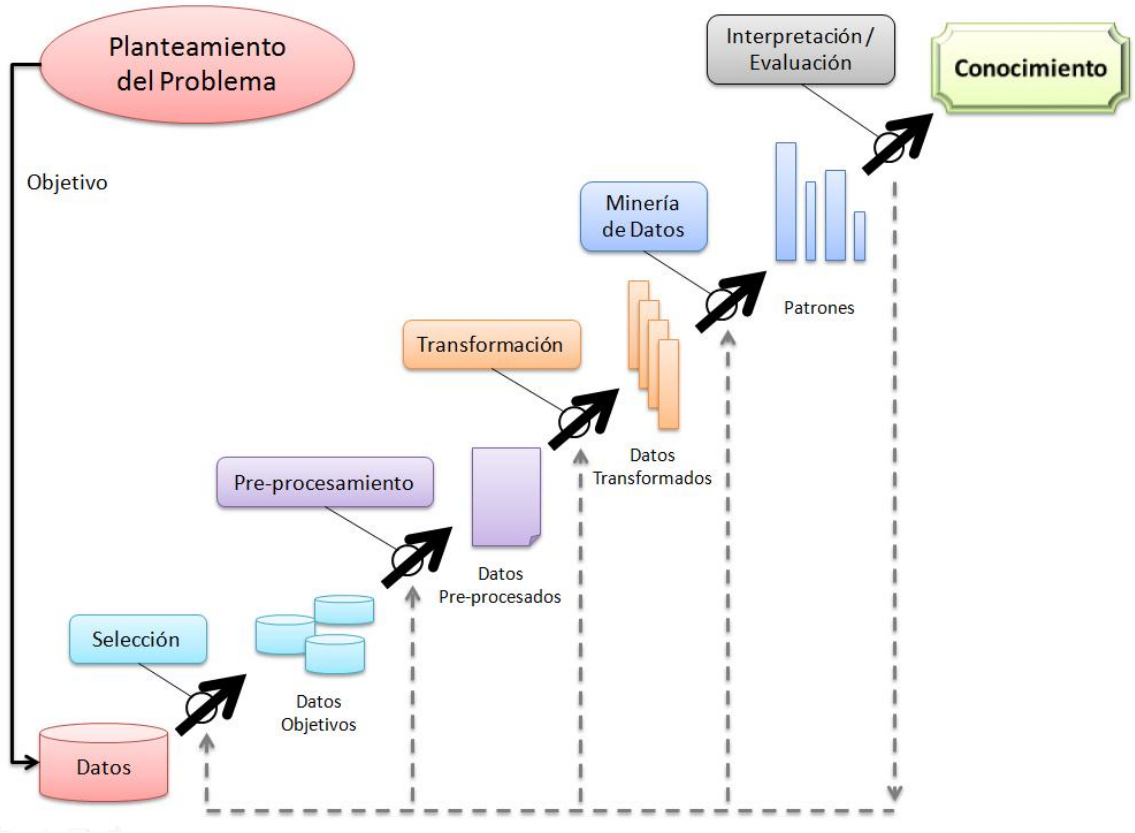


Figura 2.6. *Proceso KDD (Fayyad et al., 1996)*

Planteamiento problema. Se requiere primero comprender el dominio (contexto) de la aplicación a desarrollar a partir del conocimiento previo proporcionado por el experto; este conocimiento deberá ser consistente con el objetivo del proceso KDD, que representa la visión del analista y los requerimientos de conocimiento del usuario final.

Creación del conjunto de datos. Es necesario seleccionar un conjunto de datos, o centrarse en el subconjunto de variables o muestras de datos, en donde el descubrimiento será desarrollado.

Limpieza y procesamiento previo de datos. A través de operaciones básicas, que incluyen la remoción de ruido (en caso de ser apropiado), recolección de información necesaria para el modelo o medición del ruido y decisión de estrategias a utilizar para el manejo de datos perdidos.

Reducción y proyección de datos. Encontrar rasgos útiles para representar datos, dependiendo del objetivo de la aplicación. Con la reducción de la dimensionalidad o a través de métodos de transformación, el número efectivo de variables a considerar puede ser reducido, o bien pueden ser encontrados datos sin variación.

Identificación de método. Identificar un método o técnica analítica de la MD de acuerdo con el objetivo de la aplicación del proceso KDD.

Exploración de análisis, modelo y selección de hipótesis. Selección de algoritmo(s) y método(s) de minería de datos a utilizar en la búsqueda de patrones de datos. Este proceso incluye la decisión de cuáles métodos y parámetros son más apropiados (por ejemplo, modelos de datos categóricos son diferentes de los modelos de vectores de números reales) y la correspondencia con un método en particular de minería de datos, sobre todo el proceso KDD (por ejemplo, el usuario final está más interesado en entender el modelo que las capacidades de predicción).

Minería de datos. Búsqueda de patrones de interés, con una representación particular o un conjunto de tales representaciones, incluyendo reglas o árboles de decisión, regresión y agrupamiento. El usuario puede ayudar significativamente al método de minería de datos a través de correcciones en pasos precedentes.

Interpretación de los patrones minados. Posiblemente se tenga que regresar a cualquier paso entre el primero y el séptimo para futuras iteraciones. Este paso también implica la visualización de los patrones y modelos extraídos o la visualización de los datos dados en los modelos extraídos.

Aplicación del conocimiento descubierto. Ya sea usando el conocimiento directamente, incorporando el conocimiento en otro sistema para la acción adicional, o simplemente documentándolo y divulgándolo a los usuarios interesados. Este proceso también incluye la revisión y resolución de posibles conflictos con el conocimiento previo.

Las herramientas de MD se clasifican dependiendo del objetivo que se desea obtener al analizar los datos. Según Pérez y Santín, (2007) las clasifican en tres categorías (figura 2.7):

1) Las técnicas predictivas o aprendizaje supervisado: especifican el modelo para los datos en base a un conocimiento teórico previo. Es decir, estas técnicas tienen el objetivo de predecir o estimar los valores de un atributo.

2) Las técnicas descriptivas o aprendizaje no supervisado: no asignan ningún papel predeterminado a las variables. Estas técnicas no suponen la existencia de las variables dependientes ni independientes ni tampoco suponen la existencia de un modelo previo para los datos.

3) Las técnicas auxiliares: son herramientas de apoyo más superficiales y limitadas. Son métodos basados en funciones básicas de sistemas de administración de bases de datos como consultas e informes, enfocados hacia la verificación.

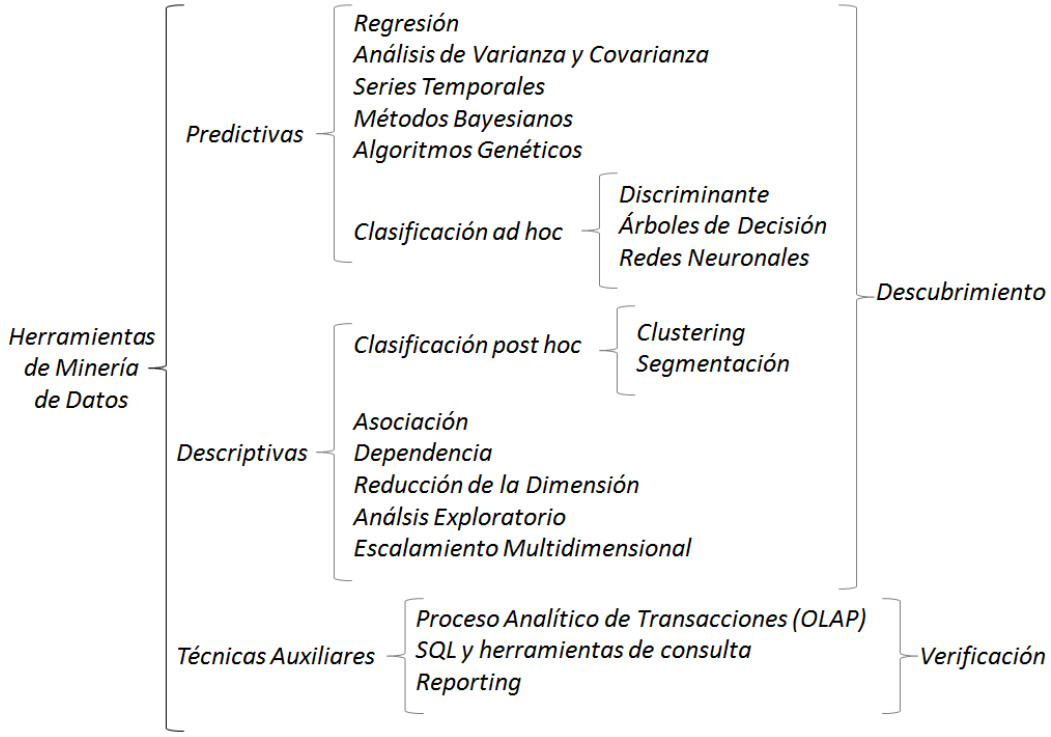


Figura 2.7. *Clasificación de Herramientas de MD (Pérez y Santín, 2007)*

Dado que en esta investigación se analiza el proceso continuo del desarrollo vegetativo de las etapas fenológicas tempranas (brotación e inflorescencia) de la producción de uva de mesa y como éste es afectado continuamente por los EM, se presentará el marco conceptual del análisis de series de tiempo (ST) como la estructura analítica que permite una explicación más precisa del comportamiento climático y el desarrollo fenológico de la uva de mesa.

2.4. Series de Tiempo

Las ST se presentan con regularidad en la práctica de una gran variedad de áreas del conocimiento como en economía (e.g. ingresos en meses consecutivos, utilidad por año de empresas), en fenómenos físicos (e.g. meteorología, geofísica), en estudios demográficos, en procesos de control, entre muchas otras áreas (Chatfield, 2005). La representación gráfica de una ST se presenta en la figura 2.8.

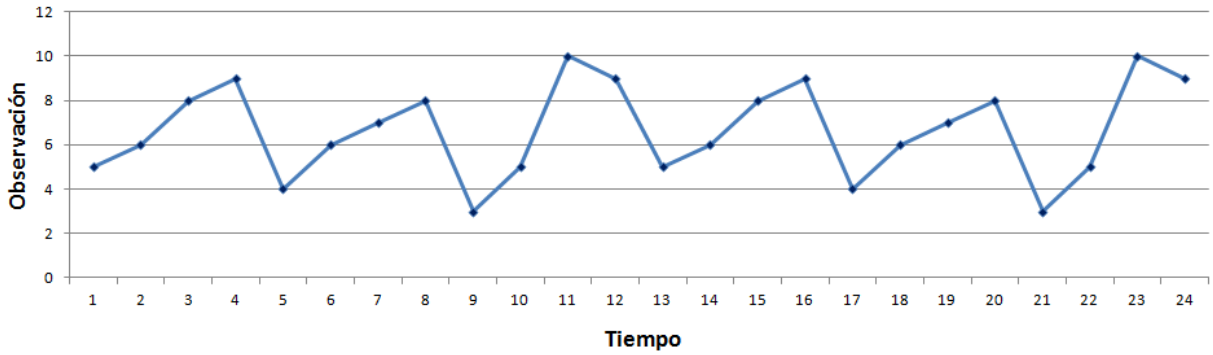


Figura 2.8. *Ejemplo de una Serie de Tiempo*

Chatfield (2005) define una ST como una colección de k observaciones realizadas secuencialmente en el tiempo t; definición con la que concuerda Rodríguez (2002). Preciado (2011) define una observación como m características que forman un vector $m - dimensional$,

$$X_k = [X_{k1}, X_{k2}, \dots, X_{km}]^T \qquad k, m \in N \qquad (ec. 2.1)$$

El análisis de las ST implica el estudio de cuatro conceptos básicos que brindan información para una explicación más precisa del comportamiento del proceso a través del tiempo (Bowerman et al., 2009):

1) Tendencia: la variación de la media en un determinado tiempo.
2) Ciclo: variación alrededor de la tendencia en un periodo fijo.
3) Estacionalidad: variación que se completa y repite año con año.
4) Fluctuaciones Irregulares: movimientos erráticos de una serie de tiempo que siguen un patrón indefinido o irregular.

Una clasificación de ST es presentada en la figura 2.9. Kitagawa (2010) clasifica una ST como continua si las observaciones son continuas en el tiempo (e.g. dispositivo análogo), o discreta si las observaciones son a intervalos de tiempo. Si las observaciones consisten en una característica única en el tiempo, la ST es considerada univariada, o multivariada si se contempla más de una característica en la observación.

Si la ST es transformada en n conjuntos de observaciones (e.g. segmentación), Morchen (2006) clasifica como intervalo de tiempo, y la clasifica como punto de tiempo si la ST es discreta y se contemplan las observaciones independientemente. El mismo autor también clasifica una ST en numérica si las observaciones son numéricas o simbólica si las observaciones son nominales.

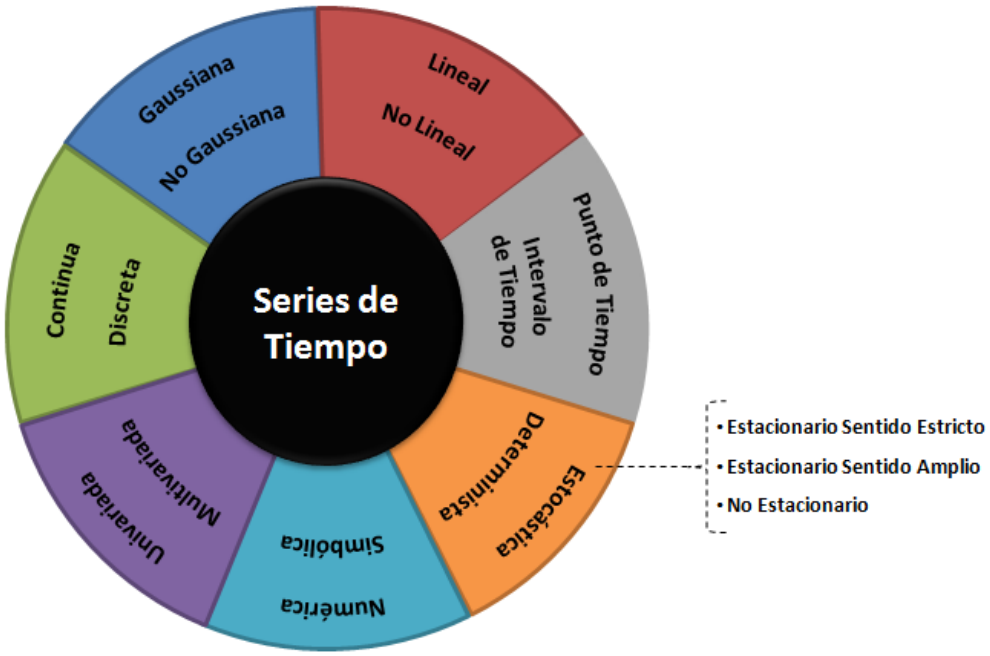


Figura 2.9. *Clasificación de ST (Adaptado de Chatfield, 2005, Morchen, 2006), Kitagawa, 2010)*

Al realizar el análisis de ST se debe considerar el hecho de que las observaciones sucesivas son usualmente dependientes. Las observaciones $X_{1j}, \dots, X_{kj}$ son independientes si para cualquier subcolección

$$X_{ij}, \dots, X_{(i+n)j}, \qquad 1 \leq i < n < k, 1 \leq j < m,$$

se cumple

$$P\left(X_{ij} \cap \dots \cap X_{(i+n)j}\right) = P\left(X_{ij}\right) \times \dots \times P\left(X_{(i+n)j}\right) \qquad (ec. 2.2)$$

Cuando las observaciones sucesivas son dependientes, valores futuros pueden ser predichos de observaciones pasadas. Una ST es determinista si se predice exactamente, o estocástica si se predice parcialmente a través de distribuciones de probabilidad (Chatfield, 2005).

Contreras (2007) define un proceso estocástico como un conjunto de observaciones aleatorias ordenadas en instantes temporales con una distribución de probabilidad conjunta. Una ST estocástica es estacionaria en sentido estricto si la distribución de probabilidad conjunta satisface

$$P(X_{1j}, \dots, X_{kj}) = P(X_{(1+\tau)j}, \dots, X_{(k+\tau)j}), \quad \forall\, X_{1j}, \dots, X_{kj}, \tau. \qquad (ec.\,2.3)$$

La ST estocástica es estacionaria en sentido amplio si su media es constante y su función de auto covarianza solo depende del retraso τ tal que

$$E[X_{kj}] = \mu_j \qquad (ec.\,2.4)$$

Y

$$Cov[X_{kj}, X_{(k+\tau)j}] = \gamma_j(\tau_j). \qquad (ec.\,2.5)$$

La ST estocástica es no estacionaria si la estructura de la ST cambia con el tiempo (Chatfield, 2005).

Por lo general, los métodos tradicionales de análisis en ST, como regresión lineal y modelos de Box-Jenkins (promedio móvil, autorregresivo y modelos mixtos), se encuentran limitados por supuestos de estacionariedad de la ST y/o normalidad e independencia de los residuales (Bowerman et al., 2009).

Las herramientas de Minería de Datos en Series de Tiempo (TSDM, por sus siglas en ingles, *Time Series Data Mining*) no se encuentran limitadas por los supuestos

anteriores, y pueden caracterizar y predecir exitosamente ST complejas, no periódicas, irregulares y caóticas. Además son aplicables a ST con características de procesos estocásticas, pero que ocasionalmente contienen patrones distintivos pero ocultos que caracterizan a los eventos (Povinelli, 1999).

La TSDM contempla dos conceptos clave: evento y patrón temporal (Povinelli, 1999). Montgomery (2003) define un evento como un subconjunto de interés del espacio muestral de un experimento aleatorio, mientras que Povinelli (1999) lo define como ocurrencia de importancia. En el contexto de ST, una estructura identificable a través del tiempo que identifica un descubrimiento de información es denominado un patrón temporal (Povinelli, 1999). Si los patrones temporales son previamente desconocidos y frecuentes en una serie de tiempo, se les denomina *motif* (Morchen, 2006). La figura 2.10 presenta el ejemplo de un *motif*; como se puede observar, tanto el segmento A como el B, a pesar de no ser completamente idénticos, comparten muchos rasgos que en común.

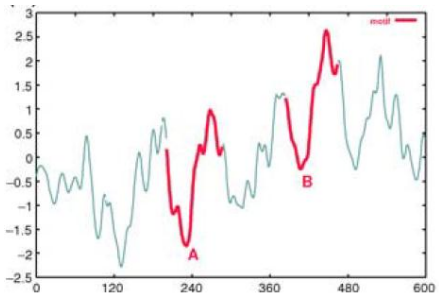


Figura 2.10. *Ejemplo de Motif (Tanaka et al., 2005)*

Generalmente, el descubrimiento de conocimiento requiere la extracción de reglas simbólicas fáciles de interpretar denominadas reglas asociativas (Morchen, 2006). Para Hernández et al. (2004) una regla asociativa expresa patrones de comportamiento entre datos (nominales), en función de la aparición conjunta de valores de dos o más características. Chakrabarti et al. (2009) definen las reglas asociativas como un forma describir todos los elementos del espacio muestral de un

evento en particular (figura 2.11). Las reglas asociativas temporales combinan las reglas asociativas tradicionales con aspectos temporales (Morchen, 2006).

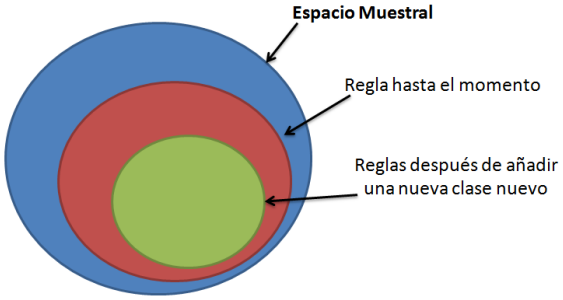


Figura 2.11. *Representación de Regla Asociativa (Chakrabarti et al., 2009)*

2.5. Metodología para Extracción de Reglas Asociativas Temporales

Pisón et al. (2005) plantean una metodología para la extracción de reglas asociativas temporales donde sobresalen seis etapas: filtrado de datos, un primer agrupamiento de los datos en función del tiempo (Segmentación de ST), un segundo agrupamiento sin contemplar el tiempo (*clustering* en inglés), extracción de reglas asociativas y su presentación en forma entendible para el usuario. Al aplicar esta metodología, se facilita la búsqueda de correlaciones temporales entre observaciones y se muestra en forma comprensible las dichas relaciones locales.

2.5.1. Filtrado de Datos

Una forma de estimar valores faltantes es a través de la interpolación lineal (Preciado, 2011). Dadas las observaciones

$$(X_t), (X_{t+1}), (X_{t+2}),$$

donde el subíndice $t+1$ representa la observación a ser estimada, la estimación de la observación desconocida está definida por (ec. 2.6),

$$X_{t+1} = X_t + \frac{\big((t+1)-t\big)(X_{t+2}-X_t)}{\big((t+2)-t\big)}. \qquad (ec.\,2.6)$$

La identificación de valores atípicos se puede realizar mediante el diagrama Caja-Bigotes (figura 2.12). Después de filtrar los datos, la siguiente etapa es la segmentación de ST, sin embargo, es necesario definir antes los agrupamientos en general.

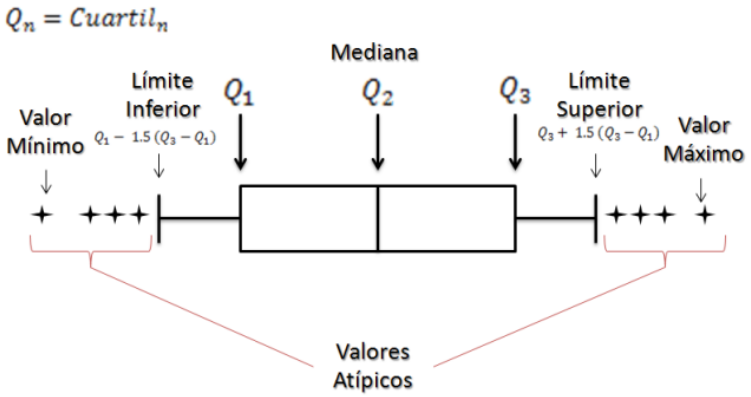


Figura 2.12. *Diagrama Caja-Bigotes (Adaptado de Montgomery y Runger, 2003)*

2.5.2. Agrupamiento

Agrupamiento (*Clustering* en inglés) es un método estadístico multivariante de clasificación automática que trata de situar las observaciones según su homogeneidad en conglomerados desconocidos previamente (Pérez y Santín, 2007). Para (Xu y Wunsch, 2009) agrupamiento es la división de un conjunto de datos en subgrupos homogéneos en base a una medida de similitud, tal que la similitud entre los datos de un subgrupo sea mayor que la similitud entre datos de diferentes subgrupos. Para Preciado (2011) la formación de los conglomerados se basa en el principio de máxima similitud al interior de los conglomerados y mínima similitud entre conglomerados.

Las medidas de similitud se clasifican en cuatro categorías (figura 2.13). 1) Las distancias son las distintas medidas entre los puntos del espacio definido por las observaciones. 2) Los coeficientes de asociación se utilizan con variables cualitativas (generalmente dicotómicas), aunque se pueden aplicar a variables cuantitativas sacrificando alguna información proporcionada por las variables; y miden la concordancia o conformidad entre los estados de dos columnas de datos; 3) los coeficientes angulares miden la proporcionalidad e independencia entre los vectores que definen los individuos; y 4) los coeficiente de similitud probabilística miden la homogeneidad del sistema por particiones o sub particiones del conjunto de los individuos e incluyen información estadística (Pérez y Santín, 2007).

Los algoritmos de agrupamiento se clasifican en base a la manera en determinar los agrupamientos (tabla 2.1). Se pueden encontrar en la clasificación de duros, si cada observación es asignada a un solo conglomerado, o difusos, si la observación puede asignarse a más de un conglomerado con un determinado grado de pertenencia. El algoritmo se considera jerárquico si se realizan una secuencia de particiones anidadas, ya sean de aglomeración (desde un conglomerado con una observación única, hasta un conglomerado con todas las observaciones) o divisiva (viceversa). En los algoritmos particionales se asignan observaciones a un conglomerado sin una secuencia de particiones anidadas. Al ser particional, el punto inicial de la n partición es aleatorio, al menos que se cuente con conocimiento previo (Xu y Wunsch II, 2009).

Los algoritmos de agrupamiento basado en redes neuronales están muy relacionados con el concepto de aprendizaje competitivo, y tratan de imitar la forma en que nuestras neuronas se comunican entre sí. Los algoritmos en base a Kernel son una transformación tal que una frontera no lineal en el espacio original puede ser linealmente separada. Los algoritmos de agrupamiento secuenciales se utilizan en multiconjuntos de puntos de tiempo (Xu, Wunsch II, 2009).

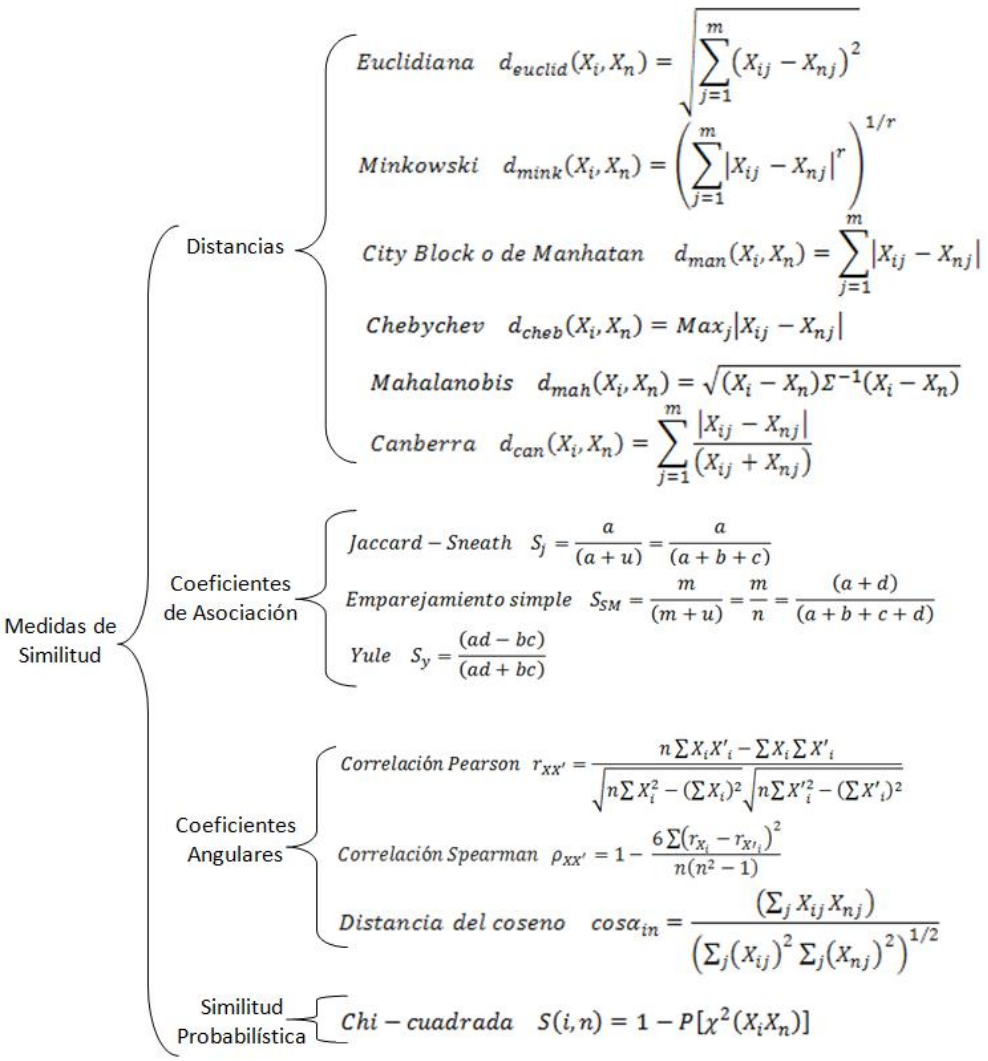


Figura 2.13. *Ejemplos de medidas de similitud (Adaptado de: Pérez y Santín, 2007)*

Los algoritmos de agrupamiento basados en búsqueda definen una función de evaluación de los agrupamientos de las observaciones y utilizan una técnica basada en búsqueda para encontrar la partición de óptimo global de las observaciones. Los algoritmos basados en grafos construyen primero un grafo o hypergrafo y luego aplican un algoritmo de agrupamiento. Los algoritmos de agrupamiento basados en cuadrículas crean una cuadrícula finita en el espacio de las observaciones, calculan

Clase	Ejemplos de Algoritmos	Ejemplos de investigaciones que han utilizado las clases de algoritmos
Difuso	k-medias difuso (Bezdek, 1974) k-modos difusos (Huang, Ng, 1999) c-medias difuso (Bezdek, 1981)	Predicciones sísmicas (Aydin et al., 2009) Predicciones de flujo redes de comunicación (Khedkar, Keshav, 1992)
Particional	k-medias (Macqueen,1967) k-medias continuo (Faber, 1994) x-medias (Pelleg, Moore, 2000) k-probabilidades (Wishart, 2002)	Predicción de desempeño de estudiantes (Oyelade et al., 2010) Monitoreo de movimiento de objetos (Zhang et al., 2008)
Jerárquico Aglomerativo	Método de Grafo: enlace simple (Florek et al. 1951) enlace completo enlace promedio de grupo (Jain y Dubes, 1988) Método Geométrico: método de Ward (Ward Jr., Hook 1963) método de centroide (Jain y Dubes, 1988) método de mediana (Gower, 1967)	Segmentación de imágenes (Maeda et al. 2008) Mejoramiento de la predicción funcionamiento de proteínas (Eisner et al., 2005) Predicción de ubicación de límites prosódicos (Ostendorf, Veilleux, 1994) Predicción de patrones climáticos Jan et al., (2008)
Divisivo	Análisis Divisivo (Kaufman y Rousseeuw, 1990)	
Redes Neuronales	k-medias en línea (Linde et al., 1980) líder-seguidor (Duda et al., 2001) teoría de la resonancia adaptativa (Carpenter, Grozzberg, 1987)	Predicción del consumo de electricidad (Abuyev et al., 2005) Predicción de actividad de Inhibidores de la proteasa VIH-1 (Andonie et al. 2005) Procesamiento de lenguajes (Borensztajn et al., 2009) Predicción de tiempo atmosférico (Paras et al., 2007)
Kernel	Kernel PCA (Muller et al., 2001) Kernel MaxEnt (Jenssen et al., 2006)	Seguimiento de video (Li et al., 2008) Predicciones estructuradas (Lampert, Blaschko, 2009)
Secuencial	Smith-Waterman (Durbin et al., 1998)	Predicciones de tiempo atmosférico (Cessna et al., 2008)
Basados en Búsqueda	Algoritmos Genéticos (Holland, 1975) Búsqueda tabú (Glover, 1989) Al-Sultan (Al-Sultan, 1995) j-medias (Hansen, Mladenovic, 2001)	Predicción de lluvia (Sen, Oztopal, 2001) Predicción de Estructura de Cristales (Luke, 2006)
Grafos	Camaleón (Karyspis et al., 1999) Cactus (Ganti et al., 1999)	Predicción de cadena proteínica lateral (Canutescu et al., 2003)
Cuadrícula	STING (Wang et al., 1997) Optigrid (Kei, Hinneburg, 1999)	Predicciones climáticas y de tiempo atmosférico (Boyd J., 2008)
Densidad	DBSCAN (Ester, 1996) BRIDGE (Dash et al., 2001) DENCLUE (Hinneburg, Keim, 1998)	Evaluación de significancia estadística de clasificación de resultados (Alnemer et al., 2010)
Modelos Probabilísticos	EM (Dempster et al., 1997) COOLCAT (Barbara et al., 2002)	Aplicación a modelos de cura mezclada (Yu, Tiwari, 2007)
Sub Espacios	PROCLUS (Aggarwal et al., 1999) DOC (Procopiuc et al., 2002)	Predicción de cáncer bio-molecular (Bertoni et al., 2004)
Series de Tiempo Completo	Abajo-Arriba Arriba-Abajo Ventana-Corrediza (Graves, Pedrycz, 2009)	Predicción de EM (Radhika, Shashi, 2009) Predicción del viento (Taylor et al., 2009) Predicción de temperatura atmosférica (Singh et al., 2011) Predicción de EM (Myers et al., 2009) Clasificación de patrones EM (Romani et al., 2010)
Subsecuencias	TSRF (Das et al., 1998)	

Tabla 2.1. *Clase de Algoritmos de Agrupamiento (*Adaptado de: Gan et al., 2007), Keogh et al., 2003, Xu y Wunsch II, 2009).

la densidad, ordena las celdas de acuerdo a su densidad e identifica los centros de los agrupamientos y sus celdas vecinas (Gan et al., 2007).

Los algoritmos basados en densidad generan los agrupamientos basados en regiones en el espacio de observaciones con alta densidad y que se encuentran separadas por regiones de baja densidad. En los algoritmos basados en modelos probabilísticos se asume que las observaciones son generadas por una mezcla finita de distribuciones de probabilidad. Los algoritmos de agrupamiento de subespacio se pueden clasificar en dos categorías generales, Arriba-Abajo que encuentra un agrupamiento inicial en todo el conjunto de las observaciones y evalúa los subespacios, o Abajo-Arriba que encuentra regiones densas en espacios de baja dimensionalidad y las combina para formar agrupamientos (Gan et al., 2007).

Los algoritmos de agrupamiento de ST se pueden dividir en dos categorías: agrupamiento completo y agrupamiento de subsecuencias. En el agrupamiento completo, dado un conjunto de ST, el objetivo es agrupar ST tal que las ST de un mismo agrupamiento tengan la mayor similitud entre si y la menor similitud con las ST de distintos agrupamientos.

En el agrupamiento de subsecuencias, dada una única serie de tiempo, se utiliza el enfoque ventana-corrediza para extraer ST traslapadas y posteriormente se agrupan de acuerdo sus similitudes (Keogh et al. 2003). Keogh et al. (2003) demuestran que el agrupamiento de subsecuencias no tiene sentido, dado que las salidas son independientes de las entradas.

Dentro de los algoritmos de agrupamiento, los algoritmos K-medias son de los más utilizados debido a su sencillez y efectividad (Jambhulkar et al., 2011). Se consideran como algoritmos K-medias los que involucran la minimización de las distancias de las observaciones del conglomerado con su centro, lo cual se consigue a través de la

actualización iterativa del conglomerado, al recolocar cada observación al conglomerado cuyo centro sea más cercano (Everitt et al., 2011). Al realizar cualquier agrupamiento, es importante hacer una evaluación de dicho agrupamiento.

La evaluación de la calidad del agrupamiento se puede realizar a través de diferentes índices de validación. Bezdek (1974b) define el Coeficiente de Partición (CP) como

$$CP = \frac{\sum_{t=1}^{k}\sum_{i=1}^{S}\delta_{ti}}{k}, \qquad (ec.2.7)$$

donde δ_{ti} es la pertenencia de la observación t en el conglomerado i. La partición óptima se obtiene maximizando CP para valores de $1 \leq S \leq k$.

Bezdek (1974a) también define la Entropía de Partición (EP) como

$$EP = \frac{-\sum_{t=1}^{k}\sum_{i=1}^{S}\left(\delta_{ti}\ log(\delta_{ti})\right)}{k}, \qquad (ec.2.8)$$

donde la partición óptima se obtiene minimizando la entropía. Tanto el índice CP como el EP presentan la desventaja de que evalúan el conglomerado considerando exclusivamente los grados de pertenencia y no la estructura geométrica (Pal y Bezdek, 1995).

El índice de partición (SC) es la relación entre la suma de la densidad (concentración) y separación entre conglomerados. Esto es, la suma individual de la medida de validación normalizada de cada conglomerado dividido a través de la cardinalidad difusa de cada conglomerado (Benzaid et al., 1996). Es determinado por

$$SC\ (c) = \sum_{i=1}^{S}\frac{\sum_{t=1}^{k}(\delta_{ti})\|X_t - C_i\|^2}{\sum_{t=1}^{k}\delta_{ti}\sum_{i'=1}^{S}\|C_{i'} - C_i\|^2}, \qquad (ec.2.9)$$

donde un menor coeficiente de esta relación nos indica una mejor partición.

El índice de validación Silhouette (S) (ec.2.10) desarrollado por Rousseeuw (1987), define a_t como la distancia promedio de la observación t con el resto de su conglomerado, b_t como la distancia promedio de la observación t con la distancia promedio del conglomerado más cercano.

$$S(t) = \frac{\big(b(t) - a(t)\big)}{max\big(a(t), b(t)\big)} \qquad (ec. 2.10)$$

Si $S(t)$ es cercano a 1, significa que la observación t se encuentra bien conglomerada, si $S(t)$ es 0, la observación puede ser asignada a otro conglomerado, y si $S(t)$ es cercano a -1, la observación fue mal conglomerada.

El índice de validación de Dunn (D) (Dunn, 1974), se basa en la idea de identificación de conglomerados compactos, bien separados y se define como

$$D = \min_{1\le i\le S}\left\{\min_{\substack{1\le i'\le S\\ i\le\neq i'}}\left\{\frac{d(c_i, c_{i'})}{\max\limits_{1\le i''\le S}\{d'(c_{i''})\}}\right\}\right\}, \qquad (ec. 2.11)$$

donde $d(c_i, c_{i'})$ es la distancia entre conglomerados c_i y $c_{i'}$ (inter conglomerados): $d'(c_{i''})$ es la distancia intra conglomerados. El número de conglomerados que maximice D es considerado el número óptimo de conglomerados.

2.5.3. Segmentación de Series de Tiempo

La segmentación de ST es la descomposición de ésta en segmentos relevantes (Keogh et al., 2001). Se han desarrollado diversos algoritmo para segmentar tanto ST

univariadas como ST multivariadas. Dentro de los clasificados como univariados entre otros están: Aproximación Agregada a Trozos (PAA, por su siglas en inglés *Piecewise Aggregate Approximation*) (Keogh et al, 2000), Aproximación Adaptativa Constante a Trozos (APCA, por sus siglas en ingles *Adaptive Piecewise Constant Approximation*) (Keogh et al. 2001), mientras que para segmentar ST multivariadas se pueden mencionar Gath-Geva Modificado (Abonyi, Feil, 2007), evolución diferencial (Graves, Pedrycz, 2009), De Abajo Hacia Arriba (Bottom-up) multivariado (McCue, Hunter, 2004).

Los algoritmos de segmentación de ST de uso más generalizado de acuerdo a la literatura se pueden dividir en tres enfoques principales: 1) Abajo-Arriba, 2) Arriba-Abajo y 3) Ventana-Corrediza (Graves, Pedrycz, 2009), los cuales son descritos a continuación.

El enfoque de Abajo-Arriba (BU, por sus siglas en ingles, *Bottom-Up*) realiza una fina segmentación de la ST, calcula el error de las combinaciones de segmentos y une los segmentos con base en el menor error hasta cumplir un parámetro preestablecido. El parámetro puede ser un determinado número de segmentos o error, dependiendo del objetivo para el que se realice la segmentación. El enfoque BU con parámetro de número de segmentos se presenta en la figura 2.14.

Enfoque de Segmentación BU

- *Crea número máx de segmentos de longitud mínima*
- *Calcula los erroes de las combinaciones de segmentos*

○ *while núm segmentos > núm segmentos deseados*
  ○ *Encuentra la combinación de segmentos de error ε mínimo*
  ○ *Une los segmentos cuya combinación es de menor error*
  ○ *Recalcula los errores de las combinaciones de segmentos*

end

Figura 2.14. *Enfoque BU (*Adaptado de: Keogh et al., 2001)

El enfoque Arriba-Abajo (TD, por sus siglas en Ingles, *Top-Down*) comienza con la ST completa como un solo segmento, y divide el segmento de mayor error con base a la combinación de observaciones de menor error. Utiliza los mismos parámetros que BU. El enfoque TD con parámetro de número de segmentos se presenta en la figura 2.15.

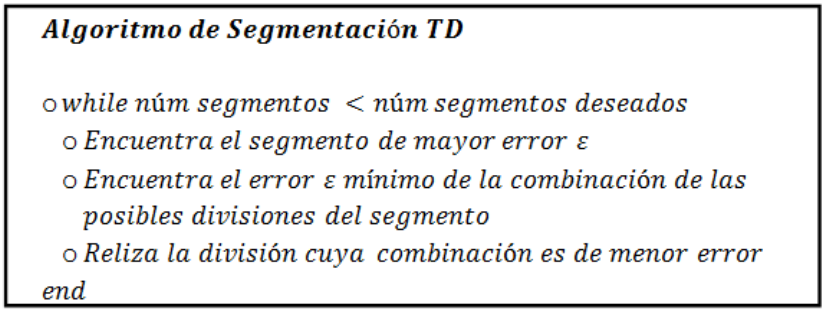


Figura 2.15. *Enfoque TD (*Adaptado de: Keogh et al., 2001)

Ventana-Corrediza (SW, por sus siglas en inglés, *Sliding-Window*) es un enfoque de segmentación que comienza con la primera observación y extiende el segmento hasta que un umbral de error es excedido. Debido a que SW es un enfoque de "tiempo real", solamente cuenta con un parámetro de error. El enfoque SW se presenta en la figura 2.16.

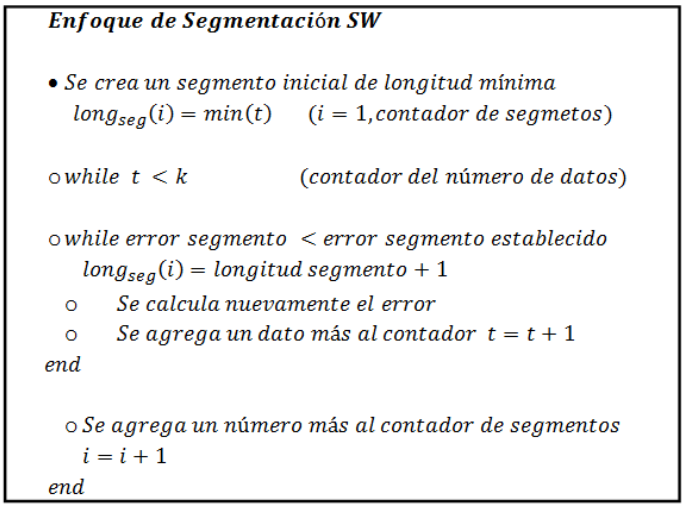


Figura 2.16. *Enfoque SW (*Adaptado de: Keogh et al., 2001)

El enfoque SW carece de la capacidad de segmentar con base a un determinado número de segmentos, mientras que los enfoques BU y TD si pueden. Sin embargo, SW es capaz de segmentar un flujo interminable de datos, mientras que BU y TD requieren el análisis de todos los datos para segmentar. Por esta razón, han surgido combinaciones de enfoques como el Ventana-Corrediza-Abajo-Arriba (SWAB, por sus siglas en ingles *Sliding Window and Bottom-Up*) (Keogh et al., 2001), para aprovechar las ventajas de diversos enfoques.

Sin embargo, más allá de obtener una buena segmentación, es importante la compresibilidad de los resultados obtenidos a través del uso de las herramientas visuales que faciliten la comprensión para el usuario.

2.5.3.1. Representaciones de Series de Tiempo

La representación de los datos es la clave para una solución eficiente y efectiva. Se han propuesto algunas representaciones de ST, incluyendo Transformadas de Fourier (Gore y Bhosle, 2011) que permiten a una señal de dominio de tiempo expresarla como dominio de frecuencia, *Wavelet* (Wu et al., 2010) que permiten descomponer una función en una suma ponderada de funciones, Mapeo Simbólico (Das et al., 1998) o Representación Lineal por Tramos (PLR, por sus siglas en ingles *Piecewise Linear Representation*) (Keogh et al., 2001) que permite obtener una representación lineal mediante mínimos cuadrados. Dentro de las representaciones de ST, una de las más utilizadas es PLR (Keogh et al., 2001; Zhu et al., 2007). PLR es una representación de la ST de k observaciones con S líneas rectas donde $1 \leq S \leq k$ (Keogh et al., 2001). Tres representaciones de ST de PLR son APCA, regresión lineal simple (RLS) y PAA, las cuales son descritas a continuación.

2.5.3.1.1. APCA

APCA ofrece una representación más fácil de interpretar, pero con mayor error (Junkui y Yuanzhen, 2007). Ésta realiza una representación lineal de longitud

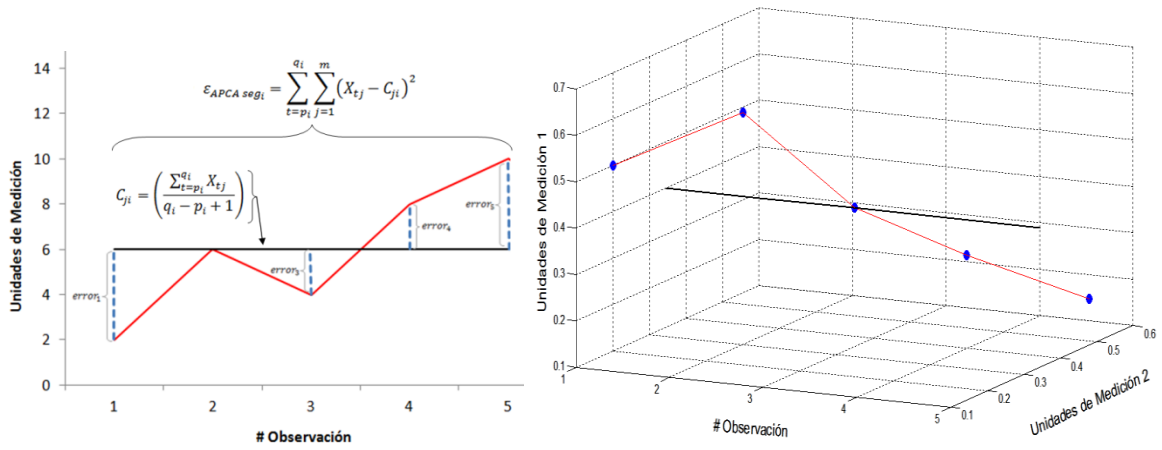
variable con base en promedio de las observaciones de cada característica. La forma de calcular el error metodológico generado por APCA se presenta en la ecuación 2.12, para $1 \leq i \leq S$,

$$\varepsilon_{APCA\ seg_i} = \sum_{t=p_i}^{q_i} \sum_{j=1}^{m} \left(X_{tj} - C_{ji}\right)^2, \qquad (ec.\ 2.12)$$

donde C_{ji} representa el promedio por característica j del segmento i, tal que

$$C_{ji} = \left(\frac{\sum_{t=p_i}^{q_i} X_{tj}}{q_i - p_i + 1} \right), \qquad (ec.\ 2.13)$$

donde $p_i\ y\ q_i$ representan las X_{tj} observaciones inicial y final del segmento i respectivamente. La figura 2.17 presenta un ejemplo de segmentación lineal mediante APCA.



a) **b)**

Figura 2.17. *Representación lineal APCA.*

a) cinco observaciones de una característica $(\mathrm{m} = 1, \mathrm{p} = 1, \mathrm{q} = 5)$

b) cinco observaciones de dos características $(\mathrm{m} = 2, \mathrm{p} = 1, \mathrm{q} = 5)$

2.5.3.1.2. Regresión Lineal Simple

Otra forma de representar linealmente una ST es a través de RLS, donde $\hat{X}_{tj}$ (ec. 2.15) es la estimación obtenida de X_{tj} mediante regresión lineal simple (Bowerman et al., 2009),

$$\hat{X}_{tj} = \alpha_{ji} + \beta_{ji} t + \varepsilon_{tj} \qquad (ec. 2.14)$$

donde:

1. $\hat{X}_{tj}$ es el valor medio estimado de la observación X_{tj}.
2. α_{ji}, β_{ji} (ec. 2.16 y 2.17 respectivamente) son estimaciones de los parámetros de regresión que relacionan el valor medio de X_{tj} para el segmento i de la característica j,

$$\alpha_{ji} = \frac{\sum_{t=p_i}^{q_i} X_{tj}}{q_i - p_i + 1} - \beta_{ji} \frac{\sum_{t=p_i}^{q_i} t}{q_i - p_i + 1} \qquad (ec. 2.15)$$

y

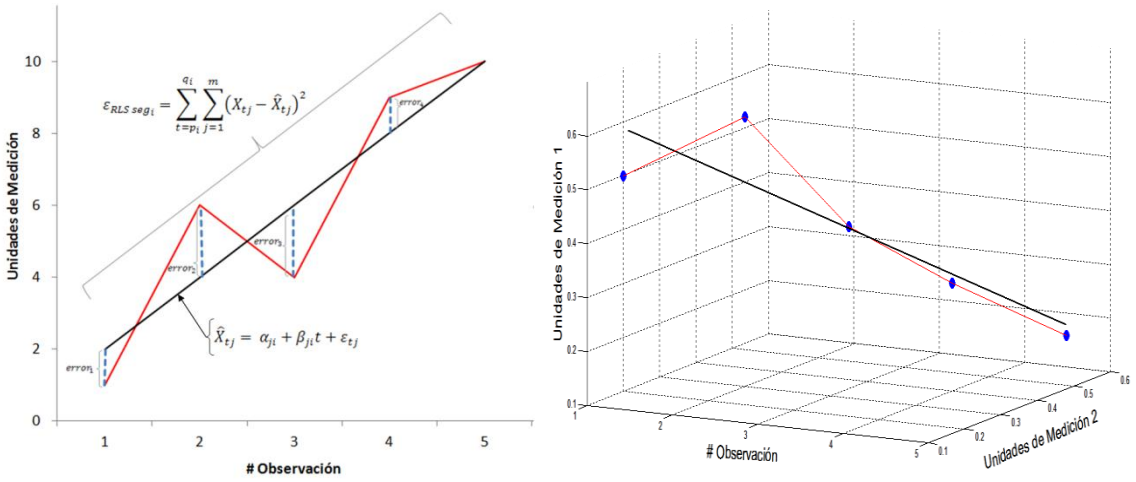
$$\beta_{ji} = \frac{\sum_{t=p_i}^{q_i} \left(t - \frac{\sum_{t=p_i}^{q_i} t}{q_i - p_i + 1} \right) \left(X_{tj} - \frac{\sum_{t=p_i}^{q_i} X_{tj}}{q_i - p_i + 1} \right)}{\sum_{t=p_i}^{q_i} \left(t - \frac{\sum_{t=p_i}^{q_i} t}{q_i - p_i + 1} \right)^2} \qquad (ec. 2.16)$$

3. ε_{tj} es un término de error que describe los efectos sobre $\hat{X}_{tj}$.

El error de RLS se calcula mediante la ecuación 2.14,

$$\varepsilon_{RLS\ seg_i} = \sum_{t=p_i}^{q_i} \sum_{j=1}^{m} \left(X_{tj} - \hat{X}_{tj} \right)^2, \qquad (ec. 2.17)$$

En la figura 2.18 se presenta un ejemplo de representación lineal mediante RLS.



a) **b)**

Figura 2.18. *Representación lineal RLS.*

a) cinco observaciones de una característica $(\mathrm{m}=1, \mathrm{p}=1, \mathrm{q}=5)$

b) cinco observaciones de dos características $(\mathrm{m}=2, \mathrm{p}=1, \mathrm{q}=5)$

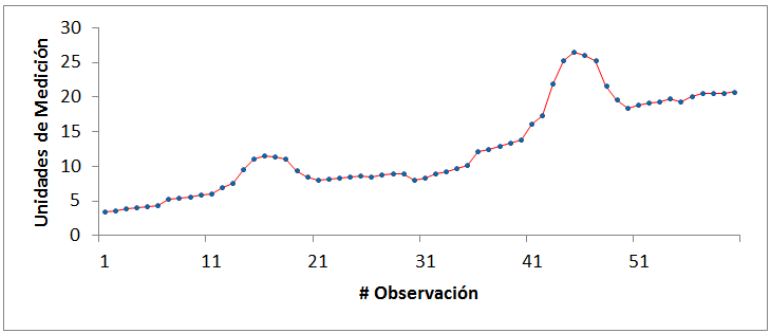
2.5.3.1.3. PAA

La representación lineal PAA se ha sido utilizada en diversas investigaciones, tales como la reducción el tiempo de búsqueda del algoritmo el k-vecino más cercano (Zhang, Glass, 2011), en el desarrollo del algoritmo de representación simbólica (Lkhagva et al., 2006), en la búsqueda de eventos financieros en línea (Wu et al., 2004).

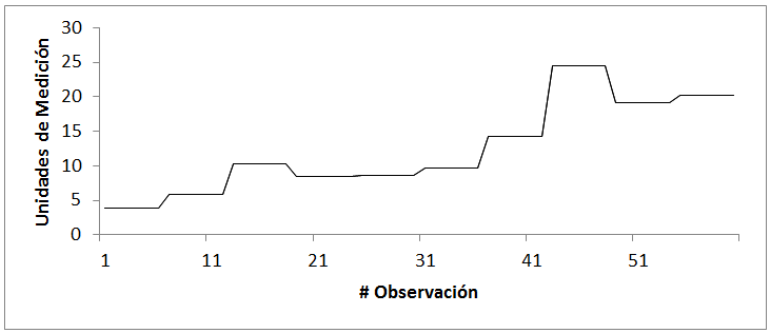
PAA es esencialmente, una proyección de coeficientes de Haar en el tiempo. Éste realiza una representación lineal de longitud constante con base en el promedio de las observaciones de cada característica (Keogh et al., 2001). Dada una ST de k observaciones es representada por S segmentos, el $i-ésimo$ segmento de S es calculado por

$$S_i = \frac{S}{k} \sum_{t=\frac{k}{S}(i-1)+1}^{\frac{k}{S}i} X_{tj}. \qquad (ec.\,2.18)$$

La figura 2.19 presenta un ejemplo de representación lineal mediante PAA.



a) 60 observaciones originales



b) Reconstrucción utilizando 10 coeficientes de Haar

Figura 2.19. *Representación lineal PAA.*

Una vez realizado el primer agrupamiento en función del tiempo ó segmentación de ST y el segundo agrupamiento (*Clustering*) independientemente del tiempo, es necesario identificar los *motifs*.

2.5.4. Identificación de *Motif*

Un motif se define como a patrón previamente desconocido que se presenta repetidamente en una serie de tiempo. Una de las principales razones para la identificación de *motifs* es la extracción de reglas asociativas y detección de anomalías de las series de tiempo (Tanaka et al., 2005, Lin et al, 2010). McGovern (2011) define *motif* unidimensional como,

$$M_g = (X_k, X_{k+1}, \dots, X_{k+h}) \qquad 1 \leq k \leq n \qquad k < h \leq n, \qquad (ec.\,2.19)$$

que consiste de una subsecuencia ordenada de X_k observaciones de una característica de la ST. Un *motif* multidimensional se define como

$$P_f = \left(M_{g_1}, M_{g_2}, \dots, M_{g_j}\right), \qquad (ec. 2.20)$$

es un conjunto de *motifs* unidimensionales de *j* características diferentes, ordenados temporalmente.

Según Tanaka et al. (2005) se pueden considerar tres criterios teóricos de información como base para la extracción de *motif*s: 1) Criterio de Información Akaike (AIC, por sus siglas en inglés, *Akaike's Information Criterion*), 2) Criterio de Información Bayesiano (BIC, por sus siglas en inglés, *Bayesian Information Criterion*) y 3) Longitud de Descripción Mínima (MDL, por sus siglas en inglés, *Minimum Description Length*). AIC estima el mejor modelo basado en la "capacidad de predicción". BIC estima el mejor modelo a través del teorema bayesiano. MDL establece que el mejor modelo es el que describe un conjunto de datos que minimiza la longitud de descripción del conjunto entero de datos.

Después de identificar los *motifs*, es necesario extraer las reglas asociativas temporales y presentarlas al usuario final en un lenguaje de su comprensión.

2.5.5. Identificación de Reglas Asociativas Temporales

Las reglas asociativas en ST de acuerdo con Morchen (2006) se pueden clasificar dos tipos: 1). intervalo de tiempo (si se modela el concepto de duración), o 2) punto de tiempo (en caso contrario). Morchen (2006) clasifica las reglas asociativas temporales en punto de tiempo de intervalo de tiempo. Las tablas 2.2 y 2.3 muestran los métodos de extracción de reglas asociativas de uso más generalizado.

Autor(s)/Año	Método	Datos del Modelo			Conceptos Temporales				
		Series de Tiempo Simbólicas Univariadas	Series de Tiempo Simbólicas Multivariadas	Series de Tiempo Simbólicas	Duración	Orden	Periodicidad	Concurrencia	Sincronicidad
Vilo (1998)	SufijoTrie	X				X			
Cohen et al. (2001)	Patrones Semánticos	X				X			
Das et al. (1998)	Asociaciones	X	X	X		X			
Oates et al. (1996)	MSDD	X	X			X			
Oates et al. (1997)	MEDD	X	X	X				X	
Mannila et al. (1995)	Episodios	X	X	X		X		X	
Harms et al. (2002)	MOWCATL	X	X	X		X		X	X
Mooney et al. (2004)	Episodios Interactivos	X	X	X		X		X	
Han et al. (1998)	Periodicidad Parcial	X				X	X		
Saetrom et al. (2003)	IQL	X		X	X	X		X	
Himberg et al. (2003)	SCM		X		X	X		X	X

Tabla 2.2. *Algoritmos para extracción de reglas asociativas temporales de punto de tiempo (Morchen, 2006)*

Autor(s)/Año	Método	Datos del Modelo			Conceptos Temporales			
		Series de Tiempo Simbólicas Univariadas	Series de Tiempo Simbólicas Multivariadas	Series de Tiempo Simbólicas	Duración	Orden	Coincidencia	Sincronicidad
Villafane et al. (1999)	Contención		X	X			X	
Last et al. (2001)	IFN	X			X	X		
Kam y Fu (2000)	Allen/A1		X	X		X	X	X
Cohen (2001)	Allen/Fluets		X	X		X	X	X
Guimaraes y Ultsch (1999)	UTG/Tcon		X	x	X	X		X
Hoppner (2001)	Ventana/Allen		X	X	X	X	X	X

Tabla 2.3. *Algoritmos para extracción de reglas asociativas temporales de intervalo de tiempo (Morchen, 2006)*

Las reglas asociativas relacionan dos eventos en base a operadores temporales. A continuación se presentan algunos ejemplos de dichos operadores.

2.5.5.1. Operadores Temporales

Los operadores temporales son utilizados para combinar elementos de datos temporales con el propósito de expresar patrones temporales (Morchen, 2006). Los operadores temporales pueden dividirse en operadores de punto de tiempo y operadores de intervalo de tiempo. Existen al menos cuatro operadores básicos de punto de tiempo: Antes, Igual, Después y Repite. Tanto Antes como Después, deben estar acompañados por un umbral de tiempo.

Los operadores de intervalos de tiempo (figura 2.20), sólo pueden ser usados con series de intervalos. De esta manera dos intervalos, cualesquiera, están relacionados exactamente por alguno de los siguientes operadores: antes, se reúne, se

superpone, se inicia, durante, termina, es igual a; así mismo, cada uno de estos operadores tiene su respectivo inverso: después, se reunió por, solapado por, iniciado por, contiene, terminó por y es igual a (Allen, 1983).

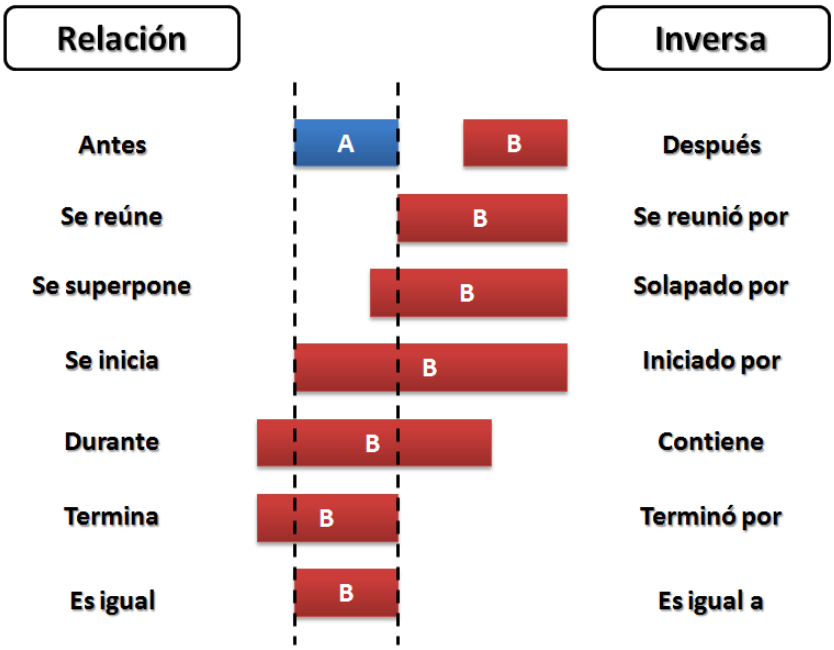


Figura 2.20. *Operadores de Intervalos de Tiempo (Allen, 1983)*

Morchen (2006) unifica los operadores temporales a través de los conceptos de duración, orden, concurrencia, coincidencia, sincronía y periodicidad (figura 2.21). La duración es la repetición de una propiedad a través de varios puntos de tiempo. El orden es la ocurrencia secuencial de los puntos de tiempo o intervalos. La concurrencia es la cercanía de dos o más eventos temporales sin un orden en particular. La coincidencia describe la intersección de varios intervalos. La sincronía es la ocurrencia de dos eventos temporales simultáneos y es un caso especial de coincidencia. La periodicidad es la repetición del mismo valor o muy parecidos en un periodo de tiempo constante. La figura 2.22 presenta el ejemplo de una regla asociativa en ST multivariada; en esta se pueden observar patrones repetidos en las tres ST.

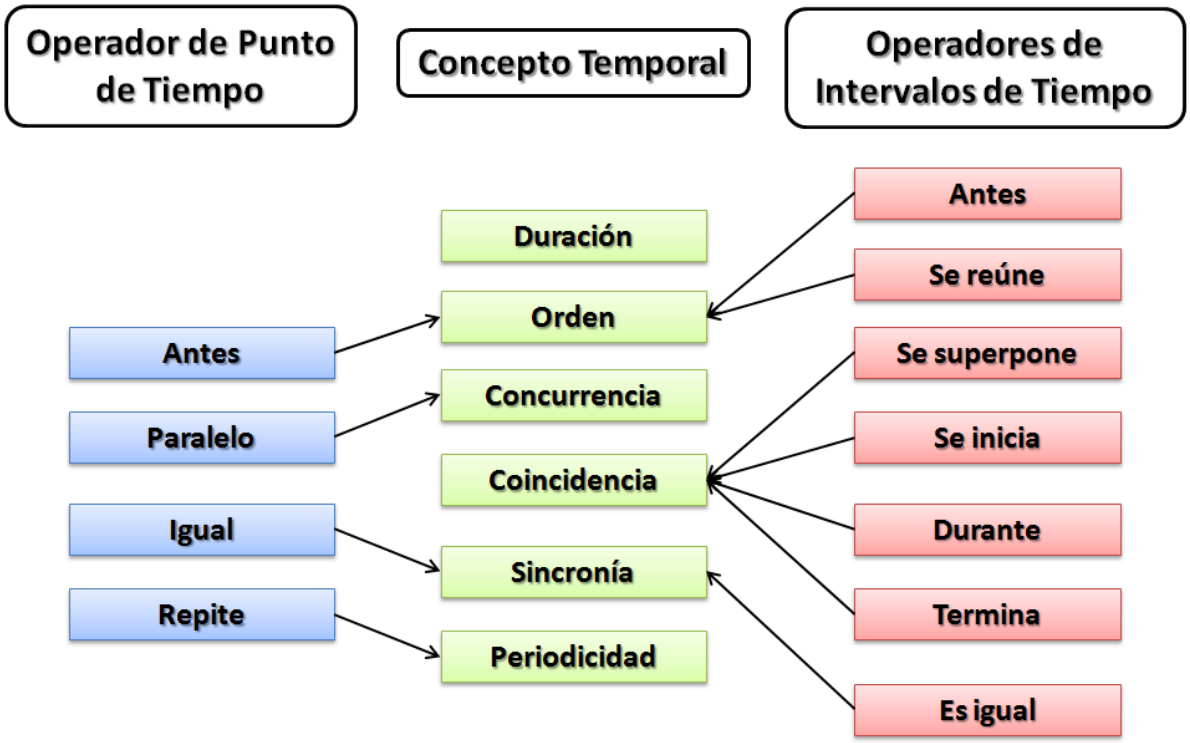


Figura 2.21. *Relación entre Operadores Temporales (Morchen, 2006)*

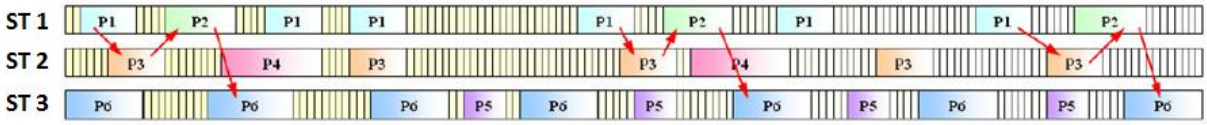


Figura 2.22. *Ejemplo de Regla Asociativa en ST Multivariada (Pisón et al., 2005)*

Una vez identificadas las reglas asociativas temporales, es necesario desarrollar un modelo de predicción de EM. A continuación se presenta el algoritmo K Vecino Más Cercano (KNN, por sus siglas en inglés *K Nearest Neighbor*) para el desarrollo de este modelo.

2.6. Modelo de Predicción de EM mediante KNN

KNN es un algoritmo ampliamente utilizado para la clasificación supervisada, estimación y predicción (Han y Kamber, 2006; Larose, 2005). Clasifica las

observaciones desconocidas en una clase predefinida, basado en observaciones previamente clasificadas (conjunto de entrenamiento). Aunque el costo computacional de KNN es alto, posee ventaja al aplicarse en datos de cambio o actualización rápida.

Para clasificar una nueva observación, KNN mide la distancia entre una observación específica y el resto de las observaciones del conjunto de entrenamiento. La distancia euclidiana es la más comúnmente utilizada. Todas las distancias d son organizadas tal que $d_i \leq d_{i+1}$ $i = 1,2,3, \dots, k$. Las k observaciones con la menor distancia a la nueva observación son conocidas como los k vecinos más cercanos, y son utilizadas para clasificar la nueva observación a la clase existente (Zahoor et al., 2008). El proceso se simplifica a continuación:

Algoritmo KNN:
Paso 1: Medir la distancia entre la nueva observación y el conjunto de entrenamiento. Dentro de las distancias disponibles en la figura 2.12, la euclidiana es la más utilizada.

Paso 2: Organizar los valores de las distancias tal que $d_i \leq d_{i+1}$, seleccionar las k observaciones de menor distancia.

Paso 3: Aplicar votación o determinar media, de acuerdo a la aplicación.

La figura 2.23 presenta de forma gráfica el funcionamiento del algoritmo KNN. En este ejemplo, para $k = 3$, la nueva observación sería de color rojo, dado que dos de las tres observaciones más cercanas son rojas. Sin embargo, para $k = 5$, la nueva observación sería color azul, dado que tres de las cinco observaciones más cercanas son azules.

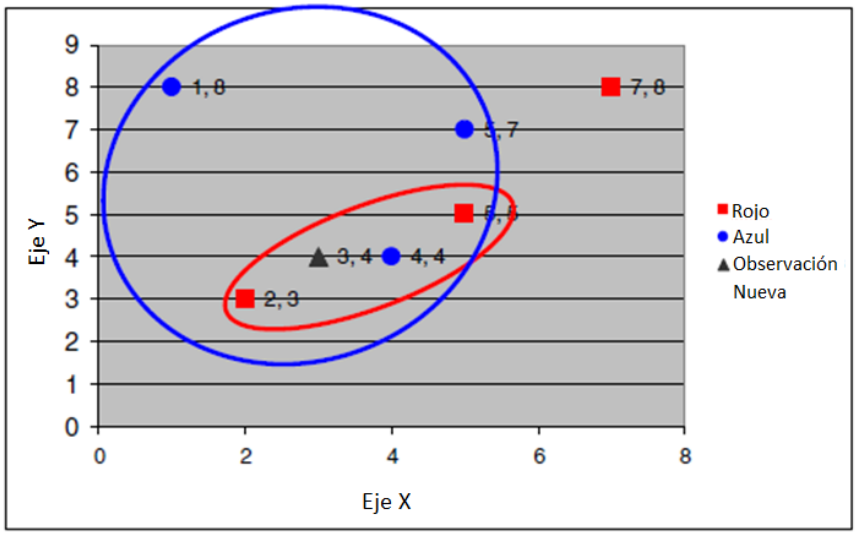


Figura 2.23. *Funcionamiento de KNN (Jan et al., 2008)*

El algoritmo KNN se ha utilizado en el desarrollo de modelos predictivos, tal es el caso de los trabajo de Chitra y Uma (2010) en ST; las predicciones de EM de Zahoor et al. (2008, quienes aplicaron este algoritmo para la predicción climática anual; y las aplicaciones de Wu et al. (2010) para el pronóstico de precipitaciones en regiones semiáridas.

A través del marco de referencia, se introdujo al lector algunas características del cultivo de la vid y como los elementos meteorológicos lo afectan. Se presentaron diferentes esfuerzos realizados a través de la historia para conocer anticipadamente los EM, así como una explicación del porqué es posible realizar una predicción de los EM. Finalmente se presentaron diversas técnicas utilizadas para desarrollar los modelos predictivos en la actualidad, así como ejemplos de investigaciones en las que se desarrollan estos modelos. Es fácil darse cuenta, con solo revisar la literatura, que no hay una técnica ideal para desarrollar un modelo predictivo. Desde el planteamiento del objetivo esta investigación, se propone realizar la extracción de reglas asociativas, y el modelo a desarrollar para lograrlo, será mediante una combinación de las técnicas presentadas anteriormente.

3. PLANEACIÓN DE IDENTIFICACIÓN DE PATRONES EN EL CULTIVO DE LA VID

En este capítulo se presenta la estructura metodológica utilizada en este proyecto, así como una descripción de distintos recursos que fueron necesarios para desarrollar el proyecto. Se describen las observaciones fenológicas utilizadas, tanto en fecha como en localización geográfica. Después se presenta la descripción de las observaciones de EM, así como del proceso para obtenerlas. Posteriormente se presenta la adaptación de la estructura metodológica de la figura 2.6.

3.1. Metodología

El proceso metodológico de este trabajo (figura 3.1) es una adaptación del modelo metodológico propuesto por Fayyad et al. (1996) (figura 2.6), en el cual, la etapa de MD, es una adaptación del modelo metodológico propuesto por Pisón et al. (2005).

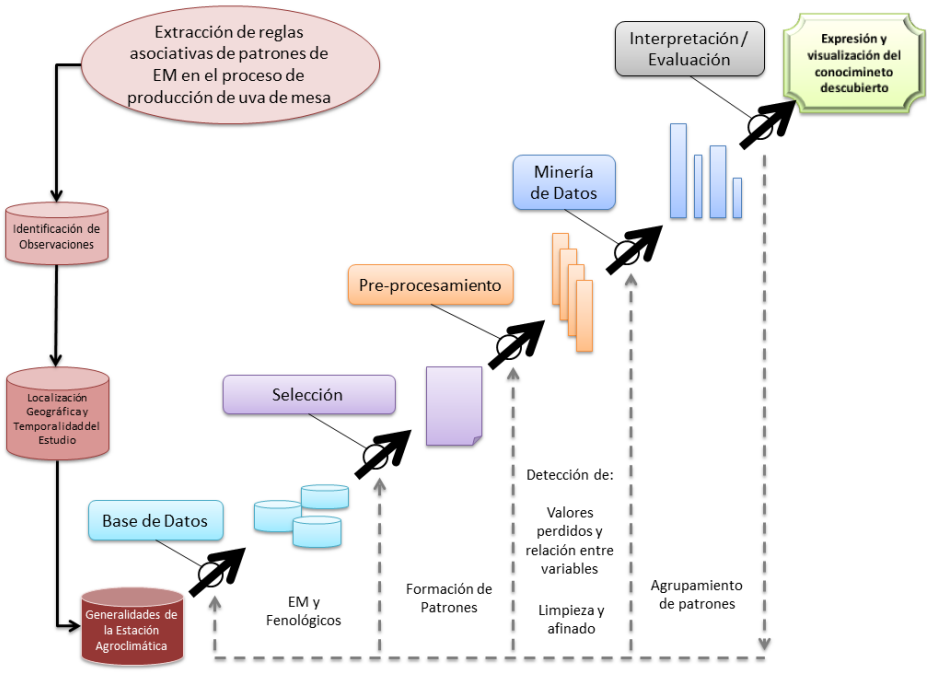


Figura 3.1. *Adaptación del KDD* (Adaptado de Fayyad et al., 1996)

A continuación se presentan los pasos de la metodología establecida para el desarrollo de este trabajo.

3.1.1. Identificación de Observaciones

En esta etapa es necesario establecer el sistema que se utilizará para determinar el crecimiento de las etapas fenológicas, así como las fases de dicho sistema. Para esto, se debe registrar la fecha del estallamiento de los brotes, separación y desarrollo de las hojas, visibilidad de inflorescencia, floración, caída de capullos, desarrollo y maduración de las bayas. El registro dependerá de las fases a contemplar en el estudio.

3.1.2. Localización Geográfica y Temporalidad del Estudio

En este, se determina la localización y la temporalidad del estudio que permitirá identificar la estación agroclimática de la cual se obtendrán las observaciones de EM. Para esto es necesario registrar las coordenadas, tanto de latitud como longitud e identificar la estación agroclimática más cercana del estado de la república en el que se realizará el estudio. El establecer la temporalidad del estudio permite acceder a datos históricos de dicha estación climática.

3.1.3. Generalidades de la Estación Agroclimática

En esta etapa se deben identificar las generalidades de la estación agroclimática que proporcionan parámetros para la investigación. Estos pueden incluir los elementos meteorológicos medibles, unidades de medición, rangos de medición y la exactitud de los sensores de la estación climática.

3.1.4. Base de Datos

Una vez identificadas las observaciones, la localización geográfica, temporalidad de estudio y las generalidades de la estación agroclimática, se debe reunir la

información que se cuente disponible tanto fenológica como meteorológica y buscar concordancias en fechas para crear la base de datos para el proyecto.

3.1.5. Selección

En esta etapa, se debe realizar una selección a detalle que permita identificar las variables fenológicas y meteorológicas necesarias para el cumplimiento de los objetivos del proyecto.

3.1.6. Pre-procesamiento

En el análisis exploratorio se busca identificar y remplazar valores atípicos y faltantes. Sin embargo, al analizar las observaciones de EM como ST, es necesario identificar valores atípicos en la secuencia de las observaciones. Por esta razón, no es posible identificar todos los valores atípicos utilizando el diagrama de caja y bigote, por lo que se utilizaron gráficas de líneas para cada EM por ciclo productivo.

Para corregir las observaciones atípicas y faltantes, se debe crear un objeto de serie de tiempo para cada EM de observaciones utilizando la función “fints” de MATLAB. Una vez creados los objetos, se utilizará la función “fillts” con el método de extrapolación cúbica para determinar valores faltantes y corregir las irregularidades en las observaciones.

3.1.7. Minería de Datos

En la etapa de MD, para la extracción de reglas asociativas, es necesario identificar una correlación entre las observaciones de EM y la duración de las fases fenológicas. En la estrategia de MD aplicada se realizó la segmentación PAA por patrón día, segmentación PAA por patrón hora, la predicción de duración de fases fenológicas y el modelo de predicción de EM. La figura 3.2 muestra el mapa conceptual de la estrategia de MD aplicada.

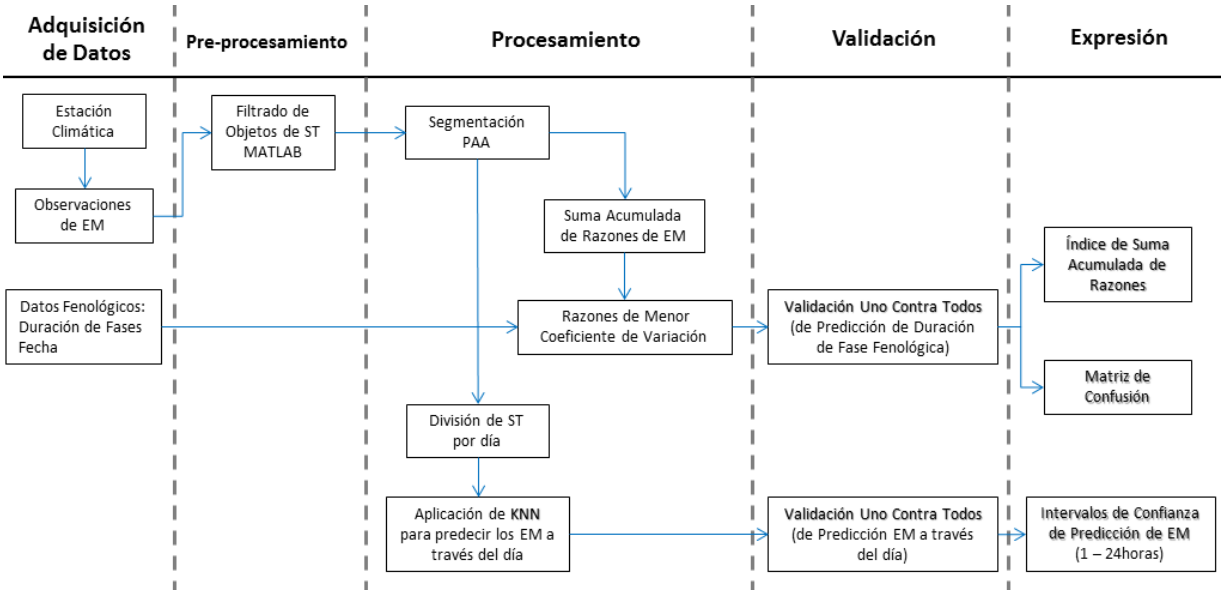


Figura 3.2. *Mapa conceptual de estrategia de MD aplicada*

3.1.7.1. Segmentación PAA Patrón Día

En este primer acercamiento se utilizó el algoritmo PAA para segmentar las ST por día mediante coeficientes de Haar. La segmentación se realizó en base a n coeficientes de Haar para obtener segmentos de longitudes constantes, y se calcula la suma acumulada de razones de EM para identificar una correlación entre los días que dura cada fase fenológica con los EM en base al menor coeficiente de variación (CV).

3.1.7.2. Segmentación PAA Patrón Hora

En la segmentación PAA patrón hora, de las fases fenológicas identificadas con el menor CV, se repite el análisis pero con segmentos de 1 hora de longitud.

3.1.7.3. Predicción de Duración de Fases Fenológicas

En este paso se utiliza la validación uno contra todos para la predicción de la duración de las etapas fenológicas.

3.1.7.3. Modelo de Predicción de EM

Para la elaboración del modelo de predicción de EM se determinaron las fechas mínimas y máximas de las fases fenológicas identificadas en el análisis anterior. Se adquirieron todas las observaciones de EM disponibles de la base de datos del SIA, del año 2002 a la 2011, filtraron y segmentaron mediante el algoritmo PAA para obtener segmentos de 1 hora y se dividió la ST por día. El modelo se creó utilizando el algoritmo KNN con distancia euclidiana y la base de datos generada por día para predecir los EM desde 1 a 23 horas a partir de la primera hora del día. Finalmente, se determinaron los intervalos de confianza para la el error promedio de la predicción de EM de 1 a 23 horas.

4. ANÁLISIS Y MODELACIÓN DEL EM PARA EL CULTIVO DE LA VID

El objetivo de este capítulo es abordar la relación entre la duración de las fases fenológicas de brotación, inflorescencia y florescencia; y el comportamiento de los EM para determinar reglas asociativas entre estas variables. Así, se presentan los resultados obtenidos a partir de la propuesta metodológica expuesta en el capítulo anterior. En la primera parte, se describen las observaciones utilizadas en el análisis. Después se presentan las segmentaciones de las ST y se analizan las sumatorias acumuladas de razones de EM. Posteriormente se determinan los intervalos para las predicciones de la duración de las etapas fenológicas y se desarrollan matrices de confusión de dichas predicciones. Las reglas asociativas son extraídas en base a las matrices de confusión y las sumatorias acumuladas de razones. Una vez extraídas las reglas, se desarrolla el modelo de predicción de EM mediante KNN utilizando distancia euclidiana.

4.1. Observaciones Fenológicas de la Uva de Mesa

Para la identificación de las etapas fenológicas, se utilizó el sistema de Eichhorn y Lorenz modificado por Coombe (1995) (fig. 2.2). A partir de este modelo, se seleccionaron de las etapas de brotación e inflorescencia, las fases: 5 (brotes reventados, punta de hoja visible), 7 (primera hoja separada de la punta del brote), 9 (dos o tres hojas separadas, brote de 2-4 cm), 11 (cuatro horas separadas), 12 (5 hojas separadas e inflorescencia claramente visible, largo del brote cercano a 10 cm), y 14 (7 hojas separadas); así como de la etapa de floración, las fases: 21 (30% capullos caídos), 23 (floración completa, 50% de los capullos caídos, 17-20 hojas separadas) y 25 (80% de los capullos caídos). Se registró la clave correspondiente a la fase fenológica, así como la fecha en que ésta se presentó.

4.2. Localización Geográfica y Temporalidad del Estudio

Los datos fenológicos provienen de la bitácora de la exploración diaria del viñedo, en la cual se anotó la fecha y el número de etapa fenológica correspondiente al desarrollo fisiológico de la planta, según la inspección visual del responsable del viñedo, por quien fueron proporcionados (Preciado, 2011). Dichos datos fenológicos corresponden a la variedad de uva de mesa *Flame Seedless*, plantada en un viñedo comercial localizado al norte de la ciudad de Hermosillo, Sonora, en las coordenadas Latitud N 29º18’15’’, Longitud W 110º 55’21’’. El estudio comprendió los ciclos 2001-2002, 2002-2003, 2003-2004 y 2004-2005.

Los datos de EM se obtuvieron de la estación agroclimática “La Cuesta” situada al norte de la ciudad de Hermosillo, en la región de Pesqueira, Sonora, México. Esta estación pertenece al Sistema de Información Agroclimática (SIA), situado en las coordenadas Latitud N 29º17’15’’, Longitud W 110º55’21’’.

4.3. Generalidades de la Estación Agroclimática

Los datos de EM obtenidos del SIA, son recolectados a través de una estación remota ADCON A733 ADDWAVE, la cual cuenta con un conjunto de sensores (tabla 4.1). En la figura 4.1 se muestra el proceso de transferencia de las mediciones de los EM hasta la publicación en el portal WEB. Los datos son colectados cada minuto y enviados de forma automática a la central receptora cada 15 minutos; ésta los decodifica y los trasmite a una computadora, misma que los convierte en información meteorológica por medio del software computacional addVANTAGE. Este software crea una base de datos agroclimática histórica la cual es publicada en la WEB a través del portal del SIA.

Sensor	Unidad de Medición	Rango de Medición	Exactitud
Dirección del Viento	grado	0 a 360 °	± 2%
Humedad Relativa	%	0 a 100%	± 3%
Presión Barométrica	Kilo Pascales	0.05 a 5.55	± 2%
Radiación Solar	KW/m^2	0 a 2 KW/m^2	± 0.15%
Temperatura	°C	-39.8 °C a +60 °C	± 0.6°C
Velocidad del Viento	m/s	0.05 a 5.55 m/s	± 2%

Tabla 4.1. *Características de la Estación Agroclimática (*Elaborado con base en Adcon Telemetry, 2010)

Figura 4.1. *Flujo de Mediciones de EM a Datos de EM*

4.4 Base de Datos

Del sistema de identificación de las etapas fenológicas de Eichhorn y Lorenz modificado por Coombe (1995) y de los registros contenidos en la bitácora diaria de campo del administrador del viñedo, se derivó la estructura de la base de datos fenológicos mostrada en la tabla 4.2.

Campo	Descripción	Tipo
Ciclo	Ciclo de producción del viñedo	Texto
Fase	Clave de la fase fenológica observada	Numérica
Fecha Inicio	Fecha de inicio de la fase fenológica	Fecha
Fecha Término	Fecha de término de la fase fenológica	Fecha

Tabla 4.2. *Estructura de los datos fenológicos*

4.5. Selección

De la estructura de los datos de EM en intervalos de 15 minutos disponibles, se seleccionaron los de la tabla 4.3; a excepción de la presión de vapor de agua promedio por intervalo de 15 minutos, la cual fue calculada mediante las ecuación 4.1 (tutiempo, 2011)

$$P_v = \left(\frac{H_R}{100}\right) P_{vs} \qquad (ec.4.1)$$

donde:

P_v: Presión de vapor

H_R: Humedad relativa

P_{VS}: Presión de saturación de vapor

Se estimó la P_{vs} para temperaturas bajo 0°C (ec.4.2) y para temperaturas sobre 0°C (ec.4.3):

$$P_{vs} = exp\left(60.433 - \left(\frac{6834.271}{Temp}\right) - \left(5.16923 * ln(Temp)\right)\right) \qquad (ec.4.2)$$

$$P_{vs} = exp\left(31.9602 - \left(\frac{6270.3605}{Temp}\right) - \left(0.46057 * ln(Temp)\right)\right) \qquad (ec.4.3)$$

donde:

T: Temperatura en grados Kelvin

El tamaño del conjunto de registros se determinó con base en las fechas de inicio y término de los datos fenológicos para los ciclos de producción comprendidos entre los años 2001 al 2005; esto arrojó un conjunto de 26,972 patrones.

Datos de EM y fenológicos		
Campo	**Descripción**	**Tipo**
Ciclo	Ciclo de producción	String
Fase	Fase fenológica	String
Fecha	Fecha del día de observación de la fase fenológica	Fecha
Hora	Hora del día relacionada a las mediciones de los EM	Numérico
T	Temperatura promedio por intervalo de 15 minutos	Numérico
HR	Humedad relativa promedio por intervalo de 15 minutos	Numérico
PV	Presión de vapor de agua promedio por intervalo de 15 minutos	Numérico
RS	Radiación solar promedio por intervalo de 15 minutos	Numérico

Tabla 4.3. *Conjunto de datos utilizados*

La base de los datos de EM, disponible para el usuario en el portal del SIA, deriva de los sensores incluidos en la estación remota y software que genera y administra esta base de datos (addVANTAGE) para datos de EM en intervalos de 15 minutos se muestra en el tabla 4.4; el dominio de cada una de estas mediciones depende directamente del rango de medición de cada uno de los sensores presentados en la tabla 4.1.

Sensor	Descripción	Unidad
t1_5	Temperatura a 1.5 metros	Grados Celsius
tmax1_5	Temperatura máxima a 1.5 metros	Grados Celsius
tmin1_5	Temperatura mínima a 1.5 metros	Grados Celsius
hr1_5	Humedad relativa a 1.5 metros	Porcentaje
Rs	Radiación solar	Kilowatt/metro^2
Vv	Velocidad del viento	Metro/segundo
Dv	Dirección del viento	Grados
dedv	Desviación estándar de la dirección del viento	Grados
Ll	Precipitación pluvial (lluvia)	Milímetros
Hh	Humedad foliar	Porcentaje

Tabla 4.4. *Disponibilidad de mediciones de EM en SIA (patrón 15 min.)*

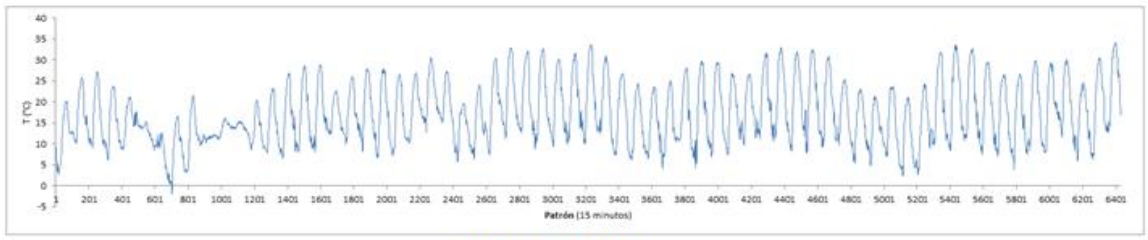
El conjunto de observaciones de EM incluye los siguientes elementos: temperatura (T), humedad relativa (HR), presión de vapor (PV) y radiación solar (RS). La distribución de las observaciones por etapa fenológica y ciclo productivo se presenta en la tabla 4.5.

	Etapa Fenológica									
Ciclo	**05**	**07**	**09**	**11**	**12**	**14**	**21**	**23**	**25**	**Total**
2001-2002	672	672	672	672	576	288	576	1,344	960	6,432
2002-2003	672	576	672	672	576	576	768	1,344	480	6,336
2003-2004	480	576	672	672	1,344	672	672	672	1,248	7,008
2004-2005	768	576	768	1,248	384	864	768	576	1,248	7,200
Total	2,592	2,400	2,784	3,264	2,880	2,400	2,784	3,936	3,936	26,976
Promedio	648	600	696	816	720	600	696	984	984	6,744
Desv. Est.	120.80	48.00	48.00	288.00	425.73	240.00	91.91	417.54	362.39	424.83
C.V.	18.64	8.00	6.90	35.29	59.13	40.00	13.21	42.43	36.83	6.30

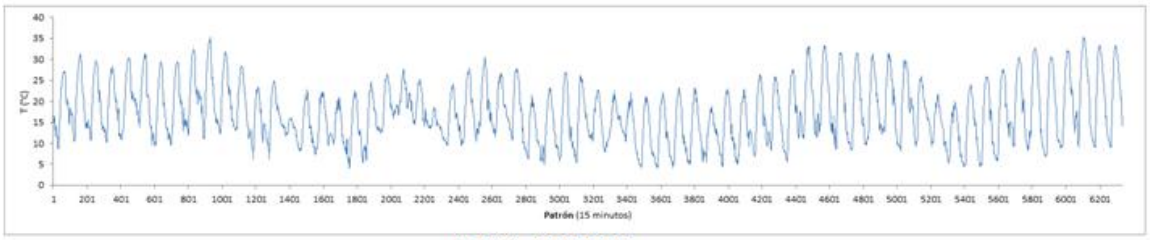
Tabla 4.5. *Número de Observaciones por Fase Fenológica (patrón - 15 minutos)*

Se puede observar en los datos en la tabla 4.5 que las fases fenológicas con los menores CV son las fases 7 y 9 con coeficientes inferiores a 10, y las fases 5 y 21 con coeficientes inferiores a 20. Estas cuatro fases son las que presentan menor variación en su duración entre ciclos. Para analizar la continuidad de las

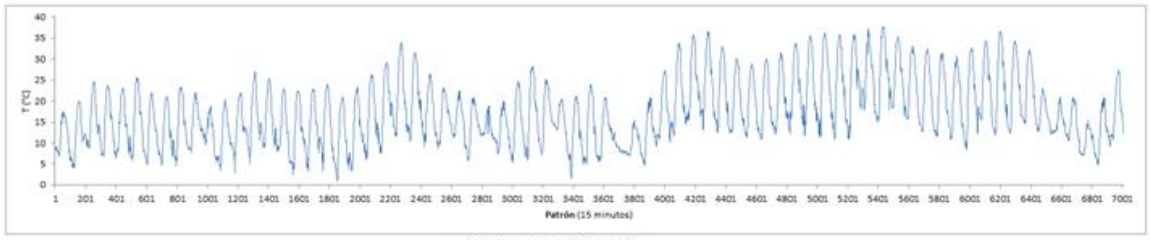
observaciones se realizaron gráficas de líneas para cada EM pro ciclo. La figura 4.2 se presenta las gráficas de líneas para la temperatura de los cuatro ciclos productivos, el resto de los EM se encuentran disponibles en el anexo A.



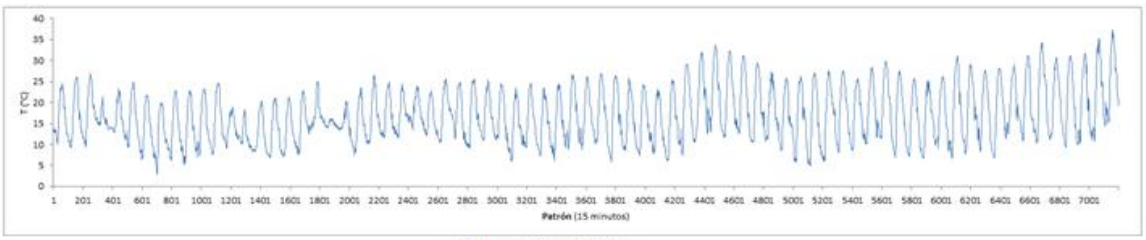
a) Ciclo 2001-2002



b) Ciclo 2002-2003



c) Ciclo 2003-2004



d) Ciclo 2004-2005

Figura 4.2. *Comportamiento de la temperatura durante el periodo de estudio*

4.6. Pre-procesamiento

Al analizar las gráficas de líneas, se identificó una irregularidad periódica en la última observación registrada por día. Se identificó que la irregularidad se debe a que la base de datos del SIA presenta la primera observación del día como la última como se muestra en la figura 4.2 de color rojo. Los datos faltantes en la base de datos del SIA se registran como -9999 como se muestra en la figura 4.3 de color verde.

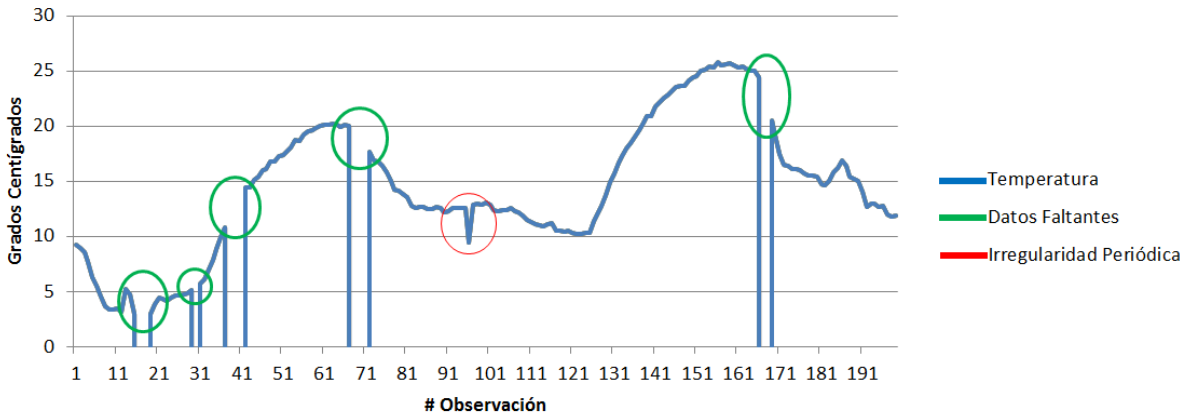


Figura 4.3. *Ejemplo de 200 observaciones de temperatura*

Nota: Se eliminaron datos de esta gráfica para la demostración del filtro utilizado para predecirlos.

Para corregir las observaciones atípicas y faltantes, se creó un objeto de serie de tiempo para cada EM de observaciones utilizando la función “fints” de MATLAB. Una vez creados los objetos, se utilizó la función “fillts” con el método de extrapolación cúbica para determinar valores faltantes y corregir las irregularidades cambio de hora en las observaciones. La figura 4.4 presenta las 200 observaciones con los filtros aplicados.

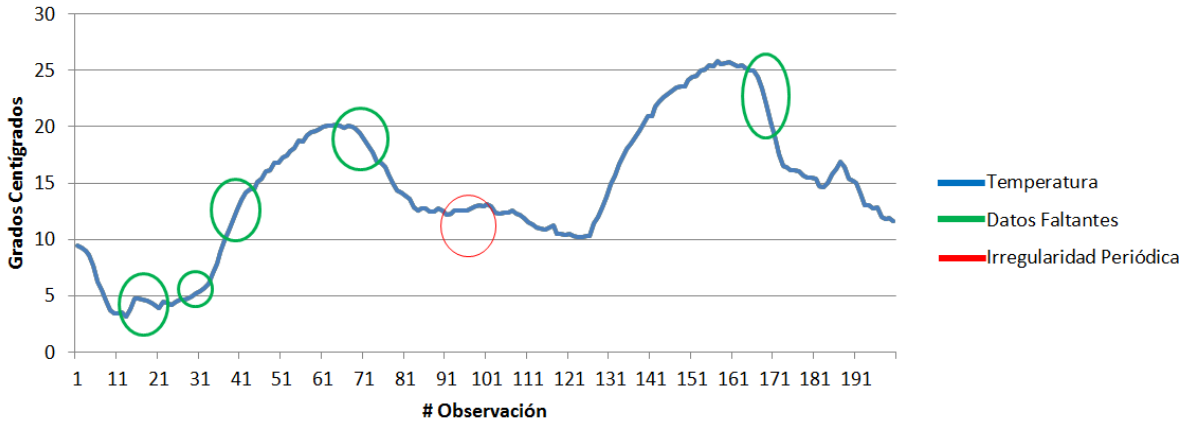


Figura 4.4. *Ejemplo de 200 observaciones de temperatura con filtro*

Además de los valores atípicos detectados anteriormente, se identificaron observaciones de radiación solar de noche en las fechas de 28/02/2002 al 20/03/2002 con promedio de 0.2 kW/m^2 y del 22/01/2003 al 06/02/2003, con promedio de 0.035 $kilowatt/metro^2$ como se muestra en la figura 4.5.

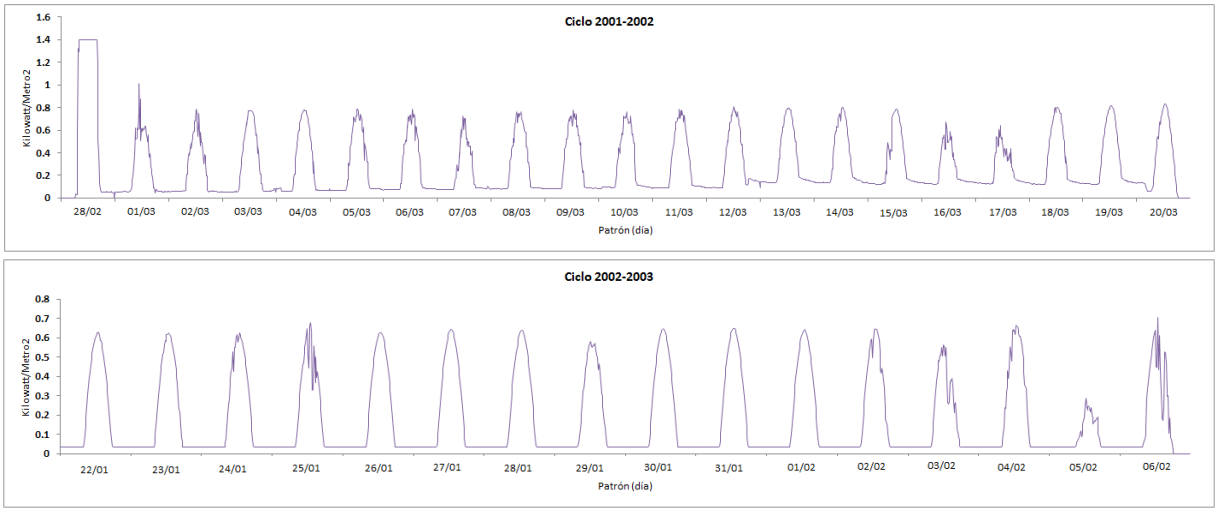


Figura 4.5. *Observaciones de radiación solar detectadas de noche*

Se analizó si había relación entre las observaciones de radiación solar registradas de noche y las fases de la luna o erupciones solares, pero no se encontró relación alguna. Considerando este factor y la gran cantidad de radiación solar en las

observaciones del 28/02/2003, se consideraron dichas observaciones como error de medición y se modificaron por el valor de cero durante la noche y se promediaron las observaciones de los días adjuntos al 28/02/2003 para predecir sus observaciones.

4.7. Minería de Datos

En la etapa de MD se identifica una correlación entre las observaciones de EM y la duración de las fases fenológicas. La estrategia aplicada es segmentar la ST utilizando n coeficientes de Haar mediante el algoritmo PAA para obtener segmentos de 1 día de longitud. Posteriormente, se utiliza suma acumulada de razones de EM para identificar una correlación entre los días que dura cada fase fenológica con los EM en base al menor coeficiente de variación (CV). De las fases fenológicas identificadas con el menor CV, se repite el análisis pero con segmentos de 1 hora de longitud. Finalmente, se utiliza validación uno contra todos para la predicción de la duración de las etapas fenológicas.

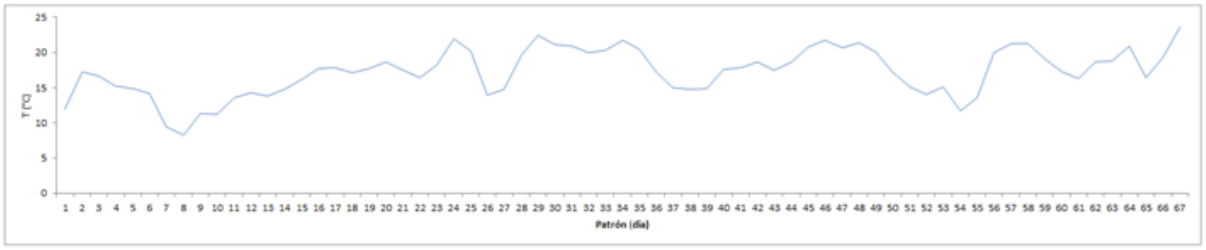
4.7.1. Segmentación PAA Patrón Día

En este primer acercamiento se utilizó el algoritmo PAA para segmentar las ST por día mediante coeficientes de Haar. Debido a que el número total de observaciones varía dependiendo del ciclo productivo, como se muestra en la tabla 4.5, fue necesario utilizar un número diferente de coeficientes de Haar por ciclo para obtener segmentos de longitud constante entre los cuatro ciclos. El número de coeficientes de Haar utilizados para cada ciclo se presenta en la tabla 4.6.

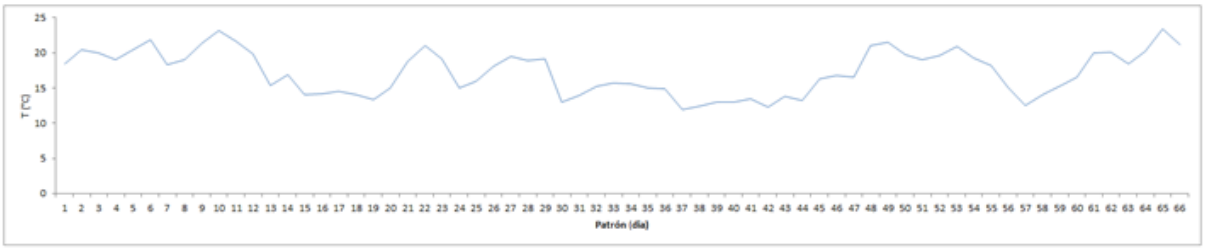
Ciclo	Núm. de Coeficientes
2001-2002	67
2002-2003	66
2003-2004	73
2004-2005	75

Tabla 4.6. *Coeficientes de Haar (patrón – día)*

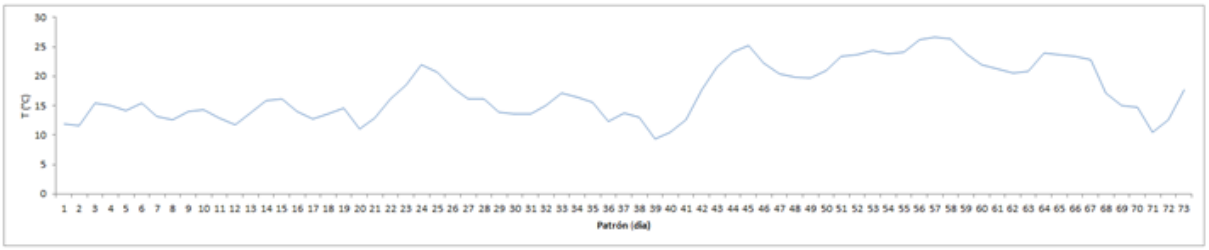
Una vez segmentadas las ST se graficaron para su análisis. En la figura 4.6 se presenta la gráfica de línea de las observaciones de temperatura segmentadas utilizando el algoritmo PAA con los coeficientes de la tabla 4.5. Las gráficas de los demás EM se encuentran disponibles en el anexo B.



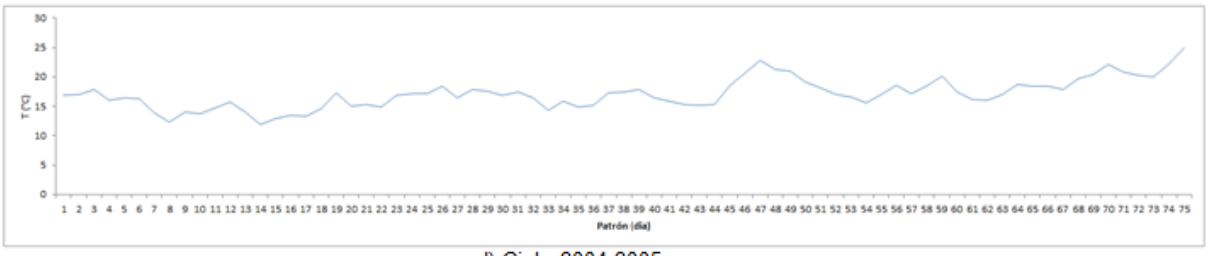
a) Ciclo 2001-2002



b) Ciclo 2002-2003



c) Ciclo 2003-2004



d) Ciclo 2004-2005

Figura 4.6. *Segmentación PAA (patrón día) – Temperatura*

Utilizando las ST segmentadas por día (anexo B), se realizaron las sumatorias acumuladas de razones entre distintos EM. Las razones analizadas se presentan en la tabla 4.7.

Razones				
Utilizando 1 EM	**Utilizando 2 EM**	**Utilizando 3 EM**	**Utilizando 4 EM**	**Otros**
T	$\frac{T}{HR}$	$\frac{T \times HR}{RS}$	$\frac{RS \times HR}{T \times PV}$	$\frac{T-10}{RS}$
HR	$\frac{T}{PV}$	$\frac{T \times PV}{HR}$	$\frac{RS \times PV}{T \times HR}$	$\frac{RS}{T-10}$
PV	$\frac{T}{RS}$	$\frac{RS \times HR}{T}$	$\frac{T \times HR}{RS \times PV}$	$\frac{RS \times HR}{(T-10) \times PV}$
RS	$\frac{HR}{T}$	$\frac{RS}{T \times HR}$	$\frac{T \times PV}{RS \times HR}$	$\frac{RS \times PV}{(T-10) \times HR}$
	$\frac{HR}{PV}$			$\frac{(T-10) \times HR}{RS \times PV}$
	$\frac{RS}{T}$			$\frac{(T-10) \times PV}{RS \times HR}$
	$\frac{RS}{HR}$			$\frac{RS^3}{\sqrt[4]{T}}$
	$\frac{RS}{PV}$			

Tabla 4.7. *Razones a analizar en segmentación (patrón – día)*

Al realizar la suma acumulada de las razones se identificaron cuatro fases fenológicas con CV de razones menores a 13. Las fases fenológicas son 5, 7, 9 y 11. Estas son las mismas fases identificadas en la tabla 4.5, las cuales presentan la menor variación en cuanto a su duración. La tabla 4.8 presenta los menores CV de variación para cada fase fenológica.

Fase	Ciclo	Suma Acumulada de Razones de EM						
		$\frac{RS}{T}$	$\frac{T-10}{RS}$	$\frac{T-10}{RS}$	$\frac{RS \times PV}{T \times HR}$	RS	T	$\frac{RS^3}{\sqrt[4]{T}}$
5	2001-2002	0.0644173	282.1483216	0.0022953	0.0098373	0.8924792	99.3666667	0.0000679
	2002-2003	0.0585560	414.8634263	0.0023131	0.0121125	1.1546563	138.4562500	0.0001279
	2003-2004	0.0571213	114.7081499	0.0014641	0.0081990	0.7748021	68.1281250	0.0000419
	2004-2005	0.0695754	403.2453868	0.0012237	0.0114393	1.0748229	126.5937500	0.0000334
C.V.		9.17	45.95	30.87	16.82	17.69	28.94	62.96
7	2001-2002	0.064067	312.833610	0.001402	0.007990	0.741000	87.210417	0.000024
	2002-2003	0.048320	369.556785	0.002163	0.010303	0.963208	120.236458	0.000123
	2003-2004	0.070378	139.418188	0.001586	0.010199	0.962438	82.388542	0.000043
	2004-2005	0.056732	191.315115	0.001229	0.008470	0.807083	84.209375	0.000034
C.V.		15.88	41.95	25.45	12.81	12.93	19.17	80.86
9	2001-2002	0.0759627	277.5799349	0.0050719	0.0136989	133.7708333	122.7500000	0.0002986
	2002-2003	0.0677262	252.4993936	0.0017088	0.0105637	341.9166667	101.8635417	0.0000532
	2003-2004	0.0874225	162.8237394	0.0034697	0.0123737	190.7708333	97.7666667	0.0001112
	2004-2005	0.0610653	544.6636884	0.0009000	0.0095484	562.3750000	116.7104167	0.0000190
C.V.		15.55	53.14	66.82	16.03	62.32	10.82	103.51
11	2001-2002	0.07038	322.87468	0.00264	0.01276	1.22500	124.97083	0.01949
	2002-2003	0.05701	469.95486	0.00103	0.01110	1.02696	127.33958	0.01477
	2003-2004	0.08399	242.55245	0.00465	0.01350	1.32101	115.97500	0.02354
	2004-2005	0.11844	659.66219	0.00195	0.02080	1.95207	216.81979	0.02654
C.V.		32.02	43.26	59.95	29.54	28.94	32.32	24.21
12	2001-2002	0.060891	282.779558	0.003942	0.013244	1.280833	126.383333	0.027257
	2002-2003	0.069916	178.984721	0.001303	0.011298	1.085104	95.813542	0.019383
	2003-2004	0.144827	251.551176	0.003449	0.022470	2.136479	203.723958	0.031507
	2004-2005	0.042648	133.442563	0.000900	0.007405	0.718615	67.633333	0.011874
C.V.		56.50	32.08	63.38	46.97	46.06	47.56	38.60
14	2001-2002	0.0350659	111.6513834	0.0011304	0.0067174	0.6057865	52.5932292	0.0120948
	2002-2003	0.0835884	130.4688334	0.0013102	0.0121665	1.1269896	82.6072917	0.0225035
	2003-2004	0.0725097	301.4900917	0.0016013	0.0143131	1.3540521	133.8364583	0.0274598
	2004-2005	0.0993233	362.6019985	0.0021401	0.0180414	1.7539583	161.2239583	0.0334407
C.V.		37.66	54.99	28.56	36.94	39.59	45.59	37.85
21	2001-2002	0.0738102	201.9920438	0.0040605	0.0136744	1.2337760	101.0843750	0.0265585
	2002-2003	0.1083590	184.0504147	0.0020588	0.0161118	1.5579479	115.2083333	0.0343581
	2003-2004	0.0723948	366.9135976	0.0018538	0.0164086	1.5683958	152.4375000	0.0365023
	2004-2005	0.0929010	307.7203288	0.0026274	0.0168987	1.6432292	142.7187500	0.0343124
C.V.		19.70	32.83	37.57	9.11	12.13	18.63	13.27
23	2001-2002	0.1822177	493.6264908	0.0050873	0.0341042	3.1557500	251.2078125	0.0808165
	2002-2003	0.1734928	519.6885122	0.0045306	0.0317441	3.0569896	252.4447917	0.0737586
	2003-2004	0.0613210	483.0401677	0.0019219	0.0170620	1.5057188	172.8742188	0.0318516
	2004-2005	0.0725345	214.0328987	0.0017978	0.0132160	1.2680313	105.3031250	0.0279945
C.V.		52.54	33.49	51.54	43.41	44.44	36.17	51.38
25	2001-2002	0.1285678	378.3308340	0.0040007	0.0247696	2.4302604	191.4791667	0.0120619
	2002-2003	0.0602394	215.0462615	0.0016441	0.0120972	1.2373854	103.2635417	0.0071932
	2003-2004	0.1290689	614.9734565	0.0028165	0.0258523	2.4085729	244.0822917	0.0104425
	2004-2005	0.1413022	617.6158968	0.0053465	0.0290607	2.7985082	260.5224638	0.0131598
C.V.		32.10	42.98	45.99	32.50	30.57	35.43	24.26

Tabla 4.8. *Razones con menor C.V. (patrón día)*

Las fases identificadas anteriormente se analizaron con mayor profundidad dado que presentan la menor variación. Para realizar el análisis, se utilizaron un número mayor de coeficientes de Haar en el algoritmo PAA.

4.7.2. Segmentación PAA Patrón Hora

Para realizar un análisis con mayor detalle se utilizó el algoritmo PAA para segmentar las ST por hora mediante coeficientes de Haar. De igual forma que en la segmentación anterior, fue necesario utilizar un número diferente de coeficientes de Haar por ciclo para obtener segmentos de longitud constante entre los cuatro ciclos productivos. El número de coeficientes de Haar utilizados para cada ciclo se presenta en la tabla 4.9.

Ciclo	Núm. de Coeficientes
2001-2002	1608
2002-2003	1584
2003-2004	1752
2004-2005	1800

Tabla 4.9. *Coeficientes de Haar (patrón – hora)*

Una vez segmentadas las ST se graficaron para su análisis. En la figura 4.7 se presenta la gráfica de línea de las observaciones de temperatura segmentadas utilizando el algoritmo PAA con los coeficientes de la tabla 4.8. Las gráficas de los demás EM se encuentran disponibles en el anexo C.

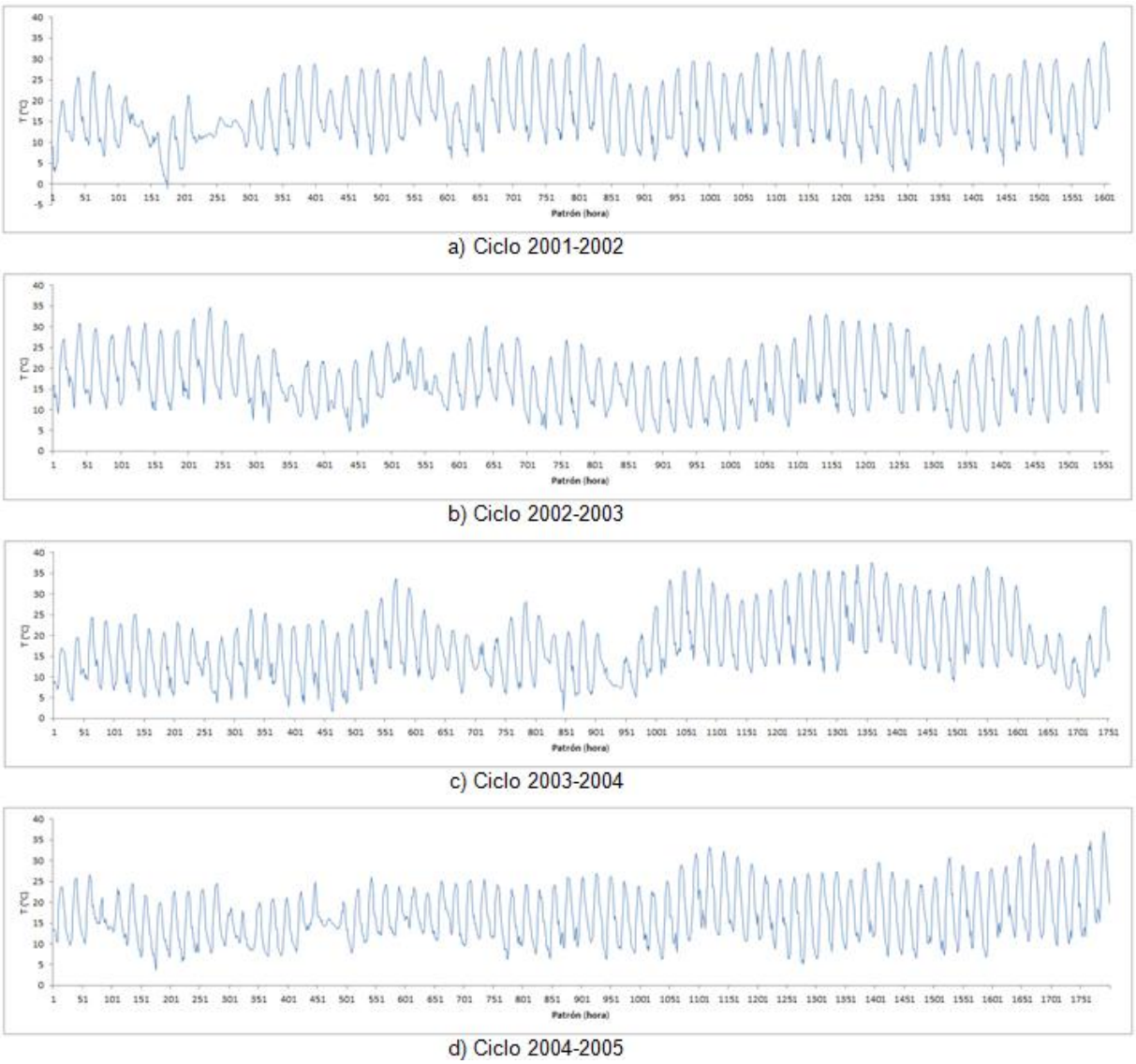


Figura 4.7. *Segmentación PAA (patrón hora) – Temperatura*

Para identificar una relación entre las observaciones de EM y la duración de las fases fenológicas seleccionadas, se realizó una matriz de gráficas de línea para cada fase fenológica. La matriz de gráficas de líneas para la fase 5 se presenta en la figura 4.8. Las gráficas de los demás EM se encuentran disponibles en el anexo D.

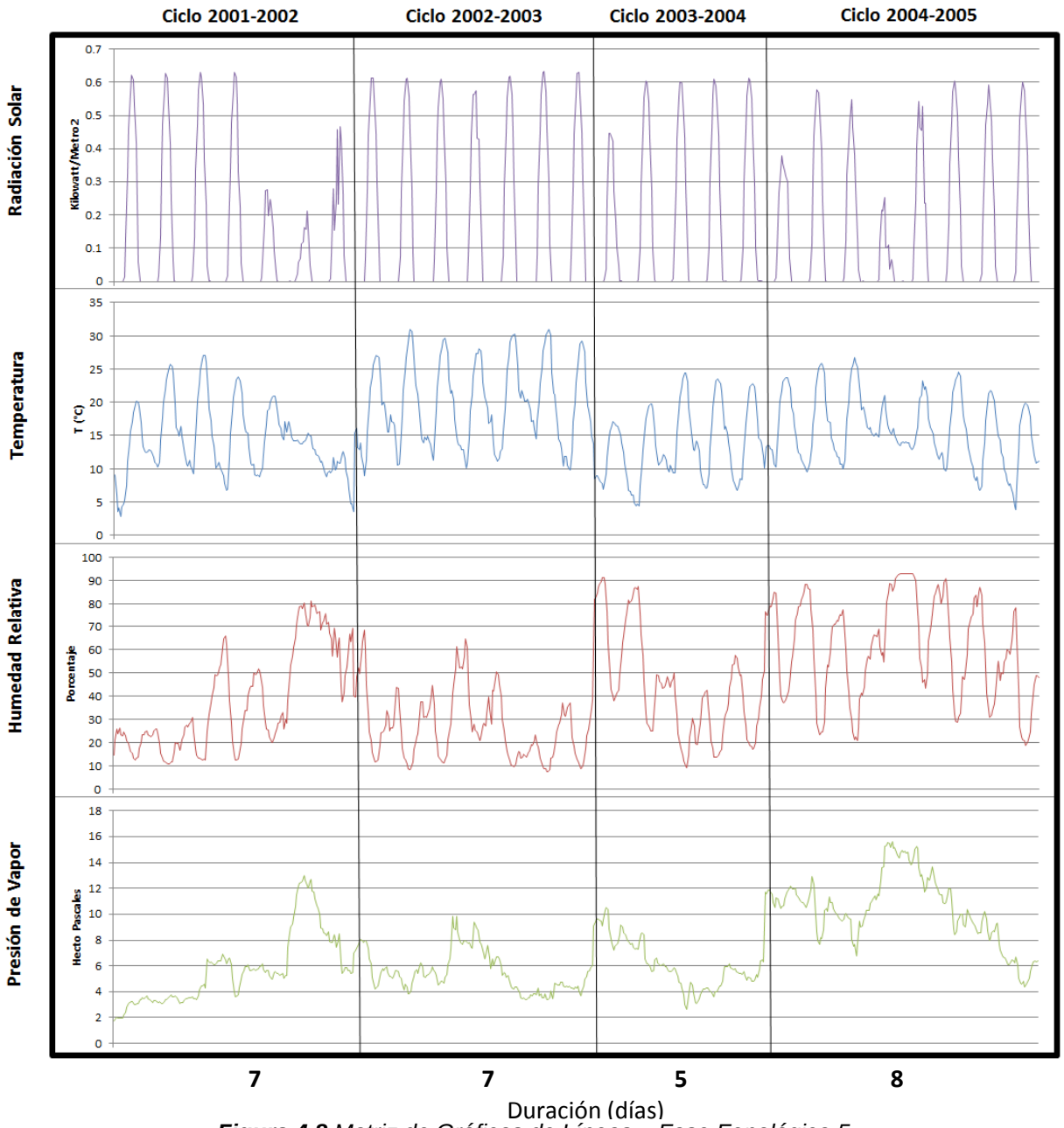


Figura 4.8 *Matriz de Gráficas de Líneas – Fase Fenológica 5*

Analizando las matrices de gráficas de líneas se identificó que en la fase 5, el primer y último ciclo tuvo radiación solar menor en varios días y sus duraciones fueron de 7 y 8 días respectivamente. Sin embargo, la radiación solar del segundo ciclo se mantuvo elevada durante todos los días y tuvo una duración de 7 días. La diferencia entre el segundo y tercer ciclo es que la temperatura del tercero fue menor. Con

estas consideraciones se decidió analizar la razón $\frac{RS^2}{T}$, para darle un peso mayor a la radiación solar alta y menor peso a la radiación solar baja.

En la fase 7 se identificó que solamente en el primer ciclo se presentaron temperaturas bajo 0°C, por lo que se analizó la razón de temperaturas mayores a 10°C. Se seleccionó el umbral de 10°C dado que es considerado la temperatura base o mínima de crecimiento para diversas variedades de vid, incluyendo la *Flame Seedless* (Centro de Información de Recursos Naturales, 1989).

En la fase 9 se identificó que en el último ciclo se presentó una humedad relativa y presión de vapor considerablemente mayor a los demás ciclos. Por esta razón se decidió incluir la razón $\frac{HR}{PV}$ en el análisis.

En la fase 21 se identificó que tanto el segundo como cuarto ciclo tuvieron humedad relativa alta, temperatura baja y su duración fue mayor, con una duración de 8 días. Las temperaturas del primer y tercer ciclo fueron menores. Estos ciclos fueron los de menor duración, con 6 y 7 días respectivamente. Considerando estas observaciones, se decidió analizar la razón $\frac{RS^2}{T\times\sqrt{HR}}$ para darle mayor peso a las radiaciones altas, menor a las radiaciones bajas y viceversa con la humedad relativa.

Las razones de menor CV de la segmentación por día y las razones identificadas en el análisis de las matrices de gráficas de líneas se presentan en la tabla 4.10. La tabla 4.11 presenta el resultado de la suma acumulada de dichas razones utilizando la segmentación por hora.

Razones		
Fase Fenológica	**Menor C.V. patrón (día)**	**Obtenidas del análisis de Matrices de Gráficas de Líneas patrón (hora)**
5	$\frac{RS}{T}$	$\frac{RS^2}{T}$
7	$\frac{RS \times PV}{T \times HR}$	$\frac{RS}{si(T > 10)}$
9	T	$\frac{HR}{PV}$
21	$\frac{RS \times PV}{T \times HR}$	$\frac{RS^2}{T \times \sqrt{HR}}$

Tabla 4.10. *Razones a analizar en segmentación (patrón – hora)*

Fase	Ciclo	Suma Acumulada de Razones de EM						
		$\frac{RS}{T}$	$\frac{RS^2}{T}$	$\frac{RS \times PV}{T \times HR}$	$\frac{RS}{si(T > 10)}$	T	$\frac{HR}{PV}$	$\frac{RS^2}{T \times \sqrt{HR}}$
5	2001-2002	1.215060	0.487377	0.256614	1.140324	2384.800000	1109.590685	0.113772
	2002-2003	1.104107	0.523208	0.360665	1.104107	3322.950000	791.847375	0.137818
	2003-2004	1.032432	0.457926	0.218967	0.989887	1635.075000	823.240835	0.093149
	2004-2005	1.304982	0.527295	0.308396	1.270701	3038.250000	1132.523213	0.088794
C.V.		10.33	6.55	21.57	10.27	28.94	18.84	20.71
7	2001-2002	1.194975	0.480639	0.209245	1.057037	2093.050000	1254.637018	0.006159
	2002-2003	0.917470	0.420402	0.306708	0.906799	2885.675000	675.957266	0.004548
	2003-2004	1.256067	0.569691	0.272312	1.198482	1977.325000	984.428072	0.005804
	2004-2005	1.022621	0.414524	0.229136	1.007799	2021.025000	952.325305	0.004133
C.V.		14.18	15.29	17.19	11.64	19.17	24.47	18.86
9	2001-2002	1.4083776	0.7312555	0.3983549	1.3973571	2946.0000000	915.5511985	0.2423807
	2002-2003	1.2525550	0.5279642	0.2780231	1.2188824	2444.7250000	1073.8250239	0.1025714
	2003-2004	1.5600279	0.7396173	0.3444218	1.4672029	2346.4000000	1146.0247268	0.1934082
	2004-2005	1.1494348	0.4285103	0.2429722	1.1293480	2801.0500000	1209.2540385	0.0603002
C.V.		13.39	25.37	21.90	11.98	10.82	11.65	55.52
21	2001-2002	1.2172874	0.6522150	0.3497548	1.2088971	2426.0250000	30.13288	0.18067
	2002-2003	1.8740151	0.9465820	0.4406028	1.7619234	2765.0000000	33.73550	0.17577
	2003-2004	1.3397579	0.7573159	0.5250875	1.3397579	3658.5000000	47.54121	0.17112
	2004-2005	1.6801246	0.8635984	0.4993489	1.6453213	3425.2500000	42.29889	0.19108
C.V.		19.81	15.89	17.15	17.33	18.63	20.65	4.76

Tabla 4.11*. Razones con menor C.V. (patrón hora)*

Una vez identificadas las sumatorias acumuladas de las razones de menor CV, es necesario calcular los intervalos de la predicción para la duración de las fases fenológicas.

4.7.3. Predicción de Duración de Fases Fenológicas

Se utilizaron las razones de menor CV para realizar la predicción de la duración de las fases fenológicas. Dado que las observaciones de duración fueron proporcionadas en días, solo es posible realizar predicciones en intervalos diarios. Para la validación del modelo se utilizó el método uno contra todos. Se utilizaron las observaciones de 3 ciclos productivos para predecir los del ciclo restante y se realizó la predicción para los 4 ciclos de la misma forma. El número que determina la duración de la fase fenológica (predictor) se calculó con la ecuación 4.4. Los intervalos para la predicción se calcularon utilizando las ecuaciones 4.5 y 4.6.

$$Predictor = \frac{\sum Razón\ de\ ciclos\ utilizados\ para\ validar}{\#\ ciclos\ utilizados\ para\ validar} \qquad (ec. 4.4)$$

$$Intervalo\ Inferior = \frac{\sum razón\ de\ ciclo\ a\ validar}{duración\ de\ fase\ en\ ciclo\ (días)} \times (num.día - 0.5) \quad (ec. 4.5)$$

$$Intervalo\ Superior = \frac{\sum razón\ de\ ciclo\ a\ validar}{duración\ de\ fase\ en\ ciclo\ (días)} \times (num.día + 0.5) \quad (ec. 4.6)$$

A continuación se presenta el ejemplo calcular la predicción de la fase fenológica 5 del ciclo 2001-2002:

$$Predictor = \frac{0.523 + 0.458 + 0.528}{3} = 0.503$$

El predictor se calcula mediante el promedio de la suma acumulada de la razón de menor CV, en este caso $\frac{RS^2}{T}$, de todos los ciclos menos el ciclo a predecir o en este caso ciclo 2001-2002.

$$Intervalo\ Inferior = \frac{0.487}{7} \times (num.día - 0.5)$$

$$Intervalo\ Superior = \frac{0.487}{7} \times (num.día + 0.5)$$

Los intervalos de confianza se calcularon en base al promedio por día de la suma acumulada de la razón de menor CV del ciclo a predecir, por el número del día $\mp$ 0.5 según sea el intervalo a predecir. La constante 0.5 da una probabilidad equitativa de predicción para un determinado día con el siguiente. En la tabla 4.12 se presentan los resultados de la predicción de duración en días de la fase fenológica 5. Las tablas de predicción de los demás fases se encuentran disponibles en el anexo E.

Duración Real de Fase

	Predictor	0.503
Núm. Día	Intervalo Inferior	Intervalo Superior
1	0.035	0.104
2	0.104	0.174
3	0.174	0.244
4	0.244	0.313
5	0.313	0.383
6	0.383	0.453
7	0.453	0.522
8	0.522	0.592

Tabla 4.12. *Predicción duración fase fenológica 5 ciclo 2001-2002*

Para analizar los resultados de las predicciones se realizaron matrices de confusión para cada fase fenológica. Cabe señalar que aunque la suma acumulada de la razón con menor C.V. para la fase 9 fue la de T, no se predijo acertadamente ninguna duración por lo que se utilizó la siguiente razón de menor CV, es decir la razón $\frac{HR}{PV}$. Las matrices de confusión se presentan en la tabla 4.13.

Los datos de las matrices de confusión (tabla 4.13) muestran como la duración de las fases 5, 7 y 9 se predicen correctamente en dos de cuatro ciclos. Las predicciones incorrectas de las fases 5 y 7 varían por un día, mientras que una predicción

incorrecta de la fase 9 varía por dos días. La duración de la fase 21 es la única que se predice correctamente 3 de 4 veces y la predicción incorrecta varía por un día. En base a estos resultados se determinan las reglas asociativas las cuales se presentan en la tabla 4.14. Las sumatorias acumuladas de las razones presentes en las reglas asociativas se determinaron de los promedio de los 4 ciclos analizados.

Matriz de Confusión
Validación Uno Contra Todos

Duración en Días Fase Fenológica 5		Reales				
		5	6	7	8	Total
Predicción	5					0
	6	1				1
	7			2	1	3
	8					0
	Total	1	0	2	1	4

a) Fase fenológica 5

Matriz de Confusión
Validación Uno Contra Todos

Duración en Días Fase Fenológica 7		Reales				
		5	6	7	8	Total
Predicción	5		1			1
	6		1			1
	7		1	1		2
	8					0
	Total	0	3	1	0	4

b) Fase fenológica 7

Matriz de Confusión
Validación Uno Contra Todos

Duración en Días Fase Fenológica 9		Reales					
		5	6	7	8	9	Total
Predicción	5						0
	6						0
	7			2	1		3
	8						0
	9			1			1
	Total	0	0	3	1	0	4

c) Fase fenológica 9

Matriz de Confusión
Validación Uno Contra Todos

Duración en Días Fase Fenológica 21		Reales				
		5	6	7	8	Total
Predicción	5					0
	6		1			1
	7			1	1	2
	8				1	1
	Total	0	1	1	2	4

d) Fase fenológica 21

Tabla 4.13. *Matrices de confusión - predicción de duración de fase fenológicas*

Detonador de Regla Asociativa	Regla Asociativa	Nivel Soporte	Nivel de Confianza
Fecha Final de Fase Fenológica 4	**Si** Sum. Acum. de $\frac{RS^2}{T}$ = 0.487 **entonces** Fase Fen. 5 concluirá	%100	%50
Fecha Final de Fase Fenológica 5	**Si** Sum. Acum. de $\frac{RS}{si(T>10)}$ = 1.043 **entonces** Fase Fen. 7 concluirá	%100	%50
Fecha Final de Fase Fenológica 7	**Si** Sum. Acum. de $\frac{HR}{PV}$ = 1,806.16 **entonces** Fase Fen. 9 concluirá	%100	%50
Fecha Final de Fase Fenológica 14	**Si** Sum. Acum. de $\frac{RS^2}{T\times\sqrt{HR}}$ = 0.1800 **entonces** Fase Fen. 21 concluirá	%100	%75

Tabla 4.14. *Reglas Asociativas Extraídas*

Una vez extraídas las reglas asociativas es necesario desarrollar modelo de la predicción de EM.

4.7.4 Modelo de Predicción de EM

Para definir las observaciones a utilizar en el conjunto de entrenamiento del modelo se analizaron las fechas de cada ciclo de las fases fenológicas 5, 7, 9 y 21. Estas fechas se presentan en la tabla 4.15.

	Fase Fenológica							
Ciclo	5		7		9		21	
	de	a	de	A	de	A	de	a
2001-2002	24/01/2002	30/01/2002	31/01/2002	06/02/2002	07/02/2002	13/02/2002	02/03/2002	07/03/2002
2002-2003	22/01/2003	28/01/2003	29/01/2003	03/02/2003	04/02/2003	10/02/2003	01/03/2003	08/03/2003
2003-2004	25/01/2004	29/01/2004	30/01/2004	04/02/2004	05/02/2004	11/02/2004	11/03/2004	17/03/2004
2004-2005	23/01/2005	30/01/2005	31/01/2005	05/02/2005	06/02/2005	13/02/2005	12/03/2005	19/03/2005
Fecha para entrenamiento	22/01	30/01	29/01	06/02	04/02	13/02	01/03	19/03

Tabla 4.15. *Fechas de fases fenológicas*

Después de identificar las fechas para el modelo de predicción de EM, se utilizaron las observaciones de la base de datos del SIA de los días que contaban con más de 22 horas de observaciones disponibles de los ciclos 2005 al 2011 de las fechas indicadas en la tabla 4.14. Las fechas de observaciones a utilizar en el modelo de

predicción de EM se presentan en la tabla 4.16. Las consideraron las observaciones de temperatura, humedad relativa y radiación solar, debido a que son las necesarias para calcular las sumas acumuladas de razones. La presión de vapor se calculó con las ecuaciones 3.2 y 3.3. Las observaciones de las fechas establecidas en la tabla 4.12 fueron filtradas de la misma forma que las observaciones utilizadas para el modelo de predicción de duración de las fases fenológicas.

Para realizar las predicciones se utilizó el algoritmo KNN con la distancia euclidiana. Para realizar la segmentación de las observaciones, se utilizó el algoritmo PAA para obtener segmentos de una hora. Además, debido a la rotación y traslación de la tierra se dividieron las observaciones en ST por día. Al segmentar la ST de esta manera se trabaja con el supuesto que los EM a una determinada hora y fecha tienden a presentar similitudes a través de los años.

La validación del modelo se realizó aplicando el método uno contra todos, de igual forma que en el modelo predicción de la duración de las fases fenológicas. El modelo en sí, realiza predicciones de 1 a 23 horas en el día, es decir, si es la 1:00 hr., el modelo predecirá hasta las 24:00 hr. de ese mismo día. De igual forma si son las 6:00, 7:00,…, 23:00 hr del día.

El error de la predicción se calcula mediante el valor absoluto de la diferencia entre el valor real y la predicción como se muestra en la ecuación 4.7. Es decir, si considera la 1:00 hr. como la hora actual, se calcula el error de predicción de 1:00 a 22:00 hr. Al considerar las 2:00 hr. como la hora actual, se calcula el error de predicción de 1:00 a 21:00 hr., y así sucesivamente hasta considerar la hora actual como las 23:00 hr.

Fase Fen. 5	Fase Fen. 7	Fase Fen. 9	Fase Fen. 21		
01/22/06	01/29/07	02/04/07	04/01/06	04/01/08	04/02/11
01/23/06	01/30/07	02/05/07	04/02/06	04/02/08	04/03/11
01/24/06	02/01/07	02/06/07	04/03/06	04/03/08	04/04/11
01/25/06	02/02/07	02/07/07	04/04/06	04/04/08	04/05/11
01/22/07	02/03/07	02/08/07	04/05/06	04/05/08	04/06/11
01/23/07	02/04/07	02/09/07	04/06/06	04/06/08	04/07/11
01/24/07	02/05/07	02/10/07	04/07/06	04/07/08	04/08/11
01/25/07	02/06/07	02/11/07	04/08/06	04/08/08	04/09/11
01/26/07	01/29/08	02/12/07	04/09/06	04/11/08	04/10/11
01/27/07	01/30/08	02/13/07	04/10/06	04/12/08	04/11/11
01/28/07	02/01/08	02/04/08	04/11/06	04/13/08	04/12/11
01/29/07	02/02/08	02/05/08	04/12/06	04/14/08	04/13/11
01/30/07	02/03/08	02/06/08	04/13/06	04/15/08	04/14/11
01/22/08	02/04/08	02/07/08	04/14/06	04/16/08	04/15/11
01/23/08	02/05/08	02/08/08	04/15/06	04/17/08	04/16/11
01/24/08	02/06/08	02/09/08	04/16/06	04/18/08	04/17/11
01/25/08	01/29/09	02/10/08	04/17/06	04/19/08	04/18/11
01/26/08	01/30/09	02/12/08	04/18/06	04/20/08	04/19/11
01/27/08	01/31/09	02/13/08	04/19/06	04/01/09	04/20/11
01/28/08	02/01/09	02/04/09	04/20/06	04/02/09	
01/29/08	02/02/09	02/05/09	04/01/07	04/03/09	
01/30/08	02/03/09	02/06/09	04/02/07	04/04/09	
01/22/09	02/04/09	02/07/09	04/03/07	04/05/09	
01/23/09	02/05/09	02/08/09	04/04/07	04/06/09	
01/24/09	02/06/09	02/09/09	04/05/07	04/07/09	
01/25/09	01/29/11	02/10/09	04/06/07	04/08/09	
01/26/09	01/30/11	02/11/09	04/07/07	04/09/09	
01/27/09	02/01/11	02/12/09	04/08/07	04/10/09	
01/28/09	02/02/11	02/13/09	04/10/07	04/11/09	
01/29/09	02/03/11	02/04/11	04/11/07	04/12/09	
01/30/09	02/04/11	02/05/11	04/12/07	04/13/09	
01/22/10	02/05/11	02/06/11	04/13/07	04/14/09	
01/23/10	02/06/11	02/07/11	04/14/07	04/15/09	
		02/08/11	04/15/07	04/16/09	
		02/09/11	04/16/07	04/17/09	
		02/10/11	04/17/07	04/18/09	
		02/11/11	04/18/07	04/19/09	
		02/12/11	04/19/07	04/20/09	
		02/13/11	04/20/07	04/01/11	

Tabla 4.16. *Fechas de observaciones adicionales – conjunto de entrenamiento*

donde solamente se realiza la predicción a 1:00 hr. Los intervalos de confianza para la media del error se calcularon con la ecuación 4.8 con un nivel de confianza de $\alpha = 0.05$ y se presentan en la tabla 4.17.

$$error\ predicción = \ abs(Valor\ real - Predicción) \qquad (ec. 4.7)$$

$$\bar{x} - t_{\alpha/2,n-1}\, S/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}\, S/\sqrt{n} \qquad (ec. 4.8)$$

Predicción a **n** Horas	Intervalos de Confianza de Error Promedio de Predicción								
	Temperatura			Humedad Relativa			Radiación Solar		
	Desv. Est.	Inter. Inf.	Int. Sup.	Desv. Est.	Inter. Inf.	Int. Sup.	Desv. Est.	Inter. Inf.	Int. Sup.
1	0.914	0.921	1.004	4.827	4.393	4.829	0.010	0.003	0.004
2	0.996	0.931	1.023	4.391	4.014	4.419	0.009	0.003	0.004
3	0.999	1.009	1.103	4.864	4.337	4.796	0.009	0.003	0.004
4	1.107	1.047	1.154	5.242	5.084	5.592	0.014	0.003	0.005
5	1.098	1.013	1.123	5.935	4.771	5.361	0.026	0.004	0.006
6	1.422	1.142	1.288	7.135	5.357	6.086	0.035	0.005	0.008
7	1.935	1.710	1.913	8.546	6.973	7.871	0.042	0.014	0.018
8	2.032	2.012	2.232	8.597	6.911	7.843	0.048	0.034	0.040
9	2.105	2.097	2.333	9.037	7.081	8.092	0.064	0.048	0.056
10	2.543	2.577	2.883	10.137	7.513	8.687	0.103	0.070	0.082
11	2.838	2.910	3.251	11.147	8.508	9.848	0.130	0.095	0.110
12	3.123	3.347	3.738	12.398	9.567	11.118	0.158	0.120	0.139
13	3.565	3.755	4.221	15.370	11.316	13.325	0.164	0.130	0.151
14	3.875	4.178	4.738	16.677	12.258	14.544	0.142	0.115	0.134
15	3.857	4.173	4.731	17.104	12.233	14.705	0.116	0.105	0.122
16	3.936	4.336	4.940	17.162	13.197	15.828	0.073	0.079	0.090
17	3.760	4.186	4.803	16.678	14.423	17.158	0.030	0.034	0.039
18	3.105	3.616	4.167	15.802	14.777	17.577	0.010	0.007	0.009
19	2.824	3.433	3.981	16.033	13.639	16.752	0.009	0.002	0.004
20	2.933	3.365	4.002	16.034	14.077	17.560	0.009	0.002	0.004
21	2.879	3.248	3.971	16.077	14.489	18.527	0.009	0.002	0.004
22	2.681	2.773	3.600	15.473	15.212	19.984	0.010	0.002	0.005
23	2.634	2.763	3.920	13.886	13.388	19.490	0.009	0.001	0.005

Tabla 4.17. *Intervalos de confianza para la media del error de predicción de EM*

La combinación de los dos modelos desarrollados en este trabajo, el modelo de predicción de la duración de las fases fenológicas y el modelo de predicción de EM

brindarán apoyo a la toma de decisiones en la agricultura. El realizar una predicción de EM con determinadas horas de anticipación permitirá a la parte interesada tomar la decisión de tomar acción para modificar el EM deseado mediante diferentes técnicas, quemadores, bloqueadores, neblina artificial, radiadores, bacterias, para controlar las sumatorias acumuladas de razones en caso de ser necesario acelerar o retardar el desarrollo vegetativo de la vid. Además, estos modelos pueden ser replicados con diferentes cultivos y en diferentes etapas fenológicas.

5. CONCLUSIONES

En este trabajo se ha demostrado la capacidad de las herramientas utilizadas en la MD para la extracción de conocimiento oculto en los datos, y como puede extraerse conocimiento nuevo basado en datos antiguos. La aplicación de este marco metodológico permite aprovechar la capacidad de reusar y reciclar los datos almacenas que ya habían cubierto su propósito operacional, utilizados en el proceso de monitoreo del desarrollo de sistema de producción de la uva de mesa.

El realizar gráficas de líneas de las observaciones de los EM permitió identificar fácilmente datos atípicos, así como la irregularidad periódica debida al cambio de hora de la observación (figura 3.3) y los registros de radiación solar de noche (figura 3.5), así como los datos faltantes. Los filtros de ST ya incluidos en programas de análisis de datos como MATLAB permiten corregir grandes volúmenes de datos en forma sencilla y rápida, reduciendo el tiempo necesario para el procesamiento de datos.

Durante la investigación de este proyecto se aplicaron diversas técnicas de MD para la extracción de reglas asociativas temporales. Como primer acercamiento a la extracción, se desarrolló la metodología presentada en el punto 2.5. Se evaluaron los tres enfoques de segmentación de ST presentados, BU, TD y SW. Se utilizaron las tres representaciones lineales APCA, RLS. Una vez segmentada la ST se determinaron agrupamientos utilizando los algoritmos k-medias y jerárquico aglomerativo con la distancia euclidiana, chebychev y manhatan. También se realizaron agrupamientos arbitrarios utilizando distintos percentiles. Los agrupamientos fueron validados con los índices Silhouette, Dunn y el Coeficiente de Partición.

Una vez agrupados los segmentos, se crea una secuencia genética con la cual se extraen las secuencias más repetidas o motif. Para la extracción de motifs se utilizó el algoritmo GreedyEM, y también se extrajeron con un algoritmo propio, el cual determinaba las secuencias de n longitudes más repetidas y la probabilidad condicional de que se cumpliera la secuencia. Se realizaron combinaciones de las distintas técnicas, algoritmos de segmentación, agrupamiento y de extracción de motifs, sin embargo no se logró identificar una relación entre los motifs extraídos con la duración de las fases fenológicas, por lo que no posible la extracción de reglas asociativas temporales.

Al analizar los resultados de la investigación, se decidió que la causa por la cual no se identificó relación alguna fue el sobre ajuste del modelo. Por esta razón se decidió utilizar el algoritmo PAA y sumatorias acumuladas de razones la extracción de reglas asociativas.

El algoritmo PAA utilizado para segmentar las ST se basa en los coeficientes de Haar, los cuales se encuentran dentro de más utilizados debido a la facilidad que cada segmento se interpreta con un dato, es decir, los segmentos cuentan con una longitud constante y no cuentan con pendiente.

La matriz de gráficas de líneas fue de gran ayuda en la investigación para la extracción de posibles razones a implementar en el análisis de CV debido a que se cuenta con las gráficas de todos los ciclos para una sola fase y permite generar una idea general de la relación entre los EM y la duración de la fase fenológica.

En cuanto a las reglas asociativas, la identificación de los detonadores fue directa debido a la forma en que se extrajeron, es decir, el final de una fase fenológica representa el inicio de la siguiente. Todas las reglas asociativas extraídas presentan un soporte de 100% debido a que durante la validación uno contra todos, se presentaron en todas las combinaciones de la base de datos generadas para realizar

las predicciones. Sin embargo, el nivel de confianza solamente fue igual a 75% en una de las reglas, mientras que en las demás se mantuvo en 50%.

El uso del algoritmo KNN en el modelo de predicción de los EM permitió realizar predicciones con un nivel de error a 5 horas cercano a un 1°C para la temperatura, 5% para la humedad relativa y 0.005 kW/m^2 para la radiación solar. Aun cuando el error pudiera considerarse bajo, el modelo no puede realizar predicciones de un día para otro, solamente dentro del mismo día.

6. INVESTIGACIONES FUTURAS

En futuras investigaciones pueden contemplar un número mayor de datos fenológicos y emplear las razones para corroborar si el nivel de confianza de las reglas asociativas varía. Además, se pueden incluir otros factores que no sean climáticos en el estudio como fertilizantes, plaguicidas, el riego, catalizadores para el crecimiento, entre otros factores. En cuanto al modelo de predicción de EM, se puede generar una base de datos utilizando el algoritmo de SW para obtener así secuencias de 24 horas para cada hora del día y realizar predicciones a plazos mayores de un día en base a la rotación y traslación de la tierra.

7. REFERENCIAS

Abonyi J., Feil B., 2007. *Cluster analysis for data mining and system identification*. Berllin: Birkhauser Verlag AG.

Abraham N., 2006. *A genetic algorithm for crystal structure prediction*. Doctorado. Universidad de Nueva York.

Abuyev R., Abiyev V., Ardil C., 2005. Electricity consumption prediction model using neuro-fuzzy system. *World Academic Science, Engineering and Technology*, (8), pp.128-131.

Addcon Telemetry, 2010. Accurate remote sensor technology. *ADCON International INC. USA, Canada, Mexico*.
Disponible en: www.adcon.at
Consultado el 04 de diciembre del 2011

Allen J., 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26 (11), pp.832-843.

Alnemer L., Wu J., Al-Azzam O., Denton A., 2010. A density-based algorithm for evaluating the statistical significance of individual classification results. *International Workshop on Data Mining in Bioinformatics,* Washington DC, USA.

Alter L., 1994. *Meteorology*. Yale-New Haven Teachers Institute
Disponible en: http://yale.edu/ynhti/curriculum/units/1994/5/94.05.01.x.html
Consultado el 11 de noviembre del 2011

Andonie R., et al. 2005. Neuro-fuzzy prediction of biological activity and rule extraction for HIV-1 protease inhibitors. En: IEEE (Institute of Electrical and Electronics Engineers), Computational Intelligence in Bioinformatics and Computational Biology, Ellensburg, USA, 14-15 Noviembre 2005.

Aydin I., Karakose M., Akin E., 2009. The prediction algorithm based on fuzzy logic using time series data mining method. *World Academic Science, Engineering and Technology*, 51, pp.91-98.

Bertoni A., Folgieri R., Valentini G., 2004. Bio-molecular cancer prediction with random subspace ensembles of support vector machines. *Neurocomputing*, 63, pp.535-539.

Bezdek J., 1974a. Cluster Validity with fuzzy sets. *Journal of Cybernetics,* 3 (3), pp.58-73.

Bezdek J. 1974b. Numerical taxonomy with fuzzy sets. *Journal of Mathematical Boilogy* , 1, pp.57-71.

Borensztajin G., Zuidema W., Bod R., 2009. The hierarchical prediction network: towards a neural theory of grammar acquisition. *Cognitive Science Society*.

Bow S., 2002. *Pattern recognition and image preprocessing*, 2da Edición, USA: Marcel Dekker Incorporation.

Bowen L., 2006. *Data mining for Information professionals.*

Bowerman B., O'Conell R., Koehler A., 2009. *Pronósticos, series de tiempo y regresión, un enfoque aplicado*. 4ta Edición, México D.F.: Learning.

Boyd J., 2008. Multiscale numerical algorithms for weather forecasting and climate modeling: challenges and controversies. *Society for Industrial and Applied Mathematics*. 41 (9).

Canutescu A., Shelenkov A., Dunbrack R., 2003. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Science*, 12 (9), pp.2001-2014.

Castellví F., Elías F., 2001. *Agrometeorología*. 2da Edición. Barcelona: Mundi-Prensa.

Cessna J., Colburn C., Bewley T., 2008. Enve: a new estimation on algorithm for weather forecasting and flow control. En: 4ta *American Institute of Aeronautics and Astronautics Flow Control Conference*, Seattle USA.

Centro de Información de Recursos Naturales., 1989. Requerimientos del suelo y el clima, frutales de hoja caduca. *Centro de Información de Recursos Naturales*, 83, pp.57-60.

Chakrabarti S., et al., 2009. *Data mining: know it all*. Morgan Kaufmann, Burlington MA.

Chatfield C., 2005. *The analysis of t.ime series, an introduction*. 6ta Edición. Chapman & Hall/CRC.

Chitra A., Uma S., 2010. An ensemble model of multiple classifiers for time series prediction*. International Journal of Computer Theory and Engineering*, 2 (3). pp.454-458.

Contreras J., 2007. *Series temporales y procesos estocásticos*. Universidad de Castilla, La Mancha.

Coombe B., 1995. Adoption of a system for identifying grapevine growth stages. Australian Journal of Grape and Wine Research, 1, pp.100-110.

Das G., et al., 1998. Rule discovery from time series. En: 4ta *International Conference on Knowledge Discovery and Data Mining,* New York, USA, Agosto 1998.

Dunn J., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Cybernetics and Systems*, 3 (3), pp.32-57.

Eisner R., et al., 2005. Improving protein function prediction using the hierarchical structure of the gene ontology. En: IEEE (Institute of Electrical and Electronics Engineers), *Computational Intelligence in Bioinformatics and Computational Biology*, Ellensburg, USA, 14-15 Noviembre 2005.

Everitt B., et al., 2011. *Cluster analysis*. 5ta Edición. John Wiley & Sons.

Fayyad U., Piatetsky-Shapiro G., Smyth P., 1996. From data mining to knowledge discovery in databases. *The American Association for Artificial Intelligence*, U.S.A. pp.37-54.

Fundación Produce Sonora (2004) Sistema de Información Agroclimática.
Disponible en: http://www.agroson.org.mx/
Consultado el 8 de agosto del 2012.

Gan G., Ma C., Wu J., 2007. *Data clustering theory, algorithms and applications. Society for Industrial and Applied Mathematics*. Philadelphia, Pennsylvania. American Statistical Association. Alexandria, Virginia.

Gore S. Bhosle U., 2011. Relative radiometric correction of multi temporal satellite imagery using fourier and wavelet transform. *Indian Society of Remote Sensing*, 40 (2), pp.201-213.

Graves D., Pedrycz W., 2009. Multivariate segmentation of time series with differential evolution. En: *International Fuzzy Systems Association World Congress and European Society for Fuzzy Logic and Technology*, Lisbón, Portugal, 20-24 Julio 2009.

Han J., Kamber M., 2006. *Data mining concepts and techniques*. Elsevier Science and Technology, Amsterdam.

Hernandez J., Ramirez M., Ferri C., 2004. *Introducción a la Minería de Datos*. Madrid: Pearson Education .

Hidalgo L., 2002. *Tratado de viticultura general*. 3ra Edición Madrid: Mundi-Prensa

Isbister W., 1918. *Meteorology a textbook on the weather, the causes of its changes and weather forecasting*. New York: The Macmilan Company .

Jambhulkar S., Borkar N., Sorte S., 2011. Comparison of k-means and adaptive k-means with VHDL implementation. *International Journal of Engineering Science and Technology*, pp.22-27.

Jan Z., et al., 2008. Seasonal to inter-annual climate prediction using data mining KNN technique. En: International Multi topic Conference 2008, *Communications in Computer and Information Science*, 20, pp.40-51.

Junkui L., Yuanzhen W., 2007. APCAS: an approximate approach to adaptively segment time series stream. *Lecture Notes in Computer Science*, 45 (5), pp.554-565.

Keogh E., et al., 2000. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems*, 3 (3), pp.263-286.

Keogh E., et al., 2001. An online algorithm for segmenting time series. En: IEEE (Institute of Electrical and Electronics Engineers*), International Conference on Data Mining*, 29 Noviembre – 2 Diciembre, pp. 289-296

Keogh E., Lin J., Truppel W., 2003. Clustering of time series subsequences is meaningless: implications for previous and future research. En: IEEE (Institute of Electrical and Electronics Engineers*), International Conference,* Washington DC, pp. 115-122.

Kellert S., 1993. In the wake of chaos: unpredictable order in dynamical systems. *University of Chicago Press*, pp.32.

Khedkar P., Keshav S., 1992. Fuzzy prediction of time series. En: IEEE(Institute of Electrical and Electronics Engineers), *Conference on Fuzzy Systems*, 8-12 Marzo 1992.

Kitagawa G., 2010. *Introduction to time series modeling*. Taylor & Francis Group.

Lampert C., Blaschko M., 2009. Structured prediction by joint kernel support estimation. *Machine Learning*, 77 (2), pp.249-269.

Larose D., 2005. *Discovering knowledge in data an introduction to data mining.* Wiley Chichester.

Lezaun M., 2002. Predicciones del tiempo y matemáticas. *Boletín de la Sociedad Española de Matemáticas Aplicada*, 22, pp.61-100.

Li Z., Chen J., Schraudolph N., 2008. An improved mean-shift tracker with kernel prediction and scale optimization targeting for low-frame-rate video tracking. En: ICPR (*International Conference on Pattern Recognition*), 37 (3).

Lin Y., McCool M., Ghorbani A., 2010. Time series motif discovery and anomaly detection based on subseries join. En: IANG (*International Journal of Computer Science*), 37 (3).

Lkhagva et al., 2006. Extended SAX: extension of symbolic aggregate approximation for financial time series data representation. DEWS, 4A-i8.

Lynch P., 2007. The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, 7 (34), pp.31–44.

Maeda J., et al., 2008. Perceptual image segmentation using fuzzy-based hierarchical algorithm and its application to dermoscopy images. En: IEEE(Institute of Electrical and Electronics Engineers), *Conference on Soft Computing in Industrial Applications*, 25-27 Junio 2008, Muroran, Japón.

Marro M., 2000. *Principios de viticultura guías de agricultura y ganadería*. Ediciones CEAC, cap. 4, pp.64-67, cap. 8, pp.115-124.

McCue P., Hunter J., 2004. *Multivariate segmentation of time series data*. U.K. University of Aberdeen.

McGovern A., et al., 2011. Identifying predictive multi-dimensional time series motifs: an application to severe weather prediction. *Data Mining Knowledge Discovery,* 22, pp.232-258.

Montgomery D., Runger G., 2003. *Probabilidad y estadística aplicada a la ingeniería*. 3ra Edición McGraw-Hill

Morchen F., 2006. *Time Series Knowledge Mining*, Doctorado. Marburg: Universidad Philipps.

Myers W., Linden S., Wiener G., 2009. A data mining approach to soil temperature and moisture prediction. Seventh Conference on Artificial Intelligence and its Applications to the Environmental Sciences, Enero 2009, Arizona USA.

Nimbus Weather Service, 2005. *Aprendiendo Meteorología.*
Disponible en: http://nimbus.com.uy/aprendiendo.html
Consultado el 11 de noviembre del 2011.

Ostendorf M., Veilleux N., 1994. A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics*, 20 (1), pp.27-54.

Oyelade J., Oladipu O., Obagbuwa C., 2010. Application of k-means clustering algorithm for prediction of student's academic performance. *International Journal of Computer Science and Information Security*, 7 (1).

Pal R., Bezdak J., 1995. On cluster validity of the fuzzy c-means model. En: IEEE (Institute of Electrical and Electronics Engineers), *Transactions on Fuzzy Systems*, 3 (3), pp.370-379.

Paras et al., 2007. A feature based neural network model for weather forecasting. *World Academy of Science, Engineering and Technology*, 34, pp.66-73.

Pérez C., Santín D., 2007. *Minería de datos: técnicas y herramientas*. Madrid España: Thomson

Piatetsky-Shapiro G., 1989. Knowledge discovery in real databases: a report on the IJCAI-89 workshop. *American Association for Artificial Intelligence*. AI Magazine, U.S.A.

Pisón F., Ordieres J., Pernía A., Alba F., 2005. Minería de datos en series de temporales para la búsqueda de conocimiento oculto en históricos de procesos industriales*. Actas del III Taller Nacional de Minería de Datos y Aprendizaje TAMIDA* (Taller Nacional de Minería de Datos y Aprendizaje), pp. 31-38.

Povinelli R., 1999. *Time series data mining: identifying temporal patterns for characterized and prediction of time series events*. Doctorado, Milwaukee: Marquette University.

Preciado J., 2011. *Agrupamiento de patrones correlacionados y con incertidumbre: caso patrones climáticos en la producción de uva de mesa en un viñedo de Sonora.* Doctorado. Mexicali: Universidad Autónoma de Baja California.

Radhika Y., Shashi M., 2009. Atmospheric temperature prediction using support vector machines*. International Journal of Computer Theory and Engineering*, 1 (1), pp.55-58.

Rashed R., Morelon R., 1996. *Encyclopedia of the history of arabic science*. UK: Routledge.

Reynier A., 2002. *Manual de viticultura guía técnica de viticultura.* Ediciones Mundi-Prensa, 6ta Edición, cap. 2, pp.365-375.

Rodríguez T., 2002*. Introducción a los métodos estadísticos, numéricos y probabilísticos*. Colegio Marista Cristo Rey, pp.113-136.

Romani L., et al., 2010. Mining relevant and extreme patterns on climate time series with CLIPSMiner. *Journal of Information and Data Management*, 1 (2), pp.245-260.

Rousseeuw P., 1987. Silhouettes a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computation and Applied Mathematics*, 20, pp.53-65.

Sen Z., Oztopal A., 2001. Genetic algorithms for the classification and prediction of precipitation occurrence. *Hydrological Sciences Journal*, 46 (2), pp.255-267.

Singh S., Bhambri P., Gill J., 2011. Time series based temperature prediction using back propagation with genetic algorithm technique. *International Journal of Computer Science*, 8 (5), pp.28-35.

Sivakumar M., 2004. *Predicción del clima y la agricultura*. En: organización Meteorológica Mundial, Genève, Suiza, 2004.

Sivakumar M., Hansen J., 2007. *Climate prediction and agriculture, advances and challenges*. Berlin: Springer.

Sneyers R., 1998. Climate chaotic instability: statistical determination and theoretical background. *Environmetrics*, 8 (5), pp.517-532.

Tanaka Y., Iwamoto K., Uehara K., 2005. Discovery of time-series motif from multi-dimensional data base don MDL principle. *Machine Learning*, 28, pp.269-300.

Tang Z., MacLennan J., 2005. *Data mining with SQL server 2005*. U.S.A.: Wiley Publishing Inc.

Taylor J., McSharry P., Buizza R., 2009. Wind power density forecasting using ensemble predictions and time series models. En: IEEE (Institute of Electrical and Electronics Engineers), *Transactions on Energy Conversion*, (24), pp.775-782.

Tu Tiempo, 2011. *Fórmulas por aquí, fórmulas por allá.*
Disponible en: www.tutiempo.net
Consultado el 12 de diciembre de 2011

Veenadhari S., Bharat M., Singh C.D., 2011. Data mining techniques for predicting crop productivity a review article. *International Journal of Computer Science and Technology*, 2 (1), pp.98-100.

Witten Ian H., Frank Eibe, 2005. *Data mining: practical machine learning tools and techniques*. 2da Edición. U.S.A.: Elseiver-Morgan Kaufmann Publishers.

Wu W. et al., 2004. Online event-driven subsequence matching over financial data streams. En: ACM SIGMOD (Association for Computing Machinery Special Interest Group on Management Data), Junio 2004, Paris Francia.

Wu W., et al., 2010. Application of a k-nearest neighbor simulator for seasonal precipitation prediction in a semiarid region with complex terrain. *Geophysical Research Abstracts*, 12, pp.5237.

Wu X., et al., 2010. Blind separation of speech signals based on wavelet transform and independent component analysis. *Transactions of Tianjin University*, 16 (2), pp.123-128.

Xu R., Wunsch D., 2009. *Clustering*. Canada: John Wiley & Sons.

Yu B., Tiwari R., 2007. Application of EM algorithm to mixture cure model for grouped relative survival data. *Journal of Data Science*, 5, pp.41-51.

Zahoor J., et al., 2008. Seasonal to inter annual climate prediction using data mining KNN technique. *International Multi Topic Conference CCIS*, 20, pp.40-51.

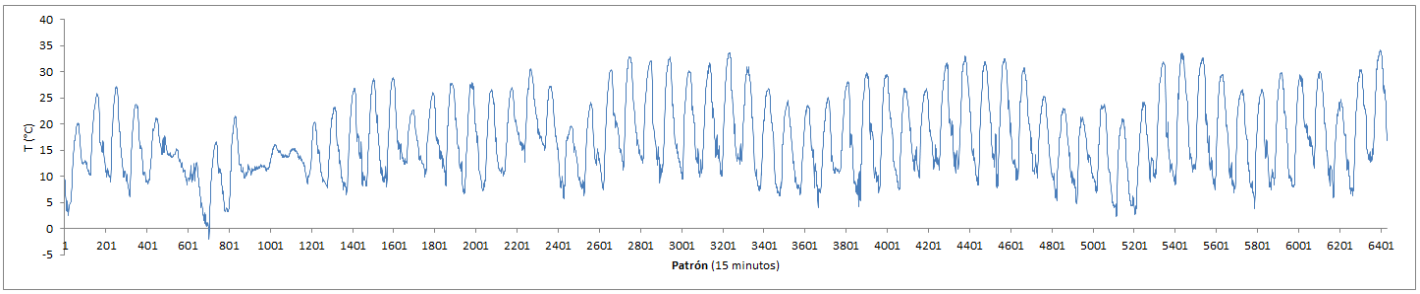
Zhang Z., et al., 2008. Continuous k-means monitoring over moving objects. En: IEEE (Institute of Electrical and Electronics Engineers), *Transactions on Knowledge and Data Engineering,* 20 (9), pp.1205-1216.

Zhang, Glass, 2011. A piecewise aggregate approximation lower-bound estimate for posteriorgram-based dynamic time warping. En: ISCA (International Speech Communication Association), Agosto 2011, Florencia Italia.

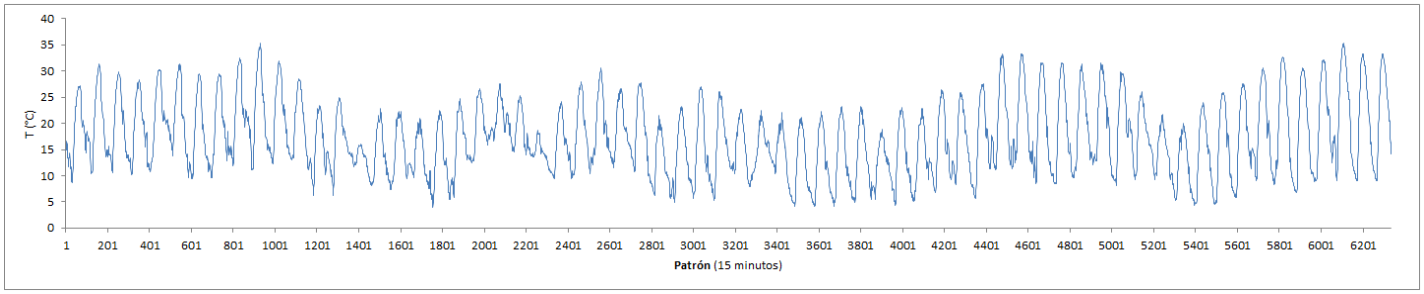
Zhu Y., De W., Li S., 2007. A piecewise linear representation method of time series based on feature points. En: *Knowledge Based Intelligent Information and Engineering Systems and the XVII Italian Workshop on Neural Networks*, 2007.

Anexo A

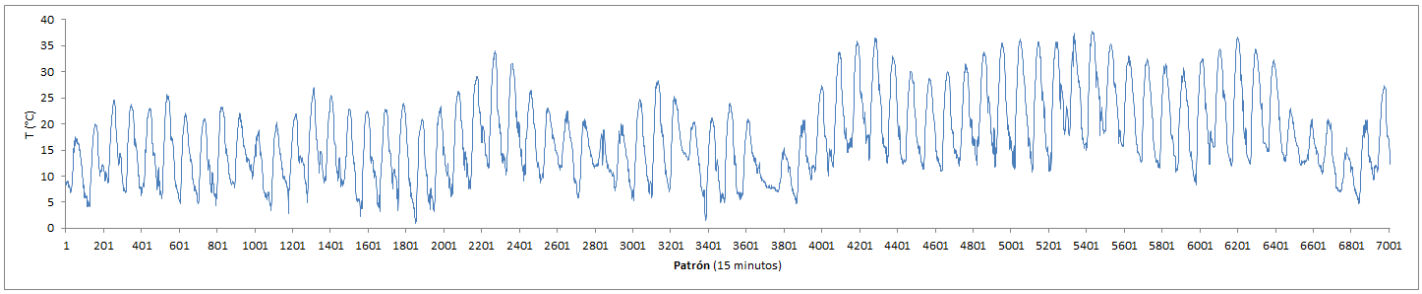
Comportamiento histórico de las observaciones de EM en intervalos de 15 minutos.



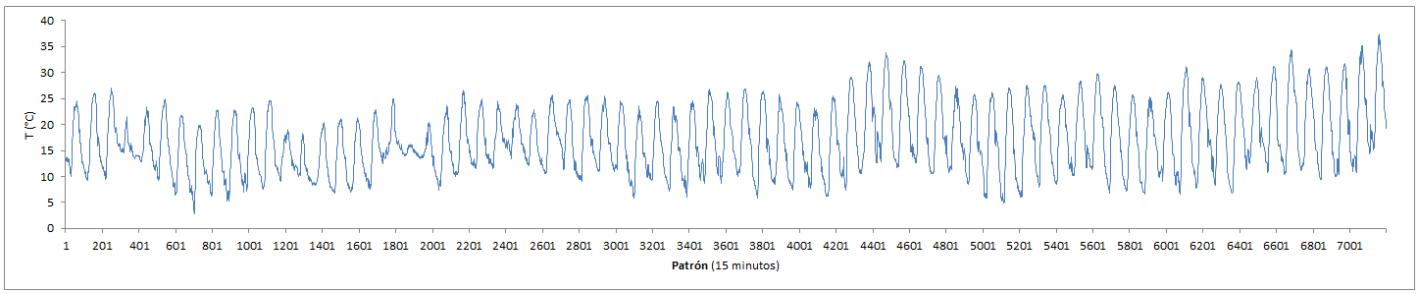
a) Ciclo 2001-2002



b) Ciclo 2002-2003

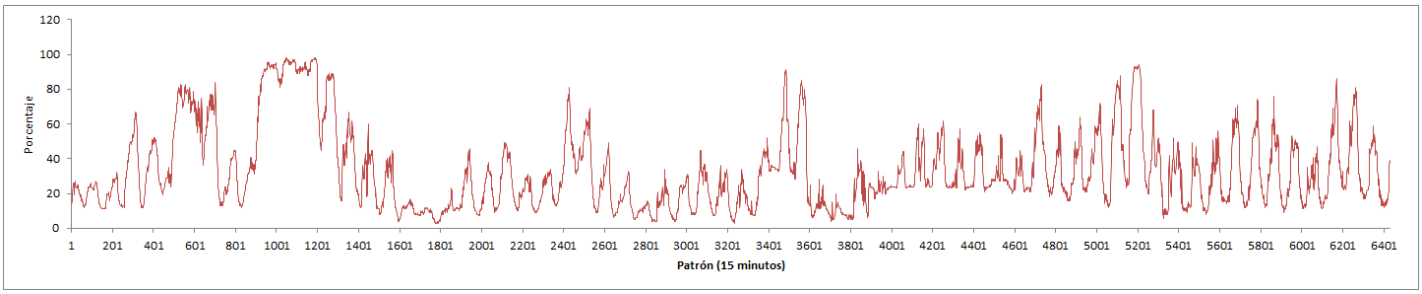


c) Ciclo 2003-2004

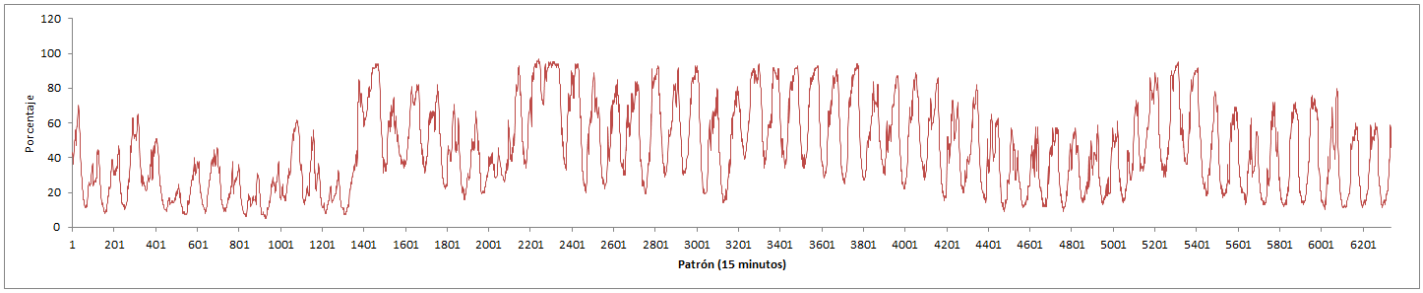


d) Ciclo 2004-2005

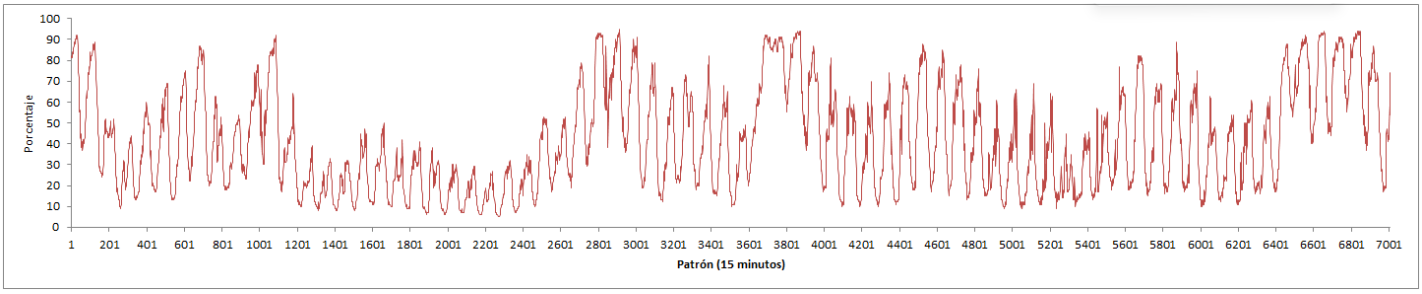
Figura A.1 *Comportamiento de la temperatura durante el periodo de estudio*



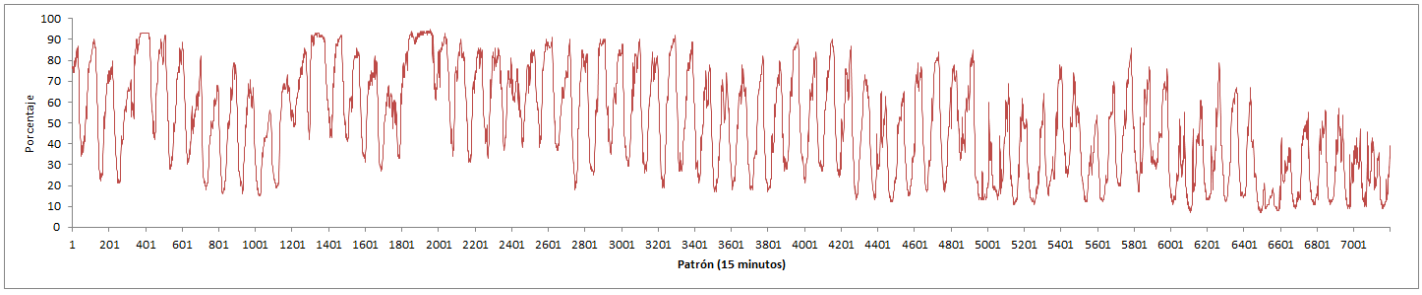
a) Ciclo 2001-2002



b) Ciclo 2002-2003

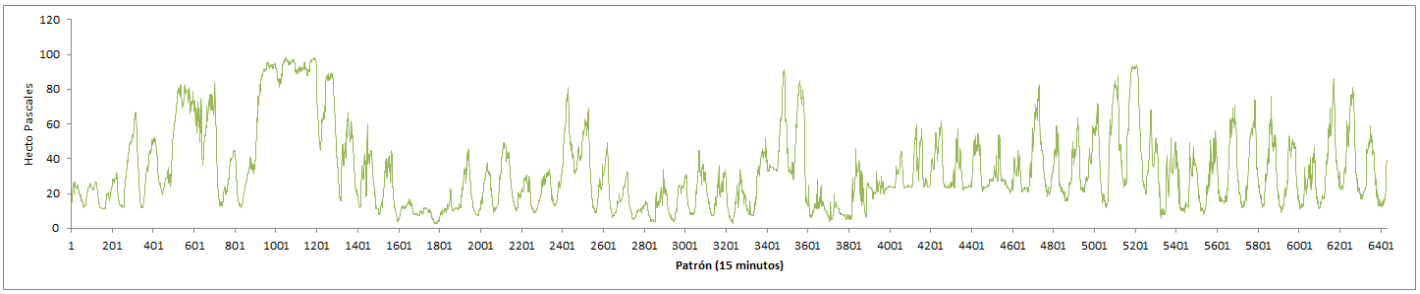


c) Ciclo 2003-2004

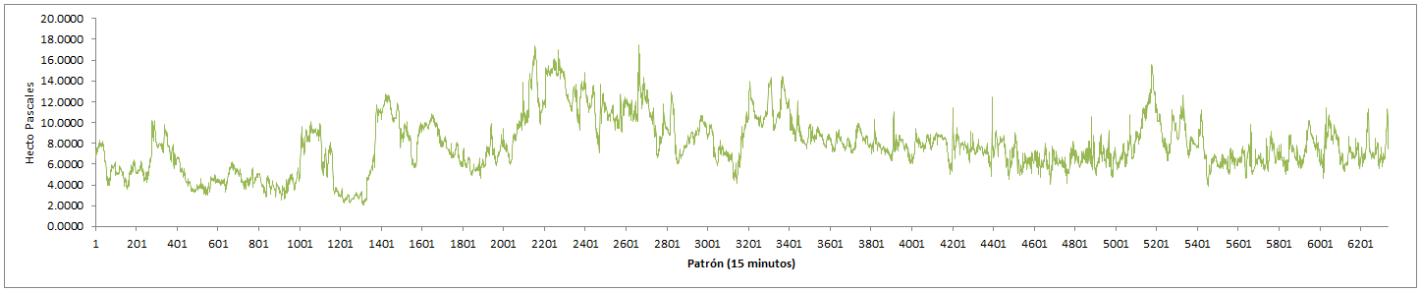


d) Ciclo 2004-2005

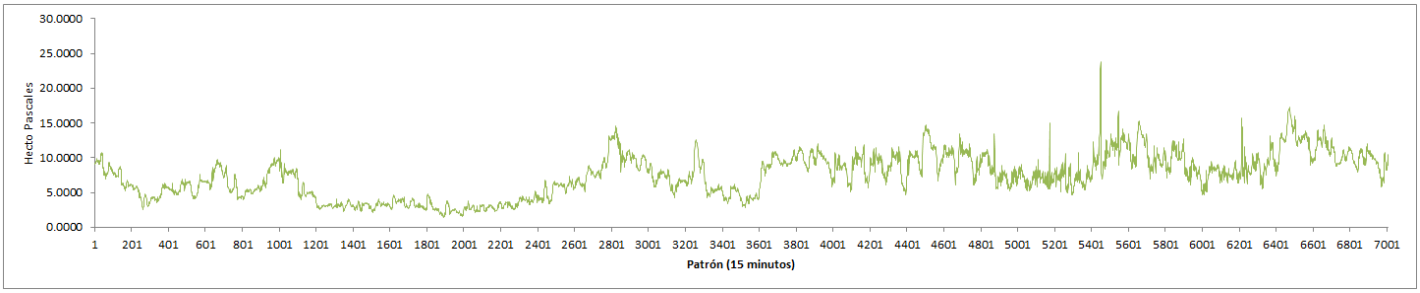
Figura A.2 *Comportamiento de la humedad relativa durante el periodo de estudio*



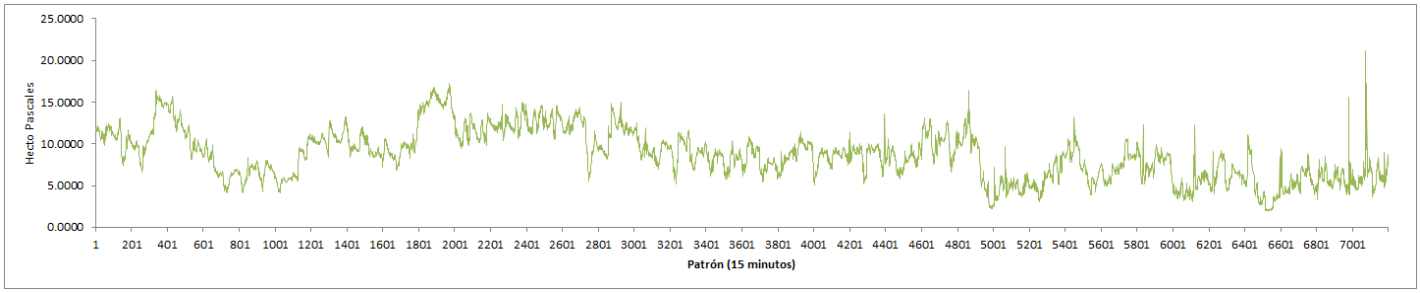
a) Ciclo 2001-2002



b) Ciclo 2002-2003

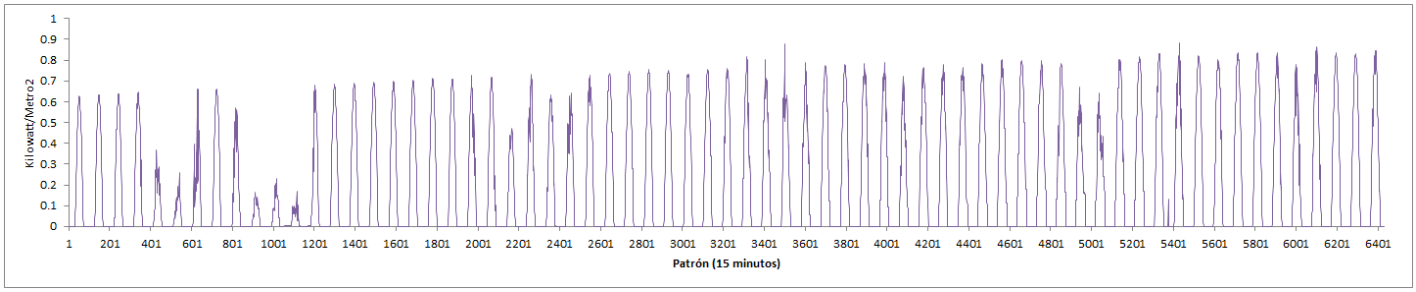


c) Ciclo 2003-2004

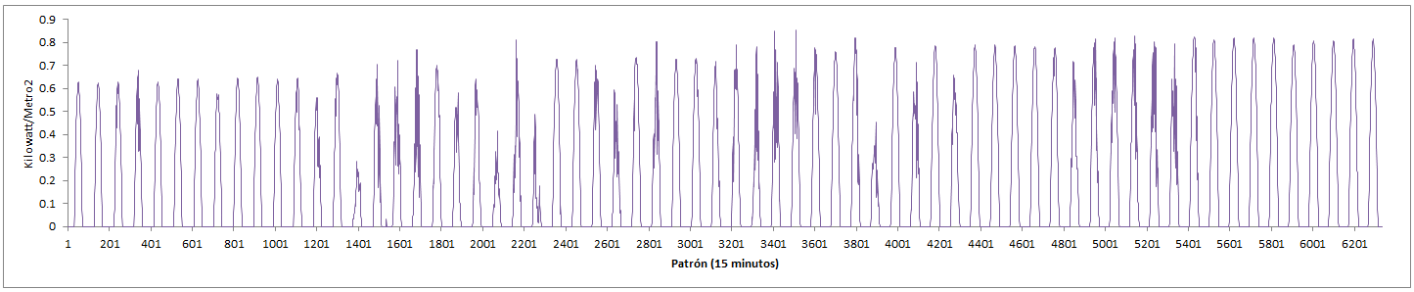


d) Ciclo 2004-2005

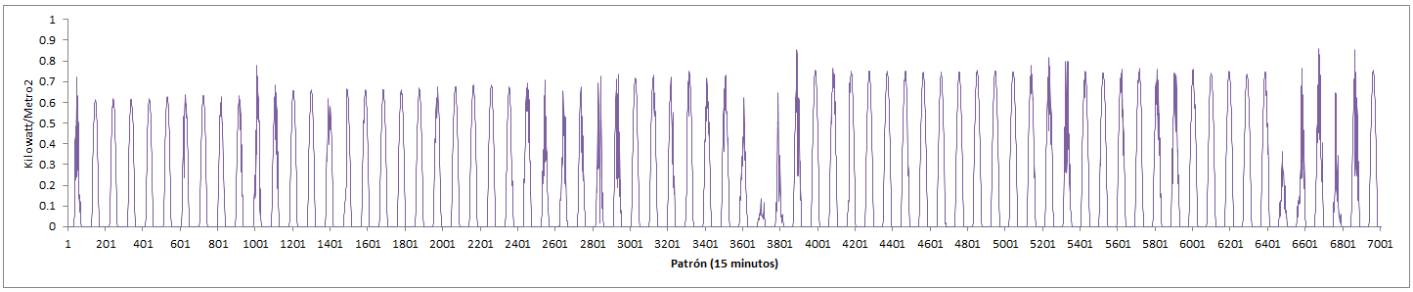
Figura A.3 *Comportamiento de la presión de vapor durante el periodo de estudio*



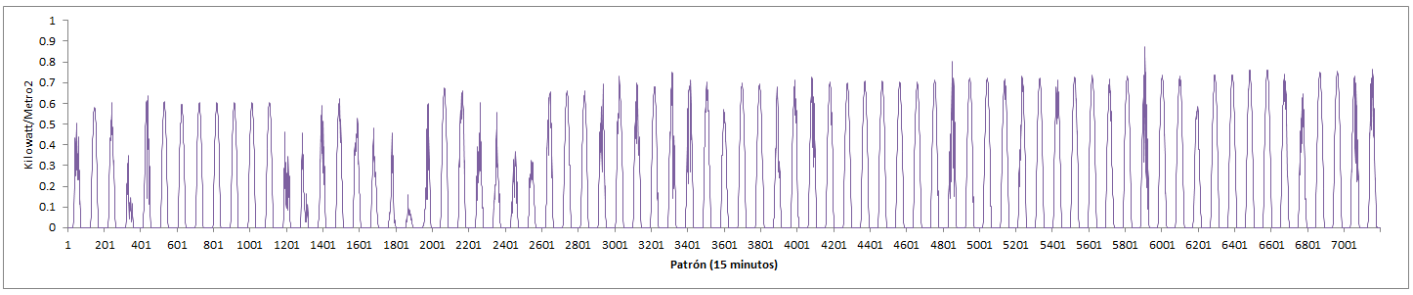
a) Ciclo 2001-2002



b) Ciclo 2002-2003



c) Ciclo 2003-2004

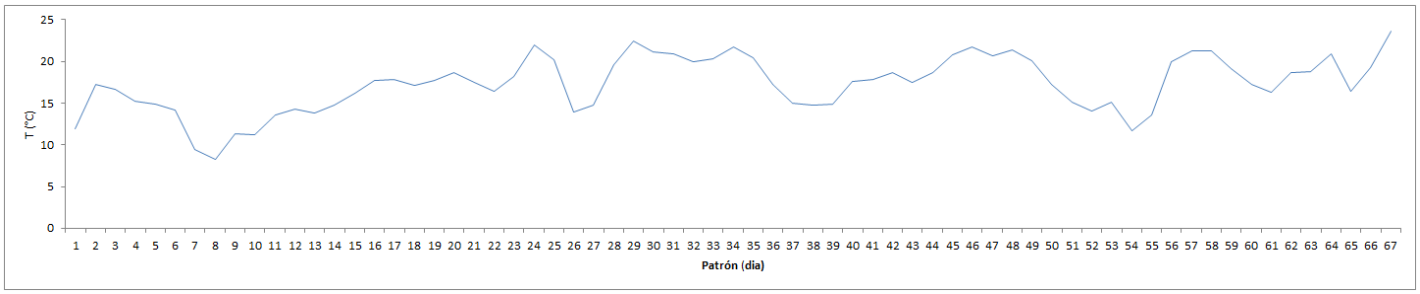


d) Ciclo 2004-2005

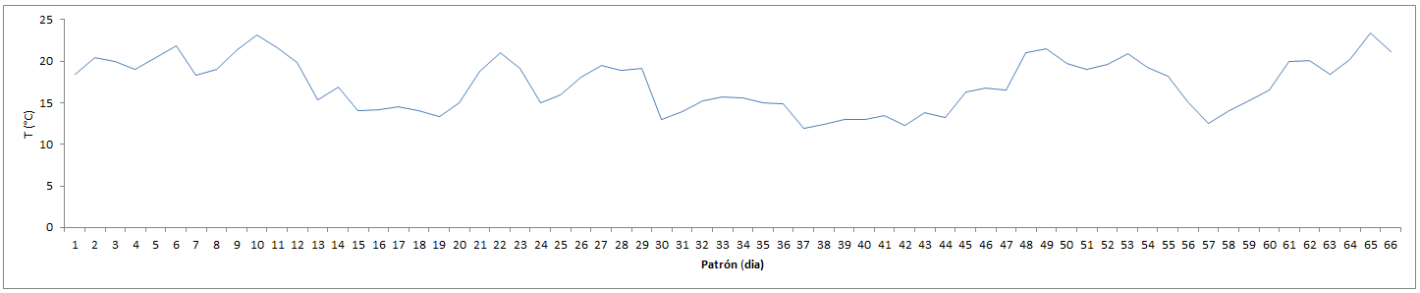
Figura A.4 *Comportamiento de la radiación solar durante el periodo de estudio*

Anexo B

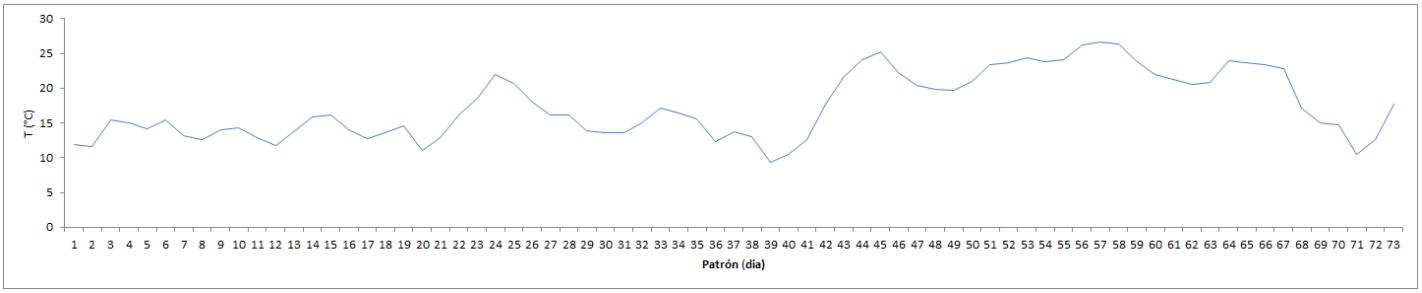
Segmentación PAA de las observaciones de EM (patrón día).



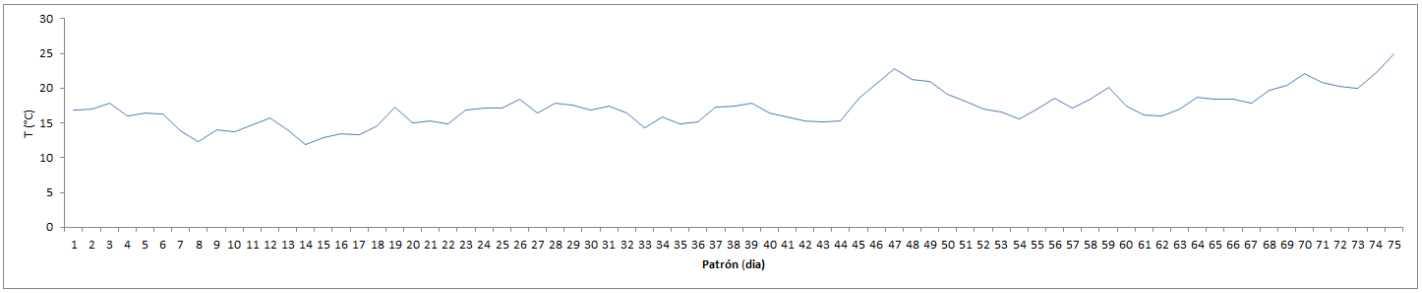
a) Ciclo 2001-2002



b) Ciclo 2002-2003

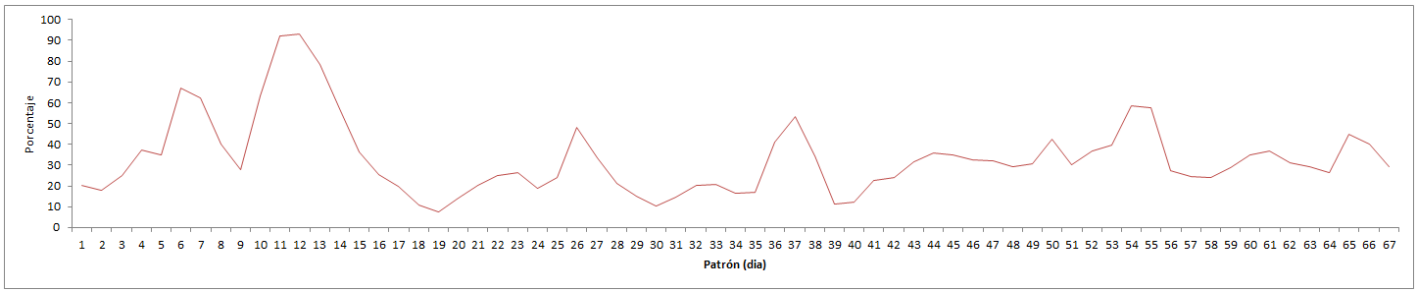


c) Ciclo 2003-2004

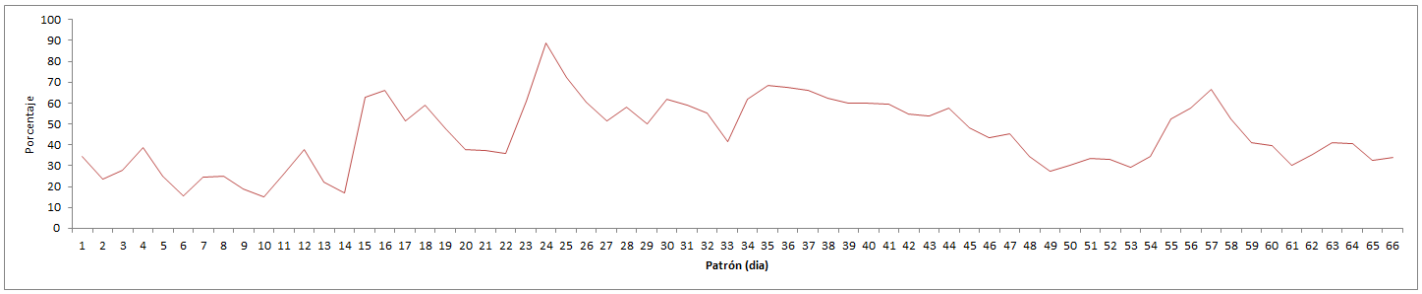


d) Ciclo 2004-2005

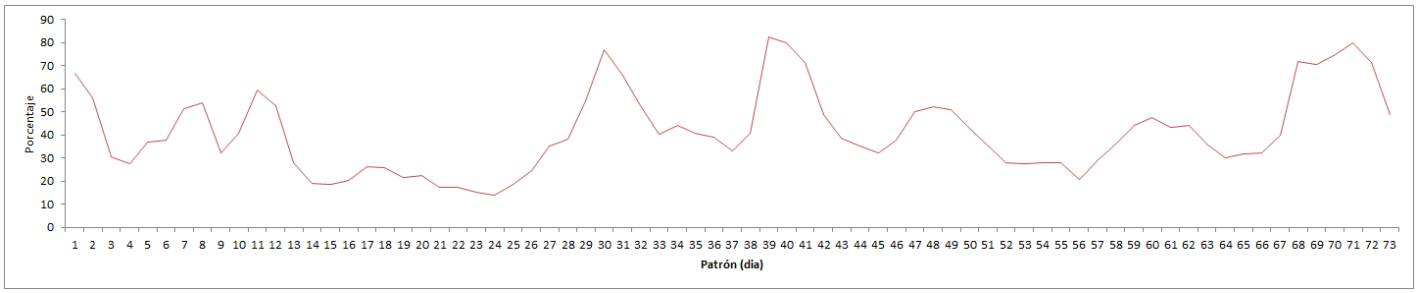
Figura B.1 *Segmentación PAA (patrón día) – Temperatura*



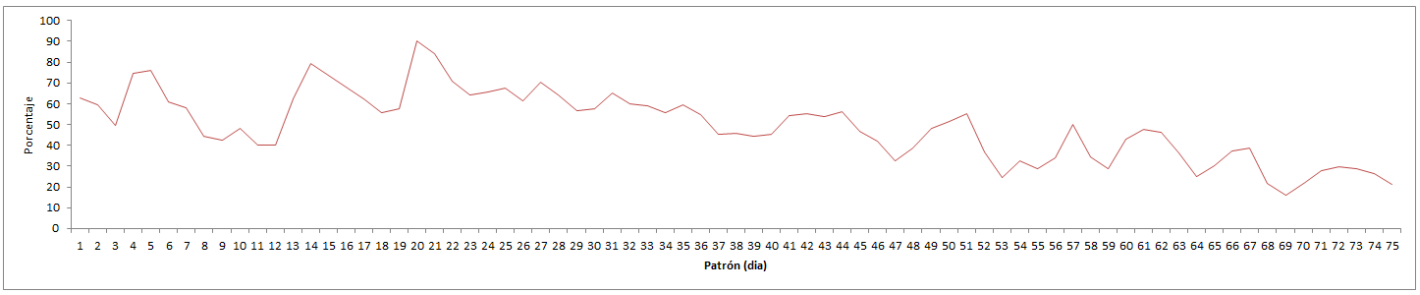
a) Ciclo 2001-2002



b) Ciclo 2002-2003

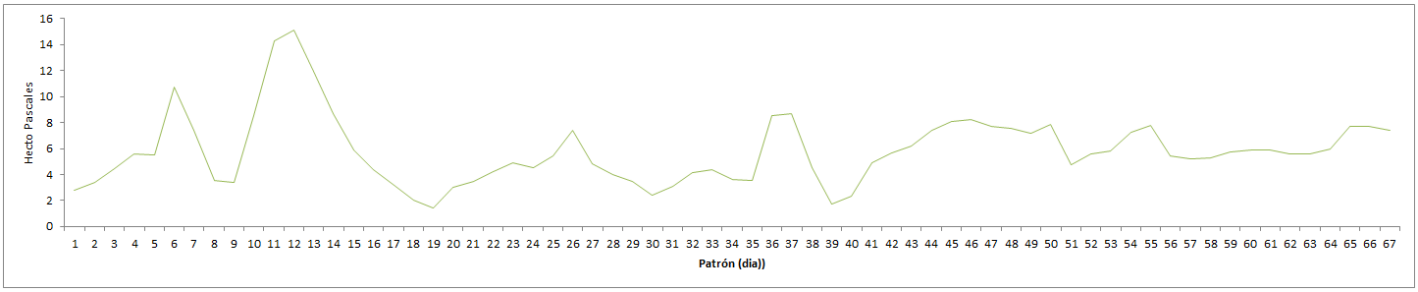


c) Ciclo 2003-2004

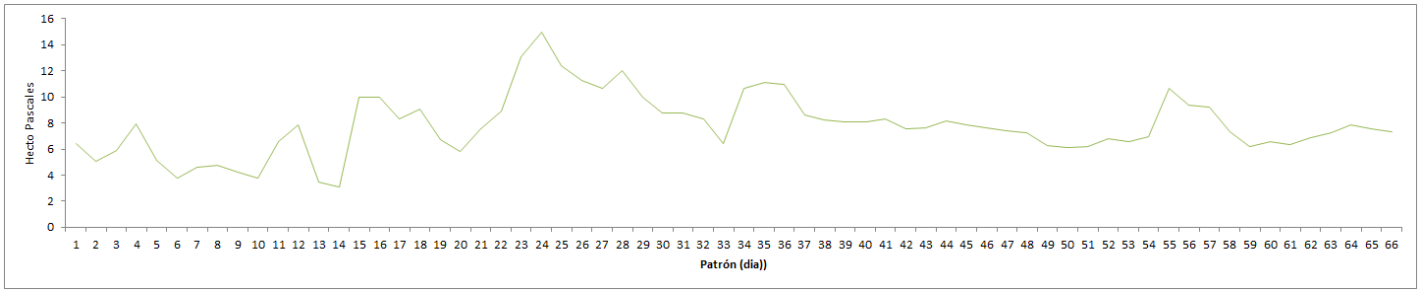


d) Ciclo 2004-2005

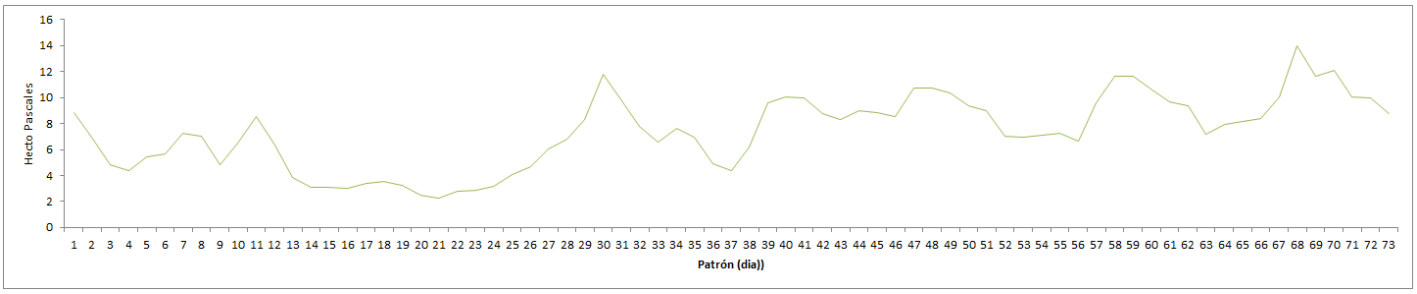
Figura B.2 *Segmentación PAA (patrón día) – Humedad Relativa*



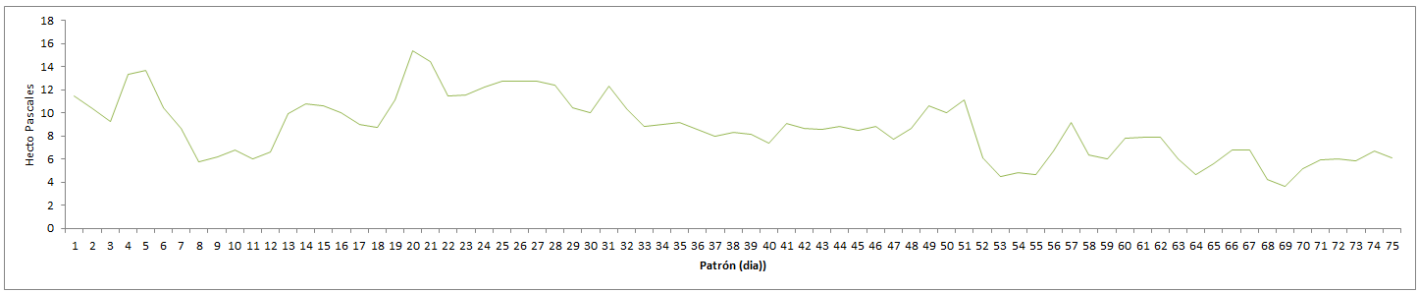
a) Ciclo 2001-2002



b) Ciclo 2002-2003

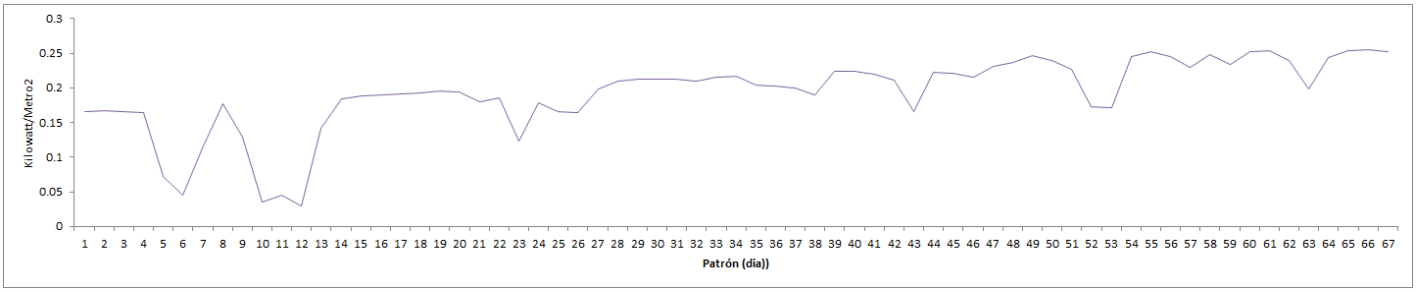


c) Ciclo 2003-2004

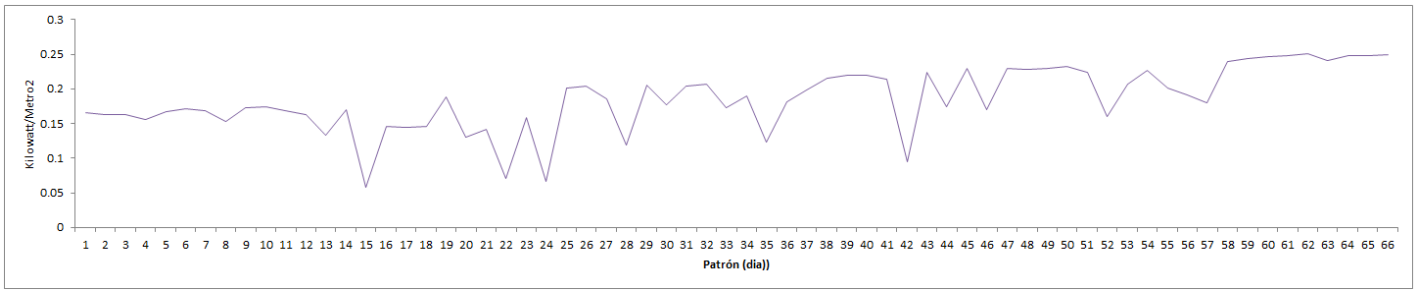


d) Ciclo 2004-2005

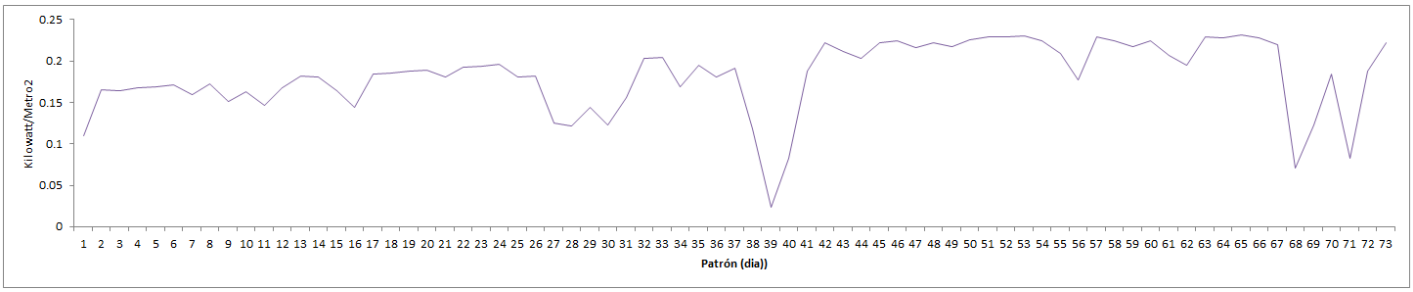
Figura B.3 *Segmentación PAA (patrón día) – Presión de Vapor*



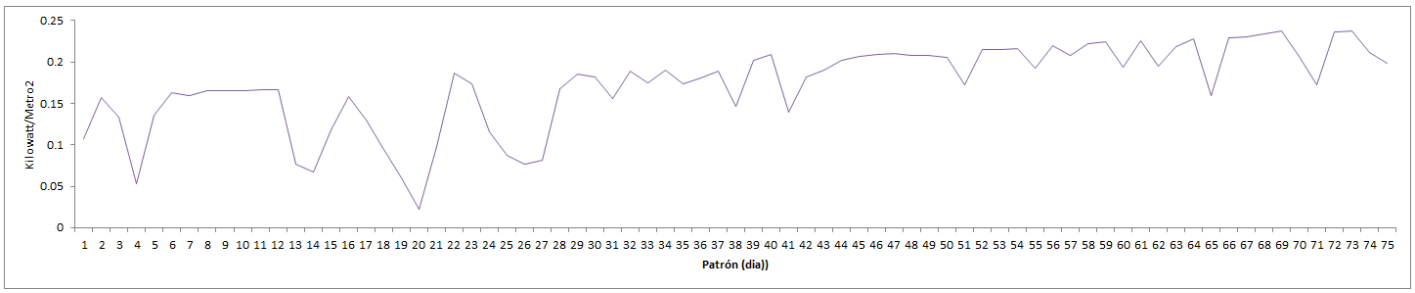
a) Ciclo 2001-2002



b) Ciclo 2002-2003



c) Ciclo 2003-2004

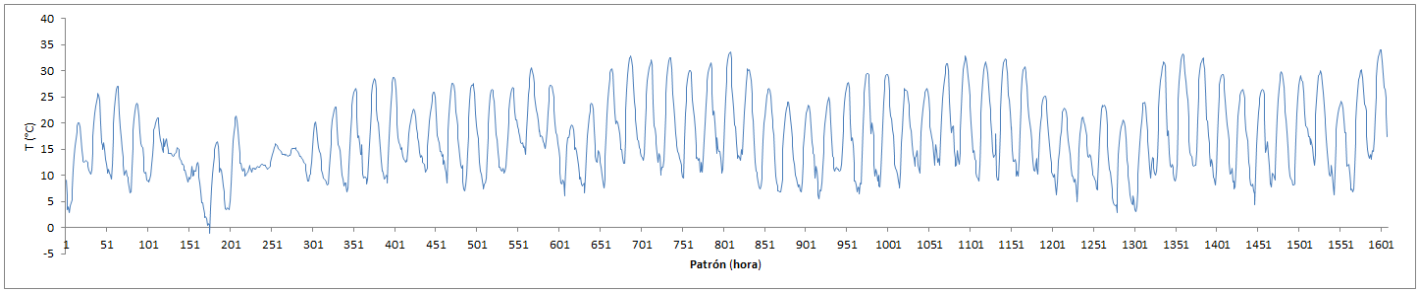


d) Ciclo 2004-2005

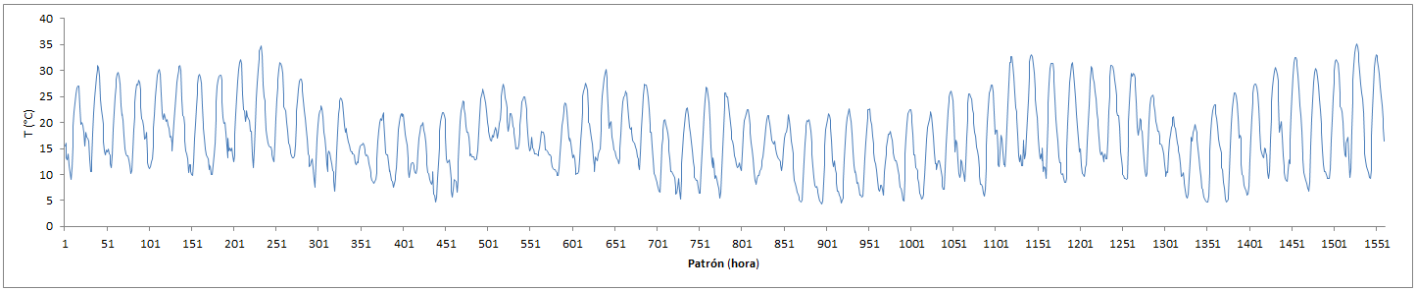
Figura B.4 *Segmentación PAA (patrón día) – Radiación Solar*

Anexo C

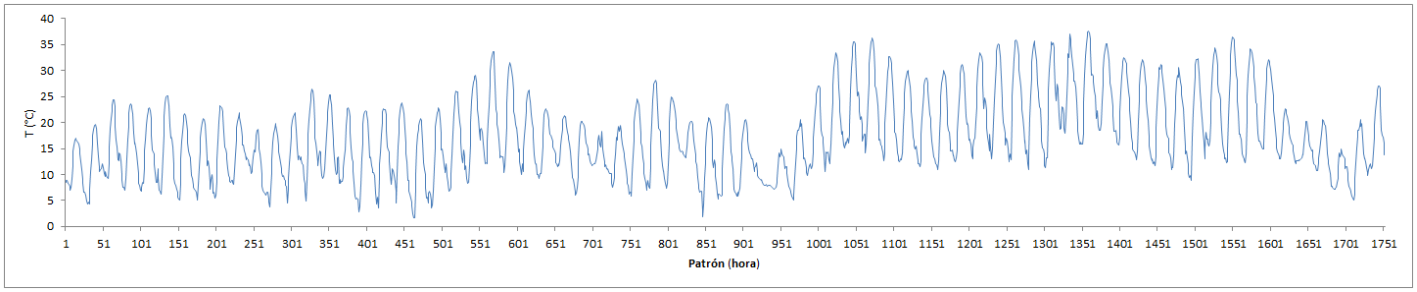
Segmentación PAA de las observaciones de EM (patrón hora).



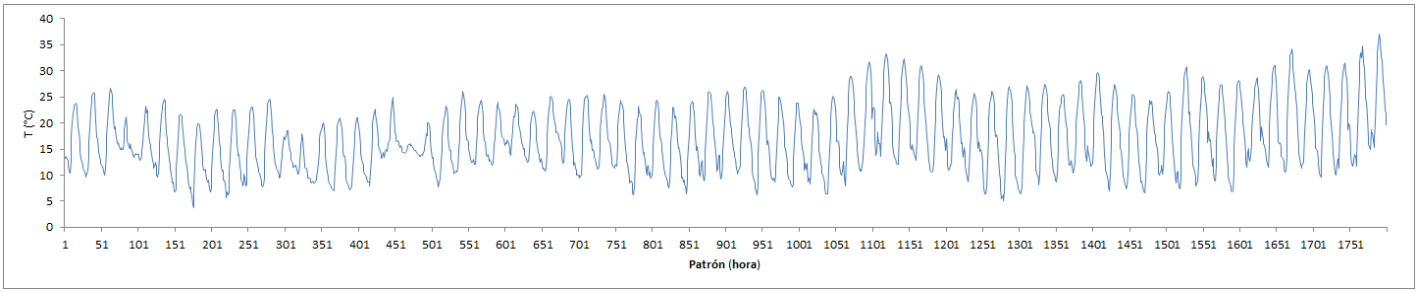
a) Ciclo 2001-2002



b) Ciclo 2002-2003

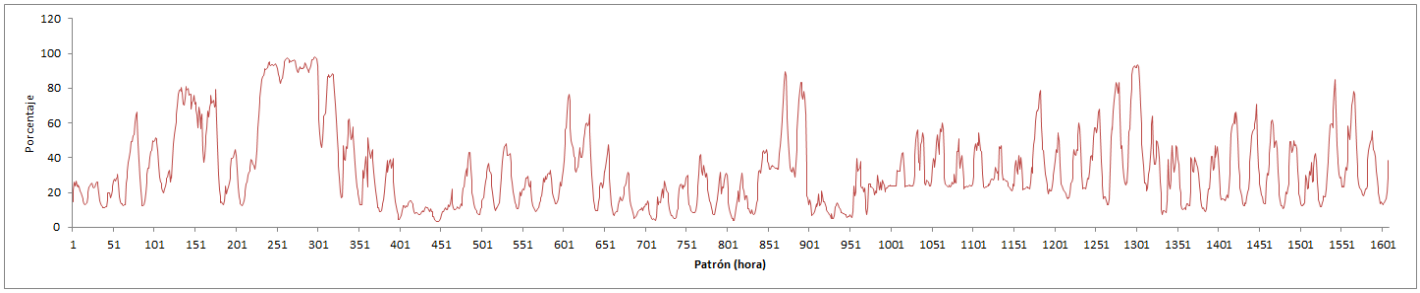


c) Ciclo 2003-2004

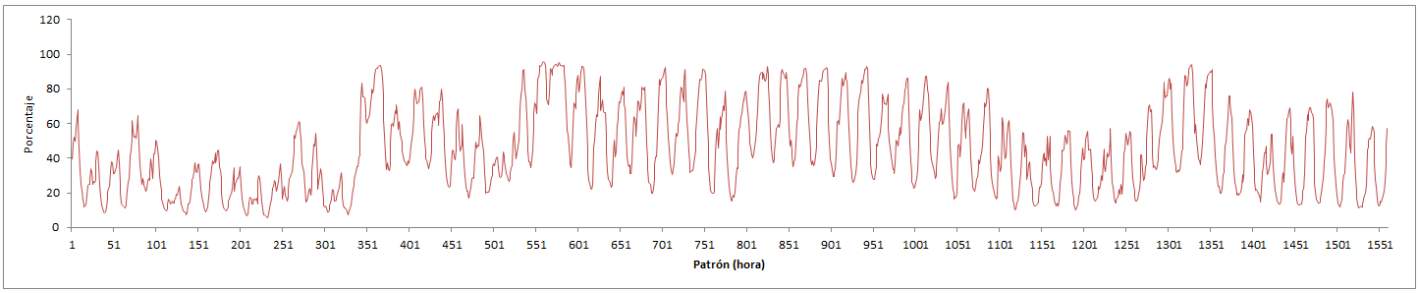


d) Ciclo 2004-2005

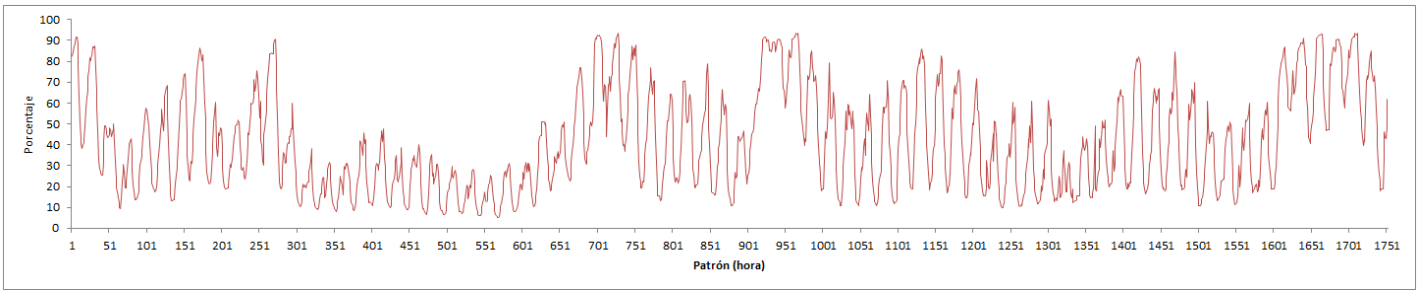
Figura C.1 *Segmentación PAA (patrón hora) – Temperatura*



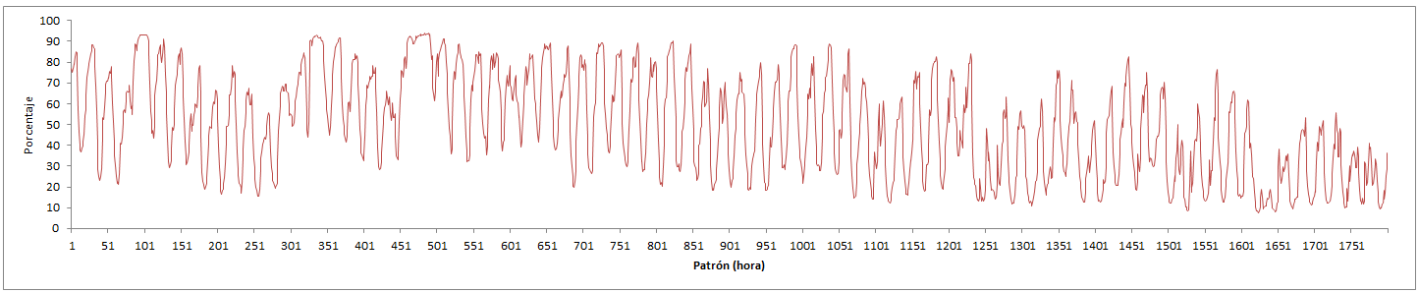
a) Ciclo 2001-2002



b) Ciclo 2002-2003

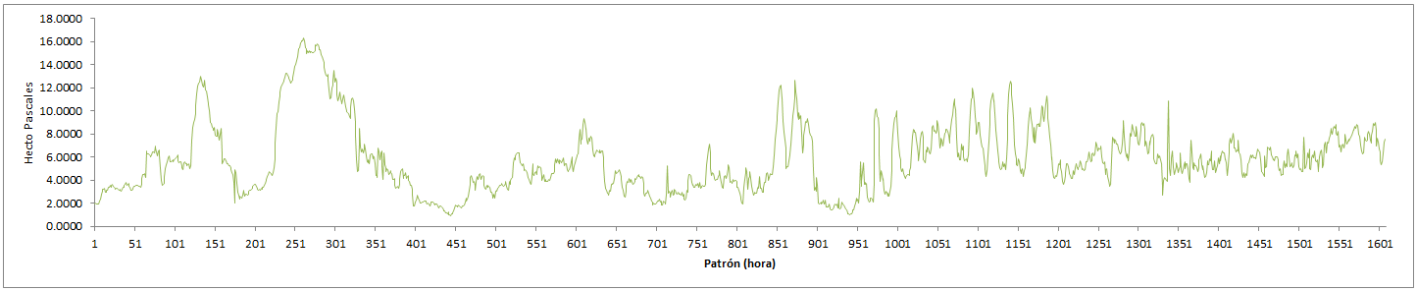


c) Ciclo 2003-2004

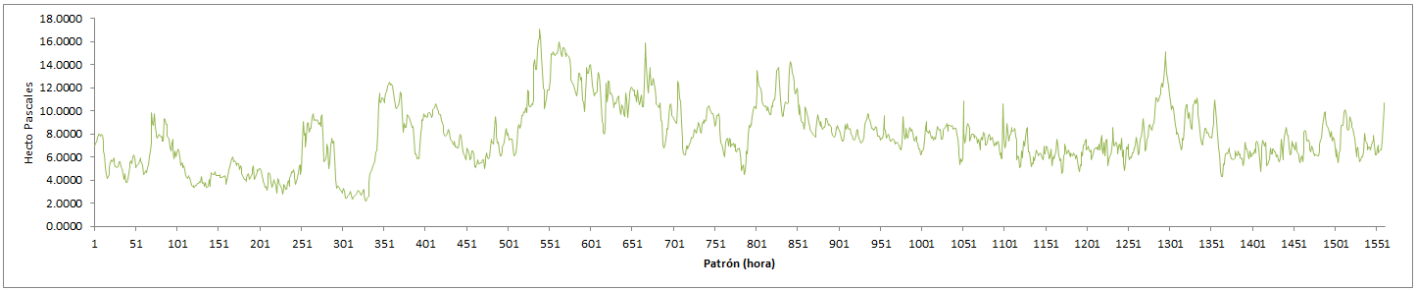


d) Ciclo 2004-2005

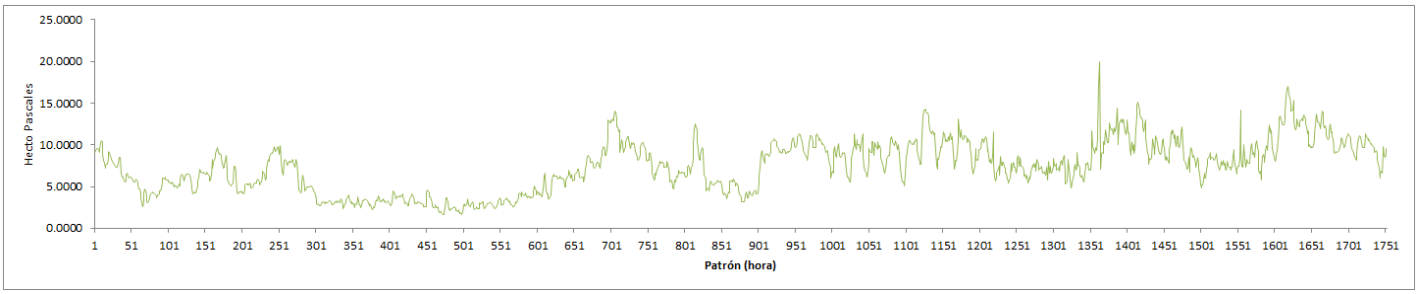
Figura C.2 *Segmentación PAA (patrón hora) – Humedad Relativa*



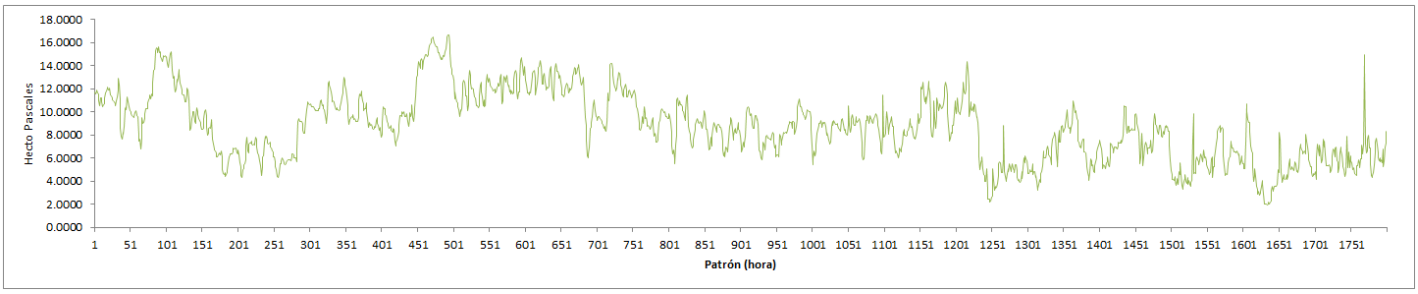
a) Ciclo 2001-2002



b) Ciclo 2002-2003

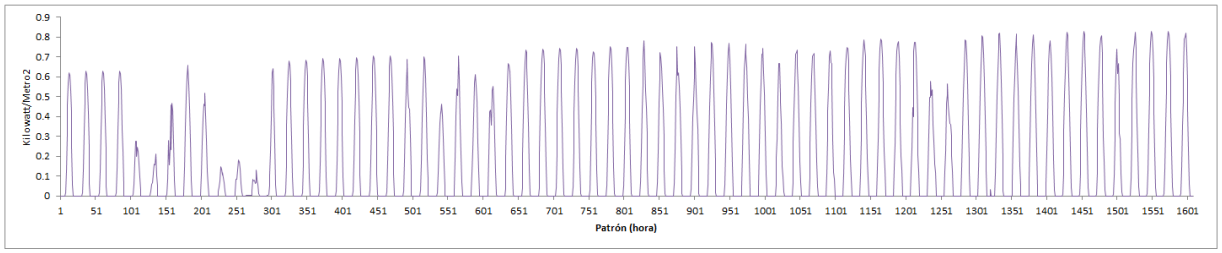


c) Ciclo 2003-2004

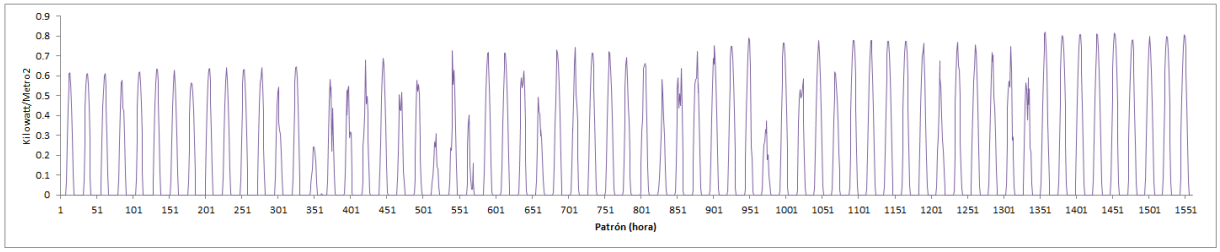


d) Ciclo 2004-2005

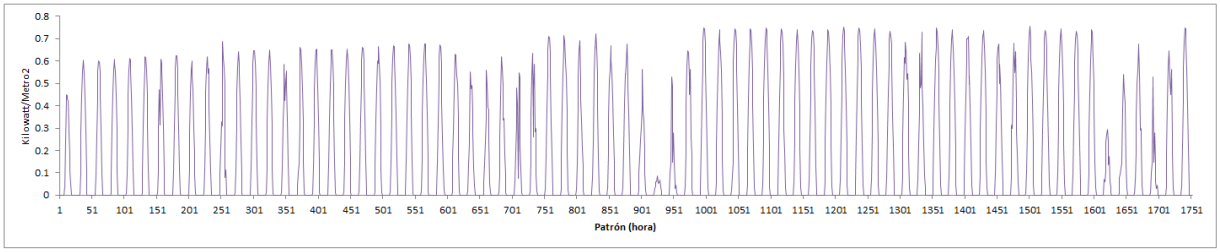
Figura C.3 *Segmentación PAA (patrón hora) – Presión de Vapor*



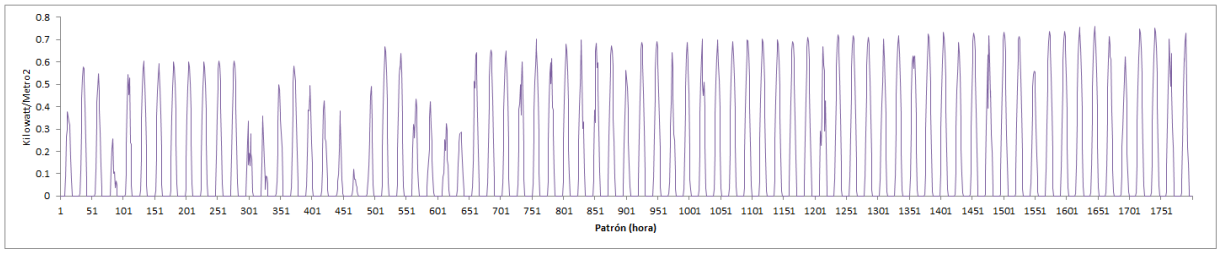
a) Ciclo 2001-2002



b) Ciclo 2002-2003



c) Ciclo 2003-2004



d) Ciclo 2004-2005

Figura C.4 *Segmentación PAA (patrón hora) – Radiación Solar*

Anexo D

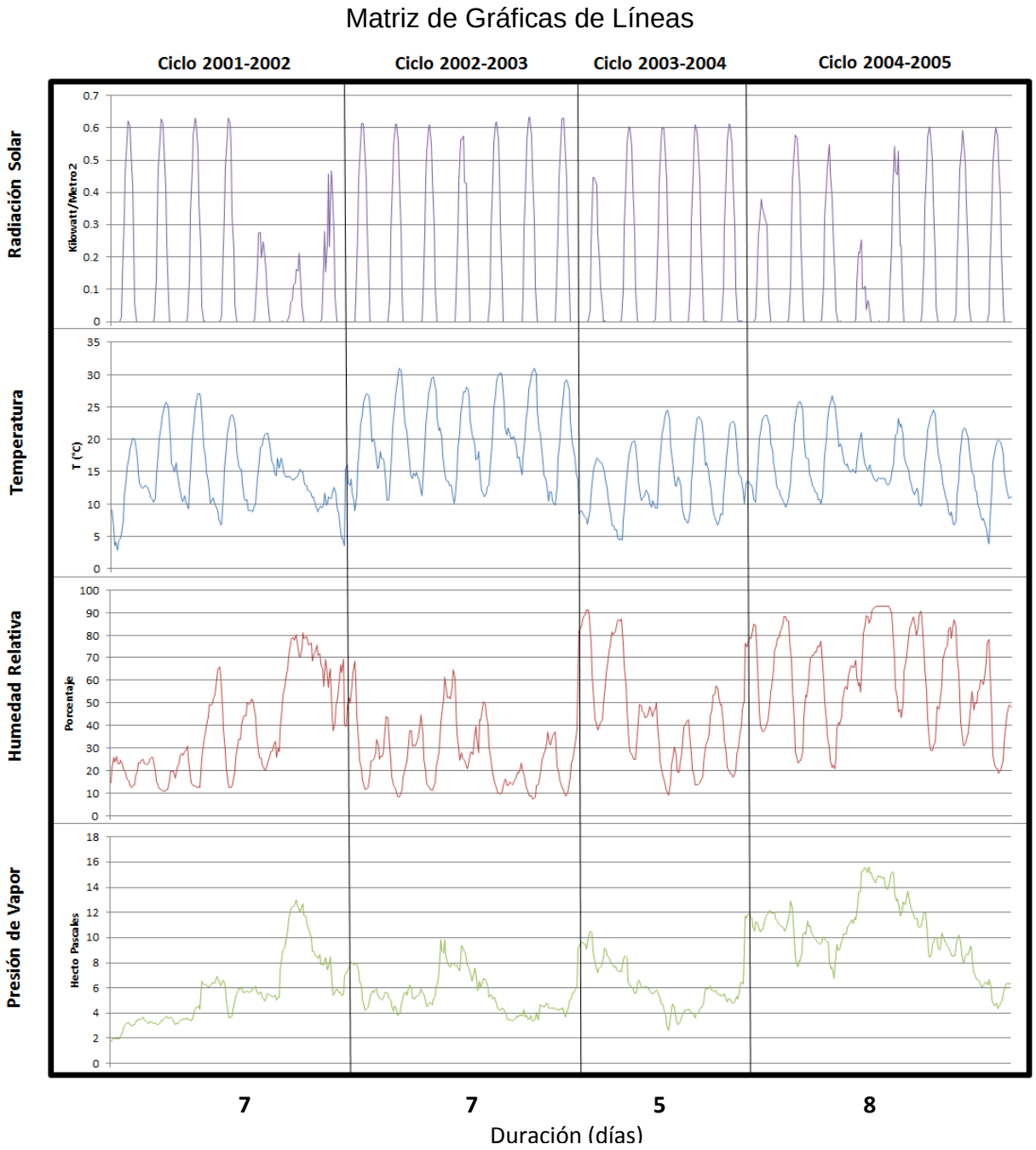


Figura D.1 *Matriz de Gráficas de Líneas – Fase Fenológica 5*

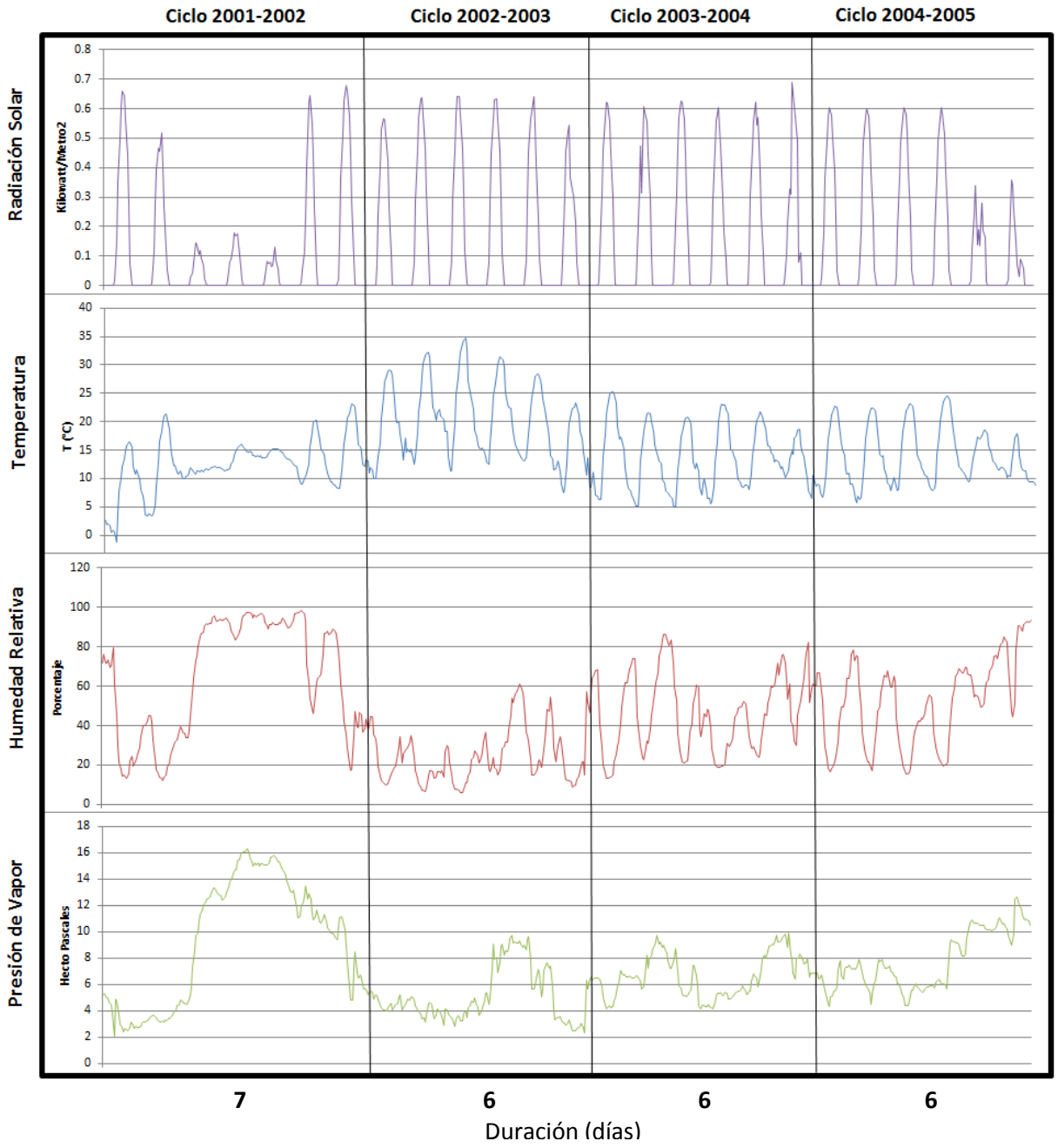


Figura D.2 *Matriz de Gráficas de Líneas – Fase Fenológica 7*

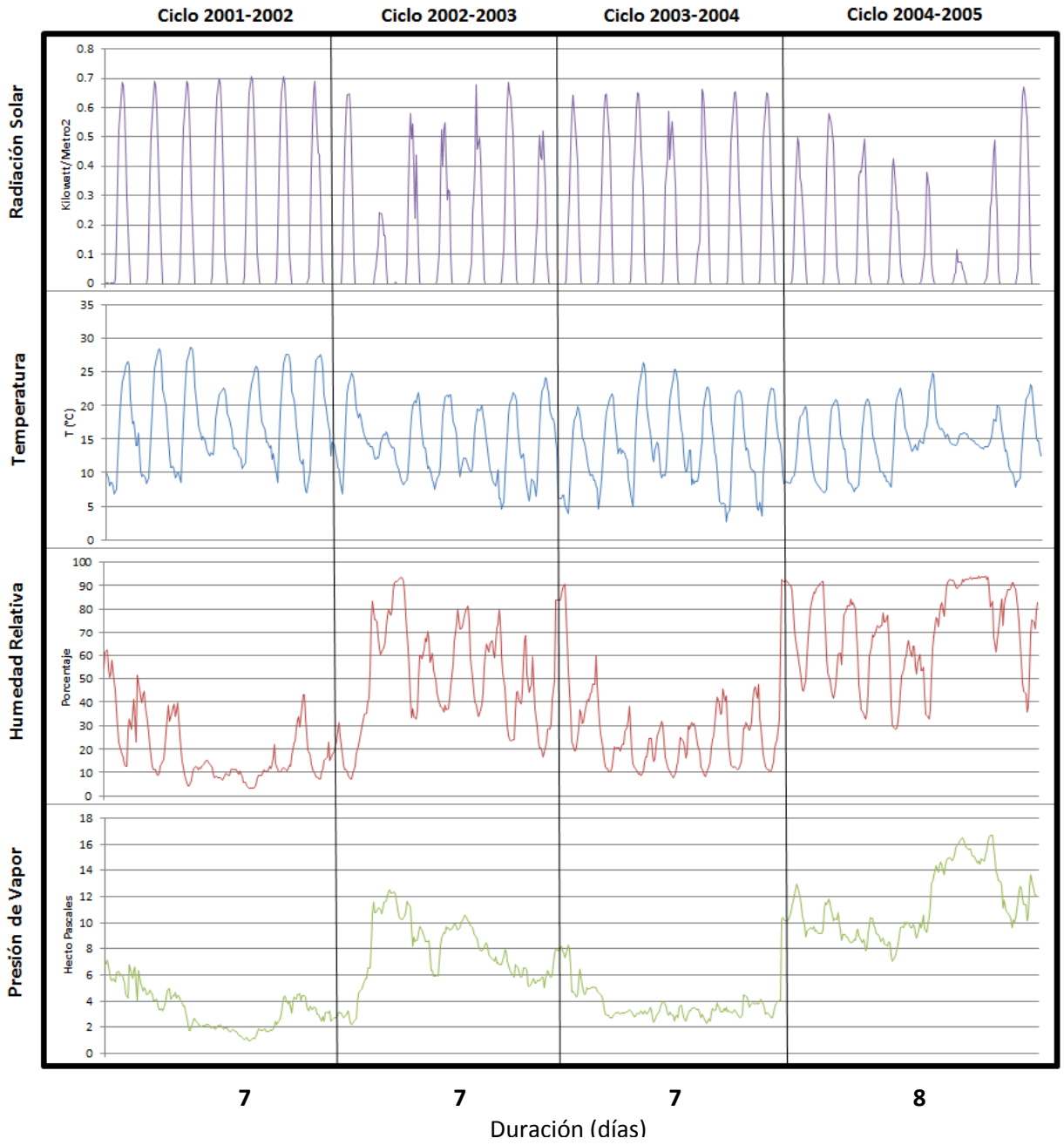


Figura D.3 *Matriz de Gráficas de Líneas – Fase Fenológica 9*

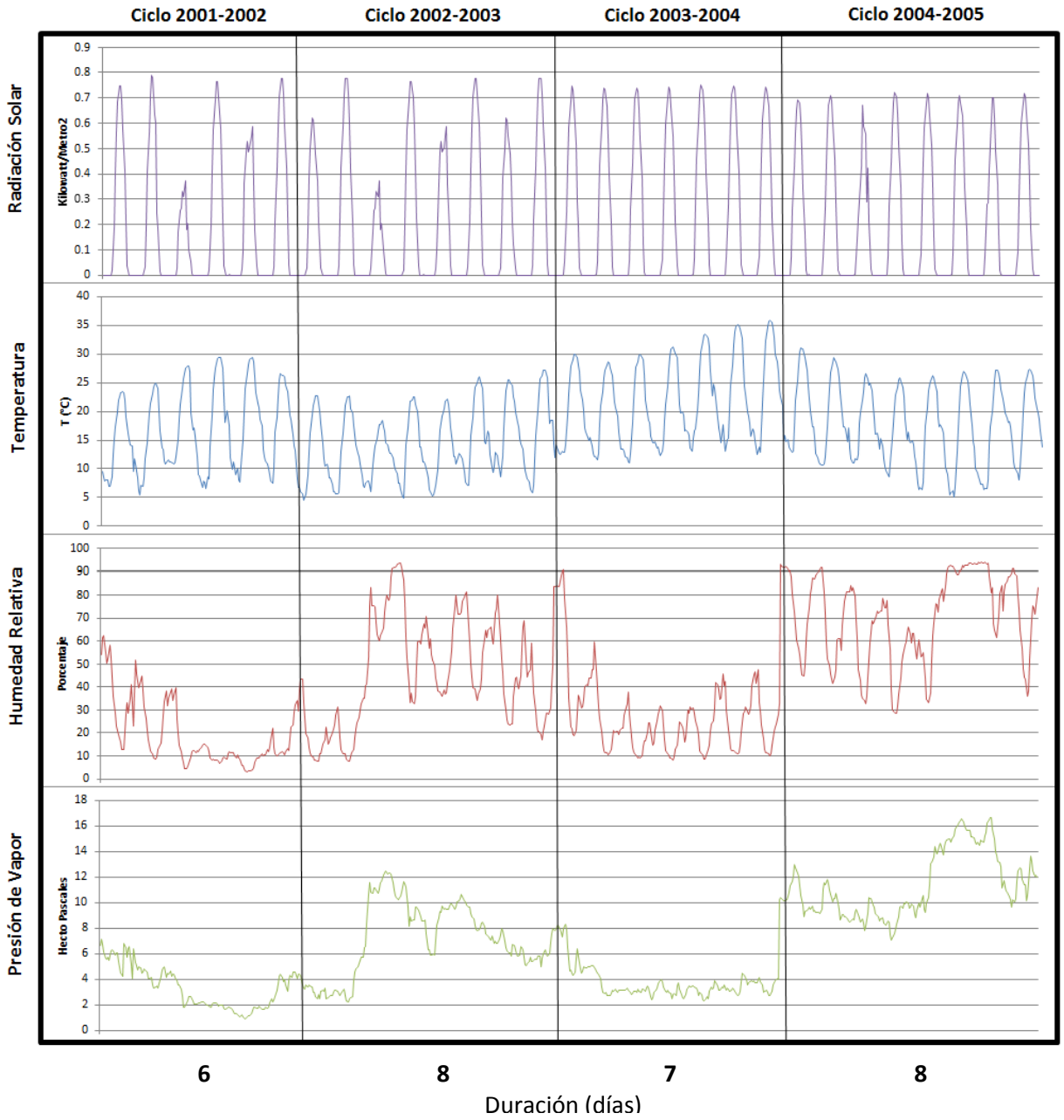
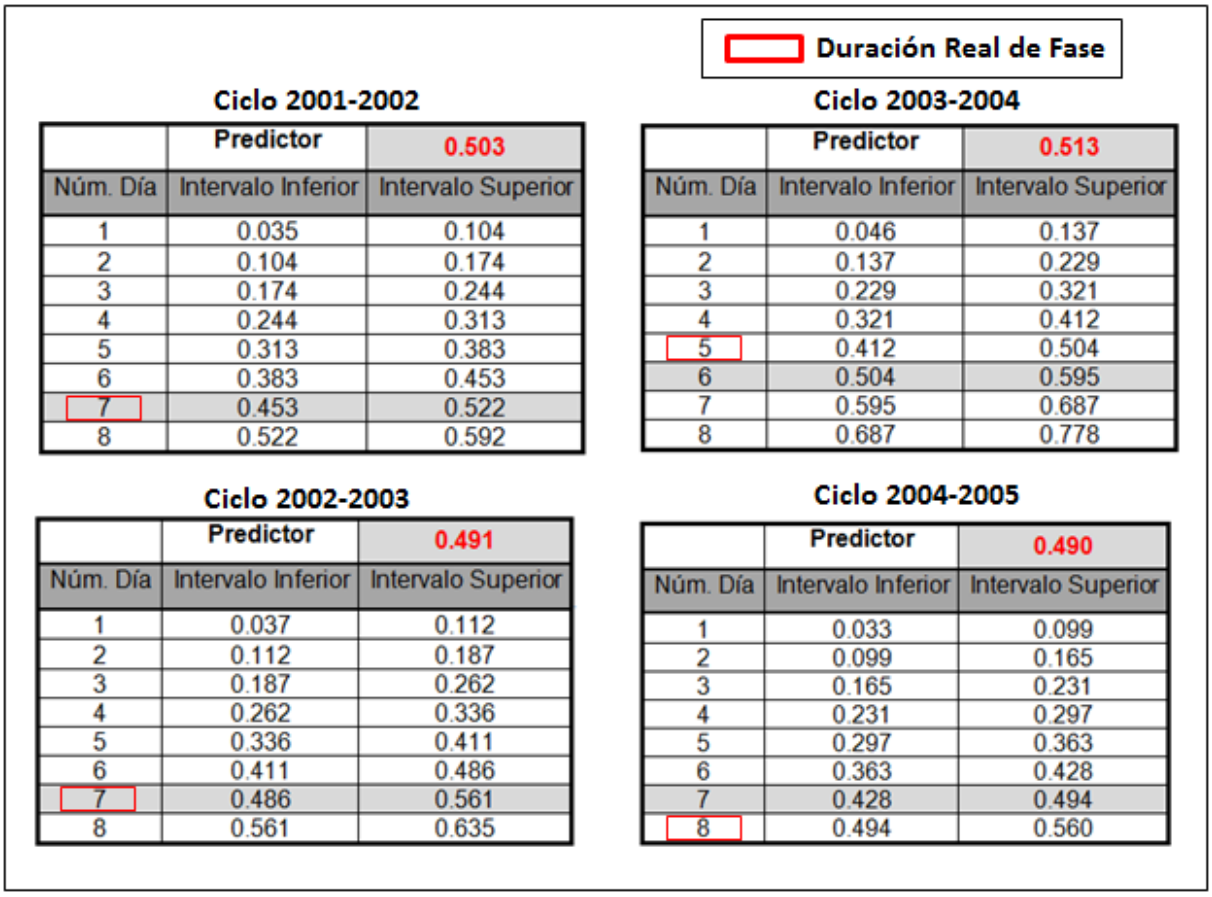


Figura D.4 *Matriz de Gráficas de Líneas – Fase Fenológica 21*

Anexo E

Tablas de Predicción - Duración de Fases Fenológicas



Duración Real de Fase

Ciclo 2001-2002

	Predictor	0.503
Núm. Día	Intervalo Inferior	Intervalo Superior
1	0.035	0.104
2	0.104	0.174
3	0.174	0.244
4	0.244	0.313
5	0.313	0.383
6	0.383	0.453
7	0.453	0.522
8	0.522	0.592

Ciclo 2003-2004

	Predictor	0.513
Núm. Día	Intervalo Inferior	Intervalo Superior
1	0.046	0.137
2	0.137	0.229
3	0.229	0.321
4	0.321	0.412
5	0.412	0.504
6	0.504	0.595
7	0.595	0.687
8	0.687	0.778

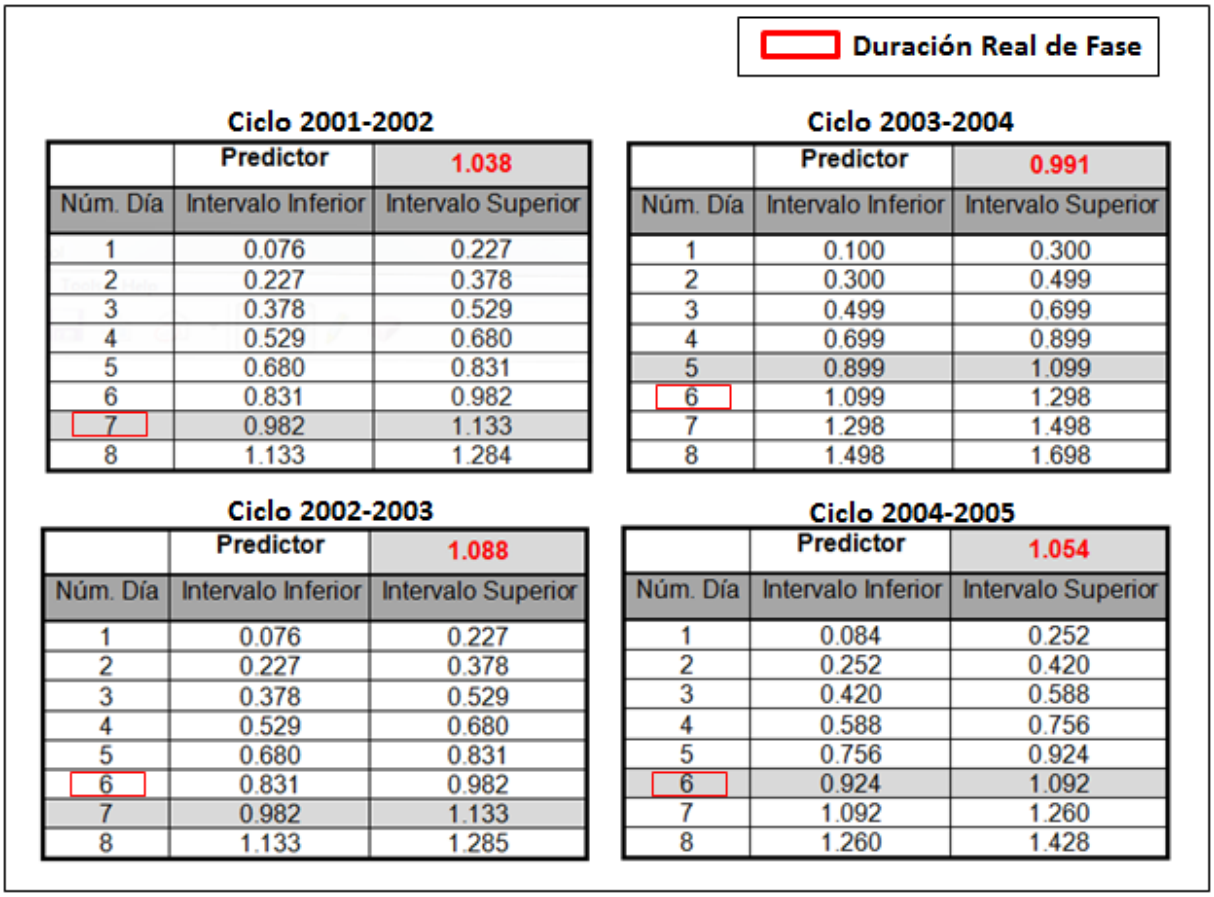
Ciclo 2002-2003

	Predictor	0.491
Núm. Día	Intervalo Inferior	Intervalo Superior
1	0.037	0.112
2	0.112	0.187
3	0.187	0.262
4	0.262	0.336
5	0.336	0.411
6	0.411	0.486
7	0.486	0.561
8	0.561	0.635

Ciclo 2004-2005

	Predictor	0.490
Núm. Día	Intervalo Inferior	Intervalo Superior
1	0.033	0.099
2	0.099	0.165
3	0.165	0.231
4	0.231	0.297
5	0.297	0.363
6	0.363	0.428
7	0.428	0.494
8	0.494	0.560

Figura E.1 *Tablas de Predicción – Fase Fenológica 5*



Ciclo 2001-2002

	Predictor	1.038
Núm. Día	Intervalo Inferior	Intervalo Superior
1	0.076	0.227
2	0.227	0.378
3	0.378	0.529
4	0.529	0.680
5	0.680	0.831
6	0.831	0.982
7	0.982	1.133
8	1.133	1.284

Ciclo 2003-2004

	Predictor	0.991
Núm. Día	Intervalo Inferior	Intervalo Superior
1	0.100	0.300
2	0.300	0.499
3	0.499	0.699
4	0.699	0.899
5	0.899	1.099
6	1.099	1.298
7	1.298	1.498
8	1.498	1.698

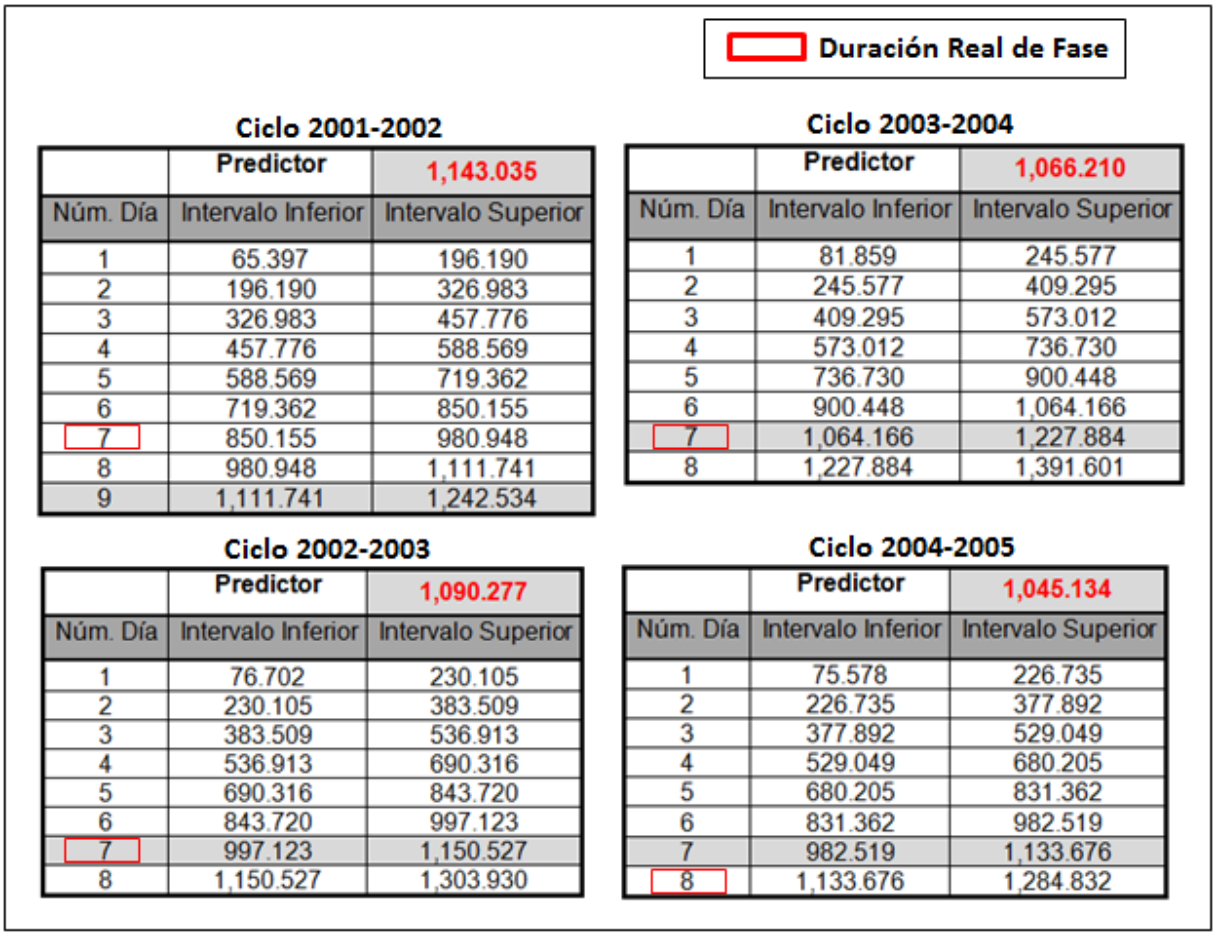
Ciclo 2002-2003

	Predictor	1.088
Núm. Día	Intervalo Inferior	Intervalo Superior
1	0.076	0.227
2	0.227	0.378
3	0.378	0.529
4	0.529	0.680
5	0.680	0.831
6	0.831	0.982
7	0.982	1.133
8	1.133	1.285

Ciclo 2004-2005

	Predictor	1.054
Núm. Día	Intervalo Inferior	Intervalo Superior
1	0.084	0.252
2	0.252	0.420
3	0.420	0.588
4	0.588	0.756
5	0.756	0.924
6	0.924	1.092
7	1.092	1.260
8	1.260	1.428

Figura E.2 *Tablas de Predicción – Fase Fenológica 7*



Ciclo 2001-2002

	Predictor	1,143.035
Núm. Día	Intervalo Inferior	Intervalo Superior
1	65.397	196.190
2	196.190	326.983
3	326.983	457.776
4	457.776	588.569
5	588.569	719.362
6	719.362	850.155
7	850.155	980.948
8	980.948	1,111.741
9	1,111.741	1,242.534

Ciclo 2003-2004

	Predictor	1,066.210
Núm. Día	Intervalo Inferior	Intervalo Superior
1	81.859	245.577
2	245.577	409.295
3	409.295	573.012
4	573.012	736.730
5	736.730	900.448
6	900.448	1,064.166
7	1,064.166	1,227.884
8	1,227.884	1,391.601

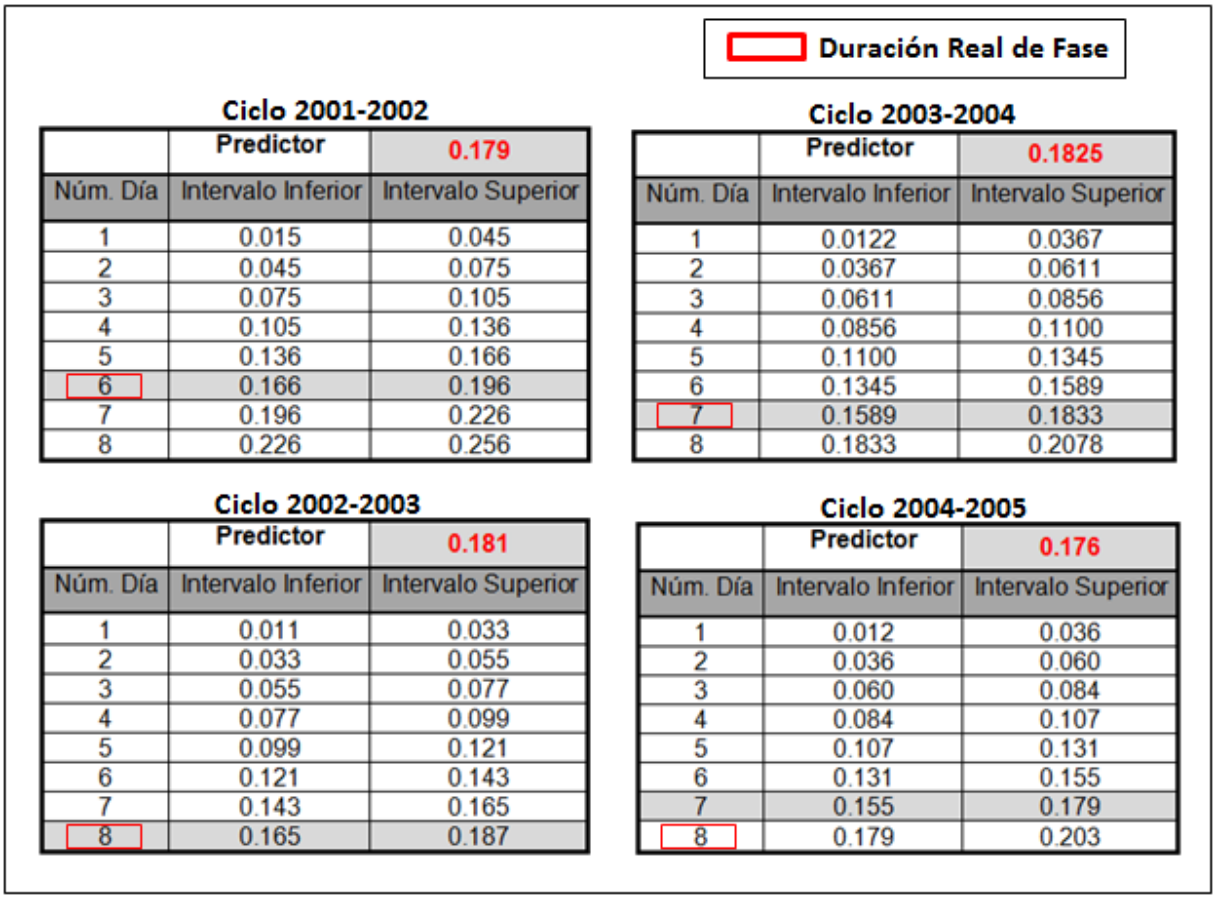
Ciclo 2002-2003

	Predictor	1,090.277
Núm. Día	Intervalo Inferior	Intervalo Superior
1	76.702	230.105
2	230.105	383.509
3	383.509	536.913
4	536.913	690.316
5	690.316	843.720
6	843.720	997.123
7	997.123	1,150.527
8	1,150.527	1,303.930

Ciclo 2004-2005

	Predictor	1,045.134
Núm. Día	Intervalo Inferior	Intervalo Superior
1	75.578	226.735
2	226.735	377.892
3	377.892	529.049
4	529.049	680.205
5	680.205	831.362
6	831.362	982.519
7	982.519	1,133.676
8	1,133.676	1,284.832

Figura E.3 *Tablas de Predicción – Fase Fenológica 9*



Ciclo 2001-2002

	Predictor	0.179
Núm. Día	Intervalo Inferior	Intervalo Superior
1	0.015	0.045
2	0.045	0.075
3	0.075	0.105
4	0.105	0.136
5	0.136	0.166
6	0.166	0.196
7	0.196	0.226
8	0.226	0.256

Ciclo 2003-2004

	Predictor	0.1825
Núm. Día	Intervalo Inferior	Intervalo Superior
1	0.0122	0.0367
2	0.0367	0.0611
3	0.0611	0.0856
4	0.0856	0.1100
5	0.1100	0.1345
6	0.1345	0.1589
7	0.1589	0.1833
8	0.1833	0.2078

Ciclo 2002-2003

	Predictor	0.181
Núm. Día	Intervalo Inferior	Intervalo Superior
1	0.011	0.033
2	0.033	0.055
3	0.055	0.077
4	0.077	0.099
5	0.099	0.121
6	0.121	0.143
7	0.143	0.165
8	0.165	0.187

Ciclo 2004-2005

	Predictor	0.176
Núm. Día	Intervalo Inferior	Intervalo Superior
1	0.012	0.036
2	0.036	0.060
3	0.060	0.084
4	0.084	0.107
5	0.107	0.131
6	0.131	0.155
7	0.155	0.179
8	0.179	0.203

Figura E.4 *Tablas de Predicción – Fase Fenológica 21*