UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

INSTITUTO DE INGENIERÍA

MAESTRÍA Y DOCTORADO EN CIENCIAS E INGENIERÍA

"Agrupamiento de patrones correlacionados y con incertidumbre: caso patrones climáticos en la producción de uva de mesa en un viñedo de Sonora"

TESIS PARA OBTENER EL GRADO DE:

DOCTOR EN INGENIERÍA

PRESENTA

JUAN MARTÍN PRECIADO RODRIGUEZ

DIRECTOR

DR. LUÍS FELIPE ROMERO DESSENS

CODIRECTOR

DRA. SARA OJEDA BENÍTEZ

Mexicali, B. C. ***Enero, 2011***

Índice

Índice de figuras

Índice de cuadros

Resumen

En la actualidad, el uso de las tecnologías computacionales se presenta de forma masiva en la mayor parte del quehacer de las organizaciones y los individuos. Cada vez más se generan y almacenan grandes cantidades de datos que representan diferentes característica de objetos concretos o abstractos, mismos que pueden ser utilizados para explorar y comprender los fenómenos de forma más aproximada a como se presentan en la vida real. No obstante del desarrollo de la computación aún existe un desfase entre la capacidad computacional para el manejo de grandes bases de datos y los métodos y técnicas para la extracción de información y conocimiento que permita una mejor comprensión de la problemática estudiada. Es a partir de este desfase que surge el Proceso de Descubrimiento de Conocimiento en Bases de Datos, con el objeto de extraer información y conocimiento no trivial, tanto de datos que se almacenaron con tal propósito, así como de aquellos que solo fueron utilizados de forma inmediata. Se aplicaron algoritmos de agrupamiento deterministas y posibilistas, para formar grupos con características homogéneas, a un conjunto de variables climáticas almacenadas en el Sistema de Información Agroclimática en Sonora, así como a datos fenológicos del desarrollo de la uva de mesa en un viñedo de Sonora. La investigación mostró que los algoritmos posibilistas tienen un mejor desempeño al momento de agrupar datos altamente correlacionados; a partir de tal agrupamiento fue posible identificar, además, características climáticas que tienen un efecto acelerador en el desarrollo vegetativo, así como otras que retardan su tiempo de duración.

Introducción

"Las computadoras nos prometieron una fuente de sabiduría y a cambio nos dejaron un océano de datos"

-Un ejecutivo MIS frustrado (Frawley et al., 1992).

El fácil acceso a los avances en el desarrollo tecnológico de la electrónica, los sistemas de comunicación y la computación, ha cubierto de forma gradual muchas de las operaciones de las organizaciones, a través de sistemas computacionales integrados, ya que éstos ofrecen la automatización completa de algunos procesos. La propagación tecnológica ha provocado la proliferación de los sistemas de procesamiento transaccionales en la mayor parte del las actividades del quehacer humano, recogiendo datos de todo aquello que pueda ser registrado. Los datos son generados automáticamente a través de una diversidad de fuentes como cajas registradoras, conexiones telefónicas, servidores Web, lectores de código de barras en almacenes o puntos de venta, escaneo de texturas e imágenes, sensores remotos, entre otras.

Cada vez se generan bases de datos más grandes ya sea por la cantidad de observaciones almacenadas y/o la cantidad de variables (atributos, mediciones) de cada observación. Bajo el principio de que lo que no es medible no es mejorable, junto con la creencia de la recolección de datos como fuente de información valiosa que proporcionará una mayor comprensión de la realidad o una ventaja competitiva, las organizaciones almacenan grandes cantidades de datos, que tienen aplicación a corto plazo y pocas veces son reutilizados para extraer información y conocimiento nuevo. A pesar de estos avances, aún se observa un vacío entre la generación de datos y las herramientas analíticas disponibles que son utilizadas en el proceso de obtención de conocimiento a partir de éstos.

En la actualidad es común encontrar empresas ricas en datos en lugar de ser ricas en información y conocimiento, debido a la gran cantidad de datos que aún siguen en espera de ser transformados. El proceso de descubrimiento de conocimiento en bases de datos, surge en respuesta a la necesidad de cubrir este vacío, ofreciendo a través de la minería de datos una serie de métodos y herramientas para el análisis de datos históricos almacenados en grandes bases de datos locales y/o distribuidas, los cuales provienen de diferentes campos de conocimiento como la estadística, inteligencia artificial, visualización, bases de datos y computación de alto desempeño.

La extracción de información y conocimiento a partir de los almacenes de datos tiene el propósito de dar soporte al proceso de toma de decisiones dentro de la organización. Tal proceso está estructurado de acuerdo al modelo lógico de la organización, por lo que la integración de la infraestructura productiva incorporada en sus procesos con las tecnologías de información y comunicación, es esencial para asegurar la calidad y el impacto de las decisiones tomadas.

La información que sustenta el proceso de toma de decisiones es generada a través del análisis de datos, ésta puede ser generada ya sea a partir de métodos y técnicas deterministas o a través de herramientas analíticas no deterministas. Cada uno de estos abordajes presenta sus ventajas y desventajas, sin embargo; de acuerdo con la naturaleza y el comportamiento de los datos crudos que serán utilizados en la generación de información, el analista deberá tener en cuenta, como uno de los criterios principales, el manejo de la incertidumbre que cada uno de estos dos enfoques ofrece, ya que es esta característica la que permitirá un mejor acercamiento a la realidad del fenómeno estudiado (Klir y Yuan, 1995).

El concepto de incertidumbre (o entropía) está íntimamente relacionado con las decisiones y la información. En la toma de decisiones diarias de cualquier individuo ordinario la incertidumbre está presente, en el futuro siempre se está incierto en cualquier decisión tomada. La selección de una acción en particular, de

entre un conjunto de acciones concebidas, se hace con base en la anticipación de las consecuencias de las acciones individuales. Sin embargo, la incertidumbre no sólo está limitada al futuro; puede también competer al pasado debido a la falta de un registro completo y consistente de los hechos ocurridos, e igualmente al presente al no contar con información relevante. De manera que cualquier fenómeno o proceso del mundo real es inimaginable sin incertidumbre.

Tradicionalmente, la incertidumbre intrínseca en un conjunto de patrones ha sido manejada a través de dos teorías: la probabilista y la posibilista. Ambas emergen de marcos teóricos diferentes; mientras que el análisis probabilista sienta sus bases en la teoría de conjuntos clásica, donde todo es falso o verdadero -tema que no es objeto de este trabajo-, el análisis posibilista se sustenta en la teoría de conjuntos difusos, donde la interpretación no es cuestión de falso o verdadero sino del grado de pertenencia del elemento al conjunto. El grado de pertenencia que puede ser definido matemáticamente para asignar a cada individuo a su universo de discurso (Vasantha et al., 2007)

El análisis de datos se desarrolló con base en el proceso de descubrimiento en bases de datos, debido a que el estudió es del corte histórico-descriptivo y comparativo, dado que los datos son eventos ya ocurridos, los cuales están almacenados en un repositorio que incluye datos de estaciones agroclimáticas distribuidas en las diferentes regiones agrícolas de todo el estado de Sonora.

Este trabajo se centra en el análisis de la capacidad que ofrecen los algoritmos de *clustering* deterministas y posibilistas en el reconocimiento de patrones, para expresar la estructura intrínseca de un conjunto de patrones altamente correlacionados y con presencia de comportamientos atípicos y extremos.

Objetivo:

Así, el objetivo de la investigación es evaluar el desempeño del reconocimiento de patrones climáticos a partir de algoritmos deterministas y difusos, así como la relación del comportamiento de los patrones climáticos y el desarrollo vegetativo en la producción de uva de mesa, medido a través de las etapas fenológicas de brotación y floración.

Los agrupamientos de patrones climáticos obtenidos a partir de la aplicación de los algoritmos K-Medias, Fuzzy C-Medias y Gustafson-Kessel son contrastados con el propósito evaluar su capacidad de expresar la estructura del conjunto de datos p-dimensionales, así mismo, se relacionan los agrupamientos con el desarrollo vegetativo de uva de mesa manifestado a través de las etapas fenológicas.

En el primer capitulo se describe los orígenes del proceso de descubrimiento de conocimiento en bases datos, el objetivo que persigue, las etapas que integran este proceso y el modelo general de su aplicación. En este mismo sentido, se presentan también los campos de conocimiento en los que se apoya, tanto aquellos que proveen las herramientas analíticas para el procesamiento de datos así como los campos que dan soporte a la infraestructura utilizada en administración de los datos y sus fuentes. El agrupamiento de datos p-dimensionales es presentado en el capítulo II, donde se aborda, el propósito de aplicación de estas técnicas y los principales criterios en que se basa la asignación de cada patrón (vector p-dimensional); así mismo, se exponen los diferentes enfoques del agrupamiento de datos: deterministas y posibilistas, se enuncian algunos campos de aplicación de este tipo de análisis.

La metodología utilizada en la investigación se expone en el capítulo III, donde se presentan los materiales y métodos utilizados para alcanzar el objetivo planteado. Así se presenta el ámbito del conjunto de patrones agroclimáticos, las fuentes de

datos y la estructura de los mismos, el proceso de acopio de datos e información, así como las herramientas analíticas que fueron aplicadas en la transformación de datos. El análisis de resultados y los principales hallazgos son presentados en el capítulo IV. Finalmente las conclusiones a las que condujo la investigación son presentadas en capítulo V, así mismo se enuncian las limitaciones de este trabajo, así como, las líneas para futuras investigaciones.

Capitulo I. Descubrimiento de conocimiento en base de datos

El desarrollo de tecnologías para el manejo y análisis de datos durante la segunda parte de la centuria pasada, muestra un vacío entre la generación de datos y las herramientas de análisis de datos disponibles utilizados en el proceso de obtención de conocimiento a partir de éstos. Cada vez se generan bases de datos más grandes ya sea por la cantidad de observaciones almacenadas y/o la cantidad de variables (atributos, mediciones) de cada observación, de manera que se dificulta cada vez más el análisis de datos a través de los métodos estadísticos tradicionales o a través de un procesamiento de datos manual desarrollado por el humano. En este capitulo se aborda el proceso de descubrimiento de conocimiento en bases de datos (KDD, por sus siglas en inglés, *Knowledge Discovery in Databases*) como la tecnología que permite llenar el vacío entre la generación y análisis de datos; se analiza su propósito, las etapas que componen este proceso, así como los paradigmas y campos de conocimiento que le dan soporte.

I.1. Proceso KDD

A pesar de los avances derivados de la convergencia de las tecnologías computacionales, tanto en hardware como software, aún permanecen muchos datos en espera de ser transformados en información y por consiguiente en conocimiento, tal como plantean Tang y McLennan (2005). En la sociedad actual donde la información es el nuevo motor de desarrollo, existe mucha información cruda en forma de datos (Witten y Frank 2005); estas colecciones de datos propician que las organizaciones se conviertan en ricas en datos y pobres en conocimiento.

Muchos de estos datos son generados por los sistemas de transacción operacional, los cuales están orientados a controlar y hacer más eficiente el proceso de operaciones del negocio. La mayor parte de estos sistemas

operacionales que generan grandes volúmenes de datos, son diseñados y puestos en marcha sin tener en cuenta algún conjunto de herramientas de análisis de datos (SAS Institute Inc., 2002, Berry y Linoff, 2005). A lo anterior se suman tres factores: primero, la baja efectividad de las mezclas ad hoc de las técnicas estadísticas y las herramientas de administración de datos para analizar grandes volúmenes de información (Mitra y Achayra, 2003); segundo, la carencia de conocimiento acerca de las herramientas estadísticas y el potencial para generar información (Giudici, 2003); y tercero, la falta de analistas entrenados con habilidades para transformar datos en conocimiento (Larose, 2005).

Han y Kamber (2002) plantean la necesidad de generar nuevas técnicas y herramientas que puedan asistir inteligentemente la transformación de datos en información y conocimiento útil. Así, las investigaciones metodológicas, particularmente en el campo de la computación y estadística, han permitido desarrollar procedimientos flexibles y escalables que pueden ser usados en el análisis de grandes volúmenes de datos almacenados (Giudici 2003). Estos requerimientos y esfuerzos han llevado al desarrollo del proceso KDD.

Frawley y otros (1992), definen KDD como el proceso de extracción no trivial de información implícita en los datos, no conocida previamente y potencialmente útil. Fayyad y otros (1996) coinciden; sin embargo, hablan de patrones en lugar de información y advierten que estos patrones deben ser entendibles, en lo cual también coinciden Pazzani y otros (1997). KDD es un proceso compuesto por varias etapas que van desde la recolección de datos hasta la generación de conocimiento; es un concepto acuñado a finales de de la década de los ochenta, que surge de varios campos del conocimiento (Piatetski-Shapiro, 1989). Su naturaleza interdisciplinaria hace converger a campos de investigación como aprendizaje automático, reconocimiento de patrones, bases de datos, métodos estadísticos, teoría de la información, inteligencia artificial (AI, por sus siglas en inglés), adquisición de conocimiento por sistemas expertos, redes neuronales,

visualización de datos y computación de alto desempeño (Fayyad, et al. 1996; Pazzani, et al. 1997).

El descubrimiento de conocimiento es un proceso interactivo e iterativo (Fayyad, et al. 1996). Brachman y Anand, (1996), enfatizan en la necesidad del conocimiento previo por parte del usuario, quien deberá dedicar el tiempo suficiente para desarrollar la compleja tarea interactiva entre él y la base de datos que guarda los hechos respecto al dominio de la problemática analizada; además, el usuario (analista) posiblemente se apoye con un grupo heterogéneo de herramientas tecnológicas. En este sentido, la expresión del conocimiento generado por el KDD, estará en función del usuario de este proceso y de los objetivos propuestos a partir del problema abordado.

El KDD muestra cuatro características básicas: el conocimiento descubierto debe ser representado en un lenguaje de alto nivel; éste no necesariamente deberá ser utilizado por el humano, sin embargo, su expresión deberá ser comprensible para él. Los descubrimientos retratan exactamente el contenido de la base de datos; el grado de exactitud de imperfecciones es expresado en medidas de certeza; el conocimiento descubierto resultante del proceso puede ser interesante según las directrices definidas por el usuario. Esto implica que los patrones son novedosos y potencialmente útiles, además de que el proceso para descubrirlos no es trivial; finalmente, el proceso de descubrimiento será eficiente, si el tiempo de ejecución para grandes bases de datos es predecible aceptable (Brachman y Anand, 1996,Frawley, et al., 1992).

De acuerdo con Frawley; et al., (1992), hay tres aspectos que se deben tomar en cuenta en el descubrimiento de conocimiento:

- Su forma; de acuerdo a su capacidad descriptiva, pueden relacionar campos numéricos a través de una ecuación matemática. Otro caso es el descubrimiento que encuentra relaciones lógicas entre campos cualitativos

y, más aún, puede ser una combinación de ambos. Sin embargo, el descubrimiento pude tomar formas más complejas, como el descubrimiento de relaciones entre reglas simples que puede llevar a modelos semánticos o dominios teóricos;

- Su representación; que estará en función del tipo de usuario final de los descubrimientos y deberá tener la forma apropiada, ya sea a través de lenguaje natural, lógica formal y modelos visuales de información. Cada una de estas formas de representar el conocimiento descubierto ofrece diferentes ventajas y limitaciones; así, el lenguaje natural y la información visual son generalmente más favorables a partir de una perspectiva humana, pero es inconveniente para la manipulación algorítmica.

- El nivel de certeza, como resultado de los valores perdidos, erróneos o el indeterminismo inherente a las causas subyacentes del mundo real. La incertidumbre hace que los patrones de datos sean más probabilistas que deterministas. La aplicación de protocolos rígidos de estandarización a los datos de entrada permite minimizar la incertidumbre.

De acuerdo con Frawley y otros (1992), rara vez una pieza de conocimiento descubierto es verdadera para todos los datos. Es en este punto donde el analista debe decidir la forma de tratar la incertidumbre que no puede ser evitada cuando trabaja con conjuntos de datos que representan procesos de la vida real (Höppner et al., 1999). De tal forma que el analista deber elegir entre cuatro alternativas: i) eliminar los valores que producen el problema, que es el caso de la eliminación de las colas; esta estrategia produce resultados estables, pero generalmente alejados de la realidad, ii) ignorar la presencia de las características presentes en los datos que provocan tal incertidumbre, lo que redundará en información incierta, iii) identificar el grupo de datos que genera la incertidumbre y analizarlos por separado, lo que sólo producirá información contextual del proceso estudiado, y iv) utilizar herramientas que le permitan trabajar con tal complejidad, como son las

técnicas estocásticas, difusas y los modelos de creencia (Amit, 2000); éstas demandan mayor conocimiento y habilidades del analista, pero permiten un mayor acercamiento a la realidad.

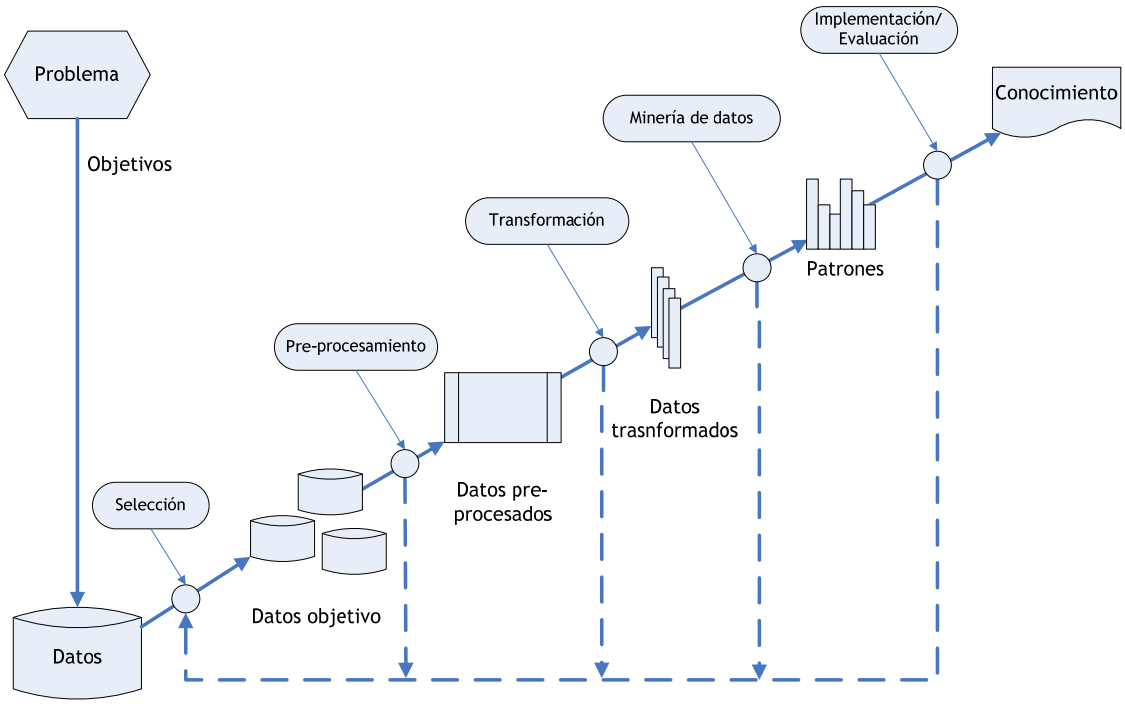
I.2. Etapas del proceso KDD

El proceso KDD ha sido descrito de diferentes formas, Fayyad y otros (1996), lo definen a través de cinco etapas: selección, procesamiento preliminar, transformación, minería de datos (DM, por sus siglas en inglés) e interpretación y evaluación (figura 1.1). Williams y Huang (1996), definen este mismo proceso a través de cuatro etapas: selección, procesamiento previo, minería y evaluación. Ambas concepciones del proceso incluyen los mismas nueve etapas propuestas originalmente por Fayyad y otros, (1996), el cual será utilizado como eje rector para el desarrollo del presente trabajo.

La serie de etapas que permiten poner en marcha el proceso KDD, son descritas a continuación. Conviene destacar que la propiedad interactiva representada por la retroalimentación de cada etapa, puede implicar el regreso a cualquiera de las etapas anteriores.

- **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.

Figura 1.1 Etapas del proceso KDD.



Fuente: Fayyad et al. (1996).

- **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, decidir qué estrategias utilizar para el manejo de datos perdidos y medir el tiempo de secuenciación de información y los cambios conocidos.

- **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.

- **Identificar un método o técnica analítica en particular, de la minería de datos, de acuerdo con el objetivo de la aplicación del proceso KDD**.

- **Exploración de análisis, modelo y selección de hipótesis.** Selección de algoritmo(s) y selección de método(s) de minería de datos que se utilizará 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

I.3. Minería de datos

Considerada como el corazón del proceso KDD, la MD consiste en la aplicación del análisis de datos y algoritmos de descubrimiento que producen un conjunto particular de patrones (o modelos) ocultos en los datos (Fayyad et al., 1996). Bigus (1996), refiere a la DM como el descubrimiento de información valiosa, no obvia a grandes volúmenes de datos. Mientras que para Hand (1998), es el proceso de análisis secundario de grandes bases de datos dirigido a encontrar relaciones inesperadas las cuales son de interés para el propietario de la base de datos. Otros autores como Tsiptsis y Chorianopoulus (2009), enuncian que ésta se orienta a la extracción de conocimiento y sabiduría a través del análisis de grandes cantidades de datos usando sofisticadas herramientas de modelación.

La DM como campo de conocimiento, se desarrolló durante la última década del siglo pasado y representa la confluencia de algunos campos bien establecidos como el análisis estadístico tradicional, la inteligencia artificial y el desarrollo de las grandes bases de datos (Bigus 1996; Fayyad et al., 1996; Hand, 1998; Nisbet et al., 2009)

Los modelos de DM son un conjunto de reglas, ecuaciones o funciones de transferencia complejas que se utilizan para identificar patrones útiles apenas visibles en los datos históricos, con el objeto de comprender o predecir comportamientos. De acuerdo con la naturaleza de la estructura del conjunto de datos y el propósito del uso de la DM, los algoritmos dentro de este ámbito pueden ser agrupados en dos clases (Fayyad et al., 1996; Bigus 1996; Hand, 1998; Mitra y Acharya, 2003; Tsiptsis y Chorianopoulus, 2009; Nisbet et al., 2009): Modelos supervisados y modelos no supervisados.

I.3.1. Modelos supervisados o predictivos

El objetivo de estos modelos es predecir o estimar los valores de un atributo numérico continuo. De acuerdo con su estructura, estos modelos cuentan tanto con campos o atributos de entrada como con un atributo o campo de salida. En este sentido, los campos de entrada o predictores son utilizados por el modelo para identificar la función de predicción del campo de salida. Esta función, que mapea las entradas y salidas, es generada por el modelo. A partir del campo de salida se supervisa el efecto que tienen los predictores sobre éste. Estos modelos se dividen en dos grupos: de clasificación y de estimación.

En los modelos de clasificación o propensión, los grupos o clases objetivo son conocidos con anterioridad; éstos tienen el objetivo de clasificar los casos dentro de tales clases. Así mismo, también calculan el puntaje de propensión que marca la posibilidad de ocurrencia de un grupo objetivo o evento.

Por otra parte, los modelos de estimación son similares a los modelos de clasificación pero su principal diferencia es que son usados para predecir un valor de un campo continuo basado en los valores observados de los atributos de entrada.

I.3.2. Modelos no supervisados

El objetivo de los modelos no supervisados o indirectos, es descubrir patrones de comportamiento en el conjunto de datos de entrada. Éstos no tienen un atributo de salida que funcione como supervisor, por lo que no hay un campo guía que indique el reconocimiento de patrones. Entre estos modelos no supervisados se encuentran:

Modelos de *clustering*: En los modelos de *clustering*, también llamados como de agrupamiento o conglomerado, los grupos no son conocidos con anterioridad. En

lugar de ello, los algoritmos analizan los patrones de datos de entrada e identifican el agrupamiento natural de los registros o casos.

Modelos de asociación o secuencia: Estos modelos también pertenecen a la clase de modelos no supervisados. Éstos no involucran predicción directa de un campo. De hecho, todos los campos involucrados tienen un doble rol, ya que actúan como entrada y salida al mismo tiempo. Los modelos de asociación detectan dependencia entre eventos discretos, producto o atributos. Los modelos de secuencia se orientan a la detección de asociaciones.

De acuerdo a con Bigus (1996) los resultados obtenidos de la minería de datos serán valiosos, si la información obtenida apoya en la adquisición de una ventaja competitiva o adhiere valor al proceso de toma de decisiones; y será eficiente si el valor de la información extraída excede al costo de procesamiento de los datos crudos.

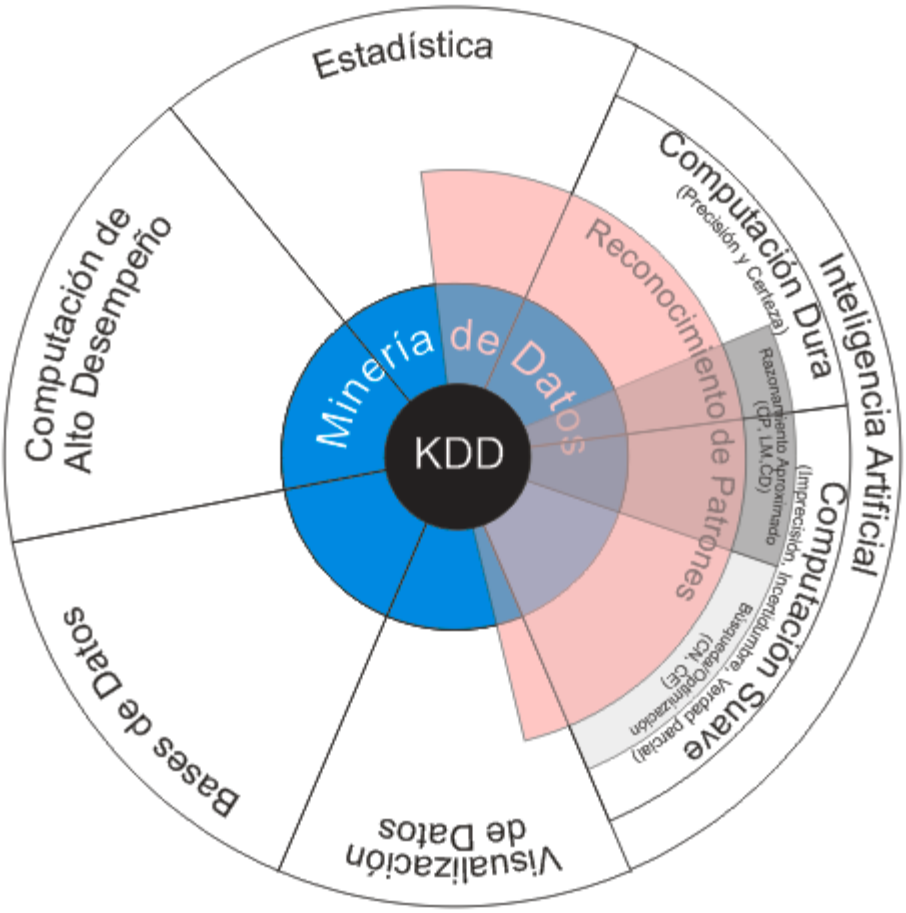
I.4. Campos de conocimiento del KDD

Para el logro de los objetivos, el KDD utiliza, a través de la DM, un conjunto de herramientas y metodologías provenientes de distintas disciplinas relacionadas con el almacenamiento, procesamiento y análisis de datos.

En la figura 1.2, se presentan las diversas disciplinas que le brindan soporte al KDD. Así, del lado izquierdo se pueden observar la Computación de Alto Desempeño, Bases de Datos y Visualización de Datos, que de alguna manera forman la infraestructura necesaria para el desarrollo del KDD y soportan los procesos de almacenamiento y procesamiento de datos, así como la presentación de resultados. Por el lado opuesto está el Reconocimiento de Patrones (PR, por sus siglas en inglés) que, a través de sus enfoques de aprendizaje, se presenta como el paradigma de aprendizaje natural, utilizado como enfoque conceptual entre la Estadística y la Inteligencia Artificial como las disciplinas proveedoras de

algoritmos de extracción de patrones. Los métodos y técnicas analíticas de ambas disciplinas apoyan a la DM en el logro de su objetivo, es la razón principal por la cual se considera a la Minería de Datos como el corazón del KDD.

Figura 1.2. Interdisciplinariedad del Proceso KDD.



CP: Computación Probabilística, LM: Lógica Multivaluada, CD: Computación Difusa,
CN: Computación Neuronal, CE: Computación Evolutiva

Fuente: Elaboración propia

Cada una de las disciplinas, desde su esencia misma, brinda soporte para la conversión de datos crudos o relacionados a conocimiento útil para el proceso de toma de decisiones. A continuación se aborda brevemente cada una de estas disciplinas y su importancia dentro del KDD.

1.4.1. Computación de alto rendimiento

La evolución tecnológica de la humanidad como reflejo de la capacidad innovadora del ser humano, ha estado regida por la capacidad para realizar cálculos matemáticos. A través del tiempo, el hombre ha desarrollado artefactos que le han permitido reducir el tiempo requerido en las operaciones numéricas que realiza. Entre algunas herramientas utilizadas hasta antes del siglo XIX para el cálculo de operaciones están el ábaco, el sistema de barras de Napier que apoyaba en la realización de multiplicaciones, y la pascalina (Kahn y Napper, 2000: Bauer 2007). Ya en el siglo XIX, aparece el motor diferencial y analítico de Babbage usado para la solución de polinomios básicos; así mismo, las herramientas mecánicas y electromecánicas para la ejecución de operaciones matemáticas (Bauer 2007). Es durante la segunda guerra mundial y la posguerra, que los instrumentos utilizados para realizar cálculos matemáticos pasan de ser completamente mecanizados a una hibridación mecánica-electrónica como fue el caso de la ENIAC[1], desarrollada en el marco de un proyecto militar secreto en la Universidad de Pensilvania y que puede ser considerada como la precursora de las computadoras modernas (Russell y Norving, 2004); esta gigantesca calculadora electrónica, incluía 1,500 relevadores electromecánicos y 17,000 tubos de vacío (Aspray, 1990).

Fue en los años sesentas, a partir del desarrollo de la microelectrónica, que aparecen las primeras minicomputadoras (Bauer, 2007, Ceruzzi, 2003). La compañía CDC presenta en 1964 la CDC6600 que operaba a una velocidad de 9 mflops ($9x10^6$ flops) [2]; así mismo, la compañía Burroughs desarrolla ILLIAC IV, primera minicomputadora en usar el diseño en paralelo; ésta contaba con 64 computadoras similares operando en paralelo. Mientras en los ochentas se alcanzan velocidades pico de 1 gflop ($9x10^9$ flops); aparecen las primeras computadoras personales con microprocesadores más flexibles y rápidos.

[1] Acrónimo de Electronic Numerical Integrator And Computer
[2] flops, es el acrónimo de Floating point Operations Per Second (operaciones de punto flotante por segundo).

Es a partir de la inquietud de grupos interdisciplinarios de investigación que en los años ochenta se identificaron problemas en la ciencia e ingeniería que requerían infraestructura de cómputo a gran escala como herramienta básica de apoyo en el descubrimiento de conocimiento científico novedoso (Berman, 2003); se requería así mismo de mayor rendimiento computacional para las aplicaciones de simulación numérica en la industria y en la investigación, además de aplicaciones como el procesamiento de consultas, minería de datos, y aplicaciones multimedia (Mohr, 2006); paralelamente aparece la computadora personal, la cual ofreció procesamiento más rápido a un menor costo, rompiendo el dominio de las minicomputadoras (Bauer, 2007, Ceruzzi, 2003). En la evolución de la capacidad de procesamiento computacional se identifican tres puntos de referencia que marca la evolución disruptiva de las eras de desarrollo de la computación, estas son: la era vectorial, la era del paralelismo masivo y la era de los multiprocesadores, donde esta última es la tendencia futura de los sistemas de cómputo (National Research Council, 2010).

La computación de alto rendimiento (HPC, por sus siglas en inglés), es el ambiente integrado de hardware y software, utilizado para resolver problemas que requieren del uso de tecnologías de cómputo a gran escala (Bacon et al., 1994; Plaza y Chang, 2008). Los sistemas de HPC proveen un mayor poder de procesamiento que el tradicionalmente disponible en una PC; va más allá del desempeño de una computadora personal de escritorio. En los últimos veinte años se ha incrementado la capacidad de procesamiento de un teraflop (10^{12} flops) a un petaflop (10^{15} flops), incluso hay quines dicen estar listos para pensar en el procesamiento exaflop (10^{18} flops) (Barder, 2008). El poder que brinda la HPC, permite en el manejo de problemas tan complejos como son la simulación de fenómenos meteorológicos, la predicción de huracanes y la de terremotos, la simulación de mutaciones de virus biológicos, el agrupamiento de expresiones génicas, entre otras aplicaciones.

La computación de alto rendimiento puede ser visualizada a partir de dos perspectivas: por un lado, la capacidad de procesar grandes cantidades de instrucciones por unidad de tiempo (flops), ya sea a través del procesamiento en paralelo trabajando en un mismo problema, o bien trabajando en varios procesadores diferentes problemas relacionados; por otra parte, puede significar una mayor rapidez de ejecución en sistemas más poderosos, a través de la utilización de múltiples sistemas simultáneamente. Otra forma de clasificar los sistemas HPC es de acuerdo con su orientación: los de alta disponibilidad, se enfocan a los sistemas tolerantes a fallos, donde las aplicaciones tienen que estar disponibles en todo momento, como sucede en algunos sistemas transaccionales en línea (servicios bancarios, supermercados, servicios gubernamentales en línea, entre otros); por otra, parte los sistemas de cómputo intensivos están orientados a resolver problemas en el menor tiempo posible, generalmente involucran la paralelización de la mayor parte del problema con el fin de resolver varias partes al mismo tiempo para después obtener la solución.

En los inicios del desarrollo de la HPC, se planteaba la necesidad de optimizar los compiladores con el propósito de reducir el tiempo de ejecución, así aparecen tecnologías como la RISC, orientada a la reducción de instrucciones de la programación y la programación en paralelo. Así mismo, surge el reto de la programación de algoritmos que requieren ser ejecutados bajo plataformas de procesamiento en paralelo.

En cuanto al hardware, la aplicación de la HPC se apoya en las tecnologías de la computación en *Clusters,* la computación *Grid* y la Computación *Cloud*[3].

La computación en *clusters*, orientada a elevar el rendimiento computacional a través del procesamiento en paralelo, surge a partir de del lanzamiento de la computadora personal. El *cluster* de computadoras es una colección de máquinas

[3] Se utilizan las palabras Cluster, Grid y Cloud, en su idioma original para evitar la confusión que pudiera generarse de la traducción al español; además, comúnmente se maneja de esta manera en los texto relacionados con la temática.

conectadas usando una red en tal forma que se comportan como una computadora sencilla (Galan et al., 2001). Morrison (2003), lo define como un tipo de sistema de procesamiento distribuido o paralelo, el cual consiste de una colección de computadoras trabajando juntas como un simple recurso integrado.

Es importante enunciar claramente la diferencia entre un sistema *cluster* y un sistema distribuido: los nodos en un sistema de cluster usan el mismo sistema operativo y pueden normalmente no ser manejados en forma individual (Rauber y Rünger, 2010). De esta manera, el incremento del poder de las computadoras personales y los avances en las redes de alta velocidad hacen más atractiva la computación en *clusters*, que actualmente muestra una tendencia creciente dentro del paradigma de la computación de alto rendimiento (Buyya R., 1999). El tipo de *cluster* de computadoras Beowulf es reconocido dentro del género de HPC (Morrison, 2003); el primer *cluster* de este tipo consistía sólo de hardware genérico (PCs) corriendo en Linux (Plaza y Chang, 2008). El procesamiento en paralelo es la propiedad de las máquinas que, idealmente, durante su ejecución debiera resultar en un desempeño más rápido; tiene como factor limitante la velocidad de comunicación (ancho de banda) y la latencia de los nodos entre el cálculo (Galan et al., 2001).

La computación *Grid*, es un término que implica diferentes tecnologías, mercados y soluciones para diferentes agentes (Stanoevska-Slabeva y Wozniak, 2010). Generalmente se le presenta como una analogía con la red de suministro de electricidad, donde los usuarios tienen acceso a la energía eléctrica a través de los conectores de pared sin necesidad de tener en cuenta por dónde o cómo es generada la energía eléctrica (Foster and Kesselman 2004; Jacob et al., 2005). Desde la perspectiva de la computación *Grid*, la computación se convierte en omnipresente y el usuario gana acceso a recursos de cómputo (procesadores, almacenamiento, datos, aplicaciones, y otros más) aun cuando tenga poco o nulo conocimiento de dónde estos recursos están localizados o cuáles son las

tecnologías subyacentes, (hardware, sistema operativo, y otros más) (Jacob et al., 2005; Bessis Nik, 2009).

El término de Computación *Cloud*, generalmente es usado como metáfora del Internet (Rittinghouse y Ransome, 2010; Velte et al., 2010). Su uso originalmente derivaba de la descripción común en los diagramas de red como una línea de una nube, usada para representar el transporte de datos a través de una columna trasportadora de una locación en el punto final en otro lado de la nube; esta representación fue popularizada por el Profesor John McCarthy a finales de los 1960s (Rittinghouse y Ransome, 2010).

En la actualidad, derivado del desarrollo de computación, su significado es más amplio. Para Velte et al. (2010), este término está relacionado a un ensamble computacional que permite el acceso a aplicaciones que residen en otra ubicación diferente a la del usuario, o al uso de otros periféricos conectados a través de Internet. Rittinghouse y Ransome (2010) coinciden en la ubicación de usuario y el proveedor, pero hacen una referencia más general a la distribución de los recursos computacionales, como un estilo de cómputo en el cual la escalabilidad masiva de las capacidades de las tecnologías de información necesarias son proveídas como un servicio a través de tecnologías de Internet a múltiples usuarios externos, coincidiendo con Gartner (2008). Para Gens (2008) se trata de un desarrollo emergente de tecnologías de información; es un modelo de despliegue y entrega, que habilitan a tiempo real la distribución de productos, servicios y soluciones mediante el uso de Internet. La computación *Cloud* se refiere tanto a la distribución de aplicaciones y servicios, utilizando como transporte la plataforma del Internet, el hardware y los sistemas de software en los centros de datos que proveen estos servicios (Stanoevska-Slabeva y Wozniak, 2010)

En la actualidad la hibridación de estas tecnologías aunado al desarrollo de las computadoras multicore, perfilan a la HPC como la plataforma ideal para la modelación y simulación de problemas complejos de gran envergadura.

1.4.2. Bases de datos (BD)

Las organizaciones siempre han generado y almacenado datos. Sin embargo, es hasta la aparición de las computadoras que su procesamiento pasa de manual a automatizado (Harrington, 2009). En la actualidad, los avances en el desarrollo de hardware y software, permiten almacenar cantidades de información en espacios muy pequeños (algunas unidades de centímetros cúbicos de espacio físico). Comparado con los segundos o minutos consumidos en el proceso de recuperación de datos en la actualidad, en épocas anteriores se hubieran requerido muchos metros cúbicos para su almacenamiento, además de bastantes días o años para su procesamiento, recuperación y manejo.

El desarrollo tecnológico de la computación y la electrónica han detonado la automatización de muchos de los procesos operacionales en las organizaciones, los cuales generan grandes cantidades de datos que automáticamente son almacenados físicamente en estructuras abstractas las cuales forman una base datos[4]. Éstas almacenan los hechos generados durante el desarrollo de los procesos dentro de una empresa u organización, al describir de alguna forma un objeto en particular; de esta manera se convierten en elementos básicos para los sistemas de información (Edmond, 1992). Más allá del almacenamiento, consulta y recuperación de datos, el principal propósito de esta herramienta es asistir en el descubrimiento de conocimiento (Revesz, 2010), razón por la cual los datos requieren ser manejados con precaución para asegurar su calidad, antes de ser trasformados en estructuras más complejas que permitan derivar información y conocimiento orientado a dar soporte al proceso de toma de decisiones, a un individuo o una organización.

La importancia de asegurar la calidad de éstos se debe a que son la base de la pirámide dentro del proceso de generación de conocimiento. De acuerdo con

[4] El concepto de base de datos, de acuerdo con Stephens (2007), se refiere a la herramienta computacional que almacena de alguna forma útil los datos y provee métodos para crear, leer, actualizar y eliminar registros.

Vercellis (2009), las fuentes de datos son la base de la arquitectura de los sistemas de inteligencia empresarial (Business Intelligence), debido a que éstos son generados por diferentes fuentes primarias y/o secundarias, a la vez que son creados por distintos medios lo que provoca que algunas veces sean heterogéneos en su origen y tipo. En la actualidad, los datos no sólo se restringen a la representación de registros numéricos o de caracteres. Hoy las tecnologías avanzadas de administración están habilitadas para integrar diferentes tipos de datos, tales como imágenes, video, texto, diferentes tipos numéricos, así como no numéricos (Mitra y Acharya, 2003; Stephens, 2007).

Generalmente una BD era referida como un conjunto de datos relacionados; sin embargo, esta concepción ha quedado rebasada debido a que muchas restricciones físicas y tecnológicas para el almacenamiento y manipulación de datos han sido resueltas (Halpin, 2001; Elmarsi y Navathe, 1997). Así mismo, cuando ésta es interpretada por el humano, puede ser vista como un conjunto de mediciones que relacionan hechos con una base de información (Halpin, 2001). Date (2001), refiere a una BD como un conjunto de datos persistentes que, una vez aceptados para ser registrados, sólo podrán ser removidos de ésta en lo sucesivo por alguna solicitud explícita del usuario. Anteriormente, se utilizaba el término “datos operacionales” en lugar de “datos persistentes”, esto debido a que originalmente la aplicación de estas herramientas estaba orientada a registrar las actividades operacionales o de producción, las cuales respondían al registro de rutinas altamente repetitivas durante el desarrollo de los procesos cotidianos de la empresa. Sin embargo, en la actualidad la tendencia en el uso de esta tecnología se expande a otro tipo de aplicación como lo son los sistemas de soporte a al proceso de toma de decisiones (DSS, por sus siglas en inglés).

Desde una perspectiva histórica, la evolución de los modelos de BD fue parte del modelo de red de datos (network data model), diseñado por Charles Bachman a principios de los años sesenta del siglo pasado, que fue estandarizado por

CODASYL[5], modelo que influenció fuertemente éste tipo de sistemas durante aquella década. Hacia finales de los sesenta, IBM desarrolla el Sistema de Administración de Información (IMS), el cual constituyó un marco base alternativo para la representación de datos llamado modelo jerárquico de datos, que aún es utilizado en grandes corporaciones, como son los caso del sistema SABRE de American Airlines e IBM (Ramakrishnan y Johannes, 1999). En 1970 la presentación del modelo relacional (en Date, (2001); Ramakrishnan y Johannes, (1999)) marca un rompimiento en el desarrollo de estas herramientas, debido a que este modelo está sólidamente fundamentado en la lógica y en las matemáticas, lo que facilita la visualización y comprensión de la relación entre las tablas de datos así como, los atributos incluidos en cada una de estas tablas.

El grado de calidad de los datos almacenados es el reflejo del desarrollo del proceso involucrado desde el diseño hasta la ejecución de la BD, que determina directamente la integridad de los datos almacenados, además de proteger los volúmenes almacenados, de daños accidentales o mal intencionados, lo que asegura la calidad de la información generada a partir de éstos (Stephens, 2001; Silberschatz, et al., 2002; Ramakrishnan y Johannes, 1999). En el argot de BD, el término integridad de datos, es un concepto referido a la exactitud o corrección de los datos dentro de una base de datos. El tener en mente este conocimiento al momento del diseño, desarrollo y aplicación de una base de datos, permite minimizar el proceso de la validación y modificación, ya sea por errores de captura a través del uso de restricciones de integridad como las restricciones de dominio, la integridad referencial y los asertos (Silberschatz, et al., 2002). Tales limitantes son derivadas directamente de las reglas de operación de los procesos del negocio que están representados en una base de datos (Date, 2001), que deberán ser tomadas en cuenta al momento de crear el diccionario de datos.

El diseño, desarrollo y aplicación de los sistemas de administración de bases de datos (DBMS), se apoya en la información contenida en diccionario de datos ya

[5] CODASYL: Conference on Data Systems Languages.

que esta por un lado describe las entidades, esquemas, permisos, atributos, etc. y por otra parte detalla los flujos de información a través del sistema (Caverlee, 2009; Teory, et al., 2006).

Más allá de asegurar la calidad de los valores almacenados, el diccionario de datos facilita la comprensión de los elementos de la información a través de la descripción semántica de los atributos almacenados. Este diccionario además de incluir las restricciones de integridad, describe información acerca de los datos (Ramakrishnan y Johannes, 1999), así mismo, clasifica las definiciones de los elementos, tipos de datos, sus flujos y otras convenciones que son usadas en un sistema de información (Caverlee, 2009). En este sentido, (Neto, et al, 2009) define el modelo completo de las clases, asociaciones, atributos y operaciones de una base de datos.

De acuerdo con Harrington (2009), entre la información que generalmente se encuentra en un diccionario de datos está: la definición de las columnas que forman cada tabla, restricciones de integridad de lugares y relaciones, los derechos sobre el tipo de operaciones que en ciertos datos los usuarios pueden desarrollar, además de definiciones de elementos de otras bases de datos tales como visitas y dominios de usuarios definidos.

Conocer el detalle de los elementos que componen la estructura de la base de datos, facilita de desarrollo de las operaciones de recuperación, consulta y transformación. Es precisamente el diccionario de datos el que provee esta información requerida; de hecho, éste es ampliamente utilizado en la optimización de consultas (Ramakrishnan y Johannes, 1999).

En síntesis, la utilidad de la aplicación de esta tecnología va más allá del simple proceso de almacenamiento de datos. Para que las organizaciones realmente obtengan el rendimiento óptimo de su inversión en cuanto a calidad y cantidad de información y conocimiento que descubran o construyan a partir de los datos

almacenados, deberán diseñar, desarrollar e integrar una plataforma de bases de datos acorde con sus objetivos de éstas, además de documentar detalladamente la cadena de generación, acopio, almacenamiento, integración y administración de los datos que minimice la deficiencia en la generación y procesamiento de datos.

1.4.3. Visualización dentro de KDD

El progreso en tecnología de hardware ha permitido a los sistemas computacionales almacenar grandes cantidades de datos. Se estima que 1.5 millones de terabytes son generados anualmente, la mayor parte de los cuales se encuentra en formato digital. Los datos son recolectados por la creencia de que constituyen una fuente potencial de información valiosa, que proporciona mayor comprensión de la realidad o bien una ventaja competitiva.

Encontrar información oculta en los datos es una tarea difícil y más aún explorarlos adecuadamente. Sin embargo, la visualización de información y el análisis visual de datos puede ayudar en el manejo adecuado del flujo de datos, ya que involucra directamente al usuario en el proceso de extracción de información. Ello permite identificar la utilidad potencial de éstos, así como discriminar la proporción de datos inútiles que provocan el almacenamiento de datos “basura” en las bases de datos (Keim y Ward, 2007).

Dentro del contexto del proceso de descubrimiento de conocimiento en bases de datos, la visualización de datos, información y conocimiento, es una herramienta crucial que interrelaciona estos tres conceptos y frecuentemente es utilizada para indicar diferentes niveles de abstracción, comprensión o veracidad (Chen, et al., 2009). Sin embargo, en el apoyo para la comprensión de datos, información y conocimiento, la selección errónea de los formatos y medios de visualización puede convertirse en la principal fuente de confusión (Gan et al., 2007).

La aplicación de herramientas de visualización ha demostrado ser una estrategia efectiva para apoyar a los usuarios en el manejo de escenarios complejos, ricos en información y conocimiento (Keller y Tergan, 2005; Keim y Ward, 2007). Sin embargo, para entender los alcances de la aplicación de las herramientas incluidas dentro del campo de la visualización, es necesario precisar algunos conceptos básicos, tales como: datos, información y conocimientos, para facilitar su comprensión y permitir ampliar el espectro de su aplicación en el desarrollo de las actividades cotidianas.

1.4.3.1. Datos

Los datos son la colección de símbolos que representan un hecho o declaración de eventos sin relación alguna entre uno y otro (Ackoff, 1989). Los datos simplemente existen y no tienen un significado más allá de su existencia; éstos pueden ser utilizables o no, además de existir en cualquier forma (Keller y Tergan, 2005). Desde una perspectiva computacional, los datos pueden ser definidos como la representación de modelos y atributos de entidades reales o simuladas (Chen, et al., 2009).

1.4.3.2. Información

La información se concibe como el conjunto de datos que tiene un significado derivado de los resultados de aplicar un procesamiento manual o automatizado, cuyo descubrimiento permite interpretar las relaciones y conexiones que éstos presentan con su contexto (Ackoff, 1989; Chen, et al., 2009). Su utilización responde las preguntas “quién”, “qué”, “dónde” y “cuándo” (Ackoff, 1989). La información puede ser clasificada de acuerdo con diferentes atributos, como: su origen, estatus de manipulación cognitiva, formato, “hechos” u “opiniones”. Es en base a esta clasificación que se descubren o generan los diferentes tipos de información, tales como: información objetiva, subjetiva; información primaria,

secundaria; verbal, impresa; visual y audio-visual. Además, de acuerdo con la forma de visualizarla, la información puede ser abstracta o concreta.

1.4.3.3. Conocimiento

Por otra parte, el conocimiento se refiere a la información que ha sido integrada dentro de una estructura de conocimiento humana, después de haber sido sometida a un proceso cognitivo como percepción, aprendizaje, asociación, razonamiento o imitación de algún conocimiento adquirido por el ser humano (Chen, et al., 2009). El conocimiento es el resultado que emerge del procesamiento de datos e información; el cual permite obtener patrones que conectan y mejoran la predicción de eventos a suceder en el futuro; el conocimiento permite responder la pregunta "cómo" (Ackoff, 1989). Si éste es utilizado para sinterizar un conocimiento nuevo a partir del previamente obtenido, es posible llegar a la comprensión.

Por su naturaleza dinámica el conocimiento está en constante cambio y adaptación para potencializar su aplicación en el desarrollo de alguna tarea (Keller y Tergan, 2005). La cognición se basa en información y conocimiento; sin embargo hay una marcada diferencia entre ambos conceptos: la información se encuentra fuera del individuo (algunas veces llamada "conocimiento del mundo") y el conocimiento está dentro del cerebro del individuo ("conocimiento en la cabeza"). Este último se refiriere a diferentes tipos de representación de patrones en el cerebro, mientras que la información puede consistir en representaciones externas que reflejan aspectos del conocimiento en la cabeza, así como artefactos culturales y cognitivos que figuran como los estímulos sensoriales y entradas perceptivas, procesadas e interpretadas automáticamente por el sistema cognoscitivo en términos de conocimiento (Rumelhart y Ortony, 1977).

De acuerdo con Brasford (1979)(en Keller y Tergan, 2005), algunos aspectos del conocimiento pueden ser externalizados. Tal es el caso de su estructura, por

ejemplo, que para otro individuo sólo puede ser información que requiere ser procesada para convertirse en conocimiento, suministrando un significado e integrándola dentro de su estructura mental de conocimiento. Algunas personas externalizan sus propios conocimientos para reconstruir su significado y reintegrarlos dentro de una estructura mental acorde a una tarea en particular.

De acuerdo con su accesibilidad cognitiva el conocimiento es tipificado como: explicito y tácito. El conocimiento explicito puede ser expresado simbólicamente, esto es, en palabras, números, o gráficamente; y puede ser compartido en forma de datos, fórmulas científicas, especificación de productos, visualizaciones, manuales, principios universales, entre otros. Este tipo de conocimiento puede ser transmitido fácilmente entre personas, formalmente y sistemáticamente. Por su parte, el conocimiento tácito es muy personal y difícil de formalizar, lo que dificulta su comunicación, además de ser casi imposible su transferencia a otros individuos (Keller y Tergan, 2005; Ward y Peppard, 2002). Así, aspectos como el talento, intuición, y corazonadas, caen dentro de esta categoría de conocimiento, que consiste en creencias, percepciones, ideales, valores, emociones, y modelos mentales. Además, el conocimiento tácito está profundamente enraizado en la acción del individuo y su experiencia (Keller y Tergan, 2005). Recientemente, Simemens (2005), sugiere que el conocimiento no sólo se restringe a “saber-qué” y “saber-cómo” sino tiene que ser complementado con el “saber-dónde”, lo cual implica el comprender dónde encontrar el conocimiento, que equivaldría al concepto de fuentes de conocimiento, esto es, saber dónde encontrar información que pueda ser usada como fuente de conocimiento (Tergan, 2005).

I.4.3.4. Visualización

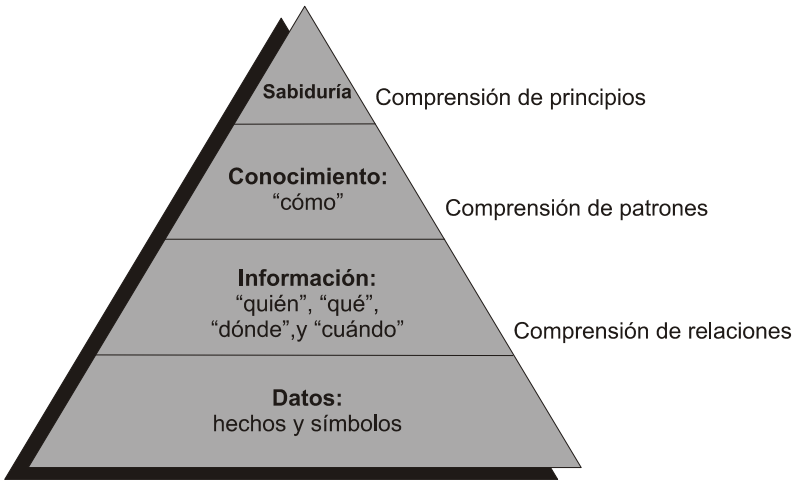
A través del uso de la visualización, el ser humano se acerca a la comprensión de los fenómenos que estudia. Así, partiendo desde la observación y medición de características clave respecto al dominio de la problemática abordada, desarrolla varias etapas de procesamiento; genera información y descubre conocimiento

que, a través de la síntesis, permite inferir los principios que definen el comportamiento de tal problemática.

Ackoff (1989), presenta la diferencia entre los conceptos: datos, información, conocimiento, comprensión y sabiduría. Los define, por asociación en términos muy sencillos: a) datos: hechos y símbolos; b) información: datos que son útiles (respuestas a "quién", "qué", "dónde" y "cuándo"); c) conocimiento: aplicación de información (respuesta a "cómo"); d) comprensión: apreciación de "por qué" y e) sabiduría: evaluación de la comprensión.

A partir de estas dimensiones el autor clasificó el contenido de la mente humana. Éstas son las dimensiones que utilizó para el diseño del modelo jerárquico DIKW[6] (figura 1.3), donde cada dimensión es temporal.

Figura 1.3. Pirámide jerárquica de conocimiento.



Fuente: Ackoff (1989)

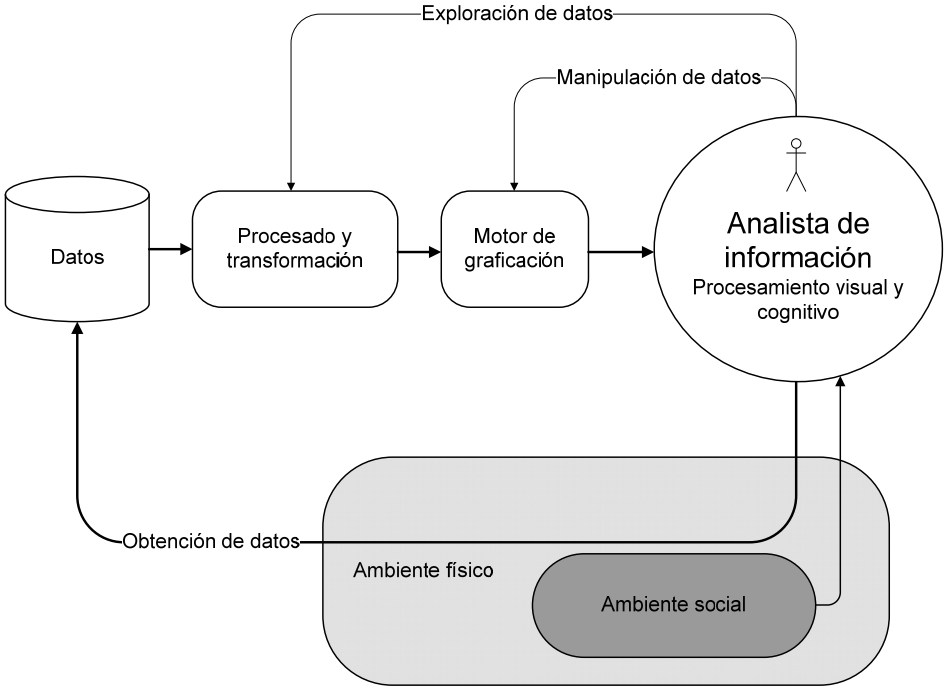
Para Ackoff, mientras que la información madura rápidamente, el conocimiento tiene una vida útil más larga; solamente la comprensión tiene una cierta permanencia y es la sabiduría la que se considera permanente.

[6] DIKW: Data, Information, Knowledge, Wisdom.

Las herramientas de visualización aprovechan las capacidades perceptivas humanas, para el análisis de grandes cantidades de datos disponibles en los sistemas de cómputo (Keim y Ward, 2007). De acuerdo con Ware (2004), este proceso es desarrollado a través de un sistema retroalimentado (Figura 1.4), el cual incluye las siguientes cuatro etapas básicas:

(1) Recolección y almacenamiento de datos; (2) Proceso previo del diseño de trasformaciones de datos para que puedan ser entendidos; (3) El hardware de exposición y los algoritmos gráficos que produzcan imágenes en pantalla; y (4) El sistema de percepción y cognición humana (el receptor).

Figura 1.4. Proceso de visualización



Fuente: Ware (2004).

En este proceso se identifican dos lazos de retroalimentación, que tienen propósitos bien definidos. Mientras que el lazo de la obtención de datos involucra tanto el ambiente físico (que es la fuente de datos) y el ambiente social (que determina en forma sutil y compleja lo que debe ser recogido y cómo es

interpretado); el otro lazo controla el preprocesado computacional que tiene lugar antes de la visualización.

1.4.3.4.1. Visualización de datos

La visualización de las estructuras inherentes en grandes cantidades de datos puede apoyar en la comprensión de las relaciones entre los elementos de la información y en la búsqueda visual de información relevante. Sin embargo, para el uso eficiente de los datos incluidos en la visualización, se requiere que el usuario comprenda la información a visualizar (Keller y Tergan, 2005). Esta información es expresada a partir de imágenes creadas en dos o tres dimensiones, que emergen de los datos procesados y que permiten detectar con mayor precisión y detalle el mundo físico; además, apoya al usuario en la modelación y simulación de fenómenos físicos complejos. En el desarrollo de estas tareas la visualización es soportada por el desarrollo de computación científica y la computación gráfica, la capacidad de transformar lo simbólico en geométrico, enriqueciendo el proceso de descubrimiento científico y promoviendo la intuición e interpretación a través de nuevos métodos visuales (Hansen y Johnson, 2005).

De acuerdo con Kusiak y Shah (2006), la visualización de datos puede ser concebida como la representación de los métodos o los resultados finales del proceso de transformación numérica y textual en un formato gráfico. Ésta es utilizada con el propósito de explorar grandes cantidades de datos en forma sistémica (holística) para la comprensión de tendencias y principios. De manera similar, Prabhu y Venkastesan (2007), señalan que la visualización de datos permite al analista la obtención de un conocimiento más profundo de los datos, elevando su compresión intuitiva. Para lograrlo, se hace uso de una amplia gama de herramientas de presentación de reportes al usuario final, que va desde una simple tabla que resume un conjunto de datos, hasta una gráfica compleja, utilizando técnicas de interpretación en dos y tres dimensiones para distribuir la

información presentada. Otros métodos utilizados para la visualización de datos, son los mapas de clases preservadas, las coordenadas paralelas, mapas de árbol y visualización de datos categóricos (Gan, et al., 2007).

La visualización de datos expresada dentro de la mente como una representación gráfica de datos o conceptos, se ha convertido en un artefacto externo que asiste al proceso de toma de decisiones, debido a que permite que grandes cantidades de información fina sean rápidamente interpretadas si se les representa en la forma correcta. Entre las ventajas de la visualización de datos se tiene que provee la capacidad para comprender una gran cantidad de datos; permite la percepción de propiedades emergentes que no pueden ser anticipadas. Además, la visualización comúnmente revela cosas no sólo acerca de los mismos datos, sino de la forma cómo fueron recolectados; la visualización facilita la comprensión de los datos tanto de las características a gran escala como a baja escala. Esto es especialmente valioso ya que facilita la formulación de nuevas hipótesis a partir de la percepción de patrones ligados a características locales (Ware, 2004).

I.4.3.4.2. Visualización de información

Partiendo de la conceptualización misma de la información, la cual asigna un significado a un conjunto de datos, su visualización óptima está en función del tipo de usuario final de la información construida o descubierta. Debido a la -gama de tipos de usuario con necesidades específicas existentes, se requieren herramientas de soporte para proveer un ambiente que habilite a los usuarios en el manejo visual e interpretativo eficiente y efectivo de la información (Pham, et al. 2009).

La visualización de información es un término usado en el contexto del procesado, comprensión y retención de información en gráficas estáticas, dinámicas, animadas e interactivas, que permiten ampliar la cognición (Card, et al., 1999). La visualización surge de las ciencias de la computación y brinda un soporte

computacional, interactivo a la presentación gráfica de datos abstractos, no físicos. Cuando el conjunto de datos es demasiado grande y diverso, es de particular importancia para la recuperación de información a través de herramientas de visualización, que los datos cumplan con los requisitos mínimos de calidad tanto en su diseño, como en su procesamiento (Keller y Tergan, 2005).

Marshall (2001), refiere que debido a la carencia de comprensión de la visualización de información, generalmente es necesario incluir elementos textuales en ésta para expandir la semántica visual y asegurar su comprensión, porque los símbolos o atributos de objetos gráficos frecuentemente no reflejan el total de las características de las unidades de datos subyacentes a expresar en la visualización de información. Sin embargo, es importante cuidar la cantidad de elementos textuales, para evitar la sobrecarga cognitiva no pertinente para el procesamiento de la información; por otra parte, la escasez de estos apoyos puede limitar la comprensión de la información visualizada (Keller y Tergan, 2005).

En la actualidad, existe un número importante de técnicas novedosas de visualización de información orientadas a cubrir el vacío dejado por las técnicas comunes, limitadas a conjuntos pequeños y con baja dimensionalidad, tales como: los diagramas de dispersión XY, gráficos de líneas, e histogramas; que limitan la exploración de conjuntos de datos carentes de una semántica inherente en dos o tres dimensiones.

Estas nuevas técnicas pueden ser clasificadas con base en tres criterios: los datos a ser visualizados, las técnicas de visualización, y las técnicas de interacción y distorsión utilizadas. La figura 1.5 muestra estas tres dimensiones, mismas que pueden ser tomadas como ortogonales, lo que significa que cualquier técnica de visualización puede ser usada conjuntamente con cualquier tipo de técnica de interacción para cualquier tipo de datos (Keim, 2002),

Figura 1.5. Clasificación de las técnicas de visualización de información

Fuente: Keim (2002).

Las técnicas para la visualización de información extraída a partir de datos multidimensionales, son clasificadas por Mazza (2009) como: técnicas geométricas, que incluyen a las coordenadas paralelas, las matrices de diagramas de dispersión, tablelens[7] y conjuntos paralelos; técnicas basadas en íconos como los gráficos estrella y las caras de Cehernoff; técnicas orientadas a pixeles donde los pixeles del monitor (pantalla) son las unidades básicas de representación que al momento de aplicar se deben tomar en cuenta los siguientes factores: forma de la pantalla, mapeo visual, arreglo de pixeles, color del mapeo, espacio físico del monitor (Keim , 2000).

El éxito de las interfases de manipulación directa es indicativo del poder de la utilización de las computadoras de una manera más visual o gráfica. Con el incremento en la velocidad de cómputo y la resolución de los monitores, las interfases gráficas y de visualización de información parecen estar jugando un papel más amplio; la visualización de información abstracta tiene el poder de revelar patrones, grupos, distancias, o valores atípicos, en datos estadísticos, comportamiento comercial, directorios computacionales o colección de

[7] También nombrada Tablelens, es una herramienta de análisis visual de datos, que se basa en una representación gráfica y simbólica dentro de una visión sencilla y coherente, que puede ser fluida y ajustada por el usuario (Rao y Card, 1994).

documentos. La habilidad perceptiva del usuario, permite la revisión, el reconocimiento y la formación imágenes en memoria rápidamente, pudiendo detectar cambios en el tamaño, color, forma, movimiento y textura (Shneiderman, 1996). Estas habilidades deberán ser explotadas a través de la implementación del Mantra de Búsqueda de Información (overview first, zoom and filter, then details-on-demand); este principio rige el diseño y desarrollo de los sistemas de visualización, los cuales comúnmente implican siete tareas (Shneiderman, 1996):

- Vista general. Obtener una descripción del conjunto completo.
- Acercamiento. Enfocarse a los objetos (datos) de interés.
- Filtrado. Separar los datos sin interés.
- Detección de demanda. Seleccionar objetos o grupos y obtener detalles cuando sean requeridos.
- Relacionar. Ver relaciones entre objetos.
- Historia. Mantener el historial de las acciones realizadas aun cunado se hayan eliminado, para realizarlas nuevamente en el refinamiento progresivo.
- Extracción. Permitir la extracción de subgrupos y parámetros de consulta.

I.4.3.4.3. Visualización de conocimiento

Como etapa subsecuente a la visualización de información, la visualización de conocimiento, surgida en el área de las ciencias sociales, particularmente en el campo del aprendizaje y la enseñanza (Keller y Tergan, 2005), es una combinación de campos como la visualización de información, el diseño gráfico y la ciencia cognitiva; a través de una representación diagramática ilustra conceptos como nodos en una gráfica interrelacionada, donde las ligas representan relaciones proposicionales entre dichos nodos (Saad y Zaghlou, 2002). El campo de la visualización de conocimiento examina el uso de las representaciones visuales para mejorar la creación y transferencia de conocimiento entre al menos dos personas, procurando trasferir intuición, experiencia, actitudes, valores,

expectativas, percepciones, opiniones, y predicciones (Burkhard y Michael, 2005, Eppler y Burkhard, 2006).

La transferencia de conocimiento entre individuos juega un papel importante en el desarrollo de cualquier organización, al apoyar el proceso de toma de decisiones y la resolución de problemas (King, 2006). Sin embargo, el elemento clave en este proceso de transferencia es la representación del conocimiento. Ésta permite estudiar cómo el conocimiento del mundo real puede ser representado y qué tipo de razonamiento puede realizarse con este conocimiento. Desde una perspectiva de la inteligencia artificial, Simons (1984), plantea que la representación de conocimiento es una combinación de estructuras de datos y procedimientos interpretativos que pueden estar disponibles en programa computacional para exhibir el comportamiento de conocimiento. En tal sentido, es importante tomar en cuenta el formato o la combinación de formatos que serán utilizados para lograr la transferencia de conocimiento. Burkhard y Michael (2005), presentan seis tipos de formatos diferentes: bosquejos heurísticos, diagramas conceptuales, metáforas visuales, conocimiento animado, mapas de conocimiento y estructuras de dominio. Todos estos formatos no sólo capturan en forma descriptiva los hechos o números, sino la perspectiva y el pronóstico intuitivo, así como los principios y relaciones involucrados (Eppler y Burkhard, 2006).

De esta manera, la visualización de conocimiento facilita la resolución de los tres problemas principales a los que se enfrentan las organizaciones al momento de transferir el conocimiento: a) la complejidad del conocimiento, esto es, su nivel de abstracción, así como la interrelación holística que éste presenta; b) las necesidades y antecedentes de los receptores del conocimiento transferido; y c) la sobrecarga de información en contraste con la limitada capacidad para absorber información nueva.

En este sentido, de acuerdo con Burkhard y Michael (2005), es necesario considerar el objetivo de la transferencia de conocimiento, el contenido del mismo,

las características de la audiencia a la cual se dirige, y los medios utilizados para esta transferencia. Además del tipo de conocimiento a transferir, que puede ser: conocimiento declarativo (saber-qué; ejemplo, los hechos), conocimiento del procedimiento (saber-cómo; ejemplo, el proceso), conocimiento experimental (saber-por qué; ejemplo, las causas), y conocimiento orientativo (saber-quién; ejemplo, experto), esto permitirá desarrollar una trasferencia efectiva de conocimiento.

Las etapas sucesivas de la visualización, desde la generación y acopio de datos hasta el descubrimiento y/o construcción de conocimiento, requieren estar centradas en las necesidades del usuario, de tal manera que se facilite la comprensión de la representación de datos, información y conocimiento. La visualización de datos permite observar la calidad de éstos en cuanto a su estructura e inventario en la base de datos; en cuanto a la información, su visualización se orienta a facilitar el entendimiento de las estructuras inherentes en conjunto de datos en dos, tres o más dimensiones. En cuanto a la visualización de conocimiento, ésta tiende a la representación de conceptos de conocimiento, su contenido y sus fuentes, que deben ser integradas de forma holística coherente, en un proceso iterativo y retroalimentado.

En el mismo orden de importancia para el KDD, se hallan los otros campos de conocimiento que dan soporte a este proceso a través de la proveeduría de herramientas analíticas que permiten la generación de información y conocimiento a partir de los datos. Tanto la estadística como la inteligencia artificial proveen técnicas y algoritmos para los modelos de minería supervisados y no supervisados. A continuación se abordan brevemente estas dos disciplinas.

I.4.4. La estadística

La estadística[8], como campo de conocimiento, ha proveído desde su inicio herramientas analíticas que han permitido la extracción de información y conocimiento a partir de un conjunto de datos. Folks (2007) refiere como los antecedentes de la estadística "la Aritmética Política", "la Teoría de Probabilidad", y "científicos experimentales del siglo XIX" (Folks, 1981; Fienberg 1992).

La estadística moderna como hoy se conoce, se construyó a partir de los trabajos de Carl Federick Gauss, quien, usando la distribución normal probó que la distribución del error en un sistema de ecuaciones lineales era equivalente a la distribución del error utilizando el método de mínimos cuadrados. Tiempo después, Laplace utiliza esta distribución normal del error para justificar sus resultados en el teorema del límite central. Esto fortaleció la aplicabilidad del enfoque estadístico en una amplia gama de problemas en la física. Sin embargo, Gauss-Laplace sentó las bases para la metodología del análisis de regresión, ésta se desarrolló 75 años después por Galton, quién describió primero el fenómeno de regresión y su liga a la distribución normal; así mismo, tiempo después formula el concepto relacionado de correlación. Edgeworth, relacionó los conceptos de correlación y regresión directamente en el contexto de la distribución normal multivariada, para lo cual introdujo el equivalente de la notación moderna para la matriz de correlación. Con la influencia de los trabajos de Edgewoth, Pearson desarrolló sus ideas respecto a los métodos de análisis de las curvas de sesgo.

---

[8] La palabra estadística deriva del latín moderno *statisticum collegium* ("consejo de estado"), del latín antiguo *status* ("posición", "forma de gobierno"), de la palabra italiana moderna *statista* ("estadista", "político") y del italiano antiguo *stato* ("estado"). En 1749, el alemán, Gottfried Achenwall (1719-1792) usa el término *Statistik* en su libro titulado "*Staatswissenschaft der vornehmen Europäischen Reiche und Republiken*"; este autor originalmente designó la palabra estadística para el análisis de los datos de un gobierno, y la definió como la "Ciencia del Estado". A Gottfried Achenwall se le conoce como el "Padre de la Estadística" (Vergara y Quezada, 2007).

De esta manera, Pearson y Yule conectan los desarrollos en la teoría de correlación y regresión, con la metodología temprana de mínimos cuadrados y la teoría de errores. Yule presenta los conceptos de correlación múltiple y correlación parcial y tiempo después introduce la notación moderna para el análisis de regresión que aún es ampliamente utilizada (Fienberg, (1992).

Hasta antes del siglo XX, la estadística reducía al resumen descriptivo de los datos observados y los índices comunes tales como medias, varianza, desviación, entre otros. No fue sino hasta principios del siglo pasado (en 1900), cuando Karl Perarson introduce la prueba de X^2 (chi-cuadrada o ji-cuadrada), utilizada para las pruebas de bondad de ajuste; poco después, en 1908, Gosset presenta la prueba t (t de Student) para la descripción de la inferencia de la media de una población normal. Esto cambió la perspectiva de la estadística, al pasar de ser una herramienta meramente descriptiva a un método para el procesamiento de datos que permitía realizar inferencias de los datos observados midiendo la incertidumbre en las generalizaciones realizadas a partir de una muestra (Rao, 1992).

Por sus contribuciones y críticas a otras aportaciones a la estadística, R. A. Fisher (1880 -1962), es considerado el fundador de la estadística moderna. Entre sus contribuciones al desarrollo de ésta se encuentran la distribución del muestreo exacto, el uso correcto de la X^2, la recomendación del uso de la máxima verosimilitud como un método general para la estimación. En el análisis de regresión aportó la metodología para la prueba de bondad de ajuste y la prueba de significancia de los coeficientes individuales con lo que abordó el problema de la selección de variables, así mismo recomendó el método de regresión para la desagregación de datos. Fisher introdujo una nueva área de investigación que fue la del diseño de experimentos con una riqueza de nuevas ideas en la experimentación científica, del análisis e interpretación de datos. Otras de sus aportaciones fueron la creación de las funciones discriminantes, y la cuantificación de variables categóricas.

El desarrollo de la inferencia estadística, ha sentado sus bases en tres paradigmas: bayesiano, fisheriano y frecuencista (Efron, 1998) los cuales se diferencian esencialmente en la forma de interpretación de la probabilidad y los objetivos mismos de la inferencia estadística. Así, en la visión Bayesiana, la interpretación de la probabilidad en términos de la frecuencia es secundaria, mientras que lo central es la descripción del estado subjetivo del conocimiento, de ahí que la distribución de probabilidad sea usada como la expresión de opinión.

Por otra parte, el enfoque Fisheriano propone que la inferencia estadística debe ser basada en probabilidades con una interpretación experimental directa; esto libera al estadista de las suposiciones previas y ha de sujetarse a lo conocido y lo no informativo de los parámetros del modelo. Finalmente, el paradigma frecuencista, surge del análisis matemático detallado de algunos conceptos de verosimilitud y suficiencia desarrollados por Fisher; con esto cambia la visión de la inferencia estadística como un resumen de datos, a procedimientos de inferencia abordados como problemas de decisión[9]. Estos procedimientos de inferencia óptima deben ser identificados antes de que las observaciones de los datos estén disponibles (Efron, (1998); Fienberg (1992); Pace Luigi y Salvan Alessandra, (1997); Yáñez (2000); Young y Smith 2005).

Durante la década de los sesenta del siglo XX, bajo la influencia de estos paradigmas, se desarrollan métodos robustos orientados hacia el manejo de valores atípicos que generan ruido respecto a la distribución normal. En el mismo sentido se desarrolla la estadística no paramétrica y se libera a los analistas de la trabas de los modelos limitados, dependientes de supuestos no realistas como la distribución normal. Ya en el último cuarto del siglo XX se desarrolla el Análisis

[9] El lector puede consular la siguiente bibliografía para profundizar en la discusión de estos paradigmas; (Efron, (1998); Fienberg (1992); Pace Luigi y Salvan Alessandra, (1997); Yáñez (2000); Young y Smith 2005).

Exploratorio de Datos[10], el cual rompe con el dogma frecuentista de no revisar los datos antes de la modelación. Se desarrolló el Modelo Lineal General; éste extendió el modelo lineal clásico a una clase más amplia que incluyó modelos probabilísticos aparte de la distribución normal y modelos estructurales que no eran lineales. Surgieron los Algoritmos de Maximización de la Esperanza, orientados a resolver problemas de estimación con datos incompletos. En el área del manejo de datos no métricos surge el Modelo Log-Lineal, orientados al análisis de datos nominales.

I.4.5. Inteligencia Artificial

La Inteligencia Artificial (AI, por sus siglas en inglés), es un campo de conocimiento relativamente joven que surge después de la segunda guerra mundial, en 1945. En sus etapas iniciales de desarrollo se consideraron a la Lógica Formal, La Psicología Cognitiva y La Computación como las bases que dan soporte a esta nueva disciplina (King y Harmon, (1988). A pesar de que el razonamiento a través de la lógica formal fue desarrollado por Aristóteles (384-322 a.c.) a través del uso de silogismos, son las aportaciones de Thomas Hobbes (1588-1679) las que se toman como referencia del inicio de la automatización de la computación[11], ya que fue él quien propuso que el razonamiento era como la computación numérica y que este es realizado a través de la lógica del cálculo (Russell y Norving, 2004; Vargas y Espinoza, 2008). Ya en la posguerra, Alan Turing (1912-1954) establece las bases teóricas de la computación, a través del uso de los operadores lógicos “and”, “or” y “not”, donde éstos podrían ser utilizados en cálculos numéricos y en la manipulación de materiales simbólicos (King y Harmon, 1988). La prueba de Turing surgida en 1950, ha permitido

[10] (EDA, por sus siglas en inglés, *Exploratory Data Analysis*) es un enfoque filosófico para el análisis de datos el cual emplea una variedad de técnicas (mayormente gráficas) para maximizar la comprensión dentro del conjunto de datos; descubrir la estructura subyacente; extraer variables importantes; detección de valores atípicos y anomalías; pruebas de soporte a los supuestos; desarrollo de modelos parsimoniosos y determinación de factores óptimos de ajuste (NIST/*SEMATECH, 2007*).

[11] La palabra computación en este contexto se toma como sinónimo de la operación matemática de calcular.

determinar el nivel de inteligencia desarrollado por los sistemas computacionales (King y Harmon, 1988; Nilsson, 1998; Russell y Norving, 2004).

Al hablar del surgimiento de la AI, es obligado referir tres trabajos que marcaron su génesis: el de Thomas Hobbes (1588-1679), que dejo claro que razonar es computar; el de Alan Turing (1912-1954) que estableció las bases teóricas de la computación; y el de John McCarthy informático americano quien acuñó la expresión de Inteligencia Artificial, en 1956 (King y Harmon, 1988; Nilsson, 1988; Martínez, 2007; Russell y Norving, 2004).

La AI estudia la forma de percibir, entender y actuar del ser humano, para, con base en ello, diseñar y construir computadoras inteligentes, máquinas inteligentes capaces de comportarse como los seres humanos (Martínez, 2007; Russell y Norving, 2004). Este comportamiento inteligente involucra desde la percepción, el razonamiento, aprendizaje, comunicación y actuación en ambientes complejos (Nilsson, 1988). En su trabajo sobre el enfoque moderno de la AI, Russell y Norving (2004) presentan una clasificación de diferentes definiciones de ésta, de acuerdo con la forma de pensar, racionalizar y actuar del ser humano (cuadro 1). Estos enfoques, centrados en el ser humano, que implican el planteamiento de hipótesis a comprobar a través de la experimentación, los centrados en torno la racionalidad que implican una combinación de matemáticas e ingeniería. Sin embargo, ambos enfoques se vuelven complementarios, a pesar de que parten de puntos distintos.

De esta manera la AI presenta dos vertientes: la computación dura, que provee herramientas analíticas caracterizadas por el procesamiento de hechos precisos y con certeza; la computación suave[12], la cual provee herramientas analíticas que

[12] SC, por sus siglas en ingles.

son inmunes a ambientes con datos e información imprecisa, con incertidumbre y verdad parcial[13].

Cuadro 1. Definiciones de inteligencia artificial clasificadas de acuerdo con las formas de pensar, actuar y racionalizar.

Sistemas que "piensan como humano"	Sistemas que "actúan como humano"
"[La automatización de] actividades que vinculamos con proceso, de pensamiento humano, actividades como la toma de decisiones, resolución de problemas, aprendizaje" (Bellman, 1978). "El nuevo y excitante esfuerzo de hacer que las computadoras piensen, máquinas con mentes, en el más amplio sentido literal" (Haugeland, 1985).	"El arte de desarrollar máquinas con capacidad de realizar funciones que cuando son realizadas por personas requieren de inteligencia" (Kurzweil, 1990). "El estudio de cómo lograr que las computadoras realicen tareas que, por el momento, los humanos las hacen mejor" (Rich y Knight, 1991).
Sistemas que "piensan racionalmente"	Sistemas que "actúan racionalmente"
"El estudio de las facultades mentales mediante el uso de modelos computacionales" (Charniak y McDermott, 1985). "El estudio de los cálculos que hacen posible percibir, razonar y actuar" (Winston, 1992)	"La Inteligencia Computacional es el estudio del diseño de agentes inteligentes"> (Poole et al., 1998). "Está relacionada con conductas inteligentes en artefactos" (Nilsson, 1998).

Fuente: Russell y Norving, (2004)

[13] Aunque los algoritmos y métodos de la computación dura escapan a los propósitos de este trabajo, se sugiere la revisión de la bibliografía presentada en este apartado, para un acercamiento al tema.

I.4.6. Computación Suave

Generalmente, para muchos problemas en la vida real es difícil disponer de información precisa, carente de incertidumbre y completa, que permita derivar soluciones. Es a partir de esta problemática que Zadeh acuña el concepto Computación Suave, para hacer referencia a la forma como los humanos solucionan muchos problemas, con base en un razonamiento aproximado, muchas veces más efectivo que un enfoque preciso (Melin y Castillo, 2005; Zadeh, 1994). Este concepto contrasta con el paradigma de la computación dura tradicional, que se rige bajo la premisa de disponer de información precisa, con certidumbre (Russel y Yuhui, 2007) y que rara vez resulta adecuado para este tipo de problemas, sobre todo por el costo computacional que implicaría el cumplir tales requerimientos (Mitra y Acharya, 2003).

La rapidez del cambio en el comportamiento del ambiente o contexto, incrementa la complejidad de los problemas a tratar tanto en el campo de la investigación como en las industrias, lo que exige soluciones más allá de los modelos matemáticos sobreidealizados que en ocasiones pierden la relación con la problemática o dominio del problema original (Zadeh, 1962; Zilaouchian, y Jamshidi, 2001). Esta necesidad ha llevado al desarrollo de herramientas de análisis y diseño de soluciones eficientes para ambientes donde prevalece información ambigua, incierta e imprecisa.

La SC es un compendio de metodologías desarrolladas a través de la imitación del conocimiento, la cognición y habilidad que posee el ser humano para entender discursos con audio distorsionados, reconocimiento de caracteres escritos a mano, comprensión de matices de lenguaje natural, resumen de textos, compresión de conceptos, reconocimiento y clasificación de imágenes, manejo de vehículos en tráficos densos y más generalmente, toma de decisiones en ambientes de incertidumbre e imprecisos (Mitra y Acharya, 2003; Du y Swamy,

2006). Los medios utilizados son arquitecturas de cómputo en paralelo que simulan procesos biológicos y pueden desarrollar mapeos de datos de entradas y de salidas de manera más eficiente que las herramientas analíticas. Éstas permitir la inferencia a partir de de datos e información con imprecisión, la incertidumbre y verdad parcial (Cordón et al., 2001). La capacidad de procesar datos con tales características hacen que el SC alcance un mejor desempeño, con mayor autonomía y flexibilidad, a bajo costo computacional y mayor concordancia con la realidad (Zilaouchian, y Jamshidi, 2001; Cordón et al., 2001).

La integración sinérgica de las metodologías incluidas dentro del dominio de la computación suave permite la incorporación efectiva del conocimiento humano; trata con impresión e incertidumbre y aprende a adaptarse a ambientes desconocidos o cambiantes para un mejor desempeño. Para el aprendizaje y la adaptación, la SC requiere cómputo exhaustivo; en este sentido, comparte las mismas características que el cómputo inteligente (Jang, et al., 1997).

Las metodologías que forman el corazón de la SC pueden ser clasificadas de acuerdo con los enfoques propuestos por Bonissone et al, (1999): el enfoque de búsqueda/optimización, donde se ubican la computación neuronal y la computación evolutiva, las cuales son inspiradas en modelos biológicos; y el enfoque de metodologías basadas en razonamiento aproximado, donde se encuentran la computación probabilística, la lógica multi valuada y la computación difusa.

La Lógica Difusa (FL,[14] por sus siglas en inglés) herramienta de uso más generalizado en este tipo e aplicaciones, fue desarrollada por Lofti Zadeh en los años sesenta del siglo pasado. Ésta proporciona un marco de trabajo natural para el tratamiento de datos con incertidumbre e imprecisión; se inspira en la habilidad de la mente humana para la manipulación de información difusa y su consecuente

[14] Este concepto se desarrolla ampliamente en el apartado dedicado al tema de agrupamiento difuso.

capacidad de síntesis, característica fundamental que distingue la inteligencia humana (Mitra y Acharya, 2003).

Cabe señalar que la principal fortaleza de la teoría de conjuntos difusos, que es la base de la FL, presenta como su principal carácterístiaca la capacidad de la representación del conocimiento a través de reglas "si-entonces" (Jang, et al., 1997; Zadeh, 1994), así como el uso de variables lingüísticas cuya función principal es la granulación de variables y sus dependencias (ibídem); además de, las Redes Neuronales Artificiales (ANN, por sus siglas en inglés), desarrolladas a partir del paradigma conexionista[15] para explicar e imitar el proceso del razonamiento humano, inspirado en la red neural biológica. Se asume que la capacidad de procesamiento de información del cerebro humano se explica por las interacciones de las grandes redes de neuronas interconectadas. Las ANN se basan en esta idea y proveen algoritmos adecuados de entrenamiento que reflejan la capacidad de adaptación y aprendizaje en ambientes cambiantes (Jang, et al., 1997). McCulloch and Pitts, propusieron en 1943, el primer modelo para una ANN: el perceptron, que la representación más sencilla de una de una neurona aislada[16].

Un paradigma natural para aplicación de los modelos minería de datos, a través de las herramientas que proveen las dos disciplinas antes expuestas, es el reconocimiento de patrones (PR, por sus siglas en inglés). A pesar de ambos (DM y PR) presentan diferencias en cuanto al volumen de datos que manejan y la formulación del objetivo que persiguen, tienen similitudes, como son el reconocimiento de regularidades y la forma en que se conceptualizan los

[15] El simbolismo y el conexionismo son los dos paradigmas utilizados para explicar e imitar el proceso de razonamiento humano, en el desarrollo de máquinas inteligentes. El simbolismo ve el razonamiento como el proceso de creación y manipulación de un mapa simbólico del mundo exterior; este enfoque se basa en el uso de un sistema formal de axiomas para el procesamiento de símbolos. Axiomas, teoremas y reglas deductivas son utilizados para manipular símbolos y derivar conclusiones significativas (Kolman y Margaliot, 2009).

[16] Bishop (1995), detalla ampliamente las características del perceptrón, además de los arreglos o topologías que forman la unión de dos de éstos.

conjuntos de datos, a través de espacios p-dimensionales. Enseguida se aborda brevemente la evolución del PR, las etapas de este proceso y las estrategias de aprendizaje que sigue.

I.4.7. Reconocimiento de Patrones[17]

Hasta antes de los años sesenta del siglo XX, el reconocimiento de patrones era soportado básicamente por la teoría estadística (Theodoridis y Koutroumbas, 2003). Surge, específicamente como una extensión del análisis discriminante (Bishop, 1995; McLachlan, 2004; Pal y Pal, 2001), en esa década donde el desarrollo de las computadoras permitió su desarrollo (Theodoridis y Koutroumbas, 2003; Webb, 2002). Después de más de cuatro décadas a pesar de haber alcanzado un buen nivel de madurez, sigue creciendo su aplicación, gracias al desarrollo de otros campos como la AI, SC, la estadística, psicología entre otras, (Pal y Pal, 2001; Webb, 2002).

Durante el desarrollo del PR se han utilizado definiciones para éste. Así, Fukunaga (1972) plantea que consta de dos partes: la selección de características y el diseño del clasificador; para Duda y Kart (1973) el PR está relacionado con el reconocimiento automático de regularidades significativas en ambientes ruidoso o complejos. Por su parte, Pavilis (1977) lo define como la imitación de un ejemplo perfecto creado por un objeto dado que fue realizado después; para Gonzalez y Thomason (1978), podría considerarse como la caracterización de datos de entrada, a través de extracción de características significativas; para Bezdek (1981), representa la búsqueda de estructuras de datos; para Devijver y Kitter (1982) la importancia del los bordes muy difusos (ambiguos); Schalkoff (1992), los define como la ciencia que relaciona descripción y clasificación de mediciones.

[17] De acuerdo con Bow (2002), “un patron puede ser definido como una descripción cuantitativa o estructural de un objeto o algunas otras entidades de interés”, sólo incluye objetos físicos visibles, sino también los objetos abstractos que forman los sistemas de datos. En un sentido jerárquico una clase en PR se entiende como un conjunto de patrones que tienen en común algunas propiedades.

Fukunaga (1991), casi veinte años después, visualiza el PR como un problema de toma de decisiones, considera que se trata de un problema para la estimación de la función de densidad en un especio de alta dimensión; divide este espacio en regiones de clasificación o clases. Para Bishop (1995), éste abarca un amplio proceso de información de importancia práctica, que va desde el reconocimiento de voz y caracteres escritos a mano, hasta la detección de fallas en maquinaria y diagnóstico médico; Bezdek et al. (1999), resaltan la importancia de los enfoques numérico y sintáctico en este campo; para Dasey y Micheli (2000), éste involucra una habilidad cerebral para asignar etiquetas a objetos, sonidos, sentimientos o ideas y discriminarlos uno de otros; en sintonía con Dasey y Micheli, Pal y Pal (2001), así como Duda et al. (2001) y McLachan (2004), se refieren al PR como el proceso de automatización de muchas de la tareas de reconocimiento que los seres humanos realizan normalmente, a través del desarrollo de algoritmos y metodologías (Bishop, 2006); Marques de Sa (2001).

Michael (2002) y Bow (2002), coinciden con Schalkoff y visualizan al PR como una área fértil de investigación, con múltiples ligas a otras disciplina. A continuación se presenta las etapas del PR y sus características principales.

I.4.7.1. Etapas del Proceso de reconocimiento de patrones

De acuerdo con Bow (2002), el reconocimiento de patrones es un proceso de clasificación de datos observados o mediciones de éstos como miembros de una o varias clases o categorías. La Figura 1.6 muestra las tres etapas que forman este proceso, así como su secuencia:

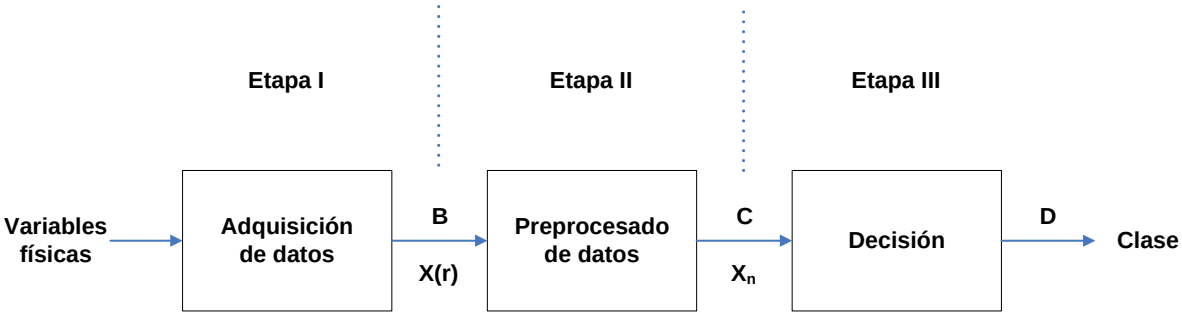
1) Adquisición de datos: los datos son obtenidos a través de un transductor y convertidos a un formato digital adecuado para el procesamiento computacional. En esta etapa, las variables físicas son convertidas en un conjunto de datos de medición X(r).

2) Los datos de medición son utilizados como entrada para la segunda etapa (procesamiento de datos) y agrupados en conjuntos de rasgos característicos x_n como salida.

3) En esta etapa se ubica la decisión a que conducirá la ejecución de conjunto de decisiones. De esta manera, los objetos pueden ser entonces clasificados a partir de un grupo de características.

Los conjuntos de salida de cada una de las etapas están dentro de un espacio o dominio específico. Así el conjunto *B* está dentro del espacio de patrones, el conjunto *C* dentro del espacio de características y *D* dentro del espacio de clasificación.

Figura 1.6. Esquema básico del proceso de reconocimiento de patrones.



Fuente: Bow (2002).

Marques de Sá (2001), sostiene que el reconocimiento de patrones es una disciplina científica que trata con modelos para la descripción y clasificación de objetos. Para desarrollar un mapeo entre el espacio de representación y el espacio de interpretación. La Figura 1.7, presenta un sistema de reconocimiento esta compuesto por cinco unidades funcionales:

1) Adquisición de patrones, que puede tomar diferentes formas: señales o adquisición de imágenes, conjuntos de datos.

2) Extracción de características, en la forma de mediciones, extracción de datos primarios, etc.

3) Pre-procesado. En algunos casos los valores característicos no son directamente alimentados dentro del clasificador o descriptor.

4) La unidad de clasificación, regresión o descripción es la unidad núcleo del sistema de reconocimiento de patrones.

5) Post-procesado. Algunas veces las salidas son obtenidas a partir de la unidad núcleo ésta no puede ser usada directamente.

Figura 1.7. Proceso de reconocimiento de patrones.

Fuente: Marques de Sá (2001).

En el mismo sentido, otros autores enfatizan ciertas características del proceso y ubican dentro de éste diferente número de etapas. Tal es el caso de Lampinen et al. (1998), quienes se refieren al reconocimiento de patrones como la mezcla de ciencia y arte que da nombres a los objetos naturales del mundo real; lo exponen a través de ocho etapas: acopio de datos, registro, pre-procesamiento, segmentación, normalización, extracción de características, clasificación y post-procesamiento. Theodoridis y Koutriymbas (2003), describen al reconocimiento de patrones como la disciplina cuyo objetivo es la clasificación de objetos dentro de un número de categorías o clases y definen la estructura de un clasificador a través de cinco elementos: detección, generación de características, selección de características, diseño del clasificador y sistema de evaluación.

De esta manera, el proceso de reconocimiento de patrones puede ser entendido como el proceso de comparación y/o clasificación de objetos, sean concretos o abstractos, donde las sensaciones percibidas de los diferentes atributos que caracterizan a tal objeto primero son convertidas a datos numéricos, para después ser ordenada en vectores homogéneos (patrones); se organizan de acuerdo con su similitud, a través de un conjunto de técnicas analíticas, que pueden ser utilizadas de forma individual o formando una combinación de dos o más de éstas.

El mapeo de los diferentes espacios conceptuales incluidos en el contexto del reconocimiento de patrones puede ser representado a través de una clasificación, una regresión o una solución descriptiva (Marques de Sá, 2001). Para la obtención de estos mapeos, el uso del proceso de reconocimiento de patrones se logra a través dos estrategias:

Aprendizaje supervisado, que implica manejo conceptual o hipótesis inductiva. Éste permite encontrar en el espacio de representación la hipótesis correspondiente a la estructura del espacio de interpretación. Éste es el enfoque de los ejemplos previos, donde, dado un conjunto de patrones se plantea una solución hipotética. El desarrollo de esta estrategia, requiere un intenso entrenamiento a través de un conjunto de prototipos, esto es, con un conjunto de patrones de entrada y su categoría de pertenencia, que en el caso de una ocurrencia de error en la salida, se llevara a cabo un ajuste en los parámetros a través de esta estrategia. De manera más concreta, en este aprendizaje existe un supervisor orientado a enseñar al sistema cómo clasificar un conjunto de patrones primarios y permitirle así seguir adelante libremente en la clasificación de otros patrones (Bow, 2002).

Aprendizaje no supervisado, que implica manejo de datos o hipótesis deductiva. Permite encontrar estructuras en el espacio de interpretación correspondientes a estructuras en el espacio de representación. El enfoque no supervisado intenta encontrar hipótesis útiles basadas sólo en relaciones de similitud. En esta

estrategia, a diferencia del aprendizaje supervisado, la clasificación no depende de información anticipada.

Como se mencionó en el apartado de minería de datos, tanto el aprendizaje directo a través de un supervisor como el aprendizaje indirecto sin la presencia de un guía, hacen que la aplicación del PR, permita predecir ambos eventos métricos como categóricos, además de describir espacios geométricos que caracterizan a un conjunto de datos p-dimensionales.

El proceso KDD ofrece a través de minería de datos, PR y sus campos de conocimiento de soporte (figura 1.2), un marco de trabajo estructurado que permite descubrir las estrategias que han desarrollado las organizaciones para mantenerse dentro de la competencia. Así mismo, el KDD habilita el reuso de los datos históricos almacenados en los repositorios de las empresas, para adherir valor al proceso de toma de decisiones.

El amplio espectro de problemas que pueden ser abordados a través del KDD, implica una variedad de métodos y técnicas que proveídas por los distintos campos de conocimiento como la estadística, la computación y la visualización de datos; que permiten procesar conjuntos de datos en diferentes escenarios: datos estructurados y sin estructura, desde precisos e imprecisos, generados en ambientes simples o ambientes ruidosos y complejos.

En el siguiente capitulo se presenta el análisis de agrupamiento de datos, herramienta a los modelos de minería de datos y PR no supervisados, de uso muy generalizado en las aplicaciones del proceso KDD.

Capitulo II. Agrupamiento de datos

El agrupamiento de datos[18] es una técnica de uso generalizado en el descubrimiento de conocimiento en bases de datos, que pertenece a los modelos de DM no supervisados. Esta técnica tiene el objetivo de revelar el agrupamiento natural de los patrones de datos, a través de la formación de grupos en un conjunto de datos, con base en el principio de máxima similitud al interior del grupo y mínima similitud entre patrones pertenecientes a grupos distintos. En este apartado se exponen las características generales y algunos conceptos básicos del análisis de agrupamiento de datos, sus criterios de agrupamiento o medidas de similitud, los tipos de algoritmos de agrupamiento y se finaliza con una descripción general de los algoritmos K-Medias, C-Medias Difuso y Gustafson-Kessel.

II.1. Análisis de agrupamiento de datos,[19]

El ser humano posee un habilidad natural para el agrupamiento de patrones (Das, et al, 2009); sin embargo, esta capacidad queda rebasada cuando se intenta agrupar, con base en más de una característica, conjuntos con un gran número de objetos abstractos (datos p-dimensionales). El análisis de agrupamiento de datos, también conocido como análisis de *clustering* (análisis de conglomerado), análisis

[18] En el presente trabajo el concepto datos implica un conjunto de mediciones de diferentes características a un mismo sujeto de investigación. En este sentido, se entiende por datos el conjunto de datos p-dimensional, donde la dimensionalidad es determinada por el número de características variables o atributos medidos al sujeto de estudio. De ahí que se usarán como sinónimos datos, patrones, vectores, observaciones y vectores. En tanto, características, variables y atributos serán tratados indistintamente como elementos de datos.

[19] Generalmente los conceptos agrupamiento y clasificación son utilizados de forma indistinta; sin embargo, son semánticamente diferentes: mientras que la clasificación es realizada a través de la existencia de clases previamente conocidas, a partir de las cuales lo objetos son asignados a clases específicas, en el agrupamiento de datos estas clases son deducidas a partir de las similitudes naturales que existen entre los objetos que forman el conjunto a agrupar (Gan, et al., 2007).

de segmentación, aprendizaje no supervisado[20], aprendizaje sin maestro (dentro del reconocimiento de patrones), taxonomía numérica (en biología y ecología), tipología (en ciencias sociales) y partición (en teoría de grafos), (Dunne 2007; Gan, et al., 2007; Theodroridis y Koutroumbas, 2006), permite revelar la organización intrínseca subyacente a un conjunto de datos no etiquetados, a través de formación de subconjuntos[21] formados por datos con características similares (Abonyi y Feil, 2007; Bow, 2002; Gan, et al., 2007; Höppner et al., 2000; Pedryckz 2005; Schenker, et al., 2006; Theodroridis y Koutroumbas, 2006). Basado en varios autores Pedryckz (2005), sintetiza al proceso de agrupamiento de datos como una metodología general y un poderoso marco conceptual y algorítmico para el análisis e interpretación de datos.

De acuerdo con Höppner y otros (2000), tales grupos deberán tener las siguientes propiedades:

Homogeneidad dentro del *cluster*; esto es, los datos pertenecientes a un *cluster* deben ser los más semejantes posible.

Heterogeneidad entre *clusters*; esto es, los datos que pertenezcan a *clusters* diferentes deberán ser lo más diferentes posible.

De manera formal, el proceso de agrupamiento de datos se determina como:

Dado un conjunto de observaciones de un proceso físico, cada observación consiste de n características medidas que forman un vector n-dimensional $x_k =$

[20] Puesto que no se requiere de un conjunto de entrenamiento para clasificar el conjunto de patrones, de acuerdo con Fukugana, (1990).

[21] En el campo del *clustering*, los subconjuntos formados a partir de este proceso han sido identificados de diferente manera. En le presente trabajo y específicamente a partir de este punto se usarán como sinónimos por su equivalencia semántica para este tipo de subconjuntos: *cluster*, clase, conglomerado, partición, segmento y grupo (Kruse et al., 2007).

$[x_{k1}, x_{k2}, \ldots, x_{kn}]^T$, $x_k \in R$. Un conjunto de N observaciones es identificado por $X = \{x_k | k=1,2,\ldots, N\}$ y es representada por una matriz $N \times n$[22].

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{N1} & x_{N2} & \ldots & x_{Nn} \end{bmatrix}$$

El objetivo es encontrar la familia de C clases que revelen la estructura oculta en los datos, a través de los centros de *cluster* representados por $v_i = \{v_{n1}, v_{n2}, \ldots, v_{nc}\}$, donde los elementos de cada vector v_i son los valores de cada una de las variables medidas en el proceso físico y caracterizan al vector tipo de cada *cluster*.

II.1. 1. Tipo de variables

Como se ha señalado, actualmente las empresas generan grandes cantidades de datos[23], los cuales son almacenados en una variedad de formatos diferentes. Los datos como elemento básico en la construcción del información y conocimiento (figura 3), requieren ser tipificados para decidir la forma en serán procesados.

De acuerdo con el tipo de valores que toman las variables que forman un patrón de datos, éstas pueden ser caracterizadas como: numéricas, que pueden asumir cualquier valor en **R**; ordinales, que toman un pequeño número de estados discretos y éstos pueden ser utilizados para hacer comparaciones, en este caso la

[22] Dentro de la terminología utilizada en el campo del reconocimiento de patrones los renglones son identificados como patrones u objetos y las columnas como características, atributos o variables y X representa la matriz de patrones o espacio de entrada.

[23] El elemento más sencillo almacenado en una base datos son los campos que componen los registros, los cuales a su vez forman las tablas de datos. Dentro del reconocimiento de patrones los campos de un registro son los atributos, variables o características que forman un patrón. Estos campos pueden provenir de una o varias tablas de datos, contenidas dentro de una o varias bases de datos.

posición del valor en la escala ordinal tiene un significado de cercanía o similitud; nominales, obtienen un pequeño número de estados, pero éstos no pueden ser usados para medir cercanía. La variables ordinales y nominales son representadas como variables discretas y para cuestiones de cálculo son codificadas en varios esquemas, tales como binarios para el caso de variables que sólo puede tomar dos estados, o en forma de columnas binarias para representar la presencia o no presencia de algún estado polinomial (Abonyi y Feil, 2007; Myatt, 2007; Pedrycz, 2005).

II.1.2. Medidas de similitud

Así como el ser humano, con base en las similitudes percibidas, toma la decisión de clasificar objetos en cierta clase, el agrupamiento automático de datos requiere de una medida numérica que permita determinar qué tan similares o disímiles son los datos a comparar; una alternativa útil para tal propósito es utilizar el cálculo de la distancia geométrica entre los datos p-dimensionales que requieren ser agrupados (Myatt, 2007). El concepto de disimilaridad es un componente esencial en cualquier agrupamiento que permite la navegación a través del espacio de datos y de los grupos de datos; permite saber que tan cerca están dos patrones y con base en esta cercanía, asignar a éstos al mimo grupo (Pedrycz, 2005).

La disimilaridad entre **x** e **y** es una función de dos argumentos denotada por $d(\mathbf{x}, \mathbf{y})$ que satisface las siguientes condiciones:

$$d(\mathbf{x}, \mathbf{y}) \geq 0 \text{ para cada } \mathbf{x} \text{ e } \mathbf{y}$$
$$d(\mathbf{x}, \mathbf{x}) = 0 \text{ para toda } \mathbf{x}$$
$$d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$$

Existen diferentes funciones matemáticas que permiten determinar la distribución geométrico-espacial de los patrones en un espacio p-dimensional, las cuales responden a requerimientos topológicos específicos. La similitud entre patrones

frecuentemente es interpretada en términos de una función de distancia *d*(***x, y***) en R^n, tal que:

$d(\boldsymbol{x}, \boldsymbol{y}) = \| \boldsymbol{x}, \boldsymbol{y} \| \epsilon R^n$,

donde $\| \boldsymbol{.} \|$ es una norma. La distancia métrica es un concepto restrictivo, que requiere satisfacer la desigualdad triangular; esto es, para cualquier patrón **x**, **y**, y **z** se tiene que:

$d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \geq d(\mathbf{x}, \mathbf{z})$

Se han desarrollado varias funciones para obtener la distancia métrica entre dos patrones, que implican diferentes visualizaciones de los datos derivadas de su geometría (Abonyi y Feil, 2007; Höppner et al, 2000; Pedrycz, 2005). El uso de estas funciones para cuantificar la distancia, hace posible que a partir de la aplicación de los algoritmos de agrupamiento se revelen tanto las diferentes figuras geométricas, como su tamaño y densidad, tal como se muestra en la figura 2.1. En ésta se observa la relación espacial entre los *clusters*, que puede tomar la forma de subgrupos bien separados, conectados continuamente uno con otro o traslapados (Abonyi y Feil, 2007; Höppner et al, 2000). A continuación se presentan las funciones de distancia de uso más común.

Distancia Minkowski

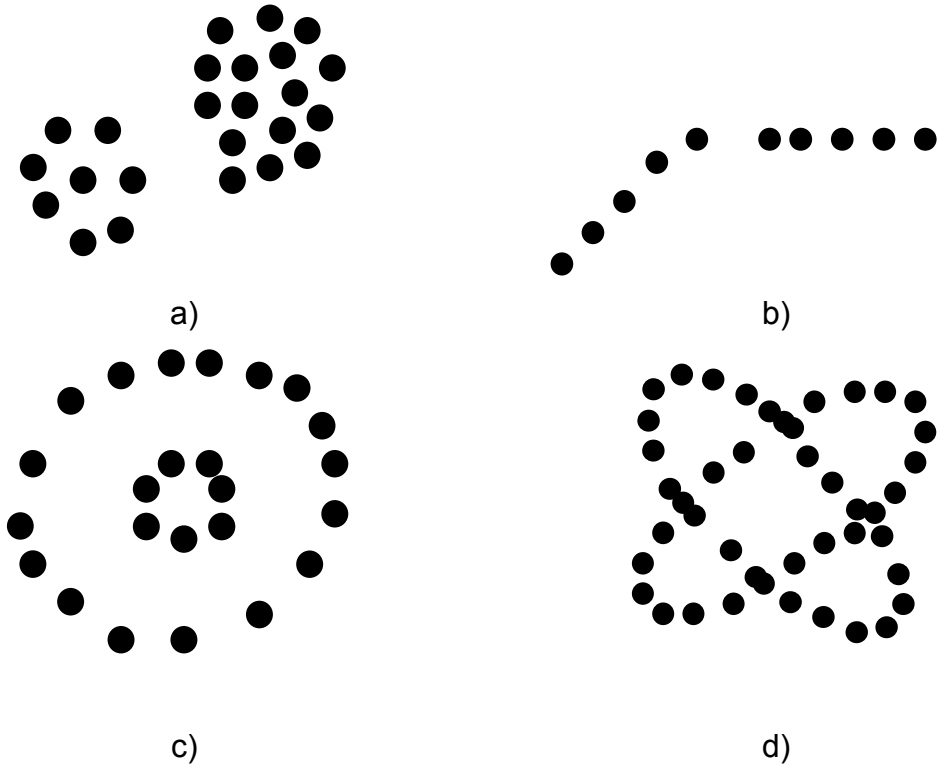
La distancia Minkowski comprende una familia infinita de distancias, incluyendo algunas distancias muy conocidas y comúnmente utilizadas como son Hamming y Euclidiana (Pedrycz, 2005).

Se representa de la siguiente forma:

$$d_{\min}(x, y) = \left[\sum_{j=1}^{d} \left| x_j - y_j \right|^m \right]^{1/m}$$

donde m = 1,2,…,∞. Para el caso de m = 1 se obtiene la distancia Mahattan y cuando m = 2 se obtiene la distancia Euclidiana. Si el conjunto de datos tiene *clusters* aislados o compactos, la distancia Minkowski es adecuada; por otra parte, atributos con grandes valores tienden a dominar a los otros. Para evitarlo, se deben normalizar los atributos o usar esquemas de ponderación (Gan, et al., 2007).

Figura 2.1. *Clusters* con diferente geometría en R^2.



Fuente: Abonyi y Feil (2007).

II.1,2.1. Distancia Euclidiana

La distancia Euclidiana es la medida de similitud entre datos de uso más generalizado, Para dos puntos **x** e **y** en un espacio *d*-dimensional, la distancia Euclidiana entre estos datos se define como:

$$d_{euc}(x,y) = \left[\sum_{j=1}^{d}(x_j - y_j)^2\right]^{1/2} = \left[(x-y)(x-y)^T\right]^{1/2},$$

donde x_j e y_j son los valores del *j*-esimo atributo de ***x*** e **y**, respectivamente.

La distancia Euclidiana favorece el descubrimiento de agrupamientos (*clusters*) de datos de forma geométrica cilíndrica y/o esférica (Pedrycz, 2005, Zhang, 2009). Desde una perspectiva estadística, esto significa que cada *cluster* se generó por una distribución normal con media μ_i y matriz de varianza-covarianza $\Sigma_i = \sigma^2\mathbf{I}$, i) = 1,2,…, K, e $\mathbf{I}$ es la matriz identidad p-dimensional (Zhang, 2009).

II.1.2.2. Distancia Manhattan

La distancia Manhattan, conocida también como Hamming o *city block*, es definida como la suma de las distancia de todos sus atributos. Para dos puntos **x** e **y** en un espacio *d*-dimensional, la distancia Manhattan entre estos datos se define como:

$$d_{man}(x,y) = \sum_{k=1}^{d}\left|x_j - y_j\right|$$

La distancia Manhattan favorece el descubrimiento de figuras geométricas similares a diamantes.

II.1.2.3. Distancia Mahalanobis

La distancia Mahalanobis es una generalización de la distancia Euclidiana que permite trabajar con datos altamente correlacionados, diferentes varianzas y un rango diferente (Vercellis, 2009). Es a través de estas características que se identifican y analizan los diferentes patrones (Marmolejo y González, 2008). Esta distancia es definida como:

$$d_{mah}(x,y) = \sqrt{(x-y)\Sigma^{-1}(x-y)^T}$$

donde Σ es la matriz de covarianza del conjunto de datos.

A través de la selección de la matriz Σ de varianzas-covarianzas, se controla la geometría de los *clusters* potenciales a través de rotar la elipsoide (de las entradas diagonales de la matriz Σ) y cambiando la longitud de los sus ejes (los elementos caen en la diagonal principal de l matriz) (Pedrycz, 2005). A su vez, responde a distancia mínimas (Mitra y Acharya, 2003), lo que le permite da robustez fuerte a problemas de multicolinealidad.

La selección de la medida de similitud o disimilitud depende del tipo de variable, así como de la escala de medición de ésta (Gan et al, 2007). Otros criterios que pueden definir la selección de esta medida, son la relación de dependencia entre las variables, su dispersión y la figura geométrica que se desea revelar.

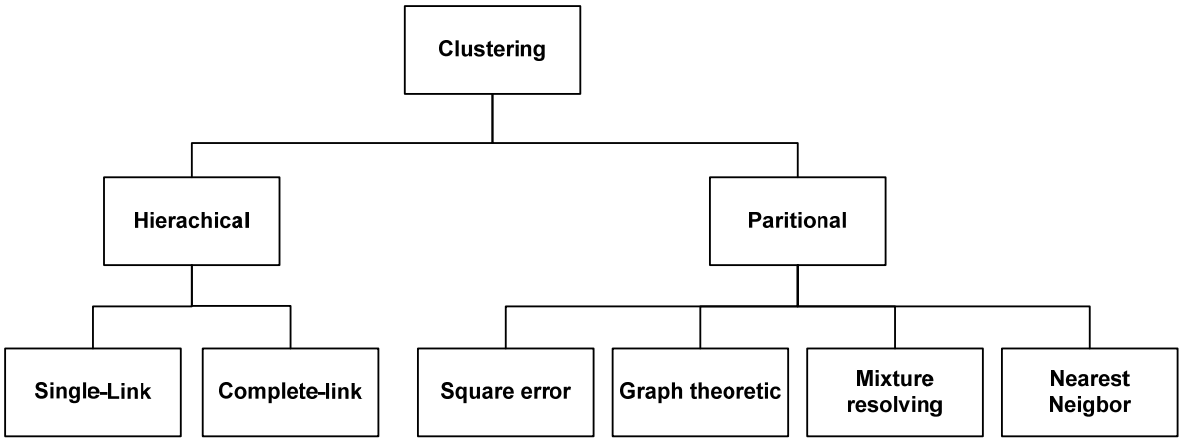
II.2. Algoritmos de agrupamiento

La variedad de técnicas de *clustering* que han sido desarrolladas durante los últimos años, ha dependido de la emergencia de técnicas de optimización, así como de nuevas metodologías y la expansión en las áreas de aplicación (Pedrycz, 2005). Esta diversidad de herramientas de agrupamiento se divide, de acuerdo al método de búsqueda del agrupamiento, en algoritmos jerárquicos, de agrupamiento basado en función objetivo, o algoritmos particionales (Abonyi y Feil, 2007; Bow, 2002; Gan, et al., 2007; Höppner et al., 2000; Pedryckz 2005; Schenker, et al., 2006; Theodroridis y Koutroumbas, 2006).

Abonyi y Feil (2007), presentan una taxonomía de clusters más detallada (figuras 2.2 y 2.3). Hacen una primera clasificación de las herramientas de agrupamiento en: jerárquicas, en la cual coinciden con Pedrycz (2005) y particionales, que dividen o particionan el conjunto de datos en *clusters* sencillos, en lugar de

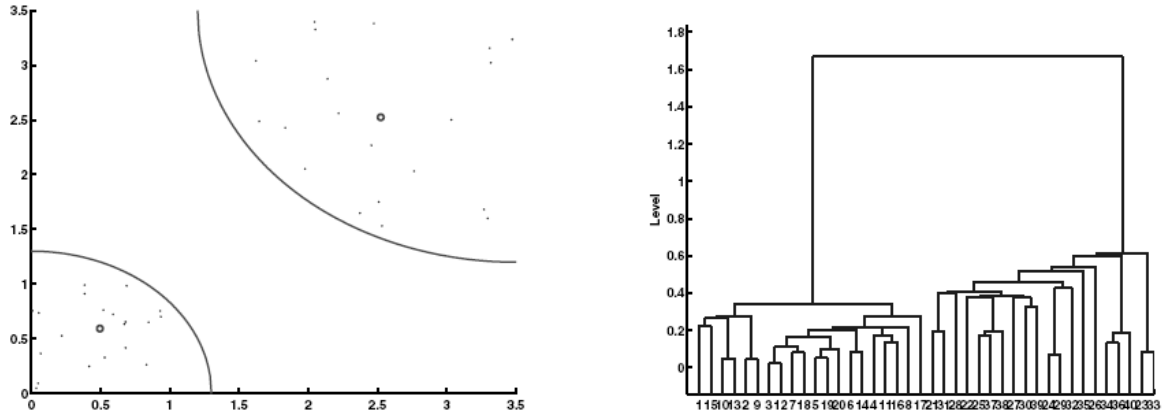
encontrar la estructura del agrupamiento como lo hace el agrupamiento jerárquico. El enfoque particional es dividio en cuatro categorías: la de soluciones mezcladas (*mixture resolving*), que asume que el conjunto de patrones a dividir proviene de uno de varias distribuciones; y su objetivos es identificar los parámetros de cada una de estas distribuciones; el agrupamiento téorico-gráfico, que se basa en la construcción de diagramas de mínima expansión (*minimal spanning tree* –MST-) de los datos (Zhan, 1971), y después eliminan los bordes de MST con las longitudes más largas para generar los *clusters*; el agrupamiento a través del vecino más cercano, partición en la que se utiliza la distancia del vecino más cercano como base de los procedimientos de agrupación. Éste asigna cada patrón no etiquetado al *cluster* del patrón más cercano etiquetado, proveyendo la distancia a la etiqueta más cercana de acuerdo a un umbral previamente definido; El criterio más intuitivo y de uso más general en las técnicas de agrupamiento particional es el cuadrado del error (Abonyi y Feil, 2007); el algoritmo K-Medias es el más sencillo y de uso más común, que emplea el cuadrado del error como criterio de asignación (Duda et al., 2001; Webb, 2002). A continuación se describen brevemente el agrupamiento jerárquico y el agrupamiento paticional, donde se exponen de forma más detallada los algoritmos K-Medias, C-Medias Difuso y Gustafson-Kessel.

Figura 2.2. Taxonomía de los enfoques de agrupamiento.



Fuente: Abonyi y Feil, (2007).

Figura 2.3. Principales categorías de algoritmos de agrupamiento.



a) Agrupamiento particional b) Agrupamiento jerárquico

Fuente: Abonyi y Feil, (2007).

II.2.1. Algoritmos de agrupamiento jerárquico

Los algoritmos de agrupamiento jerárquico forman grupos con base en una secuencia de divisiones anidadas (Bow, 2002). La forma de revelar la estructura de los datos, es a través de la representación gráfica (Duda et al., 2001). La construcción de este gráfico (llamado dendograma –figura 2.3b), puede presentarse de acuerdo a dos estrategias: arriba-abajo (*top-down*) o abajo-arriba (*bottom-up*); este último es conocido como enfoque aglomerativo y es una estrategia que inicia considerando cada elemento como un *cluster* y después, sucesivamente, junta los *cluste*rs más parecidos. El algoritmo junta en cada paso los dos *clusters* más cercanos. El proceso es repetido hasta que cada dato llega a cierto valor de umbral previamente definido.

En la estrategia arrriba-abajo, conocida como enfoque divisivo, se trabaja en el sentido opuesto; Inicia tomando todo el conjunto de datos como un solo *cluster* y se generan las particiones de éste hasta llegar a cada elemento. Dada la naturaleza del proceso, este método es computacionalmente ineficiente, con la posible excepción de patrones con variables binarias (Abonyi y Feil, 2007; Bow, 2002; Theodroridis y Koutroumbas, 2006; Duda et al., 2001; Pedricz, 2005).

La forma de diferenciar los *clusters* formados a partir de agrupamiento jerárquico, es a través del cálculo de la distancia entre estos, la cual puede se cuantificada con base en tres métodos típicos:

Método de liga-simple (*single-link*). La distancia d(A,B) se basa en la distancia mínima entre los patrones pertenecientes a A y B. La cual se calcula como:

$$d(A,B) = \min_{x \in A, y \in B} d(x,y)$$

Método de liga-completa (*complete-link*). Este método esta basado en los dos patrones más lejanos pertenecientes a dos *clusters*. Se calcula como:

$$d(A,B) = \max_{x \in A, y \in B} d(x,y)$$

Método de liga promedio del grupo. En contraste con los enfoques anteriores, donde la distancia se determina con base en los valores extremos de la función de distancia, este método considera el promedio entre las distancias calculadas entre todos los pares de patrones, uno de cada *cluster*. Se calculado como:

$$d(A,B) = \frac{1}{card(A)card(B)} \sum_{x \in A, y \in B} d(x,y)$$

II.2.2. Algoritmos de agrupamiento basados en función objetivo

Este tipo de agrupamiento también llamado agrupamiento particional (Abonyi y Feil, 2007) o funciones criterio para *clustering* (Duda et al., 2001), basa la decisión de asignar patrones a *cluster* en el desempeño de un índice o función objetivo (Pedrycz, 2005). El agrupamiento particional, a diferencia del jerárquicos, revela la estructura del conjunto de patrones a través de la partición en subconjuntos con patrones homogéneos y lo más disímiles posible a patrones de otros subconjuntos (figura 3.2a); así mismo, presenta ventaja al trabajar con grandes conjuntos de datos, para los cuales la construcción de dendogramas del agrupamiento jerárquico es computacionalmente prohibitivo (Abonyi y Feil, 2007). El problema a optimizar es minimizar la distancia entre los patrones de dentro del *cluster* al

tiempo que se maximiza la distancia entre los clusters (Das et al., 2009, Duda et al., 2001; Pedrycz, 2005; Abonyi y Feil, 2007; Höppner et al, 1999).

Höppner y otros (1999), plantean formalmente la función objetivo como:

$J: A(D,R) \rightarrow \mathbb{R},$

Donde el espacio de análisis A, tiene como elementos al espacio de datos *D* ($D \neq \varnothing$) y el espacio de resultados *R*, que es un conjunto de conjuntos ($|R| \geq 2$).

El valor de *J(f)* es entendido como un error o medida de calidad, que deberá ser minimizado y maximizado. En general, la función objetivo deberá utilizarse para comparar diferentes soluciones del mismo problema, permitiendo también la comparación indirecta para soluciones de diferentes conjuntos de datos.

Abonyi y Feil (2007), presentan el algoritmo para el agrupamiento a partir del error cuadrático:

1. Seleccionar una partición inicial de patrones con un número fijo de *clusters* y centros de *cluster*.
2. Asignar cada patrón al centro de cluster más cercano y calcular los nuevos centros de *cluster* como los centroides de los *clusters*. Repetir este paso hasta se alcance la convergencia, es decir, hasta estabilizar la pertenencia al *cluster*.
3. Combinar y dividir los *clusters* a partir de alguna información heurística, repitiendo opcionalmente el paso dos.

Por su naturaleza, las particiones generadas en cuanto a la exclusividad de los patrones que se asignan a ésta, se diferencian en partición dura y partición difusa.

II.2.2.1. Partición dura

La partición dura, también llamada partición *crisp* o partición determinista, basa sus métodos de asignación de patrones en la teoría de conjuntos clásica, en la cual un patrón sólo puede ser asignado o pertenecer a un grupo (Abonyi y Feil, 2007; Höppner et al, 1999). Entre los algoritmos de uso más generalizado esta el algoritmo K-Medias, utilizado por primera vez por James MacQueen en 1967. Donde su propósito principal era describir un proceso para particionar una población *N*-dimensional en *k* conjuntos, con base en una muestra (MacQueen, 1967).

II.2.2.1.1. K-Medias

El algoritmo de agrupamiento duro K-Medias, se, por su sencillez y bajo costo computacional, el método más utilizado para particionar una población de datos N-dimensionales en k muestras. Utiliza la distancia Euclidiana como medida de semejanza para la asignación de datos a los *clusters* en el espacio métrico; fue presentado por primera vez en el 5th Berkeley Symposium on Mathematical Statistics and Probability (MacQuee, 1967).

El algoritmo K-Medias asume que el número de clusters K es conocido. Éste trabaja iterativamente de la forma que se presenta a continuación, donde se usa t para definir el número iteración (Butenko et al., 2009):

Algoritmo K-Medias

1. Seleccionar aleatoriamente K objetos como centros iniciales de los K *clusters*, definidos como:

   $\overline{X}_1^t, \overline{X}_2^t, ..., \overline{X}_k^t; t = 0;$

2. Para cada i, calcular $s \times \left(X_i, \overline{X}_j^t\right)$ o $d \times \left(X_i, \overline{X}_j^t\right), j = 1,2,...,K$. Asignar la etiqueta del *cluster* al objeto i de acuerdo a:

   $$CL_i^t = \arg\max_j \left[s \times \left(X_i, \overline{X}_j^t\right), j = 1,2,...,K\right]$$

   o

   $$CL_i^t = \arg\max_j \left[s \times \left(X_i, \overline{X}_j^t\right), j = 1,2,...,K\right], i = 1,2,...,N$$

   donde CL_i^t define la etiqueta para el *cluster* del objeto i en la iteración t;

3. Actualización de los centros de los *K clusters* a través de:

   $$X_j^{t+1} = \frac{1}{N_{c_j}^t} \sum_{i=1}^{N} I\left(CL_i^t = j\right) X_i, j = 1,2,...,K,$$

   donde I(X) es una función de identificación tal que I(X) = 1 si X es verdadero, y I(X)=0 en caso contrario. El denominador, $N_{c_j}^t$, es el número de objetos asignados al *cluster* j en la t^{th} iteración, dado que $t=t+1$;

4. Si el criterio de convergencia es satisfecho, el algoritmo se detiene. En caso contrario, se regresa al paso dos.

II.2.2.2, Partición difusa

A diferencia de la partición dura, se basa en la lógica binaria clásica, la partición difusa sienta sus bases en la lógica difusa[24] a través del uso de los conjuntos difusos[25] que son utilizados en la expresión de las particiones del conjunto de

[24] La lógica difusa (fuzzy logic) es una lógica multivaluda, que permite valores intermedios a ser definidos entre valores convencionales tales cómo: falso/verdadero, si/no, alto/bajo, etc. Conceptos como “más alto” o “muy rápido” pueden ser formulados matemáticamente y procesados por computadora, para acercarse más la forma del pensamiento humano en la programación (Zadeh, 1984). El objetivo de la lógica difusa es proveer un modelo para el modo de razonamiento, es más aproximado que exacto. Desde esta perspectiva, la importancia de la lógica difusa deriva de que casi todo el razonamiento humano (especialmente el sentido común) es de naturaleza aproximada (Zadeh, 1990).

[25] De acuerdo con Zadeh (1965), los conjuntos difusos (fuzzy sets) son una clase de objetos con un grado de pertenencia continuo, el cual se caracteriza por una función de membresía la cual asigna a cada objeto un grado de pertenencia que puede tomar valores entre cero y uno. Donde las propiedades de los conjuntos clásicos (por partir de la teoría de conjuntos convencional) como

datos. De esta manera, la partición difusa de un conjunto de datos, que puede ser como una generalización de la partición dura (Abyi y Feil, 2007), permite que un dato pueda pertenecera más de un *cluster* simultáneamente, a partir de la asignación de un grado de membresía que asume valores de entre 0 y 1 y que define la pertenencia a cada uno de los grupos a donde fue asignado (Abyi y Feil, 2007; Bezdek, 1981; Höppner et al, 1999; Mirkin, 2005; Kruse et al., 2007).

II.2.2.2, 1. C-Medias Difuso

El algoritmo C-Medias Difuso, fue presentado por primera vez por Dunn en 1973, quien exhibió una versión difusa del algoritmo K-Medias pero fue Bezdek en 1973, quien introdujo en factor de difusidad m. Éste es una extensión del algoritmo de agrupamiento duro K-Medias al ámbito del agrupamiento difuso, que permite reconocer nubes de puntos de forma esférica en un espació p-dimensional a partir de la utilización de la distancia Euclidiana como medida de similitud entre los patrones, además de asumir que los *clusters* formados son del mismo tamaño (Bezdek 1973; Höppner et al, 1999). A continuación se describen las etapas de este algoritmo.

Dado un conjunto de datos $D=\{x_1,x_2,\ldots,x_n\}$, el algoritmo se basa en la minimización de la función objetivo:

---

inclusión, unión, intersección, complemento, relación, etc. son extendidas para este tipo de conjuntos.

Formalmente, sea ***X*** un espacio de puntos (objetos), con un elemento genérico ***X*** identificado por ***x***, tal que, ***X = {x}***.

Un conjunto difuso ***A*** en ***X*** es caracterizado por la función de membresía $f_A(x)$ la cual asocia a cada punto de ***X*** a un intervalo de números reales [0,1], con el valor de $f_A(x)$ de ***x*,** que representa el grado de ***x*** en ***A***. Tal que, si el valor de $f_A(x)$ es cercano a la unidad, ***x*** tendrá un alto grado de pertenencia al conjunto ***A***.

Ejemplo, Sea ***X*** el conjunto de números reales R^1 y sea ***A*** un conjunto difuso de números, los cuales son mayores 1. Entonces, se puede dar una caracterización precisa, aunque sea subjetiva, de ***A*** por medio de la especificación $f_A(x)$ como una función de R^1. Los valores representativos de tal función pudieran ser: $f_A(0)=0;\ f_A(1)=0;\ f_A(5)=0.01;\ f_A(10)=0.02;\ f_A(100)=0.95;\ f_A(500)=1,$ por lo tanto ***A = {(0, 0), (1,0), (5, 0.01), (10, 0.02), (100, 0.95), (500, 1), (1000,1)}***.

$$j_q(U,V) = \sum_{j=1}^{n}\sum_{i=1}^{k} u_{ij}^m d^2(x_j, v_i)$$

Con respecto a U (una k-partición difusa del conjunto de datos) y a v (el conjunto de k-prototipos), donde m es un número real mayor a 1; v_i es centro del cluster i; u_{ij} es el grado membresía del objeto x_j perteneciente al cluster i; $d^2(.,.)$ es un producto interno métrico y k es el número de *clusters*. El parámetro m controla el difusidad del *cluster* resultante (Bezdek, 1981).

1. Seleccionar los centros de cluster iniciales $v_i(i = 1,2,...,k)$;
2. Calcular la matriz de pertenencia de acuerdo a:

$$u_{ij} = \frac{\left[d^2(x_j, v_i)\right]^{-\frac{1}{m-1}}}{\sum_{l=1}^{k}\left[d^2(x_j, v_l)\right]^{-\frac{1}{m-1}}}, i = 1,2,...,k, j = 1,2,...,n;$$

3. Calcular los nuevos centros $\hat{V}_i (i = 1,2,...,k)$, a través de:

$$\hat{V}_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m},$$

y actualizar la matriz de pertenencia (u_{ij}) a $(\hat{u}_{ij})$ de acuerdo a la ecuación del paso 2;

4. Si $\max_{ij}\left|u_{ij} - \hat{u}_{ij}\right| < \varepsilon$, entonces detener; en caso contrario regresar al paso tre, donde ε es un criterio de terminación que toma valores entre 0 y 1.

II.2.2.2.2. Algoritmo Gustafson-Kessel

Gustafson y Kessel (1979), proponen la utilización de una distancia adaptativa en cada *cluster* para la detección de diferentes formas geométricas. A diferencia del planteamiento hipotético de tamaño estandarizado de *cluster* y formas

geométricas cilíndricas o esféricas, sustentado en el uso de la distancia Euclidiana como medida de similitud entre patrones del algoritmos C-Medias difuso (Bezdezk 1973; Dunn 1973), el algoritmo Gustafson-Kessel asocia en cada *cluster* su centro y su matriz de covarianza, a través del uso de la norma de distancia Mahalonobis. Estp le permite identificar clusters con espacios geométricos elípticos o elipsoidales. A partir de la introducción de la constante ρ para cada matriz de covarianza, este algoritmo permite diferentes tamaños de *clusters*; en este sentido, se generan un agrupamiento más exacto (Gustafson y Kessel 1979; Höppner et al., 1999).

Cada *cluster* tiene su propia matriz norma de inducción $\boldsymbol{A_i}$, la cual produce la siguiente norma de producto interno:

$$D^{w}_{imk,A} = (x_k - v_i)^T A_i (x_k - v_i), 1 \le i \le c, 1 \le k \le N$$

Partiendo de que $\boldsymbol{A}$ define la c-tupla de la matriz norma de inducción: $\boldsymbol{A}$= ($\boldsymbol{A_1,A_2,\ldots,A_c}$). La función objetivo del algoritmo Gustafson-Kessel está definida por:

$$J(X;U,V,A) = \sum_{i=1}^{c}\sum_{k=1}^{N} (\mu_{i,k})^m D^2_{i,k,A_i}$$

A continuación se describen las etapas de este algoritmo:

Dado un conjunto de datos X, seleccionar el número de clusters $1 < c < N$, el exponente $m > 1$, el criterio de terminación $\varepsilon > 0$ y la matriz norma de inducción A. Inicializar la matriz de partición aleatoriamente, tal que $\mathbf{U}^{(0)} \in M_{fc}$.

Repetir para l = 1,2,…

1. Calcular los centros de *clusters*.

$$V_i^{(l)} = \frac{\sum_{k=1}^{N} \left(\mu_{i,k}^{l-1}\right)^m x_k}{\sum_{k=1}^{N} \left(\mu_{i,k}^{l-1}\right)^m}, 1 \le i \le c.$$

2. Calcular la matriz de covarianza del *cluster*.

$$F_i^{(l)} = \frac{\sum_{k=1}^{N} \left(\mu_{i,k}^{(l-1)}\right)^m \left(x_k - v_i^{(l)}\right)\left(x_k - v_i^{(l)}\right)^T}{\sum_{k=1}^{N} \left(\mu_{i,k}^{(l-1)}\right)^m}, 1 \le i \le c.$$

3. Calcular las distancias.

$$D_{i,kA_i}^2\left(x_k, v_i\right) = \left(x_k, v_i^{(l)}\right)^T \left[(\rho_i \det(F_i))^{\frac{1}{n}} F_i^{-1} \right] \left(x_k - v_i^{(l)}\right).$$

4. Actualizar la matriz de partición.

$$\mu_{i,k}^{(l)} = \frac{1}{\sum_{j=1}^{c} \left(\frac{D_{i,kA_i}\left(x_k, v_i\right)}{D_{j,kA_j}\left(x_k, v_j\right)} \right)^{\frac{2}{(m-1)}}}, 1 \le i \le c, 1 \le k \le N.$$

hasta $\left\| U^{(l)} - U^{(l-1)} \right\| \le \varepsilon.$

II.2.2.3, Índices de validación de agrupamiento

El agrupamiento de patrones a partir de la utilización de algoritmos particionales, supone el conocimiento del número de particiones que mejor revele la estructura subyacente al conjunto de datos. Sin embargo, cuando los datos no pueden ser representados gráficamente, será muy difícil o casi imposible para el humano observar una determinada partición en los datos (Höppner et al., 1999). Al respecto, Aboyi y Feil (2007), presentan dos estrategias para determinar el número de *clusters* apropiado que mejor revele la estructura del conjunto de patrones: 1) Iniciar con un número grande de *clusters*, y sucesivamente reducir este número juntando los *clusters* que sean compatibles con respecto a un criterio predefinido; 2) Particionar el conjunto de datos para diferentes números de

*cluster*s; utilizar una medida de validación y evaluar la bondad obtenida de la partición.

Para evaluar la calidad de tales pariciones se han desarrollado diferentes indicadores (medidas de validación) que reflejan diferentes características de la partición obtenida. Algunos de estos índices de calidad fueron desarrollados en el ámbito crisp y después extendidos al ambiente difuso; en otros casos el sentido del desarrollo ha sido inverso.

Enseguida se presentan algunos índices de uso más común: Coeficiente de partición, Entropía de clasificación, Índice de partición, Índice de Xie and Beni, Índice de Dunn.

Coeficiente de partición (*PC*). Definido por Bezdek (1981), éste mide la cantidad de traslape entre *clusters*; valores altos para éste relejan menos ambigüedad en la partición, de tal manera que obtener *PC(c) =1*, equivale a tener particiones duras (*clusters* no traslapados). En contraste, si cada dato es asignado a cada *cluster* con el mismo grado de pertenencia se obtiene su valor mínimo, lo que indica máxima ambigüedad en la partición (Höppner et al, 1999). Este índice es calculado por:

$$PC(c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{k=1}^{N}\left(\mu_{i,k}\right)^2$$

La desventaja del *PC* es la carencia de conexión directa a alguna propiedad de los datos (Abonyi y Feil, 2007). El número óptimo de *clusters* se determina en el valor máximo.

Entropía de clasificación (*CE*). Mide sólo la difusidad de la partición y es muy semejante al *PC*. Se basa en la teoría de información de Shannon (Höppner et al., 1999). Es calculado como:

$$CE(c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{K=1}^{N}\mu_{i,k}\ln(\mu_{i,k})$$

Si *PC(c) = 1*, lo que significa que la obtención de una partición dura (*cluters* excluyentes). La entropía (esto es, la información contenida dentro de una fuente), indica que es la partición correcta *(CE(c) = 0*). En el sentido opuesto, su valor será el máximo cuando los patrones se distribuyan en cada *cluster* con el mismo grado de pertenencia. (Bezdek, 1981), probó que este índice cumple con la relación *0 ≤ 1-PC(c) ≤ CE(c)*.

Índice de partición (*SC*). Es la relación entre la suma de la densidad (concentración) y separación de *clusters*. Esto es, la suma individual de la medida de validación normalizada de cada *cluster* divido a través de la cardinalidad difusa de cada *cluster* (Benzaid et al., 1996). Es determinado por:

$$SC(c) = \sum_{i=1}^{c}\frac{\sum_{k=1}^{N}\left(\mu_{i,k}\right)^{m}\left\|x_k - v_i\right\|^2}{\sum_{k=1}^{N}\mu_{i,k}\sum_{j=1}^{c}\left\|v_j - v_i\right\|^2}.$$

Un menor cociente de esta relación nos índica una mejor partición.

Índice de Xie and Beni (*XB*). Los índices indirectos como el *PC* presentan tres desventajas: primero, en el mejor de los casos se relacionan indirectamente con cualquier *cluster* en *X*; segundo, ignoran los parámetros adicionales (tales como el conjunto de centros *cluster V*); tercero, no hacen uso del mismo conjunto de patrones *X*.

Xie y Beni definen un índice de validación de agrupamiento difuso que supera las dos últimas desventajas. Éste se orienta a cuantificar la relación entre la variación total dentro de los *clusters* y la separación de los mismos (Xie y Beni, 1991).

$$XB(c) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{i,k})^m \|x_k - v_i\|^2}{N \min_{i,k}\|x_k - v_i\|^2}.$$

El número óptimo de *clusters* deberá minimizar el valor de este índice.

Índice de Dunn (*DI*). Dunn (1973), originalmente propone este índice para la identificación de "*clusters* compactos bien-separados". Su principal desventaja es el alto costo computacional, sobre todo cuando se incrementa el número de *clusters* (*c*) y el tamaño del conjunto de patrones (*N*).

$$DI(c) = \min_{i \in c}\left\{\min_{j \in c, i \neq j}\left\{\frac{\min_{x \in c_i, y \in c_j} d(x,y)}{\max_{k \in c}\left\{\max_{x,y \in c} d(x,y)\right\}}\right\}\right\}$$

En la práctica, la selección del índice de validación está en función del objetivo del problema a resolver, del nivel de precisión con el que quieran revelarse las agrupaciones, así como de la distribución espacial de los *clusters* formados, que permitirá visualizar el nivel de traslape que presentan éstos.

El *clustering*, como una de las principales actividades desarrolladas por la minería de datos (Hand et al. 2001), puede, que por su grado de generalización, ser aplicado con diferentes propósitos, tales como: compresión de datos, recuperación de información en bases de datos y procesamiento de imágenes. Estas características han permitido el uso del análisis de *cluster* en una diversidad de campos del conocimiento humano.

En la biología, específicamente en estudios genómicos, el análisis de cluster utilizado frecuentemente en la expresión génica; En la ecología, en estudios de impacto ambiental, esta herramienta se ha utilizado para la regionalización de flujos de agua en eventos extremos, y en estudios de cobertura vegetal a partir de imágenes satelitales. En la economía, para encontrar países con indicadores de

desarrollo semejantes; así como en el agrupamiento de empresas con índices de desempeño similares, y en la caracterización de comunidades. En finanzas, para encontrar *clusters* de compañías con desempeño financiero homogéneos, conductas competitivas equivalentes. En marketing, búsqueda de grupos de clientes con comportamiento presupuestal afín, para encontrar grupos de consumidores con preferencias parecidas, canastas de mercados. En cuidado de salud, formación de grupos con síntomas iguales, búsqueda de población en riesgo por hábitos alimenticios similares. En salud mental, para la caracterización de perfiles psicológicos determinados como por ejemplo en pandillas, en estudiantes, en trabajadores, entre otras aplicaciones.

Capitulo III Materiales y métodos

En este capitulo se presenta la estructura metodológica utilizada como eje rector del trabajo empírico de esta investigación. Se describen primeramente el proceso de producción de uva de mesa y su relación con el comportamiento climático, a partir del desarrollo fisiológico de la vid expresado a través de las etapas fenológicas durante el ciclo productivo. Específicamente, se estudia el comportamiento de las etapas de brotación y floración; posteriormente se presentan la localización geográfica del área del predio del cultivo. Así como la estación agroclimática a la cual corresponden los datos analizados, así mismo, se presentan las generalidades de los sensores integrados en dicha estación; finalmente se expone la forma en que se utilizaron los métodos y herramientas analíticas aplicadas para la búsqueda del agrupamiento de las características del clima estudiadas y su relación con las etapas fenológicas de la vid de mesa observadas.

III.1. Desarrollo fenológico de la uva de mesa

El ciclo anual de producción de uva de mesa, puede ser descrito a través de la variación estacional del clima y la repetición anual del desarrollo vegetativo de las plantas de vid. De manera sucinta, el clima se define como la expresión acumulada del movimiento regular diario de la atmosfera (Trewartha y Horn, 1980); éste sigue un complejo comportamiento aleatorio dependiente de la posición geográfica donde se encuentra el territorio bajo estudio (Coonors y Loomis, 2002), que influye directamente en la producción agrícola. En tanto la repetición de los procesos biológicos de las plantas expresados con la aparición de ciertos órganos como yemas, hojas, flores y frutos, además de su relación con el comportamiento de ciertos elementos del clima, es estudiado por la fenología (Noreno, 1990).

La fenología puede ser utilizada como herramienta para determinar puntos de referencia cronológicos que permitan evaluar el desarrollo del proceso de producción de uva de mesa con el propósito de hacer una planeación eficiente de los recursos humanos, materiales y financieros orientados a elevar la productividad y competitividad del sistema de producción de uva de mesa.

A partir de la fenología del cultivo, es posible predecir la aparición de ciertos fenómenos fisiológicos en las plantas para identificar los requerimientos climáticos, así como los periodos críticos de sensibilidad a ciertos elementos del clima; planear los requerimientos materiales y culturales del cultivo; así como, prever la presencia de insectos para planear acciones que permitan controlar las plagas. Así, los datos e información fenológica brindan un soporte primordial al proceso de toma de decisiones en cada una de las fases del proceso de desarrollo de la uva de mesa.

Para la identificación de las etapas fenológicas, se utilizo el sistema de Eichorn y Lorenz modificado por Coombe (1995). A partir de este modelo se seleccionaron de las etapas de brotación e inflorescencia, las fases: 5 (brotes reventados, punta de hojas visible), 7 (primera hoja separada de la punta del brote), 9 (dos o tres hojas separada, brote de 2-4 cm), 12 (5 hojas separadas e inflorescencia claramente visible, largo del brote cercano a los 10 cm.), 17 (12 hojas separadas, inflorescencia bien desarrollada, flores individuales separadas); así como las fases de floración: 21 (30 % de los capuchos caídos), 23 (floración completa, 50% de los capuchos caídos, 17-20 hojas separadas) y 25 (80% de los capuchos caídos).

Se registraron la clave correspondiente a la fase fenológica, así como la fecha en que ésta se presentó.

III.1.2. Localización geográfica y temporalidad del estudio

Los datos climáticos se obtuvieron de la estación climática “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)[26], situado en las coordenadas Latitud N 29º17’15’’, Longitud W 110º55’21’’.

Los datos fenológicos registrados fueron proporcionados por el responsable del viñedo y 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. La estructura de estos datos se presenta en el apartado de metodología.

III.1.3. Generalidades de la estación agroclimática

Los datos climáticos obtenidos del SIA, son recolectados a través de una estación remota ADCON A733 ADDWAVE, la cual cuenta con un conjunto de sensores (descritos en el cuadro 3.1). La figura 3.1 muestra el proceso de transferencia de las mediciones de los elementos del clima hasta la computadora del investigador. 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.

Los datos fenológicos utilizados 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.

[26] Sistema de Información Agroclimática (SIA), Fundación Produce Sonora-PIEAES-INIFAP, 2004, http://www.agroson.org.mx/.

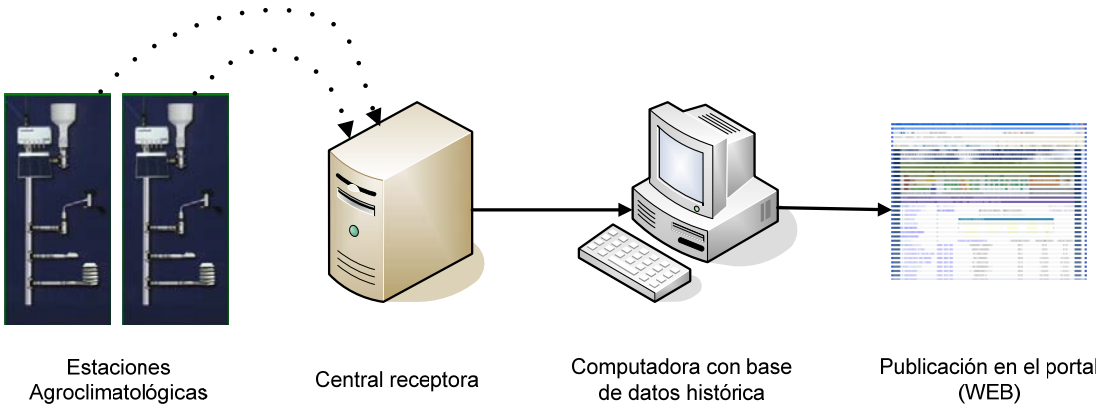
La estructura de las bases de datos climáticos y fenológicas es explicada en el apartado de metodología.

Cuadro 3.1. Características de la estación agroclimática.

Sensor	Unidad de medición	Rango de medición	Exactitud
Temperatura	°C	-39.8°C a +60°C	± 0.6°C
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/m2	± 0.15%
Velocidad del Viento	m/s	0.5 a 55.55	± 2%
Dirección del Viento	°	0 a 360°	± 2%

Fuente: Elaborado con base en Adcon Telemetry (2010).

Figura 3.1. Flujo de mediciones de elementos del clima a datos climáticos.



Fuente: Elaboración propia.

III.2. Metodología

El procedimiento metodológico seguido en este trabajo (figura 2.1), es una adaptación del modelo metodológico propuesto por Fayyad y otros (1996), en el cual la búsqueda del agrupamiento de patrones (minería de datos) y la validación

y expresión del conocimiento descubierto, se aplica a través del procedimiento del reconocimiento de patrones (figura 2.2).

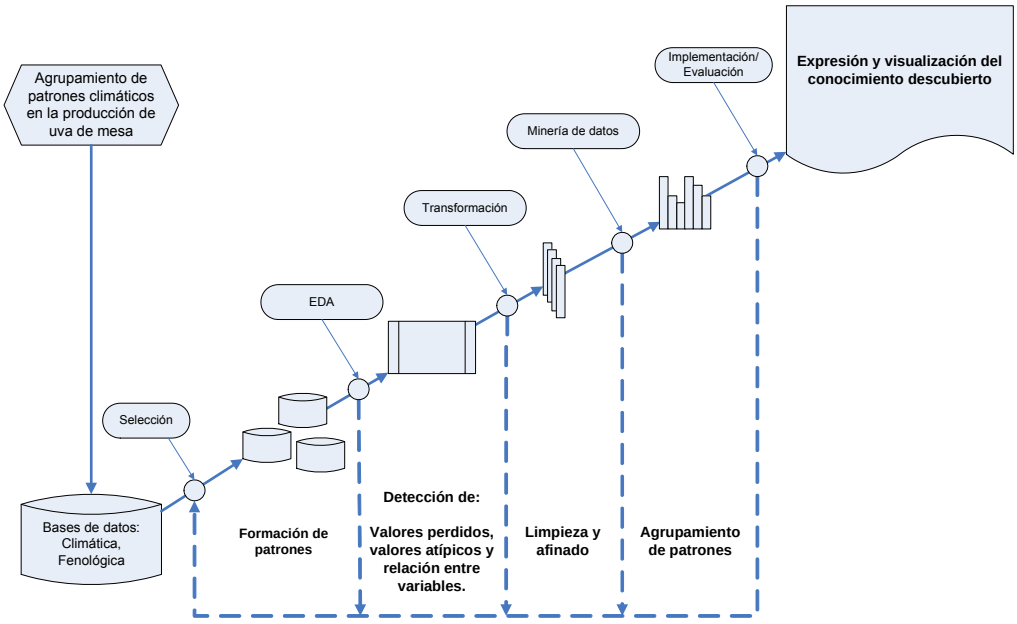
III.2.1. Bases de datos

Las bases de datos climáticos y datos fenológicos, como se ha señalado, provienen de fuentes de naturaleza muy diferente. Mientras que el proceso de medición y el almacenamiento de los elementos del clima son totalmente automatizados a través de una estación de monitoreo remota, los datos fenológicos provienen de la observación y registro directo del humano, en este caso el responsable del viñedo.

De los registros contenidos en la bitácora diaria de campo del administrador del viñedo y del sistema de identificación de las etapas fenológicas de Eichorn y Lorenz modificado por Coombe (1995) se derivó la estructura de la base de datos fenológicos mostrada en el cuadro 3.2.

La estructura de la base de los meta datos climáticos disponible para el usuario en el portal del SIA, derivada de los sensores incluidos en la estación remota y software que genera y administra esta base de datos (addVANTAGE) se muestra en el cuadro 3.3; el dominio de cada una de estas mediciones depende directamente del rango de medición de cada uno de los sensores la cual se mostró en el cuadro 3.1.

Figura 3.2 Adaptación del KDD.



Fuente: Elaboración propia, con base a Fayyad et al. (1996)

Cuadro 3.2 Estructura de los datos fenológicos.

Datos fenológico

Campo	Descripción	Tipo
Ciclo	Ciclo de producción del viñedo	Texto
Etapa	Clave de la fase fenológica observada	Numérica
FechaInicio	Fecha de inicio de la fase fenológica	Fecha
FechaTermino	Fecha de término de la fase fenológica	Fecha

Fuente: Elaboración propia con base en SIA, 2004

Cuadro 3.3 Disponibilidad de mediciones de elementos del clima en SIA.

Datos climáticos horarios

Sensor	Descripción	Unidad
tp1_5	Temperatura promedio a 1.5 metros	Grados Celsius
tmax1_5	Temperatura máxima a 1.5 metros	Grados Celsius
htmax1_5	Hora de mayor temperatura a 1.5 metros	Hora-Min-Seg
tmin1_5	Temperatura mínima a 1.5 metros	Grados Celsius
htmin1_5	Hora de menor temperatura a 1.5 metros	Hora-Min-Seg
hrp1_5	Humedad relativa promedio a 1.5 metros	Porcentaje
hrmax1_5	Humedad relativa máxima a 1.5 metros	Porcentaje
hhrmax1_5	Hora de mayor humedad relativa a 1.5 metros	Hora-Min-Seg
hrmin1_5	Humedad relativa mínima a 1.5 metros	Porcentaje
hhrmin1_5	Hora de menor humedad relativa a 1.5 metros	Hora-Min-Seg
Pvp	Presión de vapor promedio	Kilo Pascales
Pvmax	Presión de vapor máxima	Kilo Pascales
Hpvmax	Hora de mayor presión de vapor	Hora-Min-Seg
Pvmin	Presión de vapor mínima	Kilo Pascales
Hpvmin	Hora de menor presión de vapor	Hora-Min-Seg
Dpv	Déficit de presión de vapor	Kilo Pascales
Rs	Radiación solar	Kilowatt/metro2
Rsmax	Radiación solar máxima	Kilowatt/metro2
Hrsmax	Hora de mayor radiación solar	Hora-Min-Seg
Eto	Evapotranspiración	Milímetros
Vv	Velocidad del viento	Metros/seg
Dv	Dirección promedio del viento	Grados
Dsdv	Desviación estándar de la dirección del viento promedio	Grados
Vvmax	Velocidad del viento máxima	Metros/seg
Hvvmax	Hora de mayor velocidad del viento	Hora-Min-Seg
Ll	Precipitación pluvial (Lluvia)	Milímetros

Fuente: Elaboración propia con base en SIA, 2004

III.2.2. Selección

De la estructura de datos climáticos horarios disponibles se seleccionaron los datos mostrados en el cuadro 3.4. 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 6744 patrones distintos.

Cuadro 3.4 Conjunto de datos utilizado.

Datos fenológicos y climáticos

Campo	Descripción	Tipo
Ciclo	Ciclo de producción	String
Fase	Fase fenológica	String
Etapa	Etapa fenológica	String
Fecha	Fecha del día de observación de la fase fenológica	Fecha
Hora	Hora del día relacionada a las mediciones de los elementos climáticos	Numérico
Temp	Temperatura promedio por hora	Numérico
Hrelativa	Humedad relativa promedio por hora	Numérico
Presion	Presión del vapor de agua promedio por hora	Numérico
Rsolar	Radiación solar promedio por hora	Numérico
Vviento	Velocidad del viento promedio por hora	Numérico
Dviento	Dirección del viento promedio por hora	Numérico

Fuente: Elaboración propia.

De este conjunto de datos (fenológicos y climáticos), se seleccionaron las características Temp, Hrelativa, Presión, Rsolar, Vviento y Dviento. Para formar el conjunto de patrones climáticos, físicamente se almacenó en una tabla de datos, manipulada por DBMS Visual Foxpro 9.0. Es a partir de esta base de datos p-dimensionales que se desarrolló el proceso de descubrimiento de conocimiento.

III.2.3. Análisis exploratorio de datos (EDA)

Se aplicaron técnicas exploratorias como análisis de frecuencias, análisis de tabulación cruzada y herramientas visuales como histogramas y diagramas de dispersión, para asistir a la identificación de valores atípicos (outliers), valores perdidos y posibles relaciones entre variables. Esto permitió hacer el primer acercamiento a la estructura del conjunto de patrones.

Los valores atípicos y valores perdidos identificados, se suavizaron y calcularon a través de la aplicación del método de interpolación lineal, que permite estimar el valor de una coordenada desconocida a partir del conocimiento de su coordenada correspondiente, además del conocimiento de las coordenadas anterior y posterior.

Dadas las coordenadas

$$(x_0, y_0), (x_1, y_{est}), (x_2, y_2)$$

Donde *x* e *y* son el par de coordenadas, el subíndice *est* representa el valor de la coordenada *y* que será estimada, y los subíndices numéricos indican los tiempos de la valuación de *x* e *y*.

La estimación de la coordenada desconocida esta definida por:

$$y_{est} = y_0 + \frac{(x_0 - x_1)(y_0 - y_1)}{(x_0 - x_2)}$$

III.2.4. Transformación de datos

Una vez validado el conjunto de patrones de valores extremos y completados sus valores perdidos, estos fueron normalizados a un rango de valores entre 0 y 1,

para reducir o eliminar el sesgo provocado por la diferencia de escalas en los valores de cada uno de los atributos (características) que componen cada patrón climático, así como para cubrir los requerimientos de las herramientas utilizadas en la etapa de minería de datos.

Para la estandarización de los datos, se utilizó el método de normalización Mín-Máx (Myatt, 2007), el cual utiliza la siguiente formula:

$$x_{est} = \frac{(x - x_{\min})}{(x_{\max} - x_{\min})}$$

III.2.5. Minería de datos

La estrategia de minería de datos fue aplicada utilizando el reconocimiento de patrones como eje rector, tal como se muestra en la figura 2.2. Específicamente fue, dentro de la etapa de procesamiento, donde a través de un ciclo iterativo de diez repeticiones se aplicaron tres herramientas incluidas dentro la minería de datos: los algoritmos de agrupamiento de patrones determinista k-medias y los algoritmos de agrupamiento difuso c-medias difuso y Gustafson-Kessel.

El agrupamiento del conjunto de patrones o etiquetado de patrones se basa en aplicación de los algoritmos de agrupamiento. Éstos utilizan, para decidir la pertenencia de un patrón a un grupo específico, un criterio de semejanza o medida de similitud (disimilitud) que por lo general es reflejado por el cálculo de la distancia p-espacial entre patrones de p-dimensiones.

Existen diferentes ecuaciones matemáticas para el cálculo de esta distancia. Para el caso del algoritmo K-medias y c-medias difuso la medida de similitud se derivó a través de la distancia Euclidiana, la cual es definida como:

$$d_{euc}(x,y) = \left[\sum_{j=1}^{d}(x_j - y_j)^2\right]^{1/2} = \left[(x-y)(x-y)^T\right]^{1/2},$$

donde x_j y y_j son los valores del *j*-esimo atributo de ***x*** y **y**, respectivamente.

Mientras que para el caso del algoritmo de agrupamiento difuso Gustafson-Kessel, se utilizó la distancia Mahalanobis como norma de similitud/disimilitud ente patrones, misma que es definida como:

$$d_{mah}(x,y) = \sqrt{(x-y)\Sigma^{-1}(x-y)^T},$$

donde Σ es la matriz de covarinza del conjunto de datos.

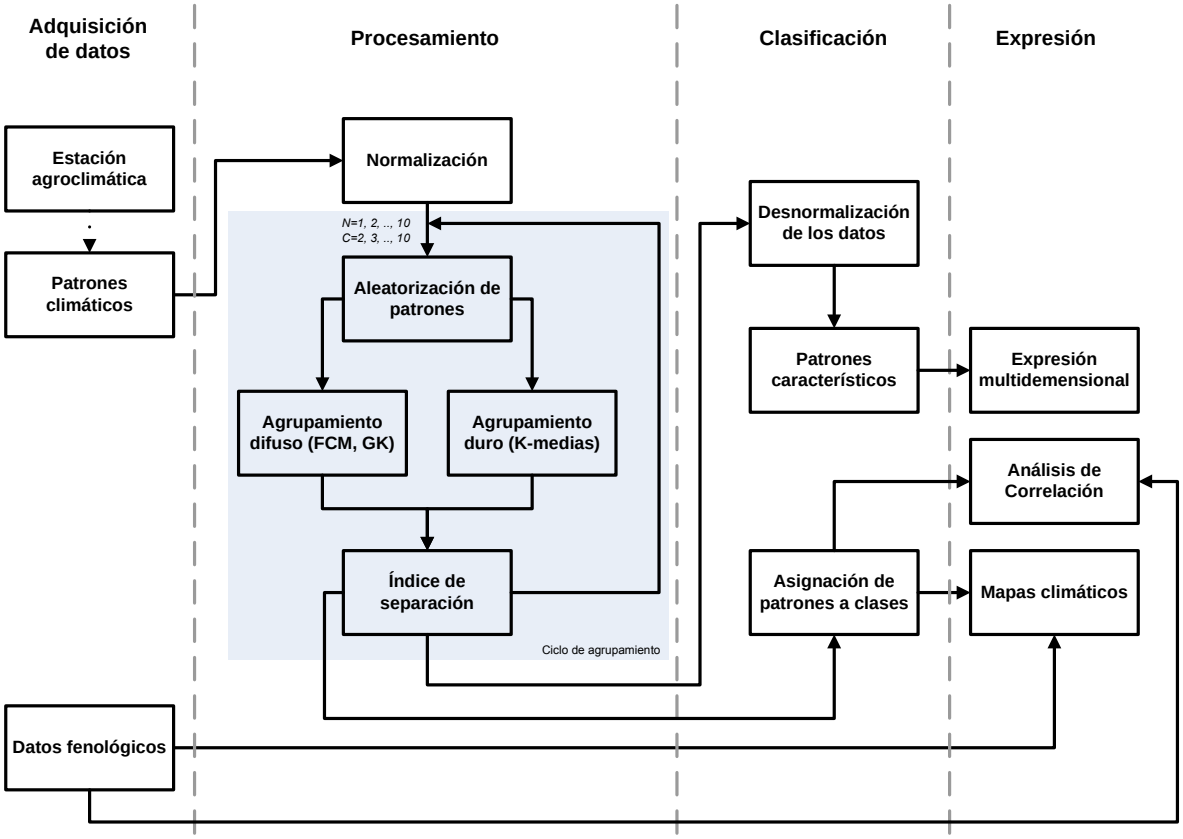
Los algoritmos k-medias, c-medias difuso y Gustafson-Kessel al ser no jerárquicos requieren como parámetro de entrada el número de clases o grupos a formar; este parámetro define el número de prototipos iniciales que son seleccionados aleatoriamente, a partir de los cuales se busca obtener grupos que minimicen la suma del cuadrado de las distancias de los centros de *cluster* (prototipos finales) para cada uno de los miembros pertenecientes a dicho grupo.

El agrupamiento de los patrones climáticos es mostrado en la etapa de procesamiento (figura 3.3). En ésta se muestra el ciclo de agrupamiento, que se ejecutó en diez ocasiones (*N*), donde en cada ciclo los algoritmos de agrupamiento se realizo nueve veces (*C*) las cuales representan el parámetro de número de *clusters* requerido por este tipo de algoritmos de agrupamiento. La idea de ordenar aleatoriamente el conjunto de patrones antes de realizar la formación de grupos es eliminar el sesgo que pudiera causar la elección de los prototipos iniciales en la aplicación de tales algoritmos.

Cabe aclarar que para el caso de los algoritmos de agrupamiento difuso FCM y GK, que son extensiones del algoritmo k-medias en el ambiente difuso (Höppner

et al., 1999), además del parámetro del número de *clusters* (*C*), se requiere la entrada de los parámetros factor de difusidad (*m*) y el criterio de terminación (*ε*), los cuales reflejan la difusidad cercana a los límites de los *clusters* y la distancia entre iteración e iteración en el proceso de agrupamiento, respectivamente (Höppner et al., 1999; Chou et al., 2004). El criterio de terminación o tolerancia de terminación constituye el umbral de referencia para que el algoritmo deje de agrupar datos. Para este caso estos parámetros se fijaron de la siguiente manera: *C = 1,2,..,10; m = 2* y *ε=0.000001.*

Figura 3.3 Minería de datos, a través del reconocimiento de patrones.



Fuente: Elaboración propia.

La determinación del mejor número de clases formadas (grupos o particiones) se observó en el cálculo del índice de separación (S), que es una adaptación del índice Xie-Beni (Siddeheswar y Rose, 1999; Chou et al., 2004), propuesto por

Siddeheswar y Rose (1999), para la determinación del número de *clusters* en agrupamiento k-medias. Este índice busca la menor relación de la distancia dentro del *cluster* y la distancia minima entre los *clusters*, definida por:

$$dentro = \frac{1}{N}\sum_{i=1}^{K}\sum_{x \varepsilon c_i} \left\| x - z_i \right\|^2$$

donde *N* es el número de patrones del conjunto, *K* es el número de *clusters*, y z_i es el centro del C_i.

$$entre = \min\left(\left\| z_i - z_j \right\|^2\right)$$

i=1, 2,…, K-1, j=i+1,…, K

Validación =dentro/entre, que para este caso será identificada también como índice de separación (S).

El número óptimo de agrupamientos difusos generados se validó a través del índice de separación (S), que es definido para el caso de este tipo de agrupamiento como:

$$S(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}\left(\mu_{ij}\right)^2 \left\| x_j - v_i \right\|^2}{N \min_{i.k}\left\| v_k - v_i \right\|^2}$$

donde μ_{ij} es el grado de pertenencia de datos x_j al grupo difuso C_i, y es uno de los elementos de la matriz de partición $U = [\mu_{ij}]$ de dimensión *cxn*. $V=(v_1, \ldots, v_n)$ que representa el conjunto de centroides para cada uno de los *clusters* generados.

Una vez obtenida la mejor partición del conjunto de patrones, los datos p-dimensionales que forman los centros de *clusters* se transformaron nuevamente

para eliminar el escalamiento generado por la normalización y regresar los patrones a su escala original.

A partir de los patrones correspondientes a los centros de *cluster* en sus escalas originales, se promediaron centros correspondientes a las diez formas aleatorias en que fue ordenado el conjunto de patrones antes de ser agrupado, para obtener los patrones característicos que explican la estructura del conjunto de patrones.

La asignación de patrones a los grupos formados para el caso del agrupamiento obtenido con el algoritmo k-medias, se realizó automáticamente por el software estadístico, mientras que para el caso del agrupamiento difuso, la asignación de los patrones a las clases se realizó de acuerdo a la clase que presenta el mayor grado de pertenencia.

La expresión de los patrones característicos obtenidos a partir de los algoritmos de agrupamiento se muestra a través de herramienta de visualización de datos multidimensionales, gráficos de coordenadas paralelas. Así mismo, para exponer la relación entre el conjunto de patrones ya etiquetados con el desarrollo vegetativo de la uva de mesa (datos fenológicos), se crearon los mapas climáticos, que son una adaptación de la visualización de patrones recursivos dentro de la exposición de densidad de pixeles (Keim, 2002). Tales mapas permiten visualizar el comportamiento de los elementos climáticos (datos p-dimensionales) hora por hora durante todo el periodo de desarrollo productivo estudiado. Finalmente, a partir del conjunto de patrones resultante de esta relación, se realizó un análisis de correlación para mostrar la relación entre la presencia de los patrones climáticos y la duración de las fases fenológicas.

Capítulo IV. Resultados y discusión

El objetivo de este capitulo es abordar la relación entre el desarrollo fenológico de la uva de mesa y el comportamiento horario de los elementos climáticos durante el desarrollo de las etapas fenológicas de brotación, inflorescencia y floración, de tal manera que se genere información y conocimiento útil, que brinde un soporte más robusto al proceso de toma de decisiones en la producción.

Así, se presentan los resultados obtenidos a partir de la propuesta metodológica expuesta en el capitulo anterior. La cual muestra las etapas desarrolladas para la transformación de datos a información y conocimiento, orientados a logro del objetivo arriba mencionado. El capitulo se divide en cuatro secciones orientadas al cumplimento de los objetivos específicos. En la primera parte se exponen las características generales del conjunto de patrones (base de datos), utilizado en esta investigación; se continúa con el análisis exploratorio de datos, con el propósito de identificar algún comportamiento atípico y las relaciones que presentan los elementos clima estudiados; como parte medular del trabajo se muestran los resultados obtenidos a través de la implementación de los algoritmos de agrupamiento; finalmente se presenta el conocimiento descubierto a partir de relacionar los datos climáticos agrupados y los fenológicos, los cuales son expresados a través de técnicas de visualización.

IV.1 Base de datos

El conjunto de patrones climáticos incluye los siguientes elementos del clima: temperatura (T), humedad relativa (HR), presión de vapor (PV), radiación solar (RS), velocidad del viento (VV) y dirección del viento (DV). Consta de 6,744 patrones (registros) que representan el valor promedio por hora de las mediciones registradas por los sensores de la estación agroclimática correspondientes a cada una de las variables climáticas.

La distribución de patrones por etapa fenológica y ciclo productivo, se presenta en el cuadro 4.1. En éste se observa que las etapas 10, 23 y 25, son en promedio, los de mayor duración (días); por otra parte, las etapas 09, 07, 21 y 05 presentan muy baja dispersión, mientras que el resto presenta un coeficiente de variación considerable, siendo más marcada la dispersión en las etapas 10, 23 y 14.

De acuerdo al total de registros de patrones climáticos que se almacenan anualmente por estación agroclimatológica (525600 registros), quedan muchos datos en espera de ser trasformados en conocimiento, lo cual es consistente con lo planteado por Tang y McLennan (2005) y Witten y Frank 2005. Además, que este tipo de datos no están orientados a un análisis estadístico especifico (SAS Institute Inc., 2002; Berry y Linoff, 2005).

Cuadro 4.1 Distribución patrones por ciclo productivo y etapa fenológica.

	Etapa fenológica									
Ciclo	**05**	**07**	**09**	**10**	**12**	**14**	**21**	**23**	**25**	**Total**
C01-02	168	168	168	168	144	72	144	336	240	1608
C02-03	168	144	168	168	144	144	192	336	120	1584
C03-04	120	144	168	168	336	168	168	168	312	1752
C04-05	192	144	192	312	96	216	192	144	312	1800
Total	648	600	696	816	720	600	696	984	984	6744
Promedio	162	150	174	204	180	150	174	246	246	
D. estándar	30.20	12.00	12.00	72.00	106.43	60.00	22.98	104.38	90.60	
C.V.	18.64	8.00	6.90	35.29	59.13	40.00	13.21	42.43	36.83	

Fuente: Elaboración propia.

IV.2 Análisis exploratorio del conjunto patrones

El análisis exploratorio del conjunto patrones se realizó a través de la estrategia *drill-down*[27], ésta permite descubrir la localización de las fuentes de variación de los datos, al indagar los datos desde lo general (resumen global) a lo particular (datos primarios). El resumen del primer acercamiento al conjunto de patrones, se expone en el cuadro 4.2. En éste se presentan tanto estimadores de tendencia central (media y la mediana), como estimadores de dispersión como la desviación estándar, varianza, rango y coeficiente de variación de Pearson.

En este primer acercamiento al comportamiento histórico de cada uno de los atributos del clima considerados en cada patrón, se observó que para el caso de T, más del 50% de los patrones varia en el rango de -0.4 a 17°C, mientras que la segunda mitad[28] de éstos varia en el rango de 17 a 37.50 °C. A partir de los valores de tendencia central y el rango de variación se observa una asimetría cargada al lado derecho, tal como se muestra en la figura 4.1a, además de una dispersión moderada de acuerdo al valor del coeficiente de variación de 40.49%. Para el caso de HR, se observan mayormente valores de humedad relativa bajos de acuerdo con la mediana (37.3 %), con un rango de variación que va desde una humedad relativa casi nula (3.0 %) hasta una saturada (97.80 %); se observó una asimetría cargada hacia alta humedad, como se muestra en la figura 4.1b. Esto refleja que la segunda mitad de los patrones está dispersa en un rango más amplio, entre el 37.3 % y el 97.8 %, con una dispersión total considerable de acuerdo con el C.V. igual al 57.67 %.

---

[27] El concepto "*Drill-Down*" de acuerdo a "*SAP for MIT Documentation on Web*", es moverse de un dato resumido a un mayor nivel de detalle. Esto es, desde un dato agregado pasar a los datos que dan soporte a dicho dato agregado. MIT, consultado el 17 de Junio del 2010 en http://web.mit.edu/sapr3/docs/webdocs/glossary/glCD.html#D.

[28] Con el propósito de facilitar la comprensión de la lectura, siempre que se utilicen los términos, primera mitad, segunda mitad, primera parte, segunda parte, o cualquier término que haga referencia a la posición dentro de un intervalo, deberá entenderse que el conjunto de patrones se ha ordenado de forma ascendente de acuerdo al(los) atributo(s) o variable(s) que se este presentando.

Cuadro 4.2 Características generales de los elementos incluidos en los patrones climáticos.

Variable	N	Media	Mediana	Desviación estándar	Varianza	Mínimo	Máximo	Rango	C. V.
T	6744	17.356	16.300	7.0266	49.373	-0.4	37.50	37.90	40.49
HR	6744	42.129	37.300	24.2950	590.247	3.0	97.80	94.80	57.67
PV	6744	0.747	0.700	0.3078	0.095	0.1	1.70	1.60	41.20
RS	6744	0.190	0.035	0.2505	0.063	0.0	0.87	0.87	131.84
VV	6744	0.920	0.700	0.807	0.652	0.0	8.00	8.00	87.72
DV	6744	226.650	244.000	81.278	6606.133	0.0	354.50	354.50	35.86

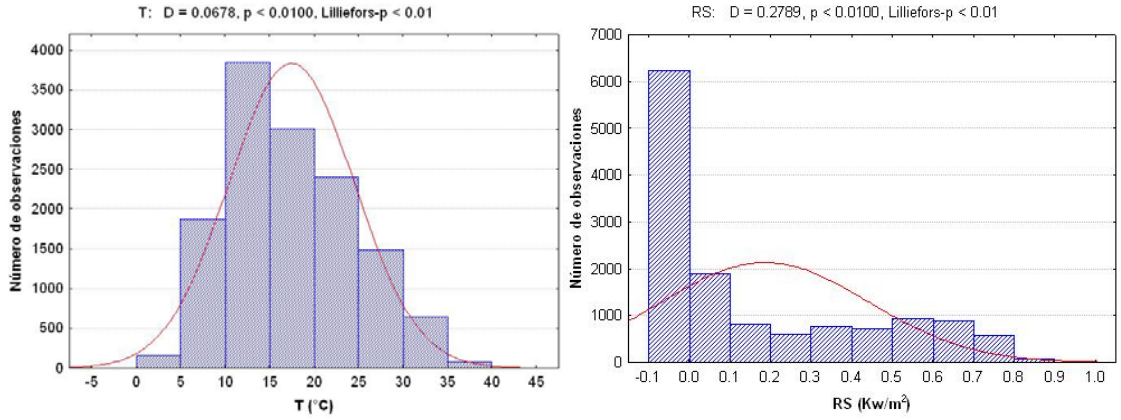
Fuente: Elaboración propia.

Para el caso de la PV, se presenta un mejor nivel de simetría, ligeramente cargado al lado derecho (figura 4.1c), que muestra una tendencia central alrededor de los 0.7 kilo-pascales de acuerdo al valor de la mediana; la segunda mitad del conjunto de patrones, está distribuida en rango más amplio en 0.7 y 1.7 kilo-pascales, así mismo, su dispersión es considerable al presentar un C.V. de 41.20 %, La RS muestra un comportamiento asimétrico demasiado sesgado hacia el lado derecho (figura 4.1d); en ésta, la distribución de la primera mitad de los patrones se encuentra en el primer 24.85 % de su rango de variación (mediana igual a 0.35 y rango entre 0.87 kw/m^2)[29], lo que refleja la mayor dispersión (C.V. igual a 131.84 %).

Respecto a VV que presenta una alta dispersión (C.V. igual a 87.72 %), presenta asimetría cargada hacia la derecha. Además, la distribución de

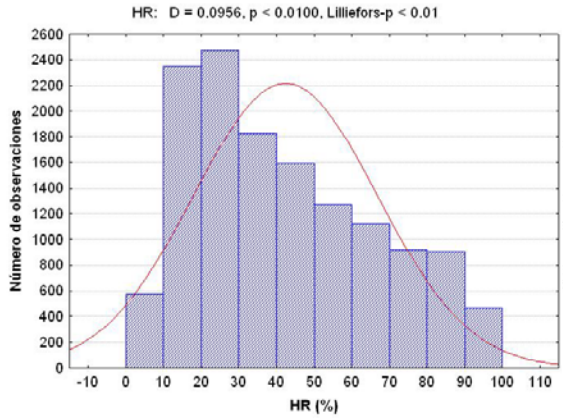
[29] Cabe aclarar que al incluir en el estudio los veinticuatro patrones generados al día, aproximadamente la mitad corresponden a lecturas nocturnas donde la radiación solar es cero, además de acuerdo al periodo del estudio (finales de febrero principios de abril), en el hemisferio norte corresponde a finales del invierno e inicio de la primavera, periodo durante el cual los días comienzan ser más largos que las noches.

Figura 4.1 Distribución de frecuencias de los elementos del clima incluidos en los patrones climáticos

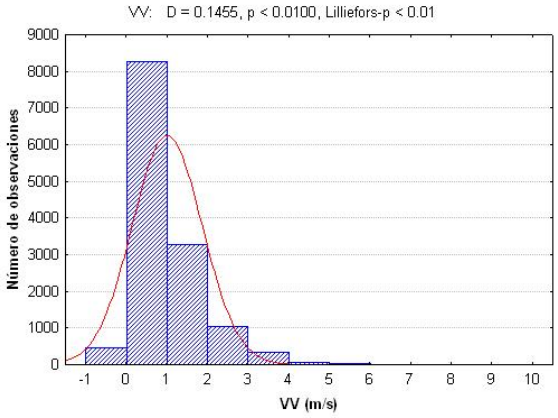


a)

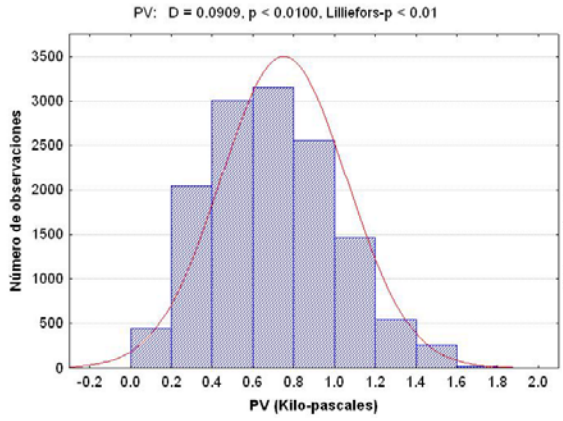
b)



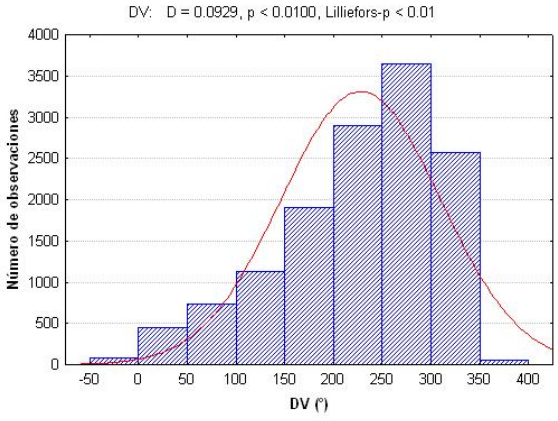
c)



d)



e)



f)

Fuente: Elaboración propia.

frecuencia mostrada en la figura 4.1d, sugiere la presencia de valores atípicos (outliers). En relación a la DV, el rango de variación indica que hubo presencia de viento en todas las direcciones (rango de 0 a 354°). Sin embargo, los valores de tendencia central, muestra asimetría cargada hacia la izquierda y la distribución de frecuencias lo que indica un predominio de vientos hacia el sur y sureste (figura 4.1f); finalmente, ésta es la variable que presenta menos dispersión de acuerdo a su coeficiente de variación igual a 35.86 %.

En este primer acercamiento al conjunto de patrones, se aplicó la prueba Kolmogorov-Simirnov a cada una de las variables dentro de este conjunto, que arrojó que en ningún caso hay evidencia para aceptar que los valores de estas variables siguen una distribución normal (figura 4.1).

IV.2.1. Detección de valores anómalos

Como complemento a este primer acercamiento, se realizó de acuerdo con Tukey (1977), quien presenta el análisis exploratorio de datos (EDA, por sus siglas en ingles) como un proceso donde el investigador examina los datos sin ninguna idea preconcebida para descubrir qué le pueden expresar acerca del fenómeno que está estudiando, a través de la indagación gráfica y numérica. El EDA se diferencia del enfoque clásico y bayesiano; éste a partir del problema, se recogen los datos, se analizan, se genera el modelo y se emiten conclusiones; mientras que el enfoque clásico se parte del problema, se recogen los datos, se modela, y después se hace análisis y se exponen las conclusiones; el enfoque bayesiano también inicia con la definición del problema, después se hace el acopio de datos, luego se modela, se continua con búsqueda de la distribución previa de los datos, analiza y finalmente formula conclusiones (NIST/SEMATEC,2007).

De acuerdo con la idea general del EDA y continuando a través de éste con el análisis más detallado de los datos, se realizó una indagación visual del comportamiento histórico de cada uno los elementos del clima incluidos en los

patrones, con el propósito de detectar anomalías. A través de la aplicación de herramientas visuales incluidas dentro del EDA, como gráficos de líneas y diagramas de caja y bigotes, se detectaron a partir de la observación del comportamiento y la distribución de los datos, la presencia de valores atípicos y extremos[30] (Hoaglin, 1986).

---

[30] Un valor se considera **atípico** si cumple con las siguientes condiciones:

UBV + o.c.*(UBV - LBV) > Valor < LBV - o.c.*(UBV - LBV)

Donde: UBV es valor superior de la caja en el diagrama de caja (esto es, la media + desviación estándar o el percentil 75).

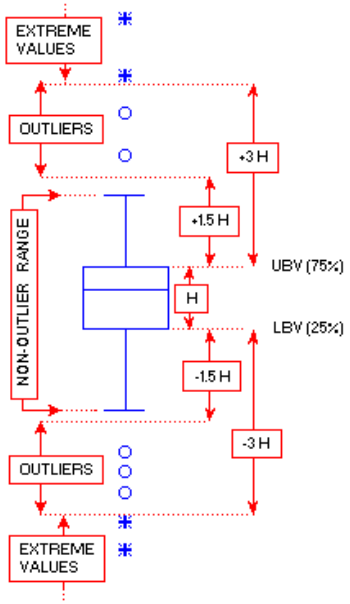
LBV es valor superior de la caja en el diagrama de caja (esto es, la media + desviación estándar o el percentil 25).

o.c. es es el coeficiente de atipicidad especificado.

Por otra parte un valor se considera **extremo** si cumple con las siguientes condiciones:

UBV + 2*o.c.*(UBV - LBV) > Valor < LBV – 2*o.c.*(UBV - LBV)

El siguiente esquema muestra los rangos de valores atípicos y valores extremos en un diagrama de caja y bigotes (StatSoft, Inc. 2007).



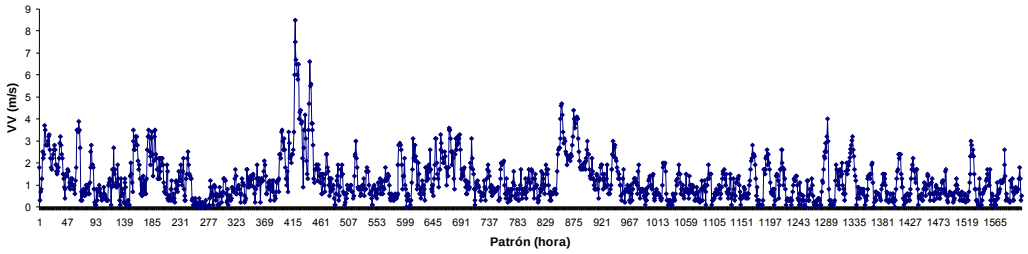
Fuente: Statistica Electronic Manual, Statistica 8.0 version.

La figura 4.2 expresa el comportamiento histórico de la velocidad del viento, a través del valor presentado en cada hora durante el periodo de estudio; en ésta se observa la presencia de picos que no siguen un patrón fácilmente detectable, y alcanzan sus valores máximos en el ciclo 2001-2002 (figura 4.2a). Los ciclos 2003-2004 y 2004-2005 (figuras 4.2c, 4.2d), presentan también picos considerables dentro de los cuales resaltan el comportamiento tan errático presentado al final del ciclo 2004-2005 (figura 4.2d). Aunque la figura 4.2b, correspondiente al ciclo 2002-2003 presenta un comportamiento más homogéneo, presenta picos con menor intensidad que el resto de los ciclos. Este comportamiento tan variable fue cuantificado por el valor del coeficiente de variación igual a 87.72 (cuadro 4.1).

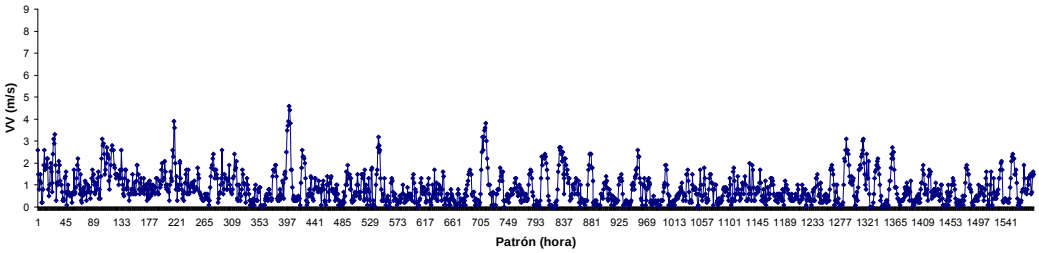
Las gráficas del comportamiento del resto de las variables, aparecen en el Anexo A, donde puede observarse que, para el caso de de la temperatura, aunque la banda de variación muestra cierta homogeneidad, resalta el comportamiento tan estable de patrones contiguos presente en los ciclos 2001-2002 (figura A.1a) y 2004-2005 (figura A.1d). En relación a la humedad relativa, su comportamiento errático refleja patrones contiguos con alta humedad en todos lo ciclos, siendo más marcado en ciclo 2001-2002 (figura A.2a), así mismo, patrones varían en bandas de baja humedad resaltado el ciclo 2003-2004 (figura A.2c). Respecto a la presión de vapor, destaca los máximos presentados al inicio del ciclo 2001-2002 (figura A.3a) y la banda de variación de baja presión que presentan algunos patrones contiguos del ciclo 2003-2004 (figura A.3c); el comportamiento de la radiación solar, exhibe la presencia de patrones contiguos con nublados al inicio del ciclo 2001-2002 (figura A.4a), y nublados menos intensos al inicio del ciclo 2004-205 (figura A.4d). De la seis variables del clima, la dirección del viento presenta menor dispersión con un coeficiente de variación igual a 35.86 (figura A.6).

La dispersión de los elementos de los patrones climáticos distribuidos por ciclo productivo se presenta en la figura 4.3, donde aparecen la dispersión de los

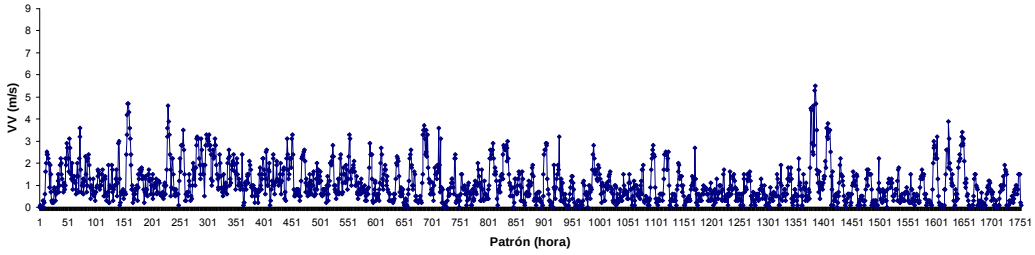
Figura 4.2 Comportamiento continuo de la velocidad del viento, para cada ciclo de producción.



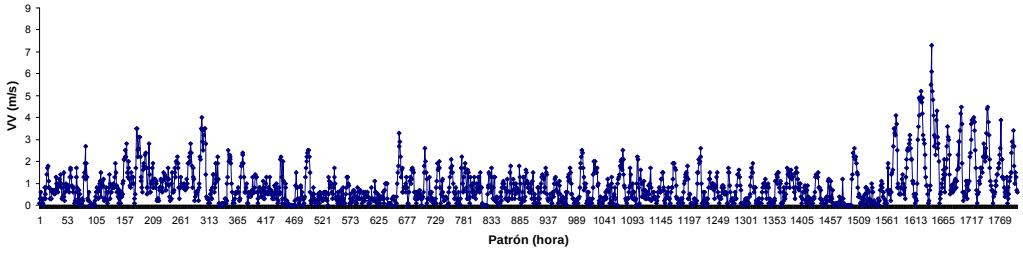
a) Ciclo 2001-2002



b) Ciclo 2002-2003



c) Ciclo 2003-2004



d) Ciclo 2004-2005

Fuente: Elaboración propia.

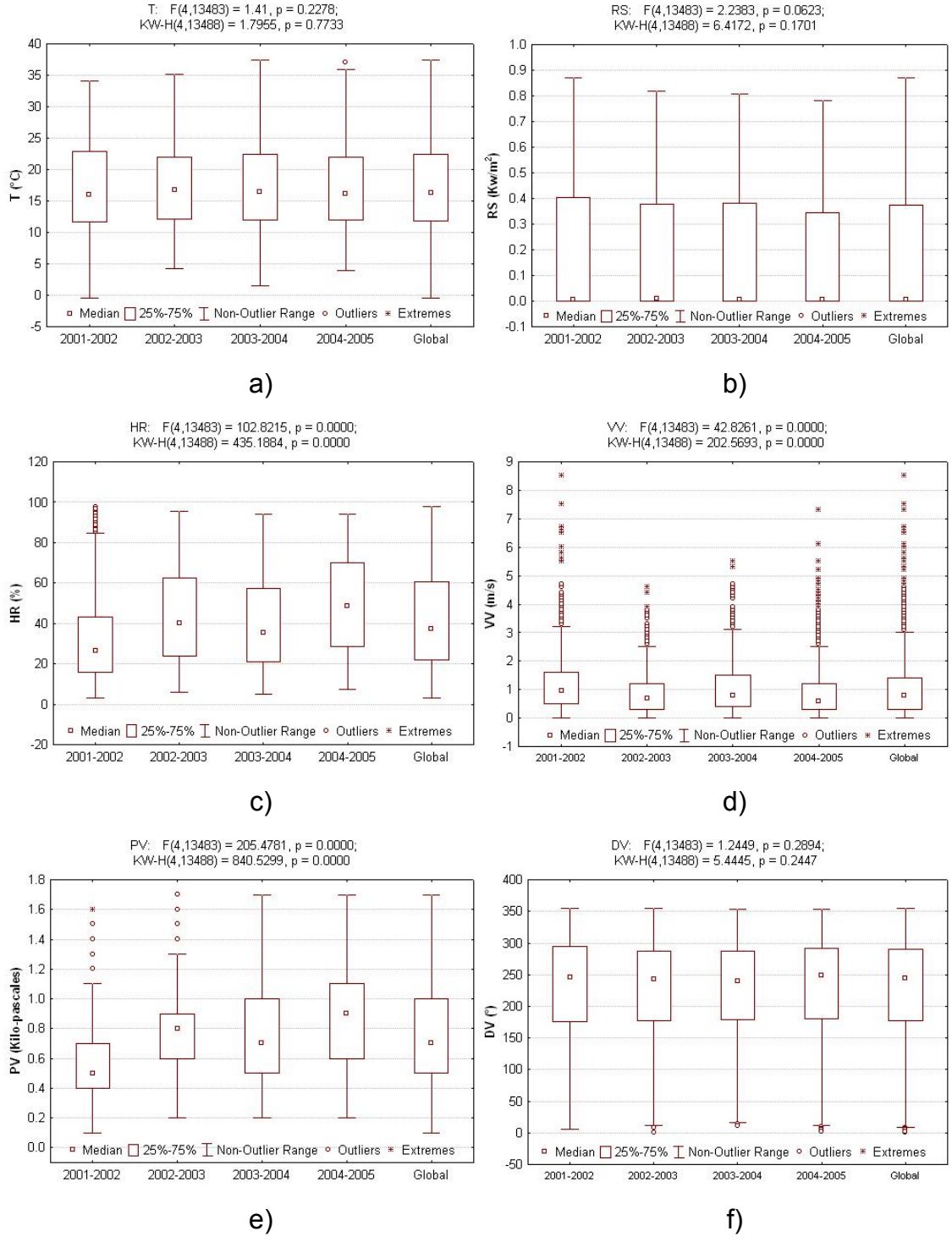
valores para cada ciclo productivo, y la dispersión para el conjunto completo de valores, la cual es representada a través del elemento llamado 'global'. Las pruebas de análisis de varianza tanto paramétrica como no paramétrica (F y

Kruskal-Wallis), reflejan que para el caso de las variables T, RS, y DV, no se encontró evidencia que permita hacer una diferenciación por ciclo productivo, mientras que la dispersión de las variables HR, PV y VV, si presenta diferencia entre los ciclos productivos.

Sólo en los casos de la VV y DV, se presentaron valores atípicos en el conjunto global, siendo más marcado el caso de la VV, donde no únicamente hubo presencia de valores atípicos, sino también de valores extremos (registros mayores a 4.5 m/s, figura 4.3d). En cuanto a distribución de estos valores anómalos entre los ciclos productivos, todas las variables, a excepción de la radiación solar, presentan estos valores, lo que resulta muy evidente es el caso de velocidad del viento, donde todos los ciclos presentan los dos tipos de valores. Otra característica que resalta, es el comportamiento alternante del valor de la mediana presentado para las variables HR, PV y VV, la cuales contrastan entre si, mientras que HR y PV presentan mínimos locales en los ciclos 2001-2002 y 2003-2004 (figuras 4.3c y 4.3e). La velocidad del viento presenta sus mínimos locales en los ciclos 2002-2003 y 2004-2005 (figura 4.3d).

En la figura 4.4, se presenta la dispersión de los valores de cada una de las variables climáticas distribuidas por etapa fenológica y de forma global. En ésta se observa como todas las variables presentan, en al menos una etapa fenológica, valores anómalos. Destacan los casos de las variables PV, que presenta valores atípicos en la etapas 05, 07, 12, 14, 23 y 25; la DV que los muestra en las etapas 10, 12, 14, 21, 23 y 25; especialmente el caso de la VV, la cual además de exhibir valores atípicos en todas las etapas, presenta valores extremos en la mayor parte de éstas. Tanto las pruebas de análisis de varianza paramétricas como no paramétricas coincidieron en evidenciar la diferencia entre las características climáticas en cada una de las etapas fenológicas.

Figura 4.3 Dispersión de las variables climáticas distribuidas por ciclo productivo.



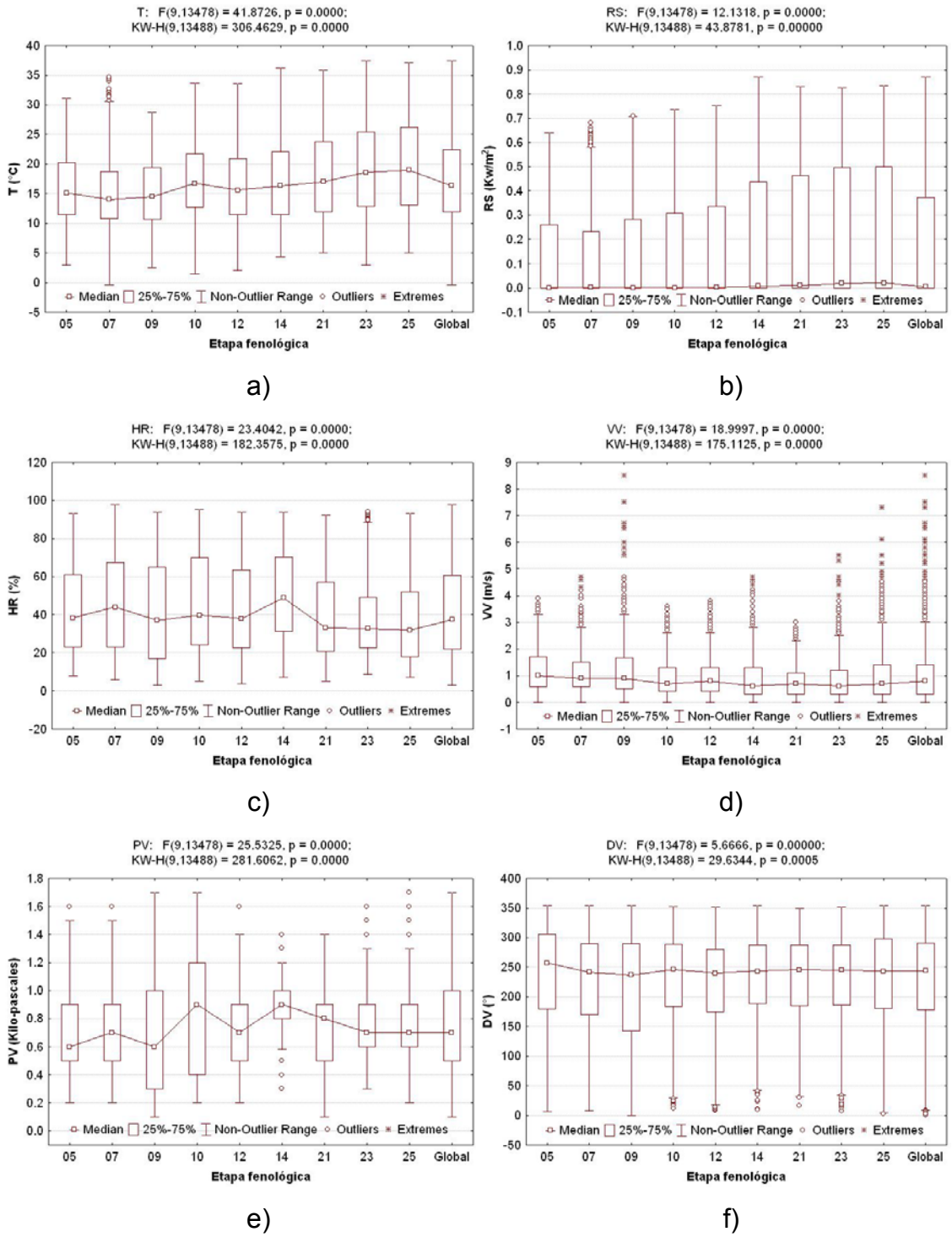
Fuente: Elaboración propia.

El siguiente paso en la exploración de las variables climáticas, permitió mostrar la dispersión de los datos, a través de las etapas fenológicas para cada uno de los ciclos productivos. Enseguida se presentan los casos de las variables HR, PV y VV. Para el resto de las variables esta exploración puede ser consultada en el Anexo B.

A pesar de que durante el análisis global de conjunto de datos, la variable HR no presenta valores atípicos, al pasar al siguiente nivel de exploración, ésta registra presencia de valores atípicos durante el ciclo productivo 2001-2002 (figura 4.3c), además, al examinar los datos a través de sólo las etapas fenológicas, ésta registra presencia de valores atípicos en la etapa 23 (figura 4.4c). Sin embargo, al combinar la exploración por etapa fenológica y ciclo productivo, la presencia de valores atípicos no sólo se da en la etapa 23, sino además en las etapas 09, 10, 14 y 21 (figura 4.5a). En el resto de los ciclos, aunque de manera global no se presentan valores atípicos (figura 4.3c), se registra la presencia de estos valores en la etapa 07 del ciclo 2002-2003 (figura 4.5b), así como en las etapas 09 y 23 de ciclo 2003-2004 (figura 4.5c) y etapa 25 del ciclo 2004-2005 (figura 4.5d). Así mismo, las pruebas de análisis de varianza no paramétrica y paramétrica muestran un comportamiento diferenciado de esta variable dentro de las etapas fenológicas para cada ciclo productivo.

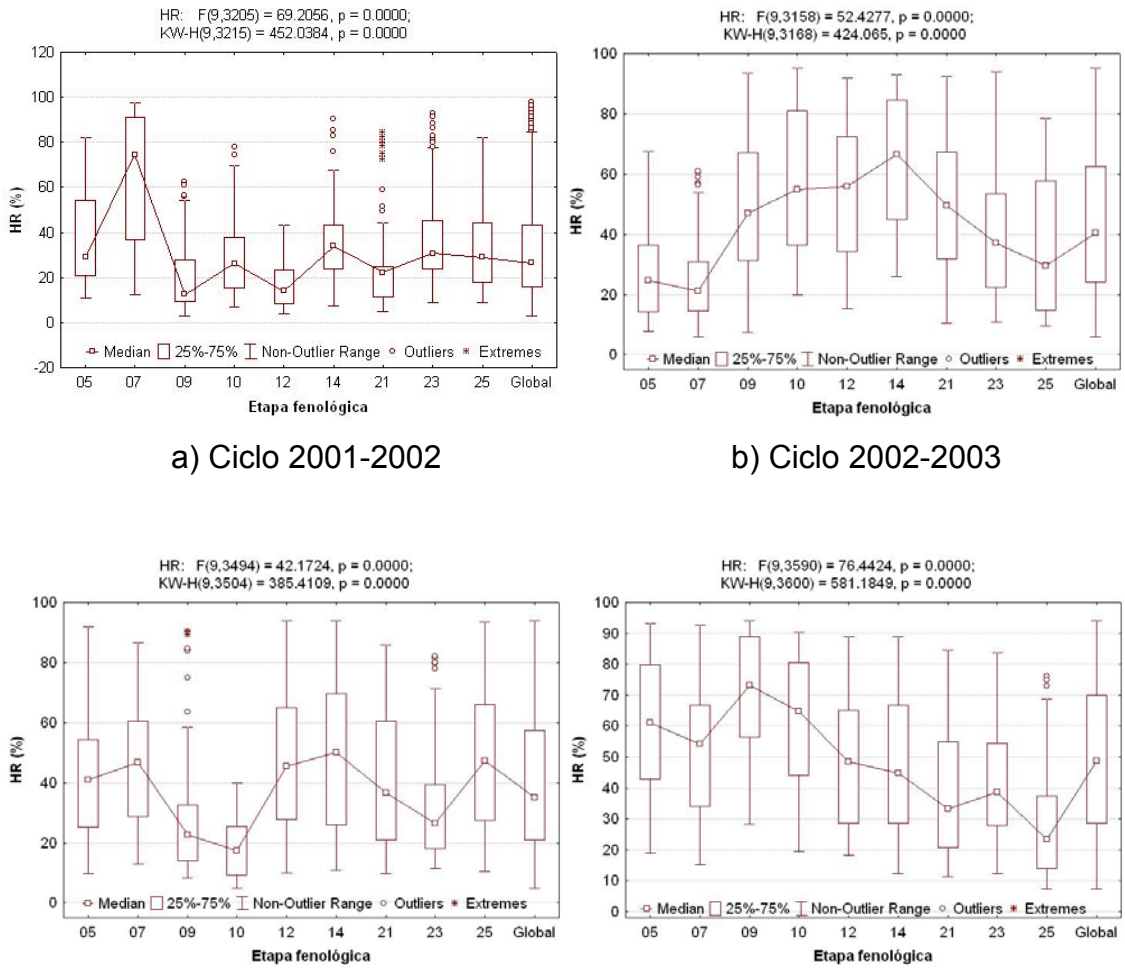
Al explorar el total de registros, la variable PV no mostró presencia de valores atípicos; sin embargo, registró presencia de estos valores para los ciclos 2001-2002 y 2003-2004 (figura 4.3e). En la exploración por etapa fenológica los valores atípicos se presentaron en las etapas 05, 07, 12, 14, 23 y 25 (figura 4.4e).

Figura 4.4 Distribución de los valores de los atributos del patrón climático distribuidos por etapa fenológica.



Fuente: Elaboración propia, a partir de los gráficos generados por el sistema.

Figura 4.5 Diagramas de caja y bigote para la variable HR distribuida por etapa fenológica, para cuatro ciclos productivos.



a) Ciclo 2001-2002

b) Ciclo 2002-2003

c) Ciclo 2003-2004

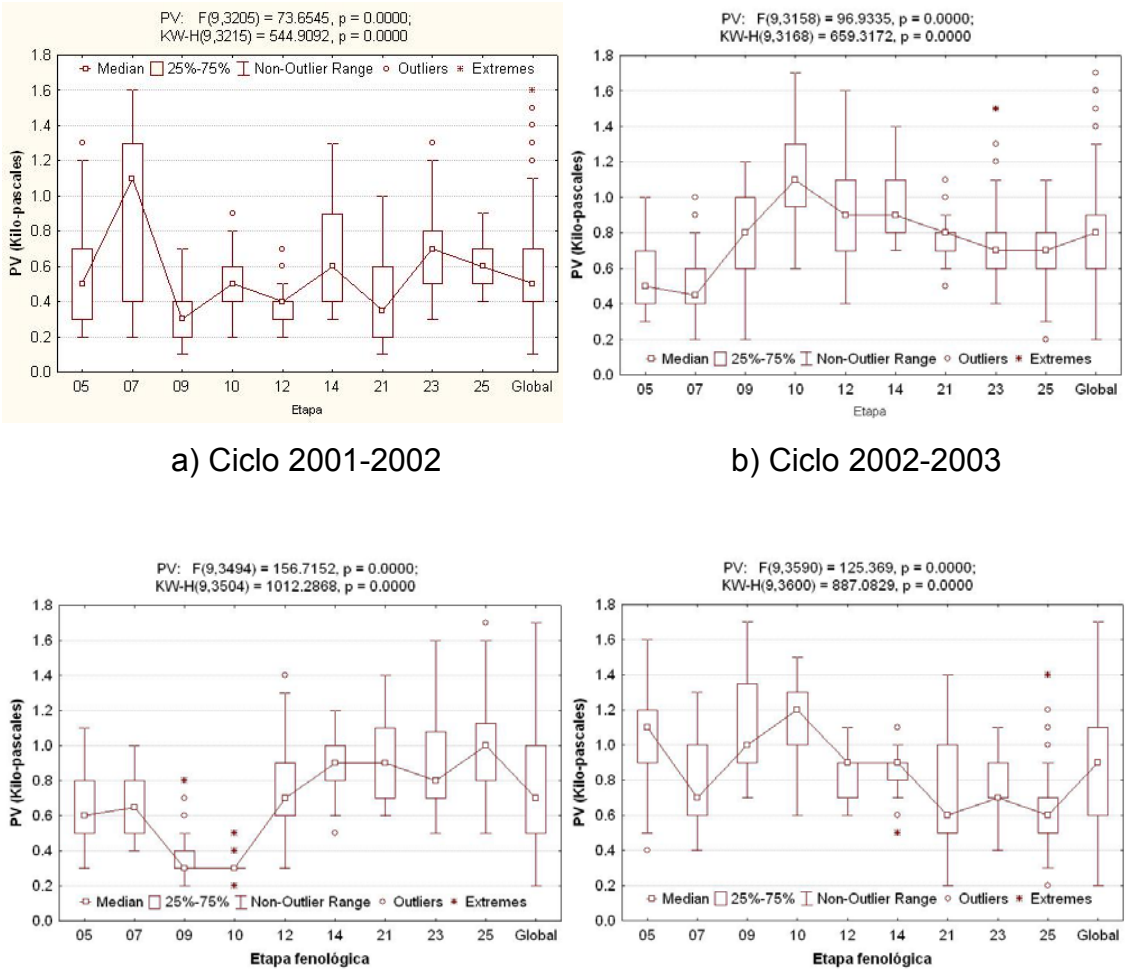
d) Ciclo 2004-2005

Fuente: Elaboración propia, a partir de los gráficos generados por el sistema.

En la figura 4.6 se muestra la distribución de este tipo de valores, al combinar los criterios de exploración por etapa fenológica y ciclo productivo. Se observa la presencia de valores anómalos, a diferencia de la no detección cuando los criterios fueron utilizados por separado (ciclos 2003-2004 y 2004-2005 figuras 4.6c y 4.6d; etapas fenológicas 09 y 10 figuras 4.6a y 4.6c). Por otra parte, las pruebas de análisis de varianza paramétrica y no paramétrica confirmaron una variabilidad

diferenciada de los datos durante las etapas fenológicas dentro de cada ciclo productivo.

Figura 4.6 Diagramas de caja y bigote para la variable PV distribuida por etapa fenológica, para cuatro ciclos productivos.



a) Ciclo 2001-2002

b) Ciclo 2002-2003

c) Ciclo 2003-2004

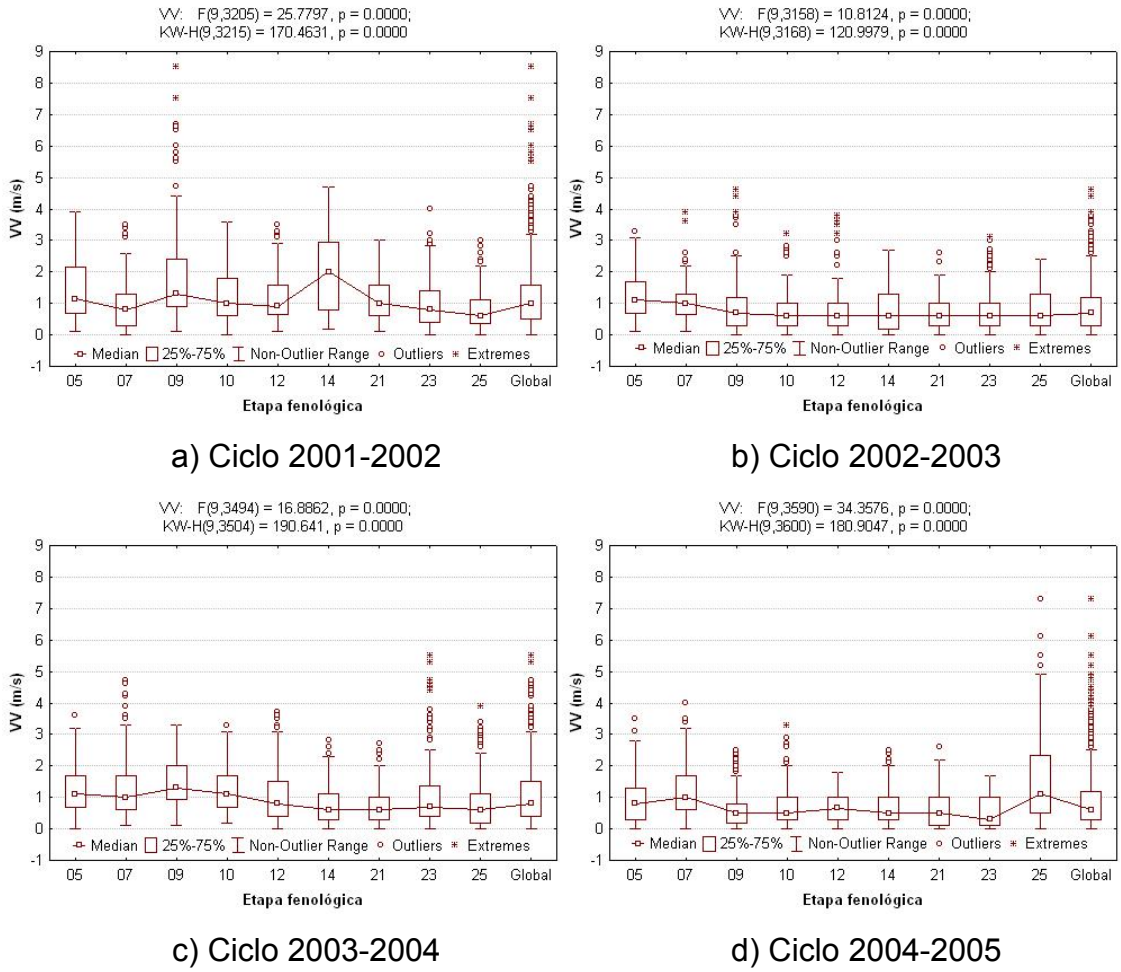
d) Ciclo 2004-2005

Fuente: Elaboración propia.

La exploración más especifica del comportamiento de la variable VV demostrado por los resultados expuestos en la figuras 4.3d y 4.4d, se realizó combinando los criterios de exploración etapa fenológica y ciclo productivo. Los resultados obtenidos se muestran en la figura 4.7, donde se observa la presencia de valores

atípicos y extremos en los cuatro ciclos productivos (columna 'Global'), siendo más marcada en los ciclos 2001-2002 y 2004-2005 (figuras 4.7a y 4.7d, respectivamente) y se presentan las rachas de viento máximas en la etapa 09 del ciclo productivo 2001-2002 (figura 4.7a).

Figura 4.7 Diagramas de caja y bigote para la variable VV distribuida por etapa fenológica, para cuatro ciclos productivos.



a) Ciclo 2001-2002

b) Ciclo 2002-2003

c) Ciclo 2003-2004

d) Ciclo 2004-2005

Fuente: Elaboración propia.

Generalmente, las fuentes de datos en el mundo real son erróneas, incompletas, o inconsistentes, debido ya sea a errores operacionales o defectos de la puesta en marcha de los sistemas (Mitra y Acharya, 2003). Una buena práctica para asegurar la calidad de la fuentes de datos, es verificar el origen de los datos

extremos, que pueden ser generados por errores en la captura, en la transferencia, por fallas en los equipos de medición o pueden representar mediciones originales (Myatt, 2007). Es rastreo de los valores atípicos y extremos encontrados en las figuras 4.3d y 4.4d, así como su distribución, se presentan en la figura 4.7. Ésta condujo al cuestionamiento sobre la originalidad de los valores extremos, por lo que hubieron de comprobarse a partir de la consulta de las mediciones de la velocidad del viento para otras estaciones agroclimáticas pertenecientes al mismo sistema de información (figura 4.8). Éstas se ubican al suroeste de la estación de referencia para el estudio (La Cuesta).

En el cuadro 4.2, se muestran las lecturas de las estaciones agroclimáticas consultas de acuerdo a los parámetros de referencia fecha y hora, de cada uno de los registros que contienen los datos extremos de la base de datos climáticos utilizada en el presente estudio (datos climáticos de la estación agroclimática, ”La Cuesta”). Se observa la correspondencia entre estas y el comportamiento de los valores extremos de la velocidad del viento mostrados en las figura 4.3d, 4.4d y 4.7. Así mismo, bajo la hipótesis de que los datos extremos, corresponden a condiciones meteorológicas presentes al momento de realizar las mediciones, se aplicó un análisis de varianza tanto paramétrico como no paramétrico, lo que permitió confirmar la sospecha acerca de la no originalidad de los datos (figura 4.9).

Hasta este punto la exploración de la dispersión de los datos realizada a través de la estrategia drill-down, permitió detectar la presencia de datos anómalos (atípicos y extremos) y su propagación conforme se avanzo en análisis más especifico. Así mismo, se validó que la presencia de este tipo de datos se debió al desarrollo natural del clima, descartándose que tales lecturas se debieran a fallas del equipo o en la captura de datos.

Figura 4.8 Localización de las estaciones agroclimáticas alternas, situads en la costa de Hermosillo.

Fuente: Elaboración propia.

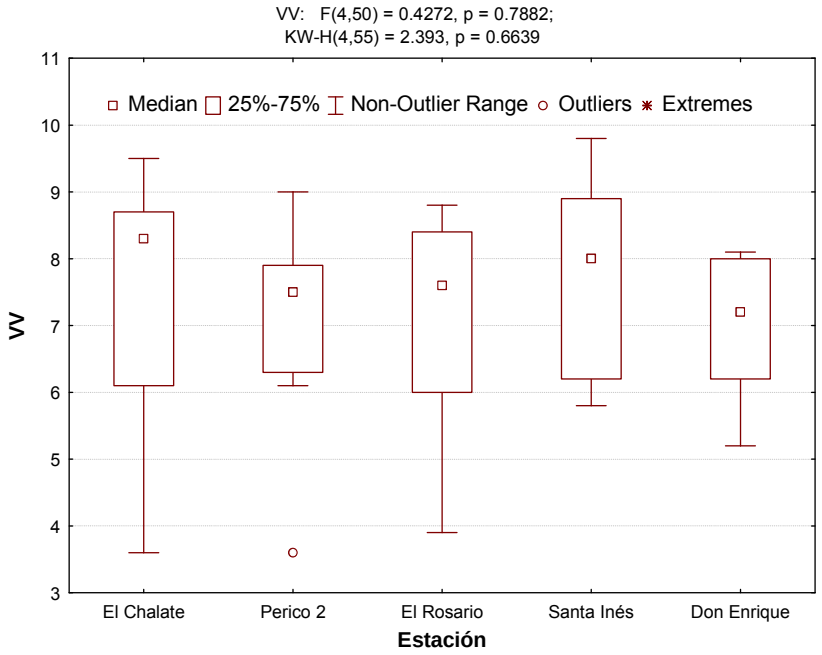
Cuadro 4.2 Lecturas de la velocidad del viento en m/s registrada en las estaciones agroclimáticas externas.

		Estación climática				
Fecha	**Hora**	**El Chalate**	**Perico 2**	**El Rosario**	**Santa Inés**	**Don Enrique**
10/02/2002	10:00:00	8.5	3.6	3.9	6.0	5.2
10/02/2002	11:00:00	9.5	6.3	6.6	7.1	7.2
10/02/2002	12:00:00	8.7	7.9	8.1	8.8	8.1
10/02/2002	13:00:00	9.4	8.8	8.8	9.8	8.0
10/02/2002	14:00:00	8.3	7.6	8.5	8.9	7.6
10/02/2002	15:00:00	8.4	9.0	8.4	9.2	8.0
10/02/2002	16:00:00	7.2	6.8	8.1	8.6	7.5
10/02/2002	17:00:00	7.4	7.1	7.6	8.0	7.1
11/02/2002	11:00:00	6.1	7.5	5.5	6.5	6.7
11/02/2002	12:00:00	4.9	7.9	6.1	6.2	6.2
11/02/2002	13:00:00	3.6	6.1	6.0	5.8	5.9

Fuente: Elaboración propia.

A partir del procedimiento descrito es posible sostener que el conjunto de patrones puede ser caracterizado como una base de datos con ruido natural generado principalmente por la variable velocidad de viento, la cual presentó sistemáticamente este tipo de valores (figura 4.3, 4.4, y 4.8). La presión de vapor, la dirección del viento y la humedad relativa (en menor medida), también presentaron estas características.

Figura 4.9 Validación de la veracidad de los valores extremos de la variable VV, a partir de la comparación con lecturas de otras estaciones agroclimáticas.



Fuente: Elaboración propia.

IV.2.2. Relación entre variables

Una vez explorada la distribución y dispersión de los datos, el siguiente paso fue buscar las relaciones de dependencia o correlación entre los elementos del clima que conforman los patrones climáticos. Con base en los resultados obtenidos del análisis de frecuencia (figura 4.1), las asimetrías encontradas y los resultados de la pruebas de normalidad Kolmogorov-Sminorv, se calcularon las matrices paramétrica de correlación de Pearson y no paramétrica de Spearman, para

identificar la relación de dependencia entre las variables, así como para los resultados a través de ambas técnicas.

Los cuadros 4.3 y 4.4, muestran que ambos coeficientes (Pearson y Spearman) convergen al presentar las matrices de correlación donde ambas exhiben prácticamente correlación significativa al $p<0.05$, para todas las parejas posibles de variables.

Cuadro 4.3 Matriz de correlación entre las variables incluidas dentro del patrón climático (Pearson).

	Media	Desv. Estándar	T	HR	PV	RS	VV	DV
T	17.351	7.024	1.000	-0.691*	-0.091*	0.690*	0.263*	-0.337*
HR	42.133	24.294	-0.691*	1.000	0.723*	-0.498*	-0.378*	0.199*
PV	0.746	0.308	-0.091*	0.723*	1.000	-0.116*	-0.247*	-0.020
RS	0.190	0.251	0.690*	-0.498*	-0.116*	1.000	0.308*	-0.476*
VV	0.917	0.807	0.263*	-0.378*	-0.247*	0.308*	1.000	-0.085*
DV	226.680	81.270	-0.337*	0.199*	-0.020*	-0.476*	-0.085*	1.000

* Correlación significativa a $p<0.05$

Fuente: Elaboración propia.

Cuadro 4.4 Matriz de correlación entre las variables incluidas dentro del patrón climático (Spearman).

	Media	Desv. Estándar	T	HR	PV	RS	VV	DV
T	17.351	7.024	1.000	-0.717*	-0.092*	0.668*	0.399*	-0.416*
HR	42.133	24.294	-0.717*	1.000	0.733*	-0.494*	-0.498*	0.231*
PV	0.746	0.308	-0.092*	0.733*	1.000	-0.082	-0.317*	-0.051*
RS	0.190	0.251	0.668*	-0.494*	-0.082*	1.000	0.365*	-0.523*
VV	0.917	0.807	0.399*	-0.498*	-0.317*	0.365*	1.000	-0.115*
DV	226.680	81.270	-0.416*	0.231*	-0.051*	-0.523*	-0.115*	1.000

* Correlación significativa a $p<0.05$

Fuente: Elaboración propia.

La matriz de correlación muestra también convergencia en los resultados obtenidos. Se observan relaciones positivas fuertes entre las parejas de variables HR, PV; T, RS, mientras que la pareja de variables T, HR presenta una relación negativa fuerte; otras relaciones positivas menos fuertes se encuentran entre las parejas (RS, VV), (T, VV) y negativas en este misma proporción (HR, RS), (RS, DV), (HR, VV), (T, DV) y (PV, VV).

Con el propósito de facilitar la visualización de la relación de dependencia entre las variables climáticas, se formó la matriz de diagramas de dispersión entre todas las posibles parejas de variables (figura 4.10). En ésta figura se presentan también el ajuste lineal de cada relación, a través de la recta de mínimos cuadrados, donde se puede observar que entre las variables PV y HR se da una la relación casi perfecta; además, la tendencia de la relación inversa entre las variables T y HR, y la relaciones débiles presentadas entre PV y DV, así como entre PV y RS. De esta manera, los resultados del análisis de correlación indican que el conjunto de patrones tiene atributos con alto grado de interdependencia.

A pesar que se sabe de antemano, que los elementos del clima son muy interdependientes, y que pertenecen a un proceso donde existe correlación entre variables, los resultados basados en el coeficiente de Pearson, así como los resultados de Spearman, no muestran coeficientes de determinación tan fuertes que permitan inferir la explicación exacta de una variable respecto a otra (coeficientes de determinación igual a 1).

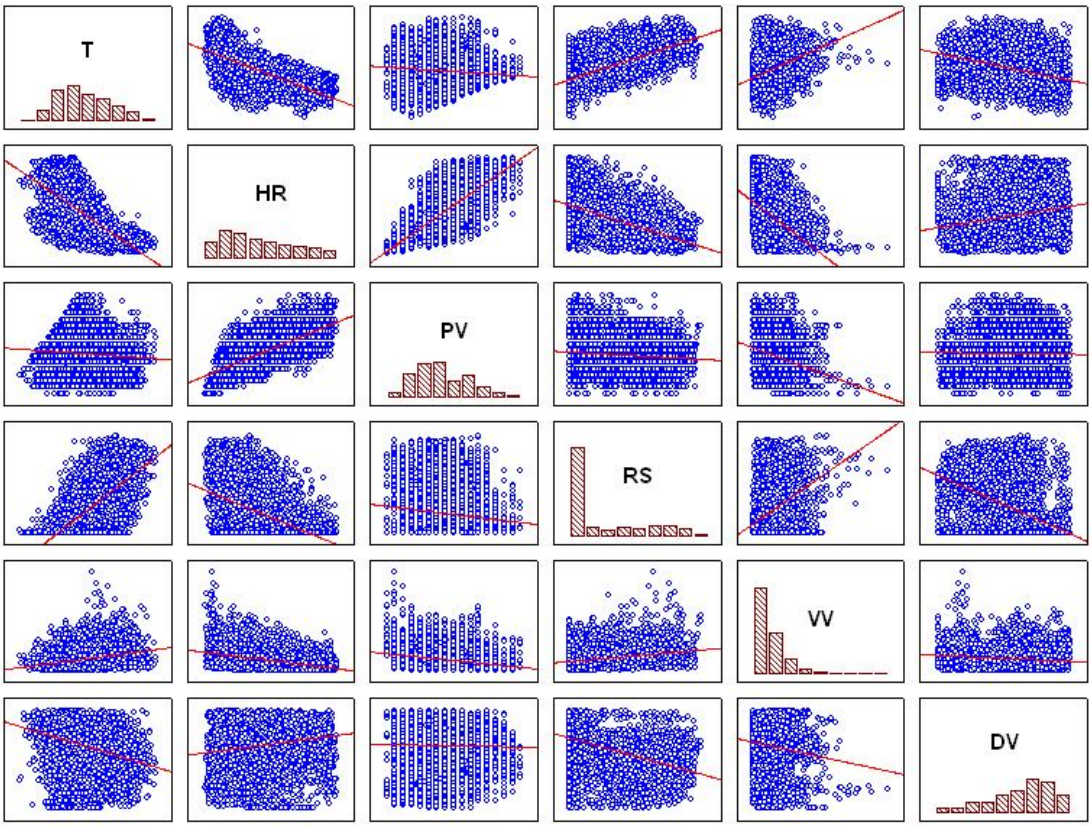
El conjunto de patrones climáticos utilizados en la comprobación empírica, se analizó a partir del EDA, que se rige a través del encadenamiento de las tareas, problematización, acopio de datos, análisis de datos, generación de modelos, y obtención de conclusiones (NIST/SEMATEC,2007). Este análisis reveló a través del análisis de frecuencia y la prueba Kolmogorov-Smirnov, que los elementos de clima incluidos en el conjunto de patrones climáticos no siguieron una distribución normal (figura 4.1). El uso de los diagramas de caja y bigote permitió la

exploración de la dispersión de los valores de las variables del clima desde lo general (figura 4.3) hasta lo particular (figura 4.7 y anexo B). Detectándose a través de la prueba paramétrica F y la no paramétrica Kruskal-Wallis, la no diferenciación de la temperatura para los cuatro ciclos productivos estudiados. En este mismo sentido los diagramas de caja, también permitieron detectar la existencia de valores atípicos y extremos, cuya veracidad para descartar errores no inherentes a un comportamiento natural fue validadas a través del contraste con los datos correspondientes cronológicamente de otras estaciones climáticas cercanas a la referenciada (figura 4.8), a través de las pruebas F y Kruskal-Wallis, las cuales confirmaron que tales valores anómalos se deben al comportamiento natural de las mediciones físicas realizadas. Por lo que de acuerdo a estos hechos, el conjunto de patrones puede ser caracterizado como una base de datos con ruido natural generado principalmente por la variable velocidad de viento la cual presentó sistemáticamente este tipo de valores (figura 4.3, 4.4, y 4.8); la presión de vapor, la dirección del viento y la humedad relativa en menor media también presentaron estas características.

Dadas las asimetrías observadas tanto en el análisis de frecuencias (figura 4.1) y las observadas en los diagramas de dispersión (figuras 4.3, 4.4, 4.5, 4.6, 4.7 y anexo B), se decidió buscar la relación entre elementos del clima incluidos dentro de los patrones climáticos, a través del análisis de correlación en base al coeficiente de Perarson (análisis paramétrico) y el basado en el coeficiente de Spearman (no paramétrico), los cuales convergieron a resultados similares. A partir de éstos, se encontraron relaciones positivas fuertes entre las variables (HR, PV), (T, RS), mientras que la pareja de variables (T, HR) presenta una relación negativa fuerte, otras relaciones positivas menos fuertes se encuentras entre las parejas (RS, VV), (T, VV) y negativas en este mismo sentido (HR, RS), (RS, DV), (HR, VV), (T, DV) y (PV, VV). Así mismo, en la complejidad reflejada del conjunto de patrones mostrada en la figura 4.10, se observa la estructura complicada de los datos, ya que ninguno de los diagramas de dispersión de dicha matriz muestra la presencia de grupos linealmente separables (Höppner, et al., 2000). Tal

caracterización climática coincide con lo sugerido por Trewartha y Horn (1980), quienes enuncian que el clima es la expresión acumulada del movimiento regular diario de la atmosfera; además, sigue un complejo comportamiento aleatorio dependiente de la posición geográfica donde se encuentra el territorio bajo estudio (Coonors y Loomis, 2002).

Figura 4.10 Matriz de diagramas de dispersión de las variables incluidas dentro del patrón climático.



Fuente: Elaboración propia.

IV.3. Formación de grupos

Los resultados obtenidos hasta aquí, caracterizan al conjunto de patrones como la expresión de un proceso que pertenece a un ambiente ruidoso y complejo, donde el uso del paradigma de reconocimiento de patrones ha demostrado ser una buena alternativa para la extracción automática de irregularidades significativas (Devijver y Kittler, 2002; Duda y Hart 1973; Duda et al., 2001; McLachlan, 2004; Pal y Pal, 2001; Dasey y Micheli-Tazanakou, 1999).

En esta sección se presentan los resultados obtenidos con la aplicación de algoritmos incluidos en el aprendizaje no supervisado (*clustering*), como el algoritmo determinista K-Medias, incluido dentro del análisis multivariado en el campo de la estadística clásica, así como de los algoritmos Fuzzy C-Medias y Gustafson-Kessel, incluidos dentro de la Computación Suave, y que han demostrado ser tolerantes a la imprecisión, incertidumbre y verdad parcial (Gordon, et al., 2001; Du y Swamy, 2006, Mitra y Acharya, 2003).

Adquiere relevancia identificar el número de grupos a formar o descubrir, al momento de ejecutar un algoritmo de agrupamiento (análisis de *clustering* o conglomerados) sobre un conjunto de datos p-dimensionales, ya que es con base en este tipo dato que los algoritmos de agrupamiento buscan optimizar o descubrir una determinada estructura para ese número de subgrupos definido a priori.

En este trabajo se utilizó el índice de validación **S**, propuesto por Xie-Beni (1991) y generalizado por Pal y Bezdek(1995), para determinar el mejor número de grupos (*clusters*, particiones o clases), que permita encontrar la estructura subyacente al conjunto de patrones climáticos utilizados. Este índice permite conocer tanto la dispersión dentro de los grupos, como la separación entre éstos; así, el índice captura el cociente entre la dispersión dentro de los grupos y la distancia entre ellos (Ray t Tury, 1999; Chou et al., 2004). Lo que se busca es obtener el menor valor para tal cociente, y de esta manera separan o cumplen con el objetivo

principal del análisis de *clustering*, que es encontrar grupos con elementos lo mas homogéneos posible entre ellos, y con heterogeneidad respecto a los elementos de otros grupos.

Para la obtención del mejor número de clases, se ejecutaron nueve veces cada uno de los algoritmos de agrupamiento en la secuencia de 2 clases hasta 10. A partir de ello se calculó el índice S, para posteriormente seleccionar el mejor número de grupos (valor mínimo del índice). Para el caso de los algoritmos de agrupamiento determinista K-Medias, en el cálculo del índice S se asignaron los dos únicos valores que puede tomar el grado de pertenencia en el caso determinista: cero y uno. Así mismo, se calcularon los índices coeficiente de partición (PC) y coeficiente de entropía, para el agrupamiento obtenido con los algoritmos difusos, con el objeto de robustecer la validación de la calidad de partición.

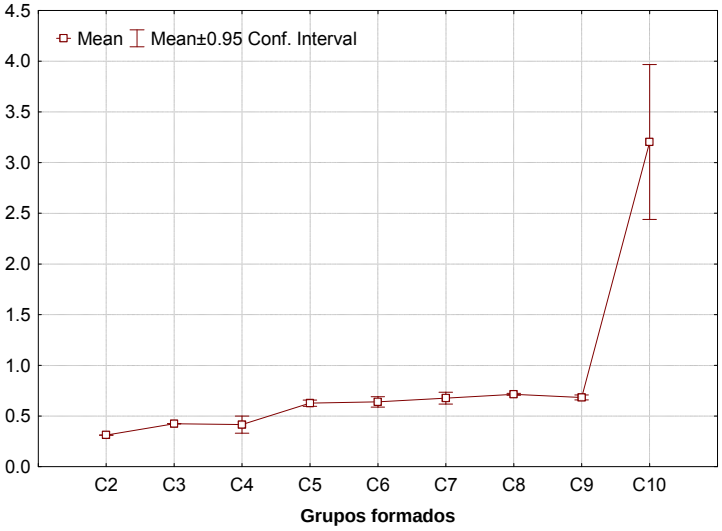
En la siguiente sección se muestran los resultados obtenidos a partir de la aplicación de los tres algoritmos de agrupamiento (K-Medias, Fuzzy C-Medias y Gustafson-Kessel). Se presenta el procedimiento para la obtención del mejor número de clases a formar, a partir del comportamiento del índice de separación S; se exhiben los patrones característicos obtenidos o centros de grupo de cada clase. Posteriormente se muestra una caracterización estadística de los patrones agrupados en cada uno de los grupos determinados; la distribución cronológica de los patrones característicos es representada en el mapa climático, que muestra la distribución de éstos durante el día para cada una de las etapas de desarrollo estudiadas. Finalmente, se presenta una caracterización textual de cada uno de los grupos en los que se dividió el conjunto de patrones estudiados. Esta estructura se replica para cada uno de los tres algoritmos de agrupamiento utilizados.

IV.3.1. Agrupamiento a través del algoritmo K-Medias

A partir de los nueve distintos agrupamientos determinados con el algoritmo K-Medias, se calcularon los índices de separación S para cada uno de los diferentes agrupamientos. El comportamiento del índice S se muestra en la figura 4.11; se observa una tendencia estable hasta la división del conjunto de patrones en nueve grupos.

El objetivo, en esta parte del proceso, es encontrar el número de grupos donde el valor del índice tomó su valor mínimo, que en este caso se obtiene cuando el conjunto de patrones se divide en dos grupos. Sin embargo, dada la naturaleza de los datos (figura 4.7), la partición del conjunto de patrones en dos grupos revela poco la estructura subyacente al conjunto de patrones, de ahí que se optó por seleccionar la segunda, que es cuando éste es dividido en cuatro grupos.

Figura 4.11 Comportamiento del índice de separación S, calculado a partir del agrupamiento realizado con el algoritmo K-Medias.



Fuente: Elaboración propia.

A partir de la determinación de que la mejor forma de agrupar el conjunto de patrones, es a través de cuatro grupos, se obtuvieron los centros de cluster para cada uno de estos grupos. Éstos son reprensados a través del patrón

característico que es el patrón de referencia que minimiza la distancia entre los elementos de cada grupo.

El cuadro 4.4, muestra los valores de los elementos climáticos de los patrones característicos en cada una las clases o grupos formados (C1, C2, C3 y C4). Se puede apreciar la diferenciación aparente de los grupos formados, los cuales siembre deberán ser visualizados como datos p-dimensionales y considerando cada variable por separado.

Cuadro 4.4 Estructura de los patrones característicos formados a partir del algoritmo de agrupamiento K-Medias.

		Patrones característicos			
Atributo	**Unidades**	**C1**	**C2**	**C3**	**C4**
T	°C	12.0539	16.1850	26.0087	18.5462
HR	%	71.9291	31.3073	21.4860	32.5681
PV	Kilo-pascales	1.0189	0.5638	0.6845	0.6618
RS	Kw/m^2	0.0325	0.0408	0.5733	0.3256
VV	m/s	0.5372	0.9707	1.3014	1.0515
DV	Grados (°)	250.9860	272.9296	193.3443	82.7049

Fuente: Elaboración propia.

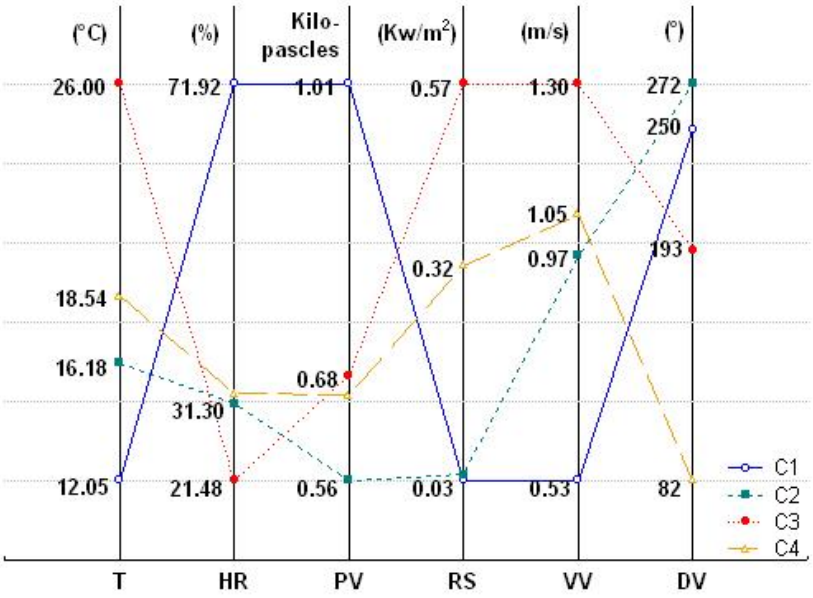
La figura 4.12, muestra de forma visual la estructura general de los patrones característicos que representan los grupos formados a partir de la aplicación del algoritmo K-Medias. En esta figura, se observa el contraste entre los grupos representados por la clase C1 y C3 las cuales pueden ser consideradas como los grupos extremos ya que, mientras que la clase C1 agrupa patrones con baja temperatura, alta humedad relativa, alta presión, baja radiación solar, y vientos leves en dirección al sur, la clase C3 agrupa los patrones completamente opuestos, a excepción de la dirección del viento.

Al observar los elementos del clima, se encontró que HR y PV diferencian de forma más extrema los grupos, mientras que el resto de las características T, RS, VV y DV, definen esta diferenciación de forma más suave.

IV.3.1.1. Caracterización estadística del agrupamiento K-Medias

El cuadro 4.4 y la figura 4.12, muestran de manera general la estructura de los grupos formados; sin embargo éstas son sólo representaciones de los centros geométrico dimensionales de los grupos formados. Para identificar la forma como se presenta la concentración de los patrones en cada cluster, es necesaria la exploración estadística de los elementos climáticos de cada uno de los subgrupos formados.

Figura 4.12 Estructura de los patrones climáticos característicos obtenidos a través del algoritmo de agrupamiento K-Medias.



Fuente: Elaboración propia.

La caracterización estadística de los cuatro subgrupos resultantes a partir de la aplicación del algoritmo K-Medias, se presenta en el cuadro 4.5. Se observa que los valores extremos del atributo VV (cuadro 4.5, figura 4.3d), quedaron incluidos

dentro de los grupos C3, C4 y los valores atípicos del mismo se repartieron en todos los grupos (cuadro 4.5, figura 4.13d). En cuanto a los valores atípicos del atributo HR (cuadro 4.5, figura 4.3c), éstos se agruparon el grupo C3 (cuadro 4.5, figura 4.13c); los valores atípicos presentes en el atributo PV quedaron distribuidos en todos los grupos.

Por otra parte, se observa que a pesar de que el atributo RS no registró valores extremos ni atípicos en el análisis exploratorio (figura 4.3b), presenta el mayor valor (extremo) del coeficiente de variación, en los grupos C1 y C2 también reflejado en el diagrama de dispersión de la figura 4.13b. Otro atributo con una distribución heterogénea en los grupos C1 y C4 fue la VV (figura 4.13d); el resto de los atributos presenta coeficientes de variación aceptables, en cada uno de los grupos formados.

Las pruebas F y Kruskal-Wallis permiten evidenciar diferencias significativas entre los cuatro grupos formados para cada una de las características climáticas que conforman cada patrón (figura 4.13). En este conjunto de gráficas se observó que para el 75 % de los patrones incluidos en cada *cluster*, las características de los elementos climáticos son:

Cluster C1. Temperaturas menores a 15°C, humedad relativa mayor al 60%, presión de vapor mayor 0.9 Kilo-pascales, radiación solar menor a 0.02 Kw/m^2, velocidad del viento menor a 0.9 m/s y dirección del viento entre 210° a 350° -de suroeste a este- (figura 4.13).

Cluster C2. Temperaturas menor a 20°C -el 50% del total de patrones en este grupo presenta temperaturas menores a 15.7°C-, humedad relativa menor a 40%, presión de vapor menor a 0.7 Kilo-pascales, radiación solar menor 0.06 Kw/m^2, velocidad del viento mayor a 0.7 m/s y dirección del viento entre 245° y 350° -de sur a este- (figura 4.13).

Cluster C3. Temperaturas mayores a 22°C, humedad relativa menor al 28%, presión de vapor menor a 0.9 Kilo-pascales, radiación solar mayor a 0.48 Kw/m^2, velocidad del viento mayor a 0.9 m/s y dirección del viento entre 25° y 240° -de noreste a suroeste- (figura 4.13).

Cluster C4. Temperaturas mayores a los 15°C, humedad relativa menor al 45%, presión de vapor menor a 0.9 Kilo-pascales, radiación solar mayor a 0.2 Kw/m^2, velocidad del viento mayor a 0.2 m/s y dirección del viento menor a los 110° -este a norte- (figura 4.13).

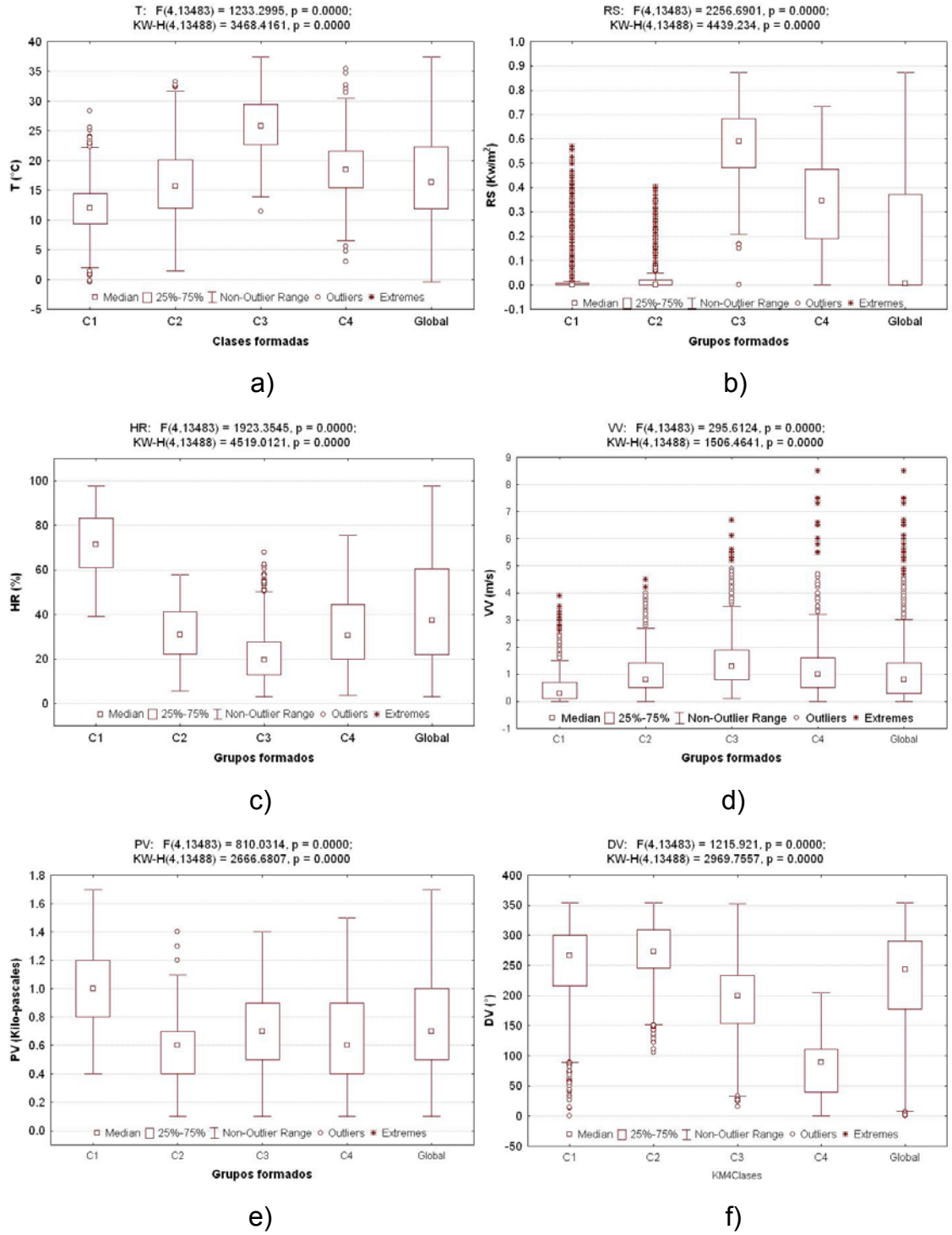
A partir del agrupamiento de los patrones climáticos a través del algoritmo K-medias, se etiquetaron cada uno de los patrones, para posteriormente organizarlos cronológicamente (columnas: horas del día y renglones días transcurridos agrupados por etapa fenológica) en un mapa climático (figura 4.14). En éste se observa que aunque en el *cluster* C1 su presencia tiene un predominio nocturno, hay días completos con el tipo de clima agrupado en este *cluster*, así mismo se observa su presencia en las primeras horas del día; la presencia del clima agrupado en el *cluster* C2, se presentó desde el ocaso hasta las primeras horas de la mañana; La características climáticas representadas por el *cluster* C4, generalmente predominaron entre las nueve de la mañana hasta el medio día; finalmente durante el día (entre las 10:00 am y las 5:00 pm), el predomino fue el clima caracterizado por el *cluster* C3.

Cuadro 4.5 Caracterización estadística del agrupamiento K-Medias.

		Características de los patrones climáticos					
Clase	**Indicador**	**T**	**HR**	**PV**	**RS**	**VV**	**DV**
C1	Frecuencia	2141	2141	2141	2141	2141	2141
	Porcentaje N	31.7467	31.7467	31.7467	31.7467	31.7467	31.7467
	Media	12.0556	71.9331	1.0188	0.0325	0.5374	251.0001
	Mediana	12.0000	71.5000	1.0000	0.0000	0.3000	266.0000
	Mínimo	-0.4000	39.3000	0.4000	0.0000	0.0000	0.0000
	Máximo	28.3000	97.8000	1.7000	0.5670	3.9000	354.0000
	Rango	28.7000	58.5000	1.3000	0.5670	3.9000	354.0000
	Desv. Estándar	3.8204	13.2470	0.2396	0.0761	0.5613	65.9061
	C.V.	31.6898	18.4158	23.5195	234.2805	104.4507	26.2574
C2	Frecuencia	2324	2324	2324	2324	2324	2324
	Porcentaje N	34.4603	34.4603	34.4603	34.4603	34.4603	34.4603
	Media	16.1844	31.3034	0.5638	0.0408	0.9703	272.9219
	Mediana	15.7000	30.8000	0.6000	0.0000	0.8000	273.0000
	Mínimo	1.5000	5.8000	0.1000	0.0000	0.0000	105.8000
	Máximo	33.2000	57.8000	1.4000	0.4020	4.0000	354.5000
	Rango	31.7000	52.0000	1.3000	0.4020	4.0000	248.7000
	Desv. Estándar	5.6234	11.8191	0.2046	0.0688	0.7221	45.4321
	C.V.	34.7457	37.7566	36.2789	168.9021	74.4175	16.6466
C3	Frecuencia	1522	1522	1522	1522	1522	1522
	Porcentaje N	22.5682	22.5682	22.5682	22.5682	22.5682	22.5682
	Media	26.0086	21.4856	0.6846	0.5733	1.3013	193.3293
	Mediana	25.8000	19.5000	0.7000	0.5915	1.2000	199.1500
	Mínimo	11.5000	3.0000	0.1000	0.1500	0.0000	15.8000
	Máximo	37.5000	67.8000	1.4000	0.8710	6.7000	353.0000
	Rango	26.0000	64.8000	1.3000	0.7210	6.7000	337.2000
	Desv. Estándar	4.5751	11.0450	0.2651	0.1408	0.8387	57.0902
	C.V.	0.1759	0.5141	0.3873	0.2456	0.6445	0.2953
C4	Frecuencia	0.6818	2.6362	55.3218	41.5231	53.7117	0.1483
	Porcentaje N	11.2248	11.2248	11.2248	11.2248	11.2248	11.2248
	Media	18.5471	32.5724	0.6617	0.3255	1.0512	82.7169
	Mediana	18.5000	30.5000	0.6000	0.3480	0.8000	89.5000
	Mínimo	3.0000	3.5000	0.1000	0.0000	0.0000	0.0000
	Máximo	35.4000	75.5000	1.5000	0.7330	8.5000	205.0000
	Rango	32.4000	72.0000	1.4000	0.7330	8.5000	205.0000
	Desv. Estándar	4.8835	16.3682	0.2919	0.1905	1.0826	49.8655
	C.V.	26.3302	50.2517	44.1135	58.5190	102.9813	60.2846
Total	Frecuencia	6744	6744	6744	6744	6744	6744
	Porcentaje N	100.0000	100.0000	100.0000	100.0000	100.0000	100.0000
	Media	17.3560	42.1287	0.7465	0.1903	0.9166	226.6497
	Mediana	16.3000	37.3000	0.7000	0.0350	0.7000	244.0000
	Mínimo	-0.4000	3.0000	0.1000	0.0000	0.0000	0.0000
	Máximo	37.5000	97.8000	1.7000	0.8710	8.5000	354.5000
	Rango	37.9000	94.8000	1.6000	0.8710	8.5000	354.5000
	Desv. Estándar	7.0266	24.2950	0.3078	0.2506	0.8073	81.2781
	C.V.	40.4851	57.6685	41.2368	131.6853	88.0737	35.8607

Fuente: Elaboración propia.

Figura 4.13 Dispersión de las variables incluidas en los patrones climáticos por grupos formados a través del algoritmo K-Medias.



Fuente: Elaboración propia

IV.3.1.2. Caracterización textual a partir del agrupamiento K-Medias

A partir de los resultados obtenidos del agrupamiento generado a través de la aplicación del algoritmo K-medias fue posible identificar cuatro grupos, mismos que se caracterizan, con base en la escala de Beaufort (Anexo C) de la siguiente manera:

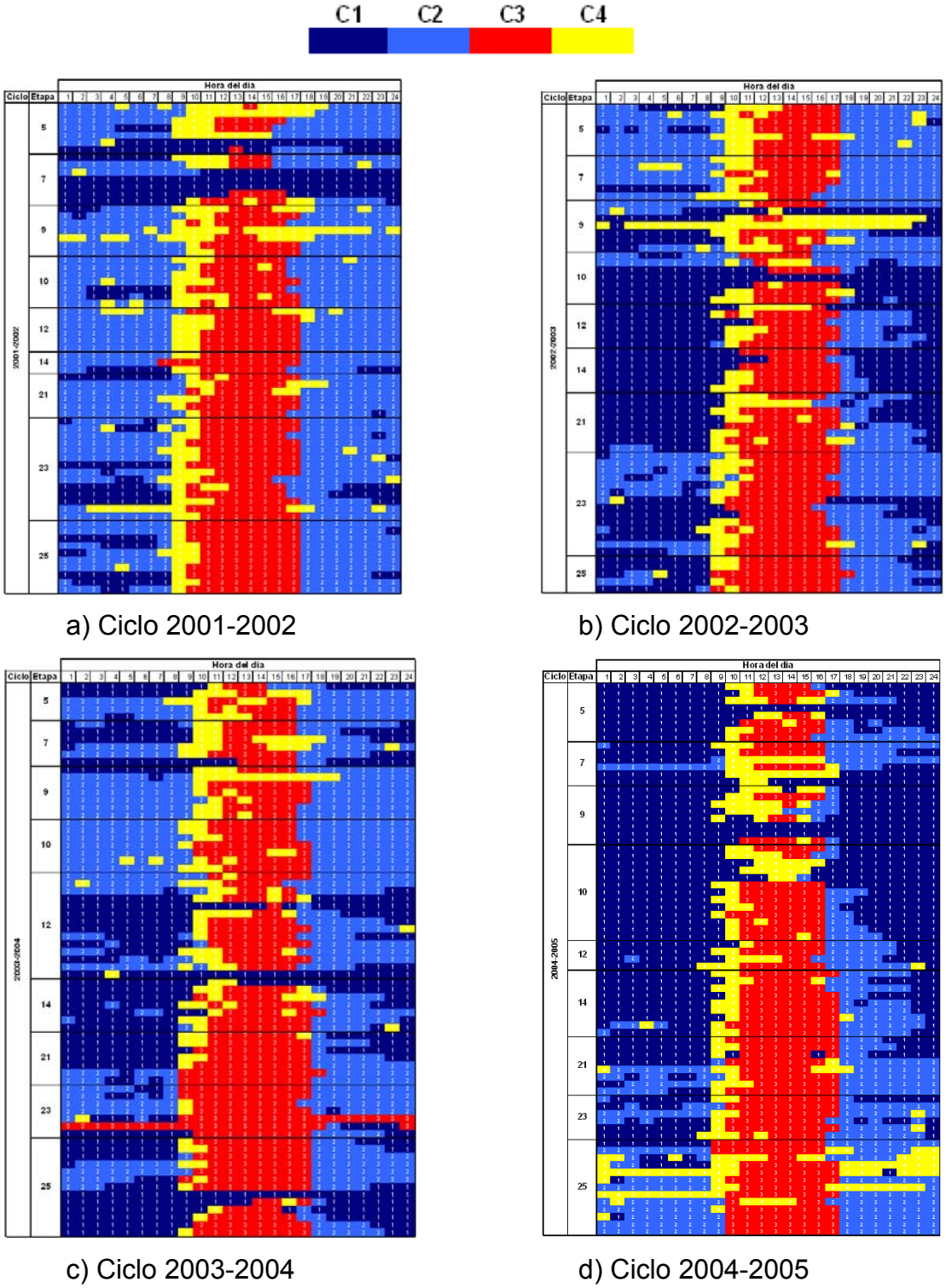
Grupo o *cluster* C1: Viento en calma con rachas de ventolina, frío, con alta humedad, proveniente del norte; generalmente se presenta por la noche o en condiciones de baja radiación solar.

Grupo o *cluster* C2: Viento en calma con rachas de ventolina, generalmente nocturnos, frescos con baja humedad, dirigidos al sureste.

Grupo o *cluster* C3: Viento en calma con rachas ligeras, generalmente presentado por la mañana, calido y seco, dirigidos hacia el oeste.

Grupo o *cluster* C4: Viento en calma con rachas ligeras, generalmente vespertino fresco y con baja humedad, en dirección al noreste.

Figura 4.14 Distribución horaria del agrupamiento K-Medias para cada ciclo productivo.



a) Ciclo 2001-2002

b) Ciclo 2002-2003

c) Ciclo 2003-2004

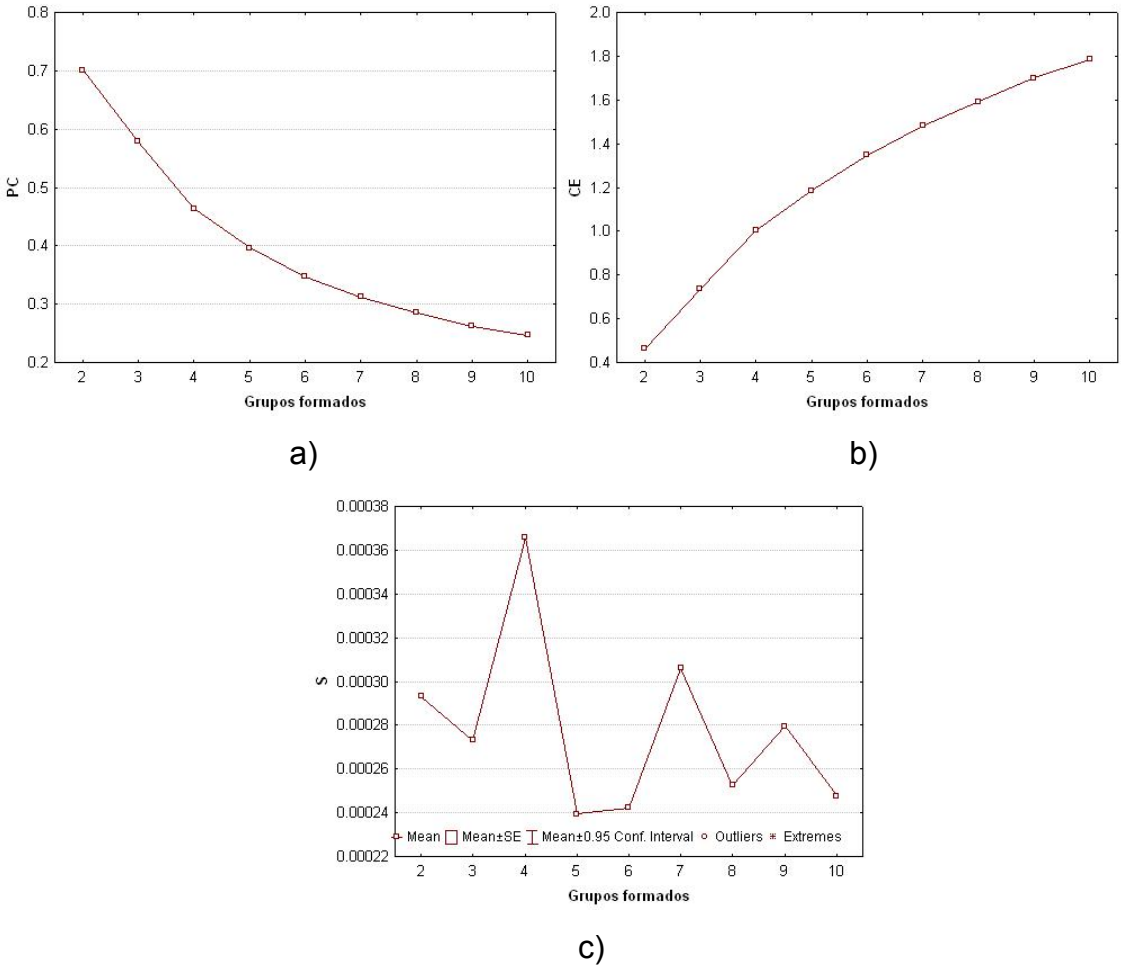
d) Ciclo 2004-2005

Fuente: Elaboración propia.

IV.3.2. Agrupamiento a través del algoritmo difuso FCM

El índice de separación S, que evalúa la relación entre la dispersión interna de los *clusters* y la separación de los mismos, presenta su mínimo local más bajo cuando el conjunto de patrones es agrupado en cinco *clusters*; mientras que los índices indirectos PC y CE, que son calculados relacionando sólo los grados de pertenencia y no el agrupamiento de los patrones, no muestran evidencia contundente para contrarrestar la validez del índice S. Por lo tanto, basados en el comportamiento del índice S, se determinan cinco *clusters* como el agrupamiento que mejor revela la estructura del conjunto de patrones.

Figura 4.15 Índices PC, CE y S, para el agrupamiento FCM.



Fuente: Elaboración propia.

A partir de la determinación de cinco *clusters*, como el agrupamiento que mejor revela la estructura del conjunto de patrones, se calcularon los valores de los elementos climáticos para los patrones característicos que representan los centros de cada uno de los cinco cluster formados, los cuales se muestran en el cuadro 4.6.

Cuadro 4.6 Estructura de los patrones característicos formados a partir del algoritmo de agrupamiento FCM.

		Patrones característicos				
Atributo	**Unidades**	**C1**	**C2**	**C3**	**C4**	**C5**
T	°C	17.37	13.17	21.40	11.79	26.14
HR	%	26.98	49.36	30.98	75.92	19.67
PV	Kilo-pascales	0.52	0.73	0.74	1.04	0.63
RS	Kw/m^2	0.05	0.03	0.39	0.02	0.59
VV	m/s	1.01	0.69	1.16	0.46	1.19
DV	Grados (°)	264.06	268.81	159.91	260.38	171.93

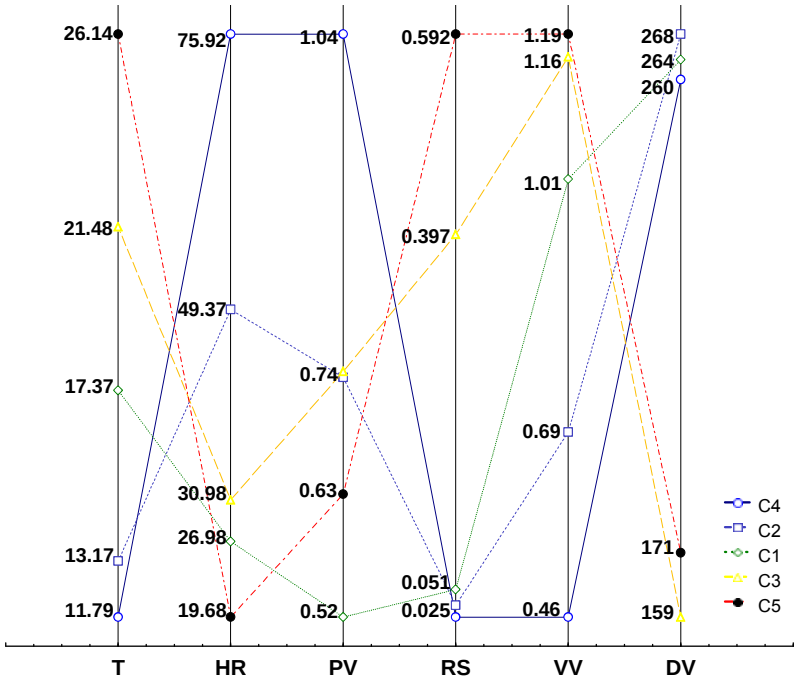
Fuente: Elaboración propia.

Con el objeto de resalar la diferenciación de los cinco grupos generados, se presenta la estructura del conjunto de patrones a través la figura 4.16, donde las coordenadas paralelas representan cada uno de los atributos (dimensiones) que forman los patrones climáticos. Éstos permiten evidenciar la separación de los patrones que caracterizan a cada grupo formado. Así, se puede observar que los grupos más contrastantes son los representados por los patrones característicos 4 y 5.

La descripción estadística de los cinco grupos formados, se presenta en el cuadro 4.7, donde se muestran indicadores de frecuencia, tendencia central y dispersión. En ésta se observa que los valores extremos presentes en las lecturas de la variable velocidad del viento (figura 4.2d), forman parte del grupo C5, mientras que

los valores atípicos de la misma quedaron repartidos en el grupo C3 y C5, los valores atípicos de la variable humedad relativa fueron integrados en el grupo 4; por otra parte, los valor atípicos de la presión de valor se diluyeron entre los grupos C2 al C5.

Figura 4.16 Estructura de los patrones climáticos característicos obtenidos a través del algoritmo de agrupamiento FCM.



Fuente: Elaboración propia.

IV.3.2.1. Caracterización estadística del agrupamiento difuso FCM

Al momento de observar la dispersión en los grupos para cada variable a través del coeficiente de variación, índice que integra los estimadores media y desviación estándar, se puede apreciar cómo la T presenta un comportamiento más homogéneo al interior de los grupos formados ya que este coeficiente varía dentro del rango del 16 a 33 %. En contraste, la variable RS presenta un comportamiento muy heterogéneo al interior de los grupos formados, cuya variación oscila dentro del rango del 18 a 244 %; a excepción del la variable VV, que presenta un menor grado de heterogeneidad al interior de los grupos formados, el resto de las

variables (HR, PV, DV) muestra un comportamiento relativamente homogéneo dentro de las grupos formados.

Cuadro 4.7 Caracterización estadística de los grupos formados a través del algoritmo de agrupamiento FCM.

		Características de los patrones climáticos					
Grupo	**Indicador**	**T**	**HR**	**PV**	**RS**	**VV**	**DV**
C1	Frecuencia	1557	1557	1557	1557	1557	1557
	Porcentaje N	23.087	23.087	23.087	23.087	23.087	23.087
	Media	17.711	23.929	0.487	0.048	1.091	264.295
	Mediana	17.600	24.500	0.500	0.001	0.900	265.800
	Mínimo	3.000	5.800	0.100	0.000	0.000	8.800
	Máximo	33.200	40.800	1.100	0.404	4.400	354.500
	Rango	30.200	35.000	1.000	0.404	4.400	345.700
	Desviación Estándar	5.624	7.874	0.188	0.076	0.765	55.607
	C. V.	31.753	32.906	38.556	160.699	70.108	21.040
C2	Frecuencia	1385	1385	1385	1385	1385	1385
	Porcentaje N	20.537	20.537	20.537	20.537	20.537	20.537
	Media	12.705	49.878	0.740	0.024	0.684	269.845
	Mediana	12.600	49.000	0.700	0.000	0.600	283.800
	Mínimo	-0.400	31.500	0.300	0.000	0.000	0.000
	Máximo	32.700	75.000	1.400	0.339	4.000	353.800
	Rango	33.100	43.500	1.100	0.339	4.000	353.800
	Desviación Estándar	4.145	8.107	0.177	0.051	0.598	56.312
	C. V.	32.627	16.253	23.895	209.434	87.414	20.868
C3	Frecuencia	1005	1005	1005	1005	1005	1005
	Porcentaje N	14.902	14.902	14.902	14.902	14.902	14.902
	Media	20.150	35.253	0.811	0.363	1.153	141.995
	Mediana	20.000	34.000	0.800	0.362	1.000	147.800
	Mínimo	8.300	4.300	0.100	0.000	0.000	0.000
	Máximo	35.700	75.500	1.600	0.721	6.500	349.000
	Rango	27.400	71.200	1.500	0.721	6.500	349.000
	Desviación Estándar	4.506	14.007	0.279	0.145	0.890	75.436
	C. V.	22.362	39.732	34.344	39.877	77.152	53.126
C4	Frecuencia	1548	1548	1548	1548	1548	1548
	Porcentaje N	22.954	22.954	22.954	22.954	22.954	22.954
	Media	11.774	77.671	1.083	0.027	0.476	245.834
	Mediana	11.800	78.000	1.100	0.000	0.300	265.900
	Mínimo	1.400	42.000	0.500	0.000	0.000	0.000
	Máximo	28.300	97.800	1.700	0.567	3.900	354.000
	Rango	26.900	55.800	1.200	0.567	3.900	354.000
	Desviación Estándar	3.548	10.497	0.227	0.065	0.525	71.927
	C. V.	30.136	13.515	20.932	243.752	110.080	29.258
C5	Frecuencia	1249	1249	1249	1249	1249	1249
	Porcentaje N	18.520	18.520	18.520	18.520	18.520	18.520
	Media	26.741	17.705	0.608	0.616	1.312	176.163
	Mediana	26.500	16.300	0.600	0.629	1.100	184.300

	Mínimo	12.500	3.000	0.100	0.266	0.000	2.500
	Máximo	37.500	47.800	1.400	0.871	8.500	353.000
	Rango	25.000	44.800	1.300	0.605	8.500	350.500
	Desviación Estándar	4.368	7.916	0.246	0.113	0.937	68.652
	C. V.	16.336	44.710	40.403	18.335	71.386	38.971
General	Frecuencia	6744	6744	6744	6744	6744	6744
	Porcentaje N	100.000	100.000	100.000	100.000	100.000	100.000
	Media	17.356	42.129	0.747	0.190	0.917	226.650
	Mediana	16.300	37.300	0.700	0.035	0.700	244.000
	Mínimo	-0.400	3.000	0.100	0.000	0.000	0.000
	Máximo	37.500	97.800	1.700	0.871	8.500	354.500
	Rango	37.900	94.800	1.600	0.871	8.500	354.500
	Desviación Estándar	7.027	24.295	0.308	0.251	0.807	81.278
	C. V.	40.485	57.668	41.237	131.685	88.074	35.861

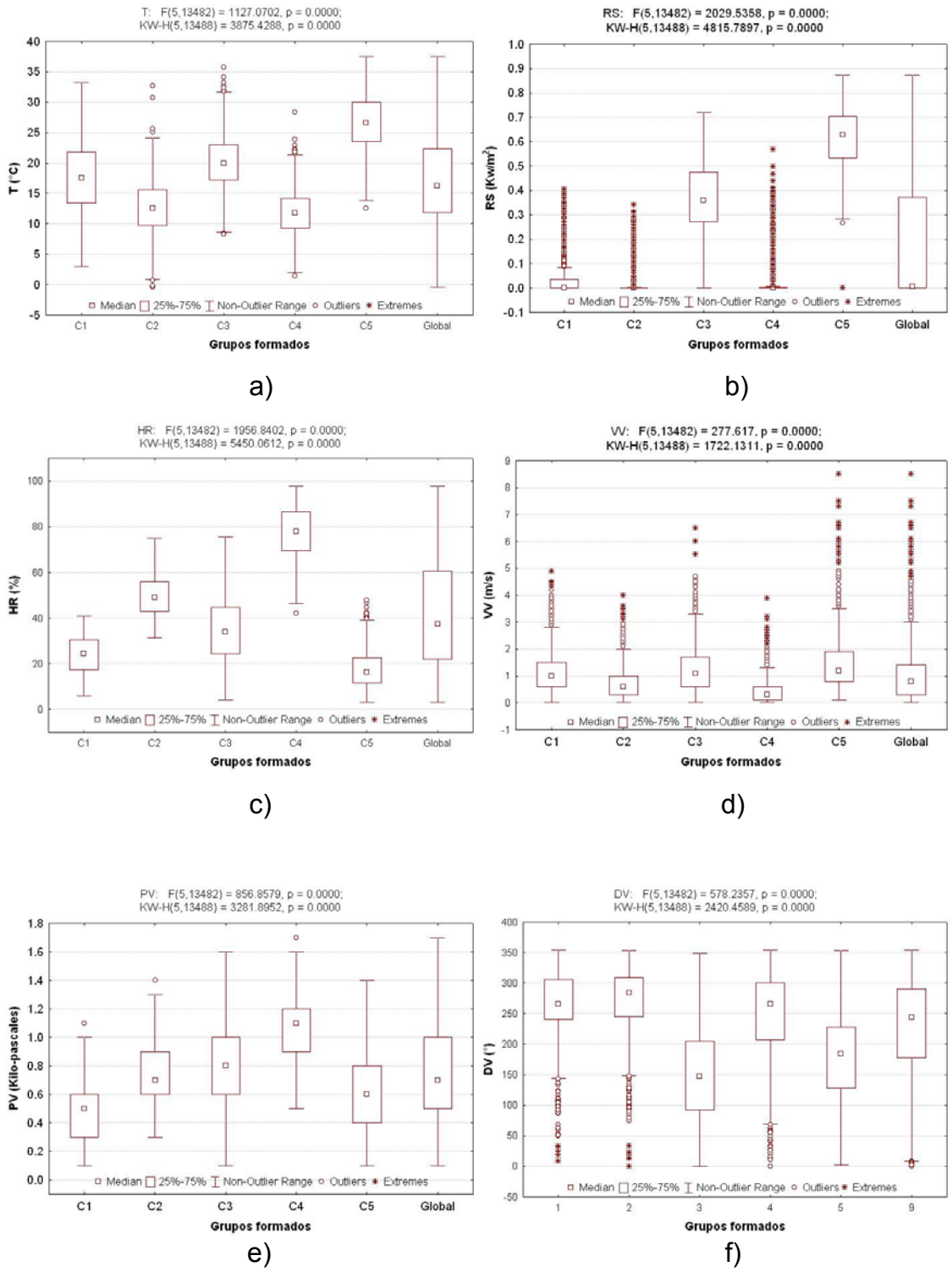
Fuente: Elaboración propia.

Las pruebas F y Kruskal-Wallis permiten evidenciar diferencias significativas entre los cuatro grupos formados, para cada una de las características climáticas que conforman cada patrón (figura 4.17). En este conjunto de gráficas se observa, que para el 75 % de los patrones incluidos en cada *cluster*, las características de los elementos climáticos son:

Cluster C1. Temperaturas mayores a los 15°C, humedad relativa menor al 31%, presión de vapor menor a 0.6 Kilo-pascales, radiación solar menor a 0.08 Kw/m^2, velocidad del viento mayor a 0.7m/s y dirección del viento entre 245° y 350° -sur a este- (figura 4.17).

Cluster C2. Temperaturas menores a los 17°C, humedad relativa mayor al 42%, presión de vapor menor a 0.9 Kilo-pascales, radiación solar menor a 0.02 Kw/m^2, velocidad del viento menor a 1 m/s y dirección del viento entre 247° y 350° ° -sur a este- (figura 4.17).

Figura 4.17 Dispersión de las variables incluidas en los patrones climáticos por grupos formados a través del algoritmo FCM.



Fuente: Elaboración propia

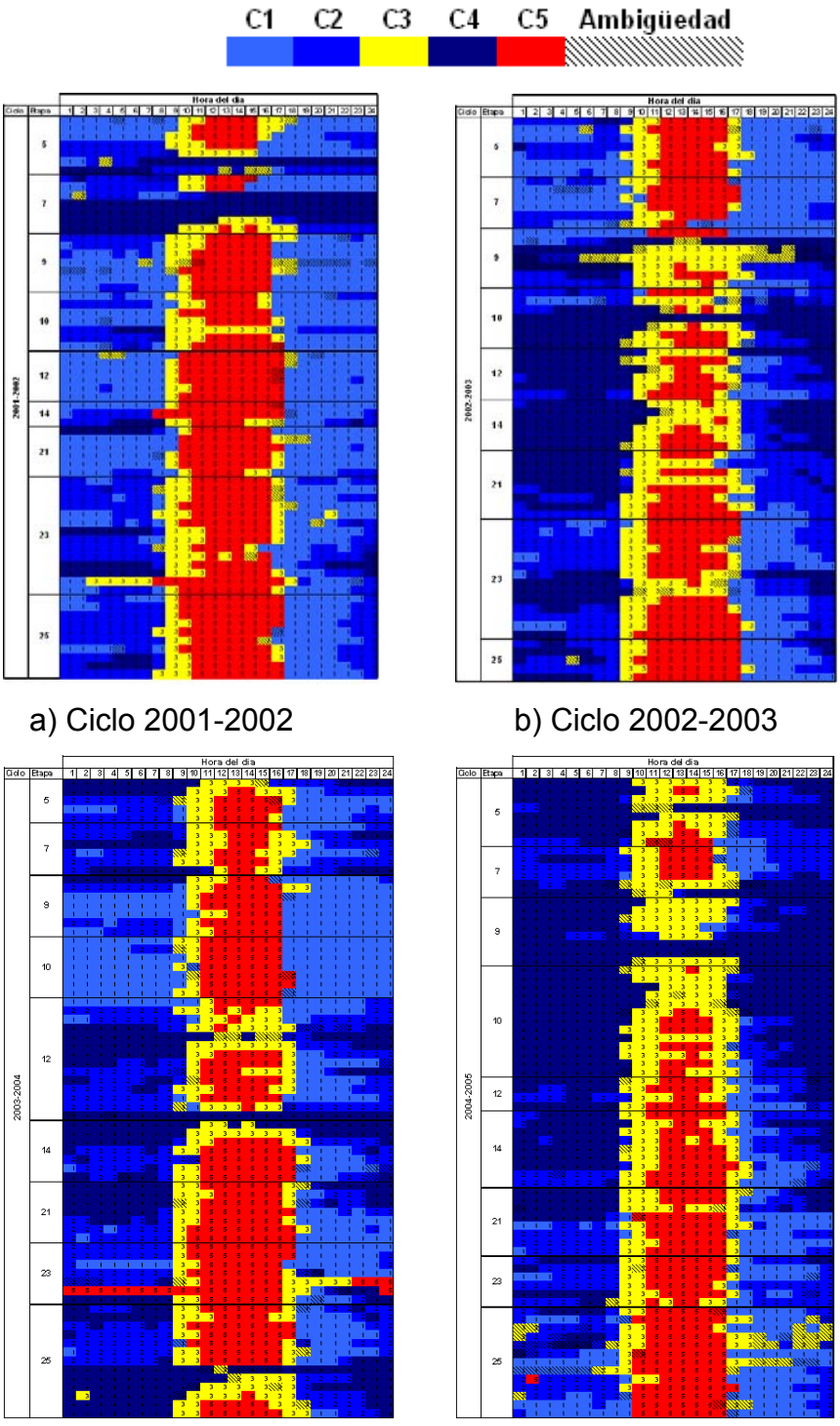
Cluster C3. Temperaturas mayores a los 17°C, humedad relativa menor al 42%, presión de vapor menor a 1 Kilo-pascal, radiación solar mayor a 0.28 Kw/m^2, velocidad del viento mayor a 0.7 m/s y dirección del viento entre los 0° y 200° -del este al suroeste- (figura 4.17).

Cluster C4. Temperaturas menores a los 15°C, humedad relativa mayor al 70%, presión de vapor mayor a 0.9 Kilo-pascales, radiación solar menor de 0.01 Kw/m^2, velocidad del viento menor a 0.5m/s y dirección del viento entre 200° y 350° -de sur a este- (figura 4.17).

Cluster C5. Temperaturas mayores a 23°C, humedad relativa menor al 24%, presión de vapor menor a 0.8 Kilo-pascales, radiación solar mayor a 0.52 Kw/m2, velocidad del viento mayor a 0.7 m/s y dirección del viento entre 0° a 225° -de este a sur oeste- (figura 4.17).

El mapa climático creado a partir de los cinco *clusters* determinados a través del algoritmo difuso FCM, se presenta en la Figura 4.18. En esta representación visual del conjunto de patrones climáticos, la etiqueta ambigüedad indica las fronteras difusas entre uno o varios *clusters,* de acuerdo a la similitud presentada por el grado de pertenencia de los patrones a uno o varios *cluster*.

Figura 4.18 **Distribución horaria del agrupamiento FCM para cada ciclo productivo**.



a) Ciclo 2001-2002 b) Ciclo 2002-2003

c) Ciclo 2003-2004 d) Ciclo 2004-2005

Fuente: Elaboración propia.

La nitidez mostrada en la distribución de patrones, permite identificar que el clima caracterizado por el *cluster* C1 persiste durante toda la noche, las primeras horas de la mañana (antes de la 9:00 am) y las horas cercanas al ocaso (después de las 5:00 pm). Respecto al *cluster* C2, su presencia fue generalmente nocturna y las primeras horas de la mañana (hasta las 8:00 am); las características climáticas agrupadas en el *cluster* C3 generalmente se presentaron entre las 9:00 y 14:00 horas (con un predominio de la 9:00 a 11:00 horas y de 15:00 a 17:00 horas). En cuanto a las condiciones climáticas representadas por el *cluster* C4, su presencia es generalmente nocturna aunque se aprecian bandas que marcan días completos con estas características; finalmente en el *cluster* C5, el clima caracterizado por este grupo se identifica claramente entre las 11:00 y 16:00 horas.

IV.3.2. Caracterización textual a partir del agrupamiento FCM

La caracterización textual de los cinco grupos de datos determinados a partir del algoritmo FCM, puede ser descrita como:

Grupo o *cluster* C1: Vientos calmos con rachas de ventolina en dirección al sur, fresco, con baja humedad relativa, presión de vapor baja y escasa o nula radiación solar.

Grupo o *cluster* C2: Vientos calmos con rachas de ventolina en dirección al sureste, generalmente nocturno, frío, moderadamente húmedo y presión moderada.

Grupo o *cluster* C3: Vientos calmos con rachas ligeras, cálido, de escasa humedad y alta presión, presente durante el día y con vientos dirigidos al noroeste; generalmente presentado entre 9:00 y 1:00 horas por la mañana, y por las tardes de 15:00 a 17:00 horas.

Grupo o *cluster* C4: Clima frío, húmedo y alta presión, generalmente nocturno y vientos calmos con rachas ligeras en dirección al sur y sureste.

Grupo o *cluster* C5: Clima cálido y seco, con presión moderada, con la mayor radiación solar y vientos calmos con rachas ligeras del este, generalmente dos o tres horas antes y después del meridiano.

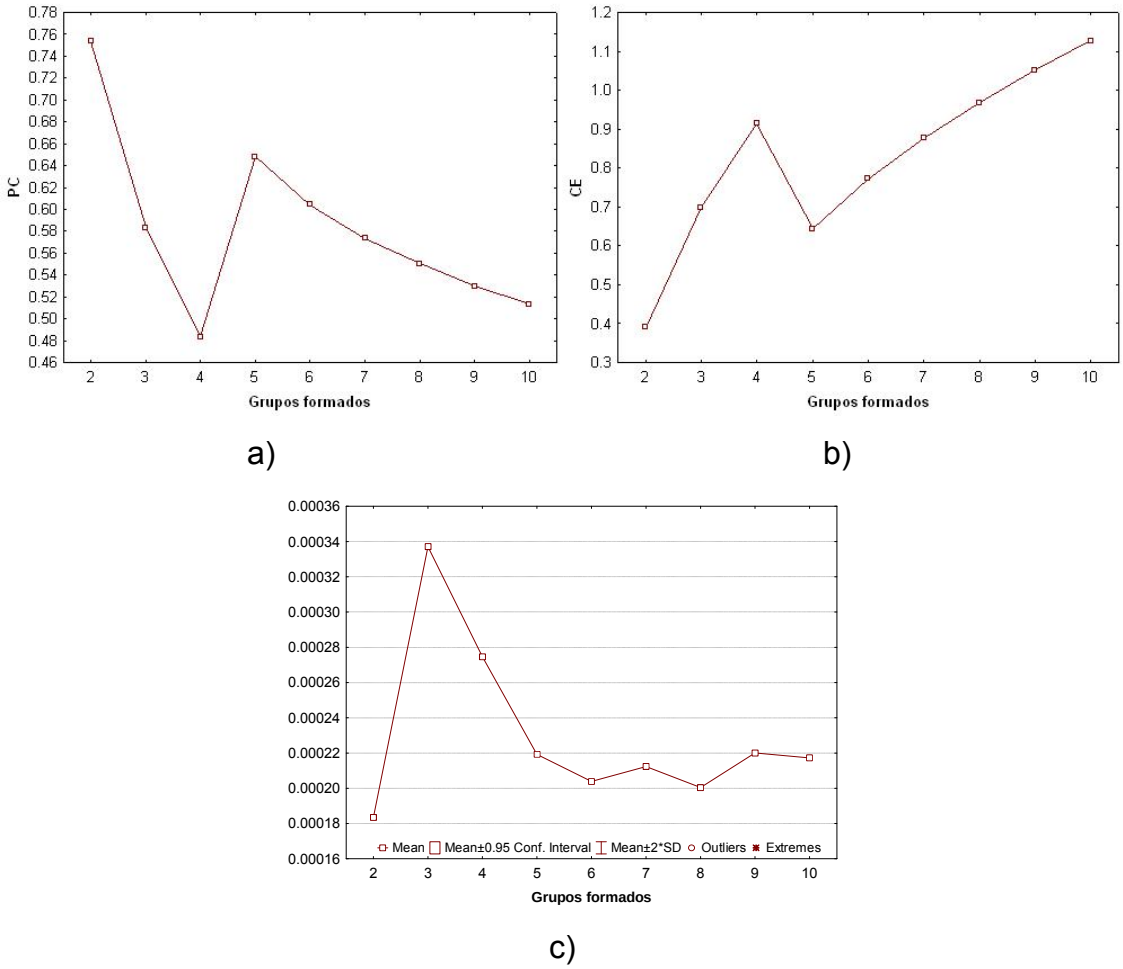
IV.3.2.1. Agrupamiento a través del algoritmo Gustafson-Kessel

Las diferentes ejecuciones del algoritmo difuso Gustafson-Kessel, fueron validadas a través de los índices PC, CE y S; el desempeño de cada uno de estos índices es presentado en la figura 4.19. Se observa que el comportamiento del índice PC, que evalúa la mejor separación de los grupos, muestra un máximo local cuando el conjunto de patrones se agrupa en cinco clusters (figura 4.19a); el índice CE, que expresa la menor ambigüedad, coincide también en cinco grupos al presentar su mínimo local en este número de grupos formados (figura 4.19b); sin embargo, el desempeño del índice S que relaciona la dispersión interna de clusters con la máxima separación de éstos, aunque no presenta su mínimo local en cinco grupos, es precisamente a partir de aquí que inicia su estabilización (figura 4.19c).

Dado el comportamiento mostrado por los tres índices, se determina que cinco *clusters* es la mejor opción para revelar la estructura subyacente al conjunto de patrones, a partir del algoritmo Gustafson-Kessel.

Una vez determinado que cinco grupos constituyen la manera más adecuada para representar la estructura del conjunto de patrones, se calcularon los centros de cada grupo, los cuales se muestran en el cuadro 4.8, a través de los valores de cada uno de los elementos incluidos en cada patrón, considerados para definir la estructura de los patrones característicos.

Figura 4.19 Comportamiento de los índices PC, CE y S, calculado a partir del agrupamiento realizado con el algoritmo Gustafson-Kessel.



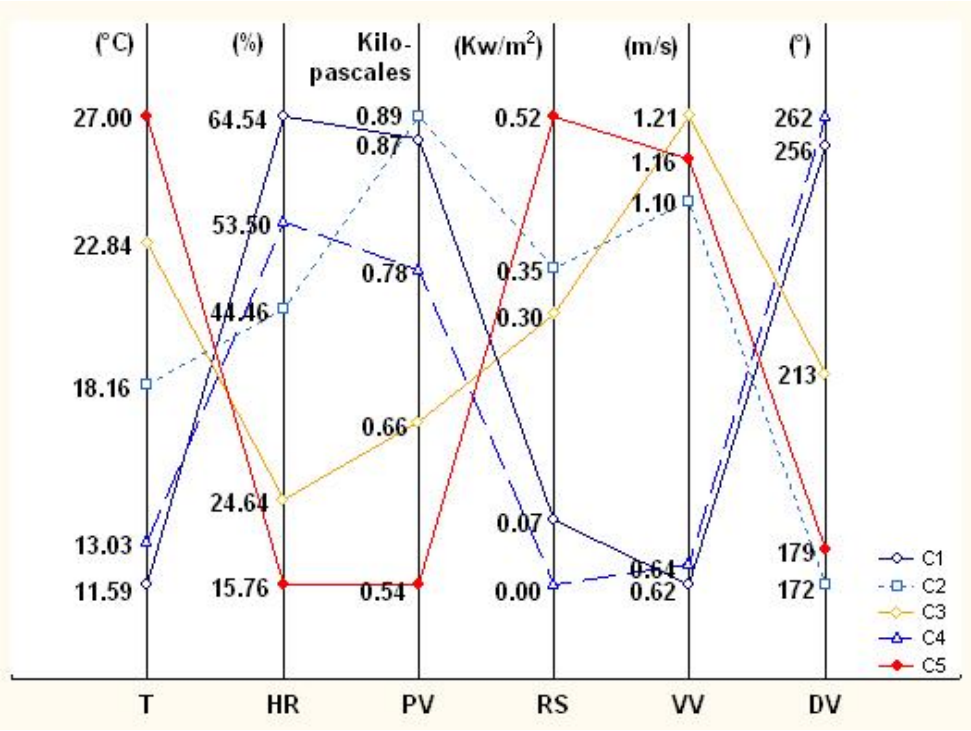
Fuente: Elaboración propia.

Cuadro 4.8 Estructura de los patrones característicos formados a partir del algoritmo de agrupamiento GK.

		Patrones característicos				
Atributo	**Unidades**	**C1**	**C2**	**C3**	**C4**	**C5**
T	°C	11.591	18.163	22.840	13.039	27.009
HR	%	64.544	44.465	24.642	53.509	15.760
PV	Kilo-pascales	0.877	0.895	0.666	0.780	0.544
RS	Kw/m^2	0.072	0.353	0.302	0	0.523
VV	m/s	0.622	1.108	1.215	0.647	1.160
DV	Grados (°)	256.799	172.925	213.267	262.400	179.696

La estructura de los patrones característicos es representada visualmente a través del gráfico de coordenadas paralelas mostrado en la figura 4.16. En ésta se observa en comportamiento inverso entre los patrones más fríos y húmedos (*clusters* C1 y C4) y el patrón más cálido y seco (C5). Resalta el comportamiento de la variable VV, la cual pude diferenciar los patrones característicos en dos grupos. En el resto de las variables, la diferenciación de éstos en menos marcada.

Figura 4.20 Estructura de los patrones climáticos característicos obtenidos a través del algoritmo de agrupamiento Gustafson-Kessel.



IV.3.3. Caracterización estadística del agrupamiento difuso G-K

La descripción estadística de los grupos formados se presenta en el cuadro 4.9. En esta tabla de datos se muestra tanto estimadores de tendencia central como de dispersión. A partir de los valores máximos de las variables incluidas dentro del patrón, se observó que el total de valores atípicos y extremos de la variable VV

(figura 4.2d), forman parte del grupo C5. Respecto a los valores atípicos de la variable HR, éstos se integraron en los grupos C1, C2 y C4.

Respecto a la explicación de homogeneidad de los grupos formados a partir de las variables climáticas estudiadas, en el cuadro 4.9 se observa que las variables RS y VV generaron un alto coeficiente de variación de 131.68 % y 88.07 % respecto al total de patrones. Mientras que por grupo la variable RS presenta altos valores este coeficiente en los grupos C1 y C3 (92.47 % y 97.73 %); variable VV generó un coeficiente de variación alto en todos los grupos (en el rango de 65.99 % a 95.44). El resto de las variables presentaron un comportamiento menos variable, con valores para el coeficiente de variación menores al 50%.

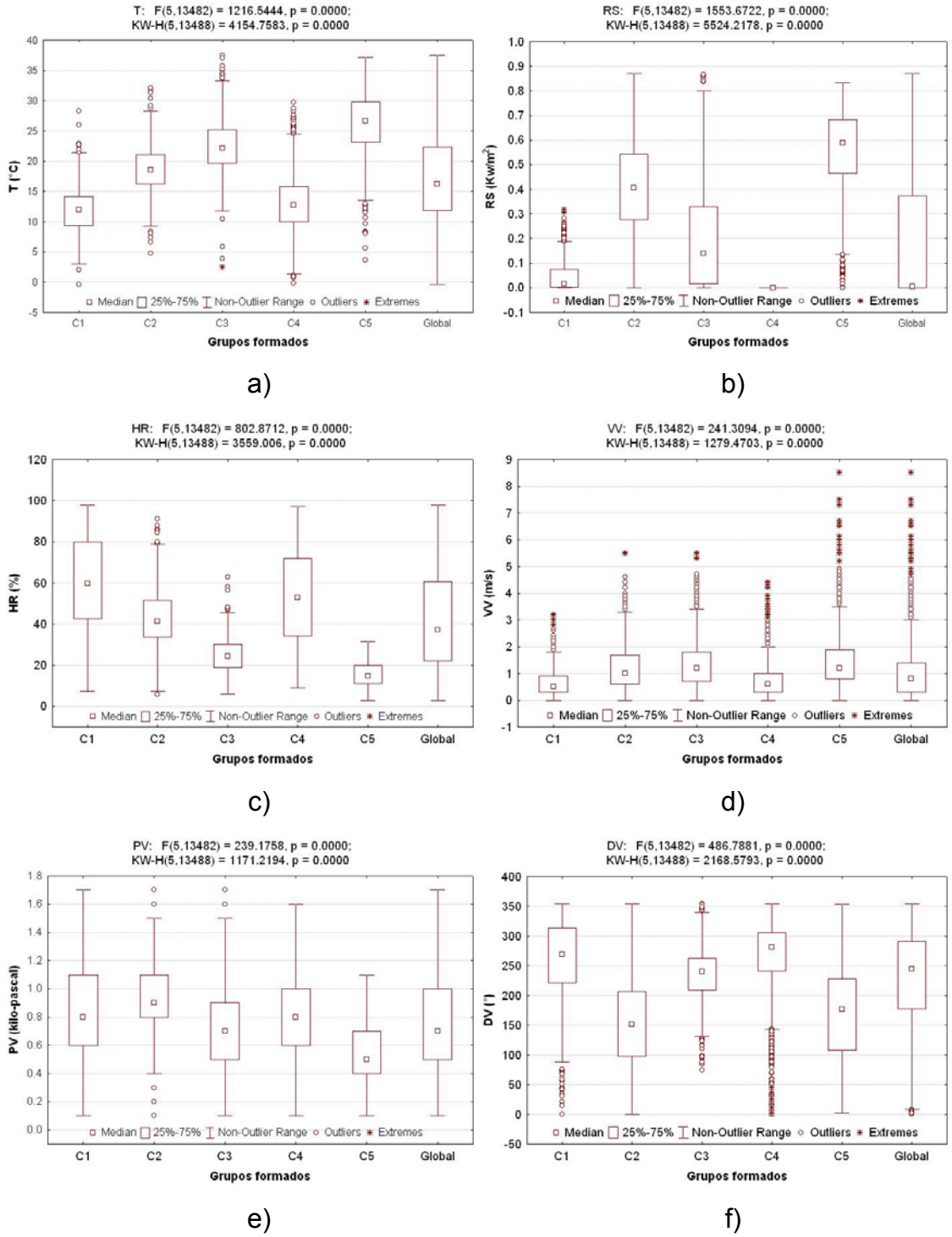
Cuadro 4.9 Caracterización estadística de los grupos formados a través del algoritmo de agrupamiento Gustafson-Kessel.

		Características de los patrones climáticos					
Clase	**Indicador**	**T**	**HR**	**PV**	**RS**	**VV**	**DV**
C1	Frecuencia	1024	1024	1024	1024	1024	1024
	Porcentaje N	15.1839	15.1839	15.1839	15.1839	15.1839	15.1839
	Media	11.8843	59.6659	0.8395	0.0647	0.6436	258.0275
	Mediana	11.9500	59.8000	0.8000	0.0430	0.5000	269.4000
	Mínimo	-0.4000	7.5000	0.1000	0.0010	0.0000	0.0000
	Máximo	28.3000	97.8000	1.7000	0.3160	3.2000	354.5000
	Rango	28.7000	90.3000	1.6000	0.3150	3.2000	354.5000
	Desv. Estándar	3.5567	22.6495	0.3541	0.0600	0.5493	69.8008
	C.V.	29.9277	37.9605	42.1799	92.7357	85.3480	27.0517
C2	Frecuencia	832	832	832	832	832	832
	Porcentaje N	12.3369	12.3369	12.3369	12.3369	12.3369	12.3369
	Media	18.6794	43.4645	0.9273	0.4069	1.1774	150.3026
	Mediana	18.5500	41.5000	0.9000	0.4155	1.0000	151.0000
	Mínimo	4.8000	5.8000	0.1000	0.0010	0.0000	0.0000
	Máximo	32.1000	91.0000	1.7000	0.8710	4.2000	354.0000
	Rango	27.3000	85.2000	1.6000	0.8700	4.2000	354.0000
	Desv. Estándar	3.8327	13.3182	0.2565	0.1872	0.8555	73.6399
	C.V.	20.5183	30.6416	27.6610	46.0064	72.6601	48.9944
C3	Frecuencia	965	965	965	965	965	965
	Porcentaje N	14.3090	14.3090	14.3090	14.3090	14.3090	14.3090
	Media	22.5786	24.7698	0.6830	0.2072	1.2911	234.9869
	Mediana	22.2000	24.3000	0.7000	0.1470	1.2000	239.5000
	Mínimo	2.4000	6.0000	0.1000	0.0010	0.0000	74.5000
	Máximo	37.5000	63.0000	1.7000	0.8651	4.7000	354.3000

	Rango	35.1000	57.0000	1.6000	0.8641	4.7000	279.8000
	Desv. Estándar	4.4932	8.3076	0.2506	0.2025	0.8521	45.7453
	C.V.	19.9003	33.5392	36.6911	97.7317	65.9980	19.4672
C4	Frecuencia	2692	2692	2692	2692	2692	2692
	Porcentaje N	39.9170	39.9170	39.9170	39.9170	39.9170	39.9170
	Media	13.0584	53.4521	0.7800	0.0000	0.6492	262.2864
	Mediana	12.7000	53.0000	0.8000	0.0000	0.5000	280.8000
	Mínimo	-0.2000	8.8000	0.1000	0.0000	0.0000	0.0000
	Máximo	29.7000	97.3000	1.6000	0.0000	4.4000	354.5000
	Rango	29.9000	88.5000	1.5000	0.0000	4.4000	354.5000
	Desv. Estándar	4.3940	22.6030	0.3019	0.0000	0.6196	66.8958
	C.V.	33.6488	42.2865	38.7051	--	95.4405	25.5049
C5	Frecuencia	1231	1231	1231	1231	1231	1231
	Porcentaje N	18.2533	18.2533	18.2533	18.2533	18.2533	18.2533
	Media	26.3173	15.4833	0.5236	0.5512	1.2589	167.6819
	Mediana	26.6000	14.8000	0.5000	0.5880	1.0000	176.8000
	Mínimo	3.7000	3.0000	0.1000	0.0010	0.0000	2.5000
	Máximo	37.2000	31.5000	1.1000	0.8340	8.5000	353.0000
	Rango	33.5000	28.5000	1.0000	0.8330	8.5000	350.5000
	Desv. Estándar	4.8842	5.7118	0.1937	0.1752	0.9627	77.4114
	C.V.	18.5589	36.8901	36.9939	31.7852	76.4715	46.1656
Total	Frecuencia	6744	6744	6744	6744	6744	6744
	Porcentaje N	100.0000	100.0000	100.0000	100.0000	100.0000	100.0000
	Media	17.3560	42.1287	0.7465	0.1903	0.9166	226.6497
	Mediana	16.3000	37.3000	0.7000	0.0350	0.7000	244.0000
	Mínimo	-0.4000	3.0000	0.1000	0.0000	0.0000	0.0000
	Máximo	37.5000	97.8000	1.7000	0.8710	8.5000	354.5000
	Rango	37.9000	94.8000	1.6000	0.8710	8.5000	354.5000
	Desv. Estándar	7.0266	24.2950	0.3078	0.2506	0.8073	81.2781
	C.V.	40.4851	57.6685	41.2324	131.6868	88.0755	35.8607

Fuente: Elaboración propia.

Figura 4.21 Dispersión de las variables incluidas en los patrones climáticos por grupos formados a través del algoritmo GK.



Fuente: Elaboración propia.

Las pruebas F y Kruskal-Wallis permiten evidenciar diferencias significativas entre los cuatro grupos formados para cada una de las características climáticas que conforman cada patrón (figura 4.21). En este conjunto de gráficas se observó, que para el 75 % de los patrones incluidos en cada *cluster*, las características de los elementos climáticos son:

Cluster C1. Temperaturas menores a los 15°C (figura 4.21a), humedad relativa mayor al 42 % (figura 4.21c), presión de vapor mayor a 0.6 Kilo-pascales (figura 4.21e), radicación solar menor a 0.1 Kw/m^2 (figura 4.21b), velocidad del viento menor a un metro por segundo (figura 4.21d) y dirección del viento entre 225° y 350° -de suroeste a este- (figura 4.21f).

Cluster C2. Temperaturas mayores a los 17°C (figura 4.21a), humedad relativa menor al 50% (figura 4.21c), presión de vapor mayor 0.8 Kilo-pascales (figura 4.21e), radiación solar mayor a 0.25 Kw/m^2 (figura 4.21b), velocidad del viento mayor a 0.7 m/s (figura 4.21d) y dirección de viento entre 0° y 200° -entre este y oeste- (figura 4.21f).

Cluster C3. Temperaturas mayores a los 19°C (figura 4.21a), humedad relativa menor al 30 % (figura 4.21c), presión de vapor mentor a 0.9 Kilo-pascales (figura 4.21e), radiación solar menor a 0.35 Kw/m^2 (figura 4.21b), velocidad del viento mayo a 0.8 m/s (figura 4.21d) y dirección del viento entre los 130° y 340° -entre el noroeste y sureste- (figura 4.21f).

Cluster C4. Temperaturas menores a 17°C (figura 4.21a), humedad relativa mayor al 35 % (figura 4.21c), presión de vapor mayor a 0.6 Kilo-pascales (figura 4.21e), sin radicación solar (figura 4.21b), velocidad del viento menor a 0.9 m/s (figura 4.21d) y dirección del viento entre los 240° y 350° -entre sur y este (figua 3.21f).

Cluster C5. Temperaturas mayores a 23°C (figura 4.21a), humedad relativa menor al 20 % (figura 4.21b), presión de vapor menor a 07 Kilo-pascales (figura 4.21e), radiación solar mayor a 0.47 Kw/m^2 (figura 4.21b), velocidad del viento mayor a 0.6 m/s (figura 4.21d) y dirección del viento entre 0° y 225° -entre este y suroeste- (figura 4.21f).

Con base en el agrupamiento logrado por el algoritmo difuso Gustafson-Kessel, se construyó el mapa climático presentado en la figura 4.22. En este mapa se aprecia como el clima caracterizado por el cluster C1, define claramente la primeras horas de la mañana (7:00 am a 9:00 am); respecto al *cluster* C2 su presencia generalmente fue durante las primer parte del día (a partir de las 9:00 am) por las tardes (5:00 pm en adelante) el clima estuvo caracterizado por el *cluster* C3; en caso del *cluster* C4 su presencia fue nocturna (entre 7:00 pm y 7:00 am); finalmente el cluster C5, se presentó entre de 9:00 a 17 horas.

IV.3.3.2. Caracterización textual a partir del agrupamiento GK

A partir de análisis estadístico de los grupos formados a través de la aplicación del algoritmo de agrupamiento difuso Gustafson-Kessel se derivo la descripción textual de estos grupos, con base en la escala de Beaufort (Anexo C) la cual se presenta a continuación.

Grupo o *cluster* C1: Clima con vientos calmos con rachas de ventolina con dirección al sur, frío con alta humedad relativa, presión de vapor alta y escasa o nula radiación solar; generalmente presentado durante las primeras horas de la mañana.

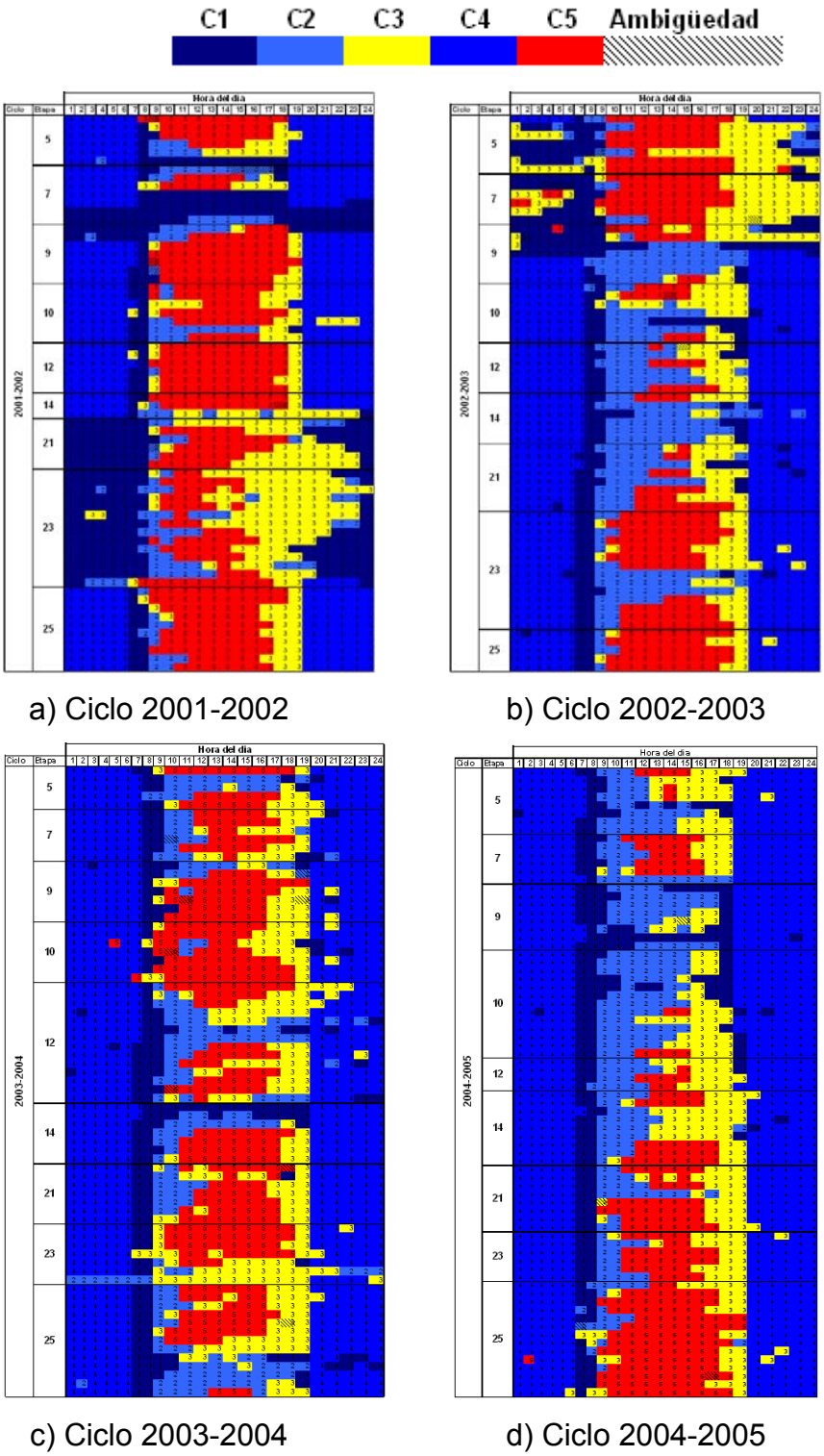
Grupo o *cluster* C2: Vientos clamos con rachas de ventolina dirigidos al sureste, fresco con humedad relativa media, presión alta y radiación solar alta, generalmente presentado durante la mañana.

Grupo o *cluster* C3: Clima calido de escasa humedad y presión de vapor media, generalmente vespertino, con presencia de vientos calmos con rachas ligeras, en direcciones que varían desde el noroeste hasta el este.

Grupo o *cluster* C4: Clima frío húmedo y alta presión, vientos nocturnos calmos con rachas de ventolina, en dirección al sur y sureste.

Grupo o *cluster* C5: Clima calido y seco con presión moderada con la mayor radiación solar y vientos calmos con rachas ligeras provenientes este, presente entre las 9:00 y 17:00.

Figura 4.22 Distribución horaria del agrupamiento GK, para cada ciclo productivos.



a) Ciclo 2001-2002 b) Ciclo 2002-2003

c) Ciclo 2003-2004 d) Ciclo 2004-2005

Fuente: Elaboración propia.

IV.3.2. Distancias entre centros de cluster

Con el propósito de cuantificar la diferencia entre los centros de *cluster* derivados de los distintos algoritmos de agrupamiento utilizados. Se calcularon las distancias Euclidianas las cuales se presentan a continuación.

En el cuadro 4.10, se observa que distancia más pequeña se presenta entre los *clusters* C3 y C4 (0.2318), los cuales muestran un comportamiento similar (figura 4.12), por los que se consideran como lo más semejantes dentro de este agrupamiento. En contraste, la máxima distancia se observa entre el *cluster* C1 y C3 (0.8824), esto se confirma a través del comportamiento completamente opuesto (figura 4.12), por lo que se consideran los dos *clusters* más disímiles.

Cuadro 4.10 Distancias entre los centros de cluster
derivados del algoritmo K-Medias

	C1	C2	C3	C4
C1	--			
C2	0.3382	--		
C3	0.8824	0.5093	--	
C4	0.5938	0.4028	0.2318	--

Fuente: Elaboración propia.

Las distancias entre los clusters derivados del algoritmo de agrupamiento FCM, son exhibidas en el Cuadro 4.11. En éste se observa que las parejas de clusters C3 y C5, además de C1 y C2, presentan un alto grado de similitud en su comportamiento (figura 4.16), de acuerdo la cercanía presentada (0.0859 y 0.0882, respectivamente); mientras que la mayor distancia se presenta entre los cluster C4 y C5, lo que coincide con el comportamiento completamente inverso observado en la figura 4.16.

Cuadro 4.11 Distancias entre los centros de cluster derivados del algoritmo FCM

	C1	C2	C3	C4	C5
C1	--				
C2	0.0882	--			
C3	0.2765	0.3538	--		
C4	0.4017	0.1188	0.5953	--	
C5	0.5178	0.7051	0.0859	1.0571	--

Fuente: Elaboración propia.

Respecto a las distancia entre los centros de cluster derivados del algoritmo G-K, éstas se presentan en el cuadro 4.12. En éste se observa que la distancia minima está entre C1 y C4, los cuales presentan un comportamiento muy semejante, así mismo C2 y C3 presenta alta similitud, como también C3 y C5 (figura 4.20). En el caso contrario se encuentran C1 y C5, los cuales presentan un comportamiento completamente opuesto (figura 4.20).

Cuadro 4.12 Distancias entre los centros de cluster derivados del algoritmo G-K.

	C1	C2	C3	C4	C5
C1	--				
C2	0.2380	--			
C3	0.3721	0.0959	--		
C4	0.0259	0.2638	0.3089	--	
C5	0.7930	0.2328	0.1003	0.7359	--

Fuente: Elaboración propia.

IV.4. Fenología a través del agrupamiento de patrones climáticos

Con el propósito de cuantificar los efectos del comportamiento de los elementos del clima incluidos dentro de los patrones climáticos y el desarrollo vegetativo de la planta que fue expresado a través de la etapas fenológicas, se realizó un análisis de correlación entre la cantidad de patrones de cada cluster y la duración en días de cada etapa fenológica, para los cuatro ciclos productivos estudiados. De encontrarse correlación positiva entre la cantidad de patrones por cluster y la duración de la etapa fenológica, esto indicará que características climáticas tienen un efecto retardador natural en el desarrollo de la planta; en el caso contrario, una correlación negativa indicará un efecto precursor de desarrollo vegetativo.

En el cuadro 4.13 se presentan los resultados del análisis de correlación entre la cantidad de patrones por *cluster* y los días de duración de la etapa fenológica para los tres algoritmos de agrupamiento utilizados en la formación de los clusters. A continuación se describen los principales resultados obtenidos[31]:

Durante el desarrollo de la etapa 05, los patrones de los *clusters* $C1_{KM}$, $C4_{FCM}$ y $C1_{GK}$, tuvieron un efecto retardador en el desarrollo vegetativo, mientras que en los patrones de los clusters $C2_{KM}$, $C5_{FCM}$ su efecto fue de precursor de desarrollo vegetativo (cuadro 4.13). En la figura 4.16, se observa que los *clusters* FCM presentan un comportamiento totalmente opuesto (cuadro 4.11); en cuanto a los patrones de los *clusters* KM, a pesar de no ser completamente antagónicos presentan una diferencia marcada respecto a los elementos HR, PV, VV y en

[31] Con la finalidad de facilitar la lectura se definen los siguientes términos que tendrán estarán presetes en esta sección:

Por etapa se deberá entender etapa fenológica; para diferenciación de los cluster la etiquetas se nombraran de acuerdo a:

C + Número + Algoritmo,
donde Algoritmo: KM, FCM, GK (K-Medias, Fuzzy C-Means y Gustafson-Kessel).

Ejemplo: $C1_{KM}$, Se refiere al cluster C1 determinado a través del algoritmo K-Medias.

menor grado T (figura 4.12); para el caso de los patrones de los *clusters* GK, que no prestan efecto precursor, los patrones de $C5_{GK}$ (que presenta un comportamiento totalmente opuesto a $C1_{GK}$) no muestran correlación respecto a la duración de etapa (cuadro 4.13).

En relación al desarrollo de las etapas 07 y 09, éstas fueron afectadas tanto por condiciones climáticas que limitaron el desarrollo vegetativo, de acuerdo con la presencia de patrones incluidos dentro de los clusters $C1_{KM}$, $C4_{FCM}$ y $C1_{KG}$, así como también por características climáticas precursoras del desarrollo, a través de la presencia de patrones pertenecientes a los clusters $C4_{KM}$, $C5_{FCM}$ y $C5_{GK}$ (cuadro 4.13). El comportamiento de estas dos grupos de *clusters* expresan estructuras completamente antagónicas (figuras 4.12, 4.16 y 4.20).

Durante el desarrollo de la etapa 10, las condiciones climáticas limitaron el desarrollo vegetativo de acuerdo con la presencia de los patrones incluidos en los *clusters* $C1_{KM}$, $C4_{FCM}$ y $C1_{KG}$; mientras que las condiciones climáticas caracterizadas en los patrones pertenecientes a los *clusters* $C2_{KM}$, $C5_{FCM}$ y $C1_{GK}$, coadyuvaron al desarrollo de esta etapa. Aunque los *clusters* derivados de los algoritmos FCM y KG, son totalmente contrastantes, los *clusters* $C1_{KM}$ y $C2_{KM}$, no son completamente opuestos; sin embargo, las características HR, PV y VV difieren bastante en ambos *clusters*.

Durante el desarrollo de la etapa 12, sólo hubo presencia de características climáticas limitantes, siendo los patrones pertenecientes a los *clusters* $C3_{KM}$, $C3_{FCM}$ y $C4_{GK}$, los que más limitaron su desarrollo, mientras que los patrones de los clusters $C1_{KM}$, $C1_{FCM}$ y $C5_{GK}$, fueron los que menos limitaron el desarrollo. Cabe hacer notar que los dos grupos de *clusters* presentan ciertas diferencias en la mayor parte de los elementos del clima incluidos en la estructura de los patrones.

Entre las condiciones climáticas que limitaron el desarrollo de la etapa 14, están los patrones de todos los clusters KM, siendo los del $C3_{KM}$ el que más retrasó el desarrollo y los incluidos dentro de $C1_{KM}$ los que menos limitaron el desarrollo. Este par de *clusters* presentan comportamiento bien diferenciado; lo mismo sucede con los patrones de los clusters $C4_{GK}$ y $C2_{GK}$ (máxima y mínima limitación respectivamente, Cuadro 4.13). Para el caso de los clusters determinados con el algoritmo FCM, los patrones incluidos en el *cluster* $C2_{FCM}$ limitaron el desarrollo, mientras que los patrones dentro del *cluster* $C1_{FCM}$ funcionaron como precursores del desarrollo vegetativo en esta etapa (Cuadro 4.13); cabe destacar que a pesar de la cercanía entre estos dos *clusters* (distancia Euclidiana de 0.0882, cuadro 4.11), presentan diferencia considerable entre ellos en las características T, HR, PV y VV (figura 4.16).

La diferenciación del efecto debido a la presencia de tipos de patrones climáticos durante el desarrollo de la etapa 21, muestra que en los patrones pertenecientes a los *clusters* $C1_{KM}$, $C3_{FCM}$ $C4_{GK}$, se expresó un retardo en el desarrollo vegetativo de la planta, mientras que los patrones incluidos en los *clusters* $C2_{KM}$, $C1_{FCM}$, $C5_{GK}$, coadyuvaron en el desarrollo de esta etapa fenológica (cuadro 4.13). Aunque los centros de los *clusters* KM, no presentan una lejanía considerable (cuadro 4.10), exhiben una marcada diferenciación en los valores de los atributos que los componen, específicamente en T, HR, PV y VV (figura 4.12); en el caso de los *clusters* FCM, la diferenciación está en los atributos T, HR, PV, RS y VV (figura 4.16). Para el caso de los *clusters* GK, sus centros de *cluster* presentan una distancia espacial considerable (0.7359, cuadro 4.12), mostrando diferencias considerables en sus atributos T, HR, PV, RS y DV (figura 4.20).

En el desarrollo de la etapa 23, no se presentaron correlaciones negativas que evidencien la presencia de patrones precursores de desarrollo; sin embargo, es en los patrones pertenecientes a los *clusters* $C1_{KM}$, $C5_{FCM}$, y $C4_{GK}$ donde se presenta un menor efecto retardador del desarrollo; el máximo efecto se registró en los patrones de los *clusters* $C2_{KM}$, $C1_{FCM}$ y $C5_{GK}$. Los *clusters* KM, a pesar de

presentar cierta cercanía p-espacial (cuadro 4.10), muestran diferencias considerables en los elementos del clima T, HR, PV y VV (figura 4.12). En cuanto a los patrones de los *clusters* FCM, a pesar de no presentar una diferencia nítida en cuanto a su efecto retardador, su distancia p-espacial es considerable (cuadro 4.11), además de mostrar diferencias en T, PV y RS (figura 4.16). Finalmente, en los patrones de los *clusters* GK, que presentan una distancia p-espacial considerable (cuadro 4.12), su efecto retardador está bien diferenciado (cuadro 4.13), siendo los patrones de $C4_{GK}$ los de menor efecto, además de presentar diferencias en T, HR, PV, RS y DV (figura 4.20).

Durante el desarrollo de la etapa 25, sólo se encontró efecto precursor en los patrones incluidos en C2FCM; en contraste, con el efecto retardante de C3FCM, a pesar de que la distancia p-espacial entre ambos no es considerable (cuadro 4.11), presentan diferencia en T, HR, PV y VV (figura 4.16). Para el caso de los *clusters* determinados por los otros algoritmos de agrupamiento, no se encontraron patrones con efectos coadyuvantes (cuadro 4.13); respecto a los *clusters* KM, el mayor efecto retardador es el causado por la presencia de patrones del C3KM, mientras que el menor efecto se debió a la presencia de patrones C1KM (cuadro 4.10), ambos *clusters* presentan centros bien diferenciados (cuadro 4.11), para todos sus atributos (figura 4.12). Por otra parte, para los *clusters* GK, el mayor efecto retardador se debió a C4GK y el menor fue C5GK (cuadro 4.13); ambos *clusters* se encuentran a una distancia considerable (cuadro 4.13) y su diferenciación se presenta en T, HR, PV, RS y DV.

En general, se observa que las etapas 05, 07, 09, 10 y 21 presentaron *clusters* con disimilaridad considerable (cuadros 4.10, 4.11 y 4.12), donde la presencia de patrones pertenecientes a *clusters* con características relacionadas a baja temperatura, alta humedad, alta presión, baja radiación solar y viento calmo, sugieren que éstos provocaron un efecto limitante del desarrollo vegetativo; en contraste, la presencia de patrones de *clusters* relacionados a temperaturas

templadas, menor humedad, menor presión, con viento y mayor radiación solar sugiere que éstos funcionaron como precursores del desarrollo.

Cuadro 4.13 Relación entre la duración de etapa fenológica y la cantidad de patrones, por tipo de agrupamiento y su efecto.

Etapa fenológica	Algoritmo de agrupamiento					
	K-Medias		FCM		G-K	
	Retarda	Acelera	Retarda	Acelera	Retarda	Acelera
05	C1(0.6421) C3(0.4033) C4(0.0652)	C2(-0.2284)	C3(0.8064) C4(0.6812) C2(0.3777)	C5(-0.4548) C1(-0.0500)	C1(0.4683) C3(0.3589) C2(0.2999) C4(0.2849) C5(0.0072)	
07	C1(0.7316)	C4(-0.8798) C3(-0.5712) C2(-0.3189)	C4(0.9865) C1(0.9372)	C3(-0.9972) C2(-0.9955) C5(-0.9539)	C1(0.9253) C4(0.0397)	C3(-0.5027) C5(-0.4941) C2(-0.1519)
09	C1(0.9396)	C2(-0.8169) C3(-0.7716) C4(-0.6160)	C4(0.9978) C3(0.9655)	C2(-1) C1(-0.9994) C5(-0.9596)	C4(0.7620) C1(0.6146) C2(0.5124)	C5(-0.6527) C3(-0.4539)
10	C1(0.9009) C4(0.8386) C3(0.7748)	C2(-0.6820)	C4(0.9997) C3(0.9695)	C1(-0.9190) C5(-0.9164) C2(-0.0850)	C3(1) C4(0.9962) C1(0.9377) C2(0.9321)	C5(-0.6016)
12	C3(0.9515) C4(0.9445) C2(0.8346) C1(0.7926)		C3(0.9484) C2(0.9216) C4(0.8373) C5(0.3144) C1(0.1257)		C4(0.9972) C1(0.9890) C3(0.9540) C2(0.8036) C5(0.3792)	
14	C3(0.9360) C1(0.8717) C4(0.8565) C2(0.2306)		C2(0.9980) C3(0.9914) C4(0.9869) C5(0.9901)	C1(-0.4620)	C4(0.9884) C1(0.8640) C5(0.6467) C3(0.4188) C2(0.3870)	
21	C1(0.8421) C3(0.5083) C4(0.4821)	C2(-0.5814)	C3(0.9966) C2(0.9894) C4(0.8466) C5(0.4478)	C1(-0.7995)	C4(0.9205) C2(0.8594)	C1(-0.7915) C3(-0.6481) C5(-0.1685)
23	C2(0.9946) C4(0.8738) C3(0.8634) C1(0.7601)		C1(0.9917) C3(0.9873) C2(0.9620) C4(0.9135) C5(0.9103)		C5(0.9879) C2(0.9854) C1(0.7206) C3(0.6066) C4(0.1706)	
25	C3(0.9800) C2(0.8518) C4(0.5445) C1(0.3569)		C3(0.9987) C4(0.6090) C1(0.0572) C5(0.0470)	C2(-0.8551)	C4(0.9983) C3(0.9036) C1(0.6911) C2(0.6628) C5(0.4617)	

Fuente: Elaboración propia.

Respecto a las etapas 12, 14, 23 y 25, no se encontraron patrones climáticos con propiedades precursoras de desarrollo de la planta; sin embargo, se observó que los niveles de máxima y mínima limitación coincidían generalmente con *clusters* con características contrastantes (cuadros 4.10, 4.11 y 4.12 y 4.13). Este comportamiento coincide con la dispersión por etapa fenológica presentada por la variable PV (figura 4.4e); por otra parte, la no presencia explicita de patrones precursores sugiere la presencia de otros factores diferentes a los aquí estudiados, que pueden ser climáticos o de otra índole, y que intervienen en el desarrollo en estas etapas.

Conclusiones

En este trabajo se ha demostrado la capacidad del proceso KDD como eje teórico metodológico interdisciplinario para extraer conocimiento de bases de datos que expresan procesos de distintos grados de complejidad, a través de una amplia gama de técnicas analíticas, provenientes de distintos campos de conocimiento. En los sistemas informáticos actuales, es posible, trasformar cualquier tipo de datos existente además de interpretar un extenso rango de fenómenos que pueden presentar desde un comportamiento determinista hasta un comportamiento caótico. El soporte analítico que brindan la estadística, la inteligencia artificial a través de las herramientas incluidas en la computación suave, así como las técnicas visuales que facilitan la expresión y transferencia del conocimiento generado al usuario final, son las características que convierten al KDD en un soporte fundamental para la toma de decisiones en la organización.

La aplicación de este marco metodológico, permitió aprovechar la capacidad de reusar y reciclar los datos almacenados, que ya habían cubierto su propósito operacional, como son los casos de los patrones climáticos y los registros fenológicos, utilizados en el proceso de monitoreo del desarrollo de sistema de producción de la uva de mesa. Esta practica, permitió la recuperación de la inversión realizada por la organización en sistemas de información para el registro de operaciones en sus procesos y/o en servicios de información proveídas por terceros, a través de la generación de información y conocimiento, utilizados para la definición de políticas orientadas a mantener o mejorar el desempeño de la organización productiva.

El iniciar el proceso de análisis a partir del problema-objetivo, posiciona por una parte al ser humano (analista o decisor) como el centro del KDD, lo que plantea los retos de la comprensión y utilidad, como atributos que deberán contener el conocimiento generado. En otro sentido, antepone el problema al método, técnica

o herramienta analítica, lo cual es metodológicamente correcto, ya que se enfatiza más en la problematización de la situación que en los medios para su solución.

La aplicación del EDA, a través del uso de los gráficos de frecuencia, los diagramas de dispersión, y la pruebas, de ajuste Lilliefors, paramétrica F y no paramétrica Kruskal-Wallis. Permitió diferenciar el comportamiento de las variables climáticas con base en los diferentes criterios: para el total del conjunto de patrones (figura 4.1), por ciclo productivo (figura 4.3), por etapa fenológica (figura 4.4) y la combinación ciclo productivo – etapa fenológica (figuras 4.5, 4.6, 4.7 y anexo B).

En este mismo sentido, los diagramas de caja permitieron detectar la existencia de valores atípicos y valores extremos en el comportamiento de la variable velocidad del viento (figura 4.3d). La utilidad de estos diagramas para descartar errores no inherentes a un comportamiento natural, fue validada a través del contraste con los datos correspondientes cronológicamente, provenientes de otras estaciones climáticas cercanas a la referenciada (figura 4.8); se utilizaron las pruebas F y Kruskal-Wallis, cuya aplicación no arrojó diferencia significativa entre los datos del conjunto climático y los datos pertenecientes a las otras estaciones cercanas, lo que confirma que tales valores anómalos se deben al comportamiento natural de las mediciones físicas realizadas.

Así, el conjunto de patrones puede ser caracterizado como una base de datos con ruido natural generado principalmente por la variable velocidad de viento, la cual presentó sistemáticamente este tipo de valores (figura 4.3d, 4.4d, y 4.8d) y en menor medida por las variables presión de vapor, dirección del viento y humedad relativa, que también presentaron estas características (figuras 4.3, 4.4, 4.5, 4.6 y 4.7).

Se observó una marcada relación entre las variables climáticas a partir del uso de los diagramas de dispersión (figura 4.10), y los resultados de los análisis de

correlación tanto paramétrico (cuadro 4.3), como el no paramétrico (cuadro 4.4). Estas características no permitieron descubrir una estructura nítida de agrupamiento.

Con base en la caracterización del conjunto de patrones, se decidió utilizar el algoritmo de agrupamiento K-Medias, que forma parte de los métodos de la estadística multivariada, así como los algoritmos de agrupamiento difusos Fuzzy C-Medias y Gustafson-Kessel incluidos dentro de la computación suave, con el propósito de descubrir la estructura implícita de dicho conjunto.

De acuerdo con la naturaleza del conjunto de patrones arriba mencionada, el algoritmo Gustafson-Kessel presentó un mejor desempeño, al mostrar los valores menores para el coeficiente de variación (cuadros 4.5, 4.7 y 4.9), lo que coincide con lo sugerido por Abonyi y Feil (2007), quienes recomiendan el uso de este algoritmo cuando la base de datos presenta variables correlacionados. Esto, debido a que al utilizar como criterio de semejanza la distancia Mahalonobis, la cual es determinada a través de la matriz de covarianza de los grupos formados, se elimina el sesgo producido por los datos correlacionados. El segundo mejor desempeño lo mostró el agrupamiento derivado del algoritmo Fuzzy C-Medias (cuadros y figuras), el cual utiliza como medida de similitud la distancia Euclidiana.

A pesar de la semejanza de las características de los agrupamientos formados por los tres algoritmos, los algoritmos difusos (posibilistas) Gustafson-Kessel y Fuzzy C-Medias, mostraron mejor desempeño, en contraste con el algoritmo determinista K-Medias. De esta manera se evidencian las ventajas del usos de algoritmos de agrupamiento difusos cuando el conjunto de datos agrupar muestran las características de un conjunto con ruido.

Se concluye así, que la ventaja que presentan los algoritmos difusos se derivan de la naturaleza de la lógica de conjuntos que subyace a los algoritmos de agrupamiento (difuso o crisp), del tipo de partición que generan (posibilista o

determinista) y de la figura geométrica que forman las medidas de similitud utilizadas (cilíndrica o esférica).

La aplicación de herramientas de visualización de información incluidas dentro del soporten del proceso KDD (figura 1.2), permitieron observar la diferenciación entre los grupos formados, objetivo principal de los algoritmos de agrupamiento (Höppner et al., 1999). Tal es el caso del uso de las gráficas de coordenadas paralelas (figuras 4.12, 4.16 y 4.20), en las cuales fácilmente se aprecia el contraste de los patrones característicos que representa cada *cluster* descubierto. Así mismo, la utilización de los mapas climáticos, que son una adaptación de las técnicas de patrones recursivos (Keim 2002, Keim y Ward, 2007), facilitó el hallazgo de comportamientos climáticos contrastantes, como fue posible al visualizar días con características completamente veraniegas durante el invierno, así como días invernales en plena primavera (figuras 4.14, 4.18 y 4.22), lo cual a su vez facilitó la transferencia del conocimiento descubierto, a través de los métodos de agrupamiento.

La correlación entre el comportamiento de los patrones climáticos y el desarrollo de la uva de mesa expresada a través de las etapas fenológicas, permitió identificar las propiedades precursoras e inhibidoras del desarrollo vegetativo en algunos grupos de patrones (Cuadro 4.13). A partir del agrupamiento formado a través del algoritmo Gusstafson-Kessel, pudo observarse que en las primeras etapas después de la brotación, las condiciones climáticas con temperaturas frías, alta humedad, alta presión y generalmente nocturnas, mostraron un efecto inhibidor del desarrollo vegetativo; en contraste, las condiciones caracterizadas como cálidas, secas, con vientos durante los periodos de mayor radiación solar, presentaron un efecto precursor del desarrollo de la vid. Sin embargo, cuando el desarrollo vegetativo se hizo más evidente (etapas de la 12 a la 25), el contraste entre condiciones precursoras e inhibidoras no fue tan evidente (Cuadro 4.13), lo que indica que los factores climáticos son menos determinantes del desarrollo de la planta en estas etapas, ya que conforme aumenta el follaje, se incrementa

también la aplicación foliar de aceleradores de desarrollo. Este resultado plantea la necesidad de examinar otros factores que afectan el desarrollo de la uva de mesa, como son las condiciones edafológicas y prácticas culturales, que permitan una mejor descripción del proceso fenológico en futuras investigaciones.

De esta manera, el modelo metodológico desarrollado en esta investigación, que presenta como eje rector al proceso KDD, presenta la propiedad de ser replicable, no sólo a problemas como el aquí planteado, sino a otros sistemas de producción agrícola y más aun, a otras problemáticas que emerjan de procesos complejos que puedan ser expresados en conjuntos de patrones cómo los aquí utilizados.

Una de las limitantes del presente trabajo es su corte descriptivo, los resultados obtenidos no van más allá de la explicación del proceso, sin embargo es precisamente esta característica la que permite definir nuevos horizontes en futuras investigaciones, las cuales se deberán centrar en el desarrollo de modelos predictivos a través de la aplicación de herramientas difusas e hibridaciones entre algoritmos y técnicas analíticas incluidas dentro de la computación suave y la computación de alto rendimiento.

Bibliografía

Abonyi János y Feil Balázs, 2007, Cluster Analysis for Data Mining and System Identification, Birkhäuser Verlang AG, Germany.

Addcon Telemetry, 2010, Acurate remote sensor technology, ADCON International INC. USA, Canada, Mexico, Homepage: www.adcon.at

Ackoff R. L., 1989, "From data to wisdom", *Journal of Applies Systems Analysis*, vol. 16, 1989, pp. 3-9.

Aspray William, 1990, John Von Newman and the Origin of Modern Computing, MIT Press, Cambridge Massachesetts, London, England.

Bauer Friedrich L., 2007, Origins and Foundations of Computing In Cooperation with Heinz Nixdorf MuseumsForum, Springer, Garching, Germany.

Berman F., Fox G., y Hey T., 2003, Grid Computing: Making the Global Infrastructure a Reality, Wiley and Sons, USA.

Berry Michael J. A. and Linoff Gordon S., 2005, Data Mining Techniques: For Marketing, Sales and Customer Relationship Management, Second Edition, Wiley Publishing, Inc., U.S.A.

Bessis Nik, 2010, Grid Technology for Maximizing Collaborative Decision Management and Support: Advancing Effective Virtual Organizations, Information Science Reference, New York, USA.

Bezdek J. C., 1981, Fuzzy Mathematics in Pattern Classification, Ph, D. Thesis, Applied Mathematical Center, Cornell University System.

Bishop, Chistopher M., 1995, Neural Network for Pattern Recognition, OXFORD University Press, USA.

Bow Sing-Tze, 2002, Pattern Recognition and Image Preprocessing, Seconf Edition, Marcel Dekker, Incorporation, USA.

Brachman, R., and Anand, T. 1996. The Process of Knowledge Discovery in Databases: A Human-Centered Approach. In *Advances in Knowledge Discovery and Data Mining,* 37–58, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Menlo Park, Calif.: AAAI Press.

Butenko Sergiy, Chaovalitwongse W. Art. Pardalos Panos M, 2009, Clustering Challenges in Biological Networks, World Scientific Publisihing, U.S.A.

Buyya R., 1999, High Performance Cluster Computing: Systems and Architecture, volumn 1, Prentice Hall PTR, New Jersey, USA.

Card S. K., Mackinlay J. D., Shneiderman B., 1999. Readings in Information Visualization; Using Vision to think, Morgan Kaufmann, Los Altos, CA, USA.

Caverlee James, 2009, “Data Dictionary”, en Liu Ling, Özsu M. Tamer, Encyclopedia of Database Systems Volume 1, Springer Science, USA.

Ceruzzi Paul E., 2003, A History of Modern Computing, The MIT Press, Cambridge, Massachisetts London, England.

Chen Min, Ebert David, Hagen Hans, Laramee Robert S., van Liere Robert, Kwan-Liu Ma, Ribarsky William, Scheuermann Gerik, Silver Deborah, 2009, **“**Data, Information, and Knowledge in Visualization**”,** IEEE Computer Graphics and Applications**,** Volume 29 , Issue 1 (January 2009), Pages 12-19, ISSN:0272-1716.

Chou C-H, Su M-C, Lai E., 2004, “A new cluster validity measure and its application to image compression”, Patern Anal Applicaction, Springer-Verlang, London.

Coombe, B. 1995, “Adoption of a system for identifying grapevine growth stages”, Australian Journal of Grape and Wine Reseach.

Coonors D. J, y Loomis R S, 2002, “Ecología de Cultivos, Productividad y manejo en sistemas agrarios”, Ediciones Mundi-Prensa, España.

Cordón, Oscar; Herrera, Francisco; Hoffmann, Frank; Magdalena, Luis, 2001, “Genetic Fuzzy Systems, evolutionary tunning and learning of fuzzy knowledge bases”, in Advances in Fuzzy Systems – Application and Theory, Vol. 19, World Scientific Publishing, USA.

Das Swagatam, Konar Amit, Abraham Ajith, 2009, Metaheuristic Clustering, Springer-Verlang, Germany.

Date C. J, 2001, Introducción a los sistemas de bases de datos, Séptima Edición, Pearson Eduación de México, S.A. de C.V., México.

Duda Richard O., Hart Peter E., Stork David G., 2001, *Pattern Classification*, 2nd edition, John Wiley, USA.

Dunn J. C., 1973, “A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact, Well Separeted Clusters”, Cybernetics and Systems, Volume 3, Issue 3 1973 , pages 32 – 57.

Edmond David, 1992, Information Modeling: specification and implementation, Prentice Hall, New York.

Eisen M. and Brown P., 1999, “DNA arrays for analysis of gene expresión”, *Methods in*
Enzymology, 303:179–205.

Elmarsi Ramez, Navathe Shamkant B, 1997, Sistemas de bases de datos: conceptos fundamentales, Addison-Wesley, USA.

Fayyad Usama, Piatetsky-Shapiro Gregory, Smyth Padhraic, 1996, “From Data Mining to Knowledge Discovery in Databases”, The American Association for Artificial Intelligence, Fall 1996, U.S.A.

Frawley William J., Piatetsky-Shapiro Gregory, Matheus Christopher J., 1992, “Knowledge Discovery in Data Base: an Overview”, The American Association for Artificial Intelligence, Fall 1992, U.S.A.

Fukugana Keinosuke, 1990, Introduction to Statistical Pattern Recognition, Second Edition, Morgan Kaufmann, Academic Press, U.S.A

Galan Manuel J., García Fidel, Álvarez Luis, Ocón Antonio y Rubio Enrique, 2001, “'Beowulf Cluster' for High-performance Computing Tasks at the University: A Very Profitable Investment”, EUNIS Proceeding DTD Version 1.0, Certified Document Server at Humbikdt-University, Berlin, Germany. http://edoc.hu-berlin.de/ conferences/eunis2001/c/Galan/HTML/.

Gan Guojun, Ma Chaoqun y Wu Jianhong, 2007, Data Clustering: Theory, Algorithms, and Applications, American Statistical Association and the Society for Industrial and Applied Mathematcis, USA.

Giudici Paolo, 2003, Applied Data Mining: Statistical Methods for Business and Industry, Wiley, England.

Grama Ananth, Gupta Anshul, Karypis George, Kumar Vipin, 2003, , Inroduction to Parallel Computing, Second Edition, Addison-Wesley, USA.

Halpin Terry, 2001, Informtion Modeling and Relational Databases: from conceptual analysis design, Academic Press, USA.

Han Jiawei and Kamber Micheline, 2000, “Data Mining: Concepts and Techniques”, Morgan Kaufman Publishers, U.S.A.

Hand, D. J., 1998, “Data Mining: Statistics and more?”. The American Statistician, 52, 2, 112-118.

Hansen Charles D. y Johnson Cris R., 2005, The visualization Handbook, Elsevier Butterworth–Heinemann, USA.

Harrington Jan L., 2009, Relational database design and implementation : clearly explained, Third Edition, Morgan Kaufmann Publishers is an imprint of Elsevier, USA.

Höppner Frank, Klawonn Frank, Kruse Rudolf y Runkler Thomas, 2000, Fuzzy Cluster Analysis Methods for Classification, Data Analysis and Image Recognition, John Wilet & Sons LTD, Toronto.

Husiak Andrew y Shah Shital C., 2006, “Data Mining and Warehousing in Pharma Industry”, en Encyclopedia of Data Warehousing and Mining, IDEA GROUP REFERENCE, USA

Jacob Bart, Brown Michael, Fukui Kentaro y Trivedi Nihar, 2005, Introduction to Grid Computing, International Business Machines, USA.

Kahn Hilary J, y Napper R. B. E., 2000, “The Birth of the Baby”, The Proceedings of the 2000 IEEE International Conference on Computer Design: VLSI in Computers & Processors.

Keim Daniel, 2000, “Designing pixel-oriented visualization techniques: Theory and applications”, *IEEE Transactions of Visualization and Computer Graphics*, 6(1).

Keim Daniel A., 2002, “Information Visualization and Visual Data Mining”, Transaction on Visualization and Computer Graphics, Vol. 7, No. 1, January-March.

Keim Daniel y Ward Matthew, 2007, ”Visualization”, en Intelligent Data Analysis an Introduction, Second Edition, Springer-Verlag Berlin Heidelberg, New Yor, USA.

Keller Tanja y Tergan S. O., 2005, “Visualization Knowledge and Information: An Introduction”, en Knowledge and Information Visualization Searching for Synergies, LNCS 3426, pp. 1-23, Springer Berlin Heidelberg New York, USA.

King William R., 2006, “Knowledge Transfer”; D. Schwartz, “Encyclopedia of Knowledge Management”, Idea Group.

Klir George J. y Yuan Bo, 1995, Fuzzy Sets and Fuzzy Logica:Theory and Applications, Prentice Hall P T R, New Jersey, U.S.A.

Kruse Rudolf, Döring Christian, y Lesot Marie-Jeanne, 2007, “Fundamentals of Fuzzy Clustering”, en Advances in Fuzzy Clustering and its Application, Valente de Oliveira José y Pedrycz Witold, Editores, John Wiley and Sons, England.

Larose Daniel T., 2005, Discovery Knowledge in Data: an introduction to data mining, Wiley Interscience, U.S.A.

MacQueen, J. B., 1967, "Some Methods for classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press.

Marmolejo D. Carlos y González T Carlos, 2008, “Valoración económica del ruido como componente ambiental en la formación del precio del mercado inmobiliario de la vivienda. *El caso de la Ciudad deBarcelona”*, Universitat Politécnica de Catalunya, España

Marshall, C.C., 2001, *The haunting question of intelligibility. Paper presented at the eleventh Hypertext ´01* (Aarhus, Denmark, August 14-18, 2001) (Online available: http://www.csdl.tamu.edu/~shipman/SpatialHypertext/SH1/marshall.pdf).

Mazza Ricardo, 2009, Introduction to Information Visualization, Springer-Verlag London.

Mitra Sushmita, Acharya Tinku, 2003, Data Mining: Multimedia, Soft Computing, and Bioinformatics, Wiley-Interscience, U.S.A.

Mirkin Boris, 2005, Clustering for Data Mining: A data recovery approach, Chapman & Hall/CRC, USA.

Mohr Bernd, 2006, “Introduction to Parallel Computing”, en Computational Nanoscience: Do it Yourself, Grotendorst J., Blügel S, Marx D. Editores, John von Newman Institute for Computing, Jülich, NIC Series, Germany.

Myatt Glenn J., 2007, Making Sencse of Data: A practical Guide to Exploratory Data Analysis and Data Mining, Wiley-Interscience, U.S.A.

Neto Ribeiro Pedro F., Barbosa Perskusich María L., Oliveira de Almeida Hyggo, Perkusich Angelo, 2009, “A Formal Verification and Validation Approach for Real-

Time Databases", en Erikcson John, Database Technologies: Concepts, Methodologies, Tools, and Applications, Information Science Reference, USA.

Nisbet Robert, Elder John y Miner Gary, 2009, Handbook of Statistical Analysis and Data Minning Applications, Elseiver Inc., U.S.A.

Pal A. y Pal S. K., 2001, "Pattern recognition: Evolution of methodologies and data mining", en Pattern Recognition: From Classical to Modern Approaches, World Scientific Publishing.

Pazzani Michael J., Mani Subramani And Shankle W. Rodman, 1997, "Comprehensible Knowledge-Discovery in Databases", Cognitive Science Conference, U.S.

Pham Binh, Streit Alex, Brown Ross, 2009, "Visualization of Information Uncertainty: Progress and Challenges", en Trends in Interactive Visualization State-of-the-Art Survey, Springer-Verlag London

Pedrycz Witold, 2005, Knowledege-Based Clustering: From Data to Information Granules, Wiley-Interscience, U.S.A.

Piatetsky-Shapiro Gregory, 1989, "Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop", American Association for Artificial Intelligence, AI Magazine, U.S.A.

Plaza Antonio J., y Chang Chein I., "High-Performance Computer Architectures for Remote Sensing Data Analysis: Overview and Case Study", en High Performance Computing in Remote Sensing, Plaza Antonio J., y Chang Chein I.. Editores, Chapman & Hall/CRC, New York, USA.

Prabhu S. y Venkatesan N., 2007, Data Mining and Warehousing, NEW AGE INTERNATIONAL (P) LIMITED, PUBLISHERS, New Delhi, India.

Ramakrishnan Raghu, Gehrke Johannes, 1999, Database Management Systems, Second Edition, McGraw Hill, USA.

Rao Romana y Card Skuart K, 1994, “The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus+Context Visualization for Tabular Information”, Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, April 1994, ACM.

Revesz Peter, 2010, Introduction to Databases: From Biological to Spatio-Temporal, Springer, USA.

Rao A. Ramachandra and Srinivas V. V., Regionalization of Watersheds: an approach based on cluster analysis, in Wather Science and Technology Library, Volume 58, Springer Science + Business Medias B. V., USA.

Rauber Thomas y Rünger Gudala, 2010, Paralell Programming for Multicore and Cluster Systems, Springer, New York, USA.

Rumelhart D.E, y Ortony, A, 1977, “The representation of knowledge in memory”, en Schooling and the acquisition of knowledge (pp. 99-133). Hillsdale, NJ: Lawrence Erlbaum Associates.

Saad Ashraf y Zaghlou A.-R., 2002, “A Knowledge Visualization Tool for Teching and Learning Computer Engineering Knowledge, Compcepts, and Skill”, 32nd ASEE/IEEE Frontiers in Educations Conference, IEEE, November 6-9, Boston, MA.

SAS Institute, 2002, Applying Data Mining Techniques: Using Enterprise Miner, Course Notes, SAS Institute Inc., 2002, U.S.A.

Shneiderman Ben, 1996, “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations”, Proceedings IEEE Symposium on Visual Languages 1996**,** September.

Siddeheswar Ray y Rose H. Turi, 1999, “Determination of Number of Clusters in *K*-Means Clustering and Application in Color Image Segmentation”, en N R Pal, A K De and J Das (eds), Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99), Calcutta, India, 27-29 December, 1999, Narosa Publishing House, New Delhi, India, ISBN: 81-7319-347-9, pp 137-143.

Siemens, G., 2005, “Connectivism. A learning theory for the digital age”, International Journal of Instructional Technology & Distance Learning, 2(1), 3-10. (En línea: 26 de enero, 2010: http://www.itdl.org/Journal/Jan_05/article01.htm).

Silberschatz Abraham, Korth Herny F., Sudaarshan S., 2002, Database Systems Concepts, Fourth Edition McGraw Hill Inc., USA.

Simons, G.L., 1984, *Introducing artificial intelligence*. Manchester: NCC Publications.

Schenker Adam, Bunke Horst, Last Mark y Kandel Abraham, 2006, “Polynomial Time Complexity Graph Distance Computation for Web Content Mining”, en Data Complexity in Pattern Recongnition. Basu Mitra and Ho Tin Lam, Editores. Springer-Verlang, USA.

Stanoesvska-Slabeva, Wozniak Thomas, 2010, “Cloud Basics – An Introduction to Cloud Computing”, en Grid and Cloud Computing: A Busniess Perspective on

Technology and Applications, Stanoesvska-Slabeva, Wozniak Thomas, Santi Ristol, Editores, Springer, Germany.

Stephens Rod, 2009, Beginning Database Design Solutions, Wiley Publishing Inc., USA.

Syed A. and Shah A., 2006, “Data, information, knowledge, wisdom: A doubly linked chain?”, In International Conference on Information and Knowledge Engineering, 2006, pages 270–278, Nevada, USA.

Tang Zhaoa Hui y MacLennan Jamie, 2005, Data Mining with SQL Server 2005, Wiley Publishing Inc., U.S.A.

Teorey Toby, Lightstone Sam, Nadeau Tom, 2009, Database Modeling & Design: Logical Design, Fourth Edition, Morgan Kaufman Publishers, Elseiver Incorporation, USA.

Terga S. O., 2005, “Digital Concept Maps for Managing Knowledge and Information”, en Knowledge and Information Visualization Searching for Synergies, LNCS 3426, pp. 1-23, Springer Berlin Heidelberg NewYork, USA.

Theodoridis Segios and Koutroumbas Konstantinos, 2003, Pattern Recognition, Second Edition, Elseiver Academic Press, USA.

Trewartha, G. T. and Horn L. H. 1980, An introduction to climate, Fifth Edition, McGraw-Hill, New York, USA.

Tsiptsis Kostantinos y Chorianopoulus Antonios, 2009, Data Mining Techniques in CRM: Inside Customer Segmentation, John Wiley & Sons, Ltd., United Kingdom.

Vasnatha Kandasamy W. B., Samrandache Florentin, Llanthenral K., 2007, Elementary Fuzzy Matrix Theory and Fuzzy Models for Social Scientists, Automaton, W. B., U.S.A.

Velte Anthony T., Velte Toby J., y Elsenpeter Robert, 2010, Cloud Computing: A Practical Apporach, McGraw Hill, USA.

Vercellis Carlo, 2009, "Business Intelligence: Data Mining and Optimization for Decision Making", John Wiley and Sons Ltd., United Kingdom.

Ward John y Peppard Joe, 2002, Strategic Planning for Information Systems, Third Edition, John Wiley & Sons, Incorporation, New York, U.S.A.

Ware Colin, 2004, Information Visualization Perception for Design, Second Edition, Morgan Kaufmann Publishers, San Francisco, CA, USA.

Webb Andrew, 2002, Statistical Pattern Recognition, Second Edition, John Wiley and Sons, England.

Williams Graham J., Huang Zhexue, 1996, "Modelling the KDD Process", Data Mining Portafolio, KDD Model, CSIRO, June, Autralia.

Witten Ian H. and Frank Eibe, 2005, Data Mining: Practival Machina Learning Tools and Techniques, Second Edition, Elseiver-Morgan Kaufmann Publishers, U.S.A.

Yeung K., Medvedovic M., and Bumgarner R., (2003), "Clustering gene-expression data with repeated measurements", *Genome Biology*, 4(5):G34.1–17.

Zadeh Lofti A. 1965, "Fuzzy sets", Information and Control, 8(3), pp. 338-353, en Adavances in Fuzzy Mathematics and Engineering, Fuzzy Sets and Fuzzy

Information-Granulation Theory, Key Selected Paper By Lofti A. Zadeh, Editado por Da Ruan y Chongfu Huang, Beijin Normal University Press, Beijin.

Zadeh Lofti A., 1984, ”Making computers think like people,” IEEE. Spectrum, 8/1984.

Zadeh Lofti A., 1990, “The Birth and Evolution of Fuzzy Logic”, International Journal of General Systems, Vol 17, Gordon and Breach Science Publishers, S.A., United Kingdom.

Zahn C.T., 1971, “Graph-theoretical methods for detecting and describing gestalt Clusters”. IEEE Trans. Comput. C, 20:68–86.

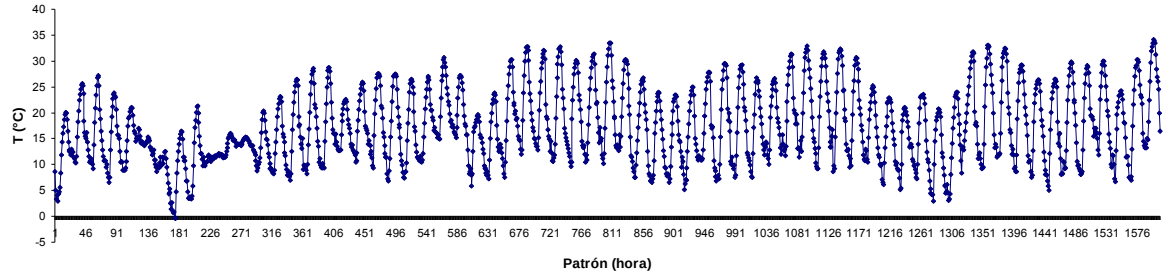
Zhang Hang, 2009,”Statistical Clustering Analysis: An Introduction”, en Butenko S., Chaovalitwongse W., Pardalos P., Clustering Challenges in Biological Networks, World Scientific Publishing, U.S.A.

Zilouchian Ali, Jamshidi Mo, 2001, Intelligent Control Systems Using Soft Computing Methodologies, CRC Press, USA.

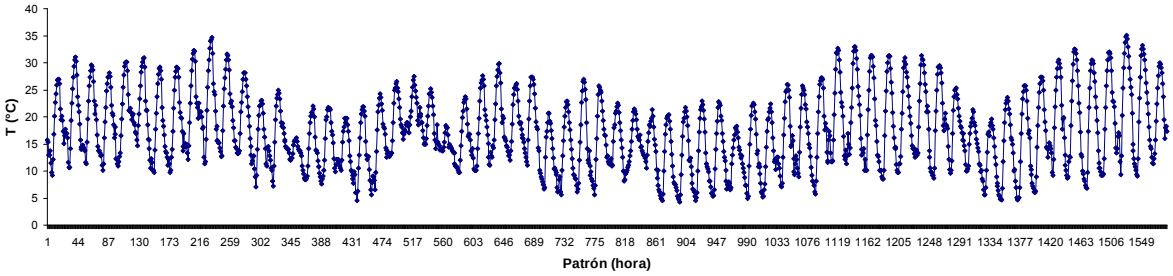
Anexo A

Comportamiento histórico de los atributos incluidos en el patrón climático

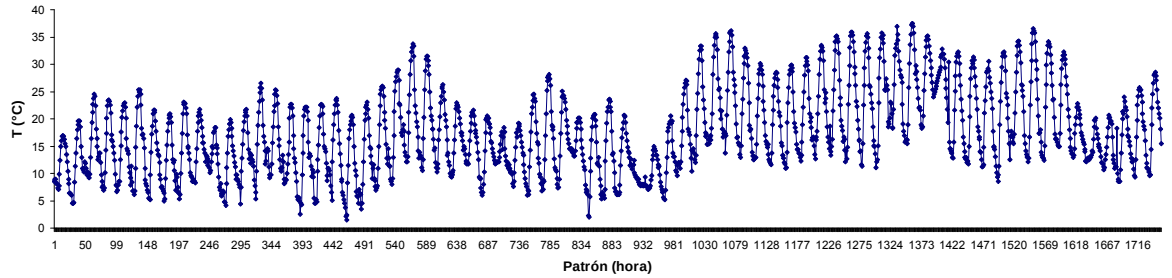
Figura A.1 Comportamiento de la temperatura durante el periodo de estudio



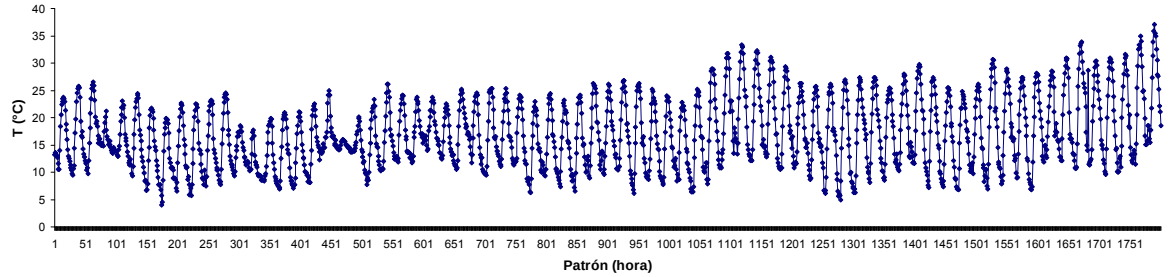
a) Ciclo 2001-2002



b) Ciclo 2002-2003



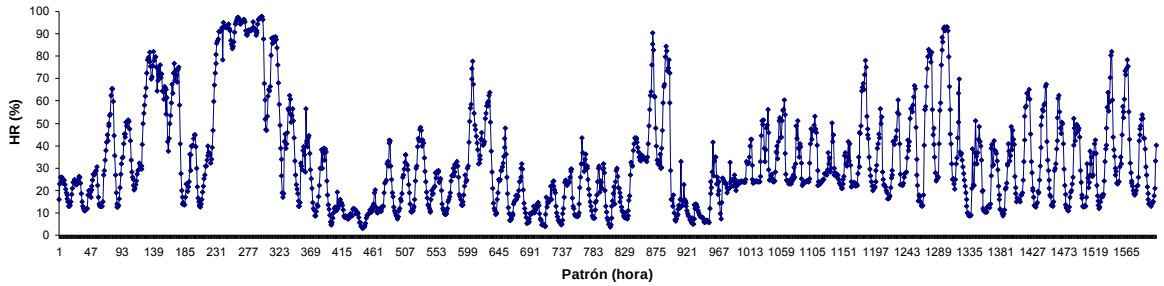
c) Ciclo 2003-2004



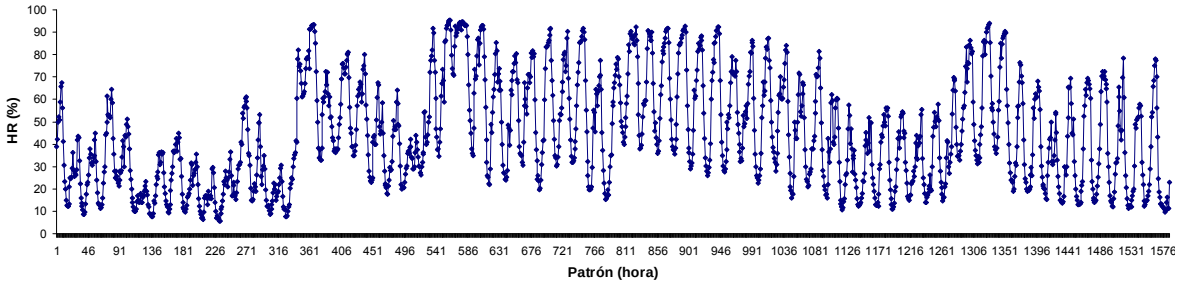
d) Ciclo 2004-2005

Fuente: Elaboración propia.

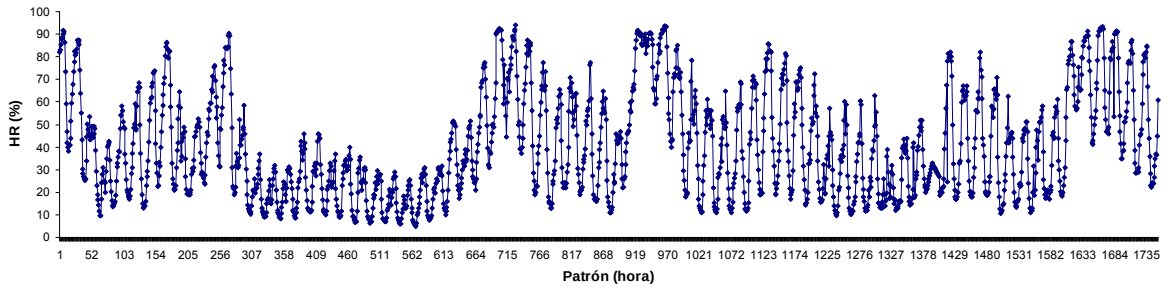
Figura A.2 Comportamiento de la humedad relativa durante el periodo de estudio



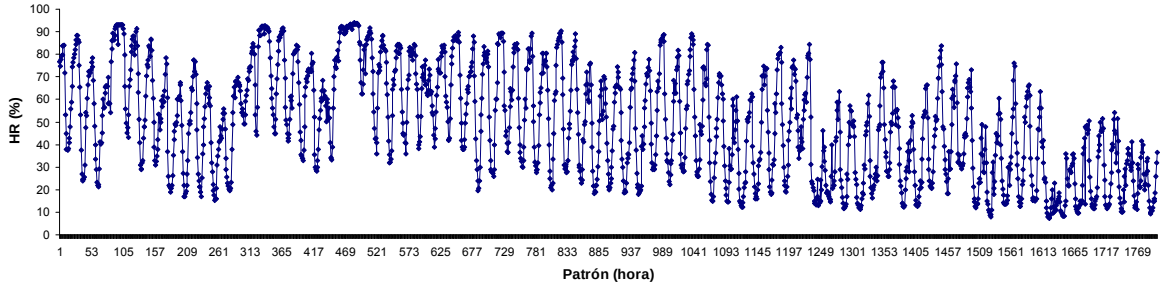
a) Ciclo 2001-2002



b) Ciclo 2002-2003



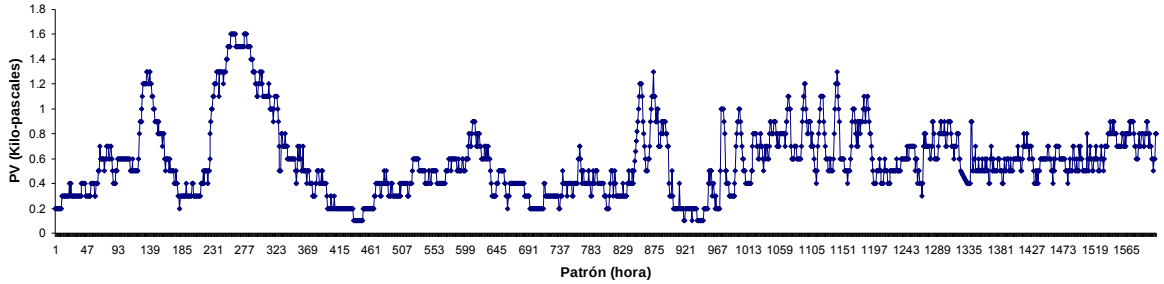
c) Ciclo 2003-2004



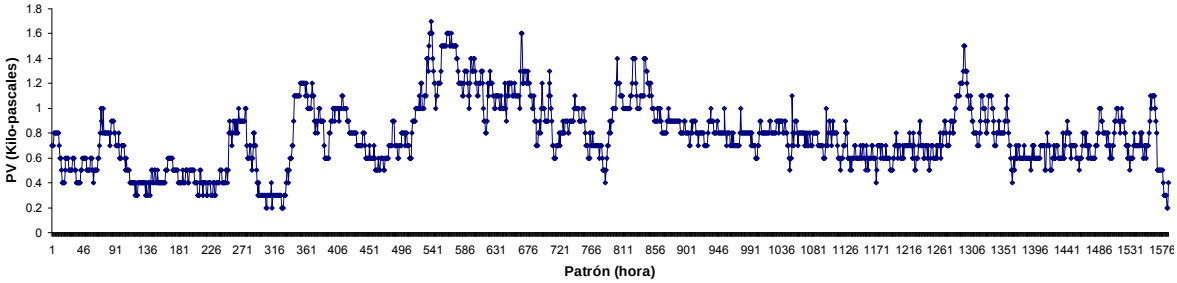
d) Ciclo 2004-2005

Fuente: Elaboración propia.

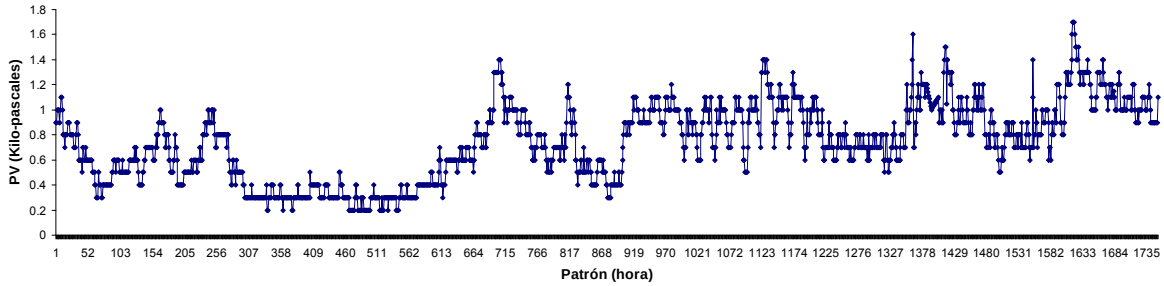
Figura A.3 Comportamiento de la presión de vapor durante el periodo de estudio



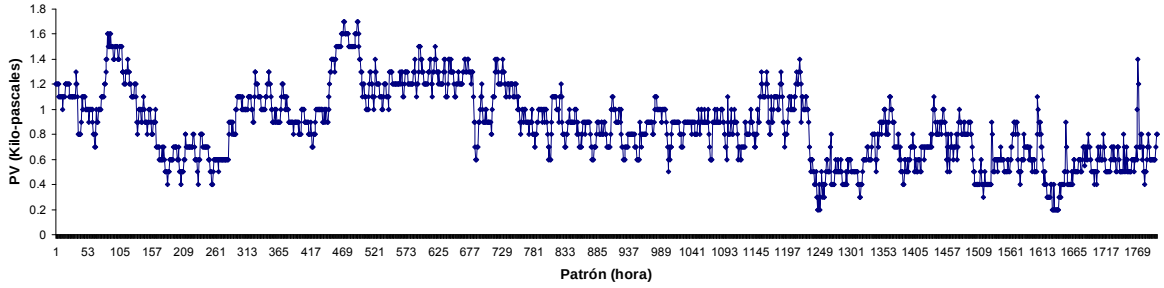
a) Ciclo 2001-2002



b) Ciclo 2002-2003



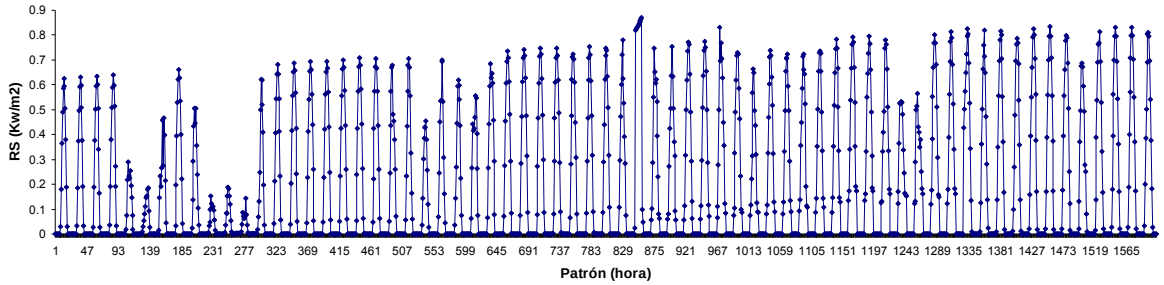
c) Ciclo 2003-2004



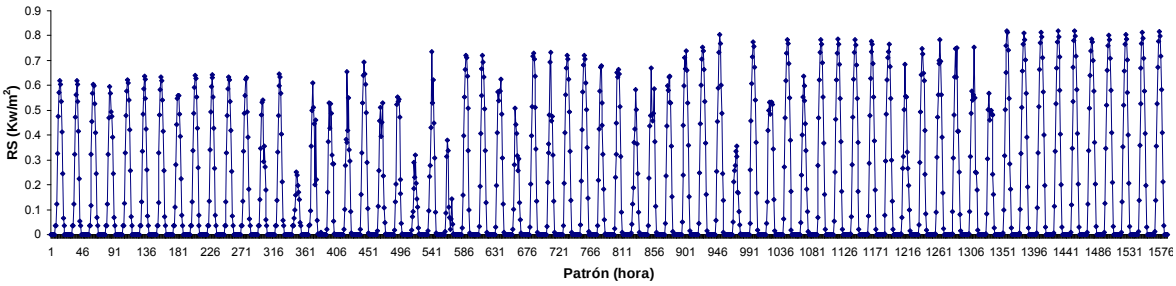
d) Ciclo 2004-2005

Fuente: Elaboración propia.

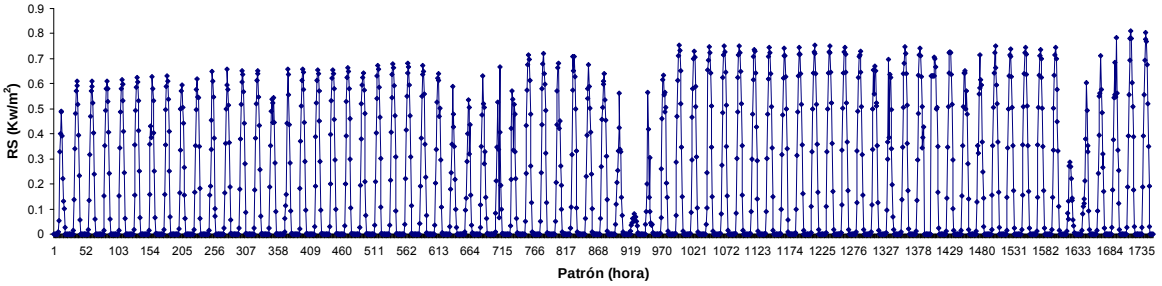
Figura A.4 Comportamiento de la radiación solar durante el periodo de estudio



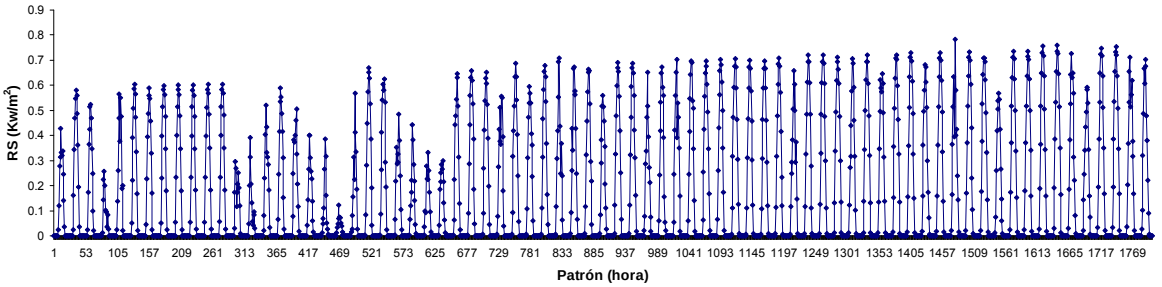
a) Ciclo 2001-2002



b) Ciclo 2002-2003



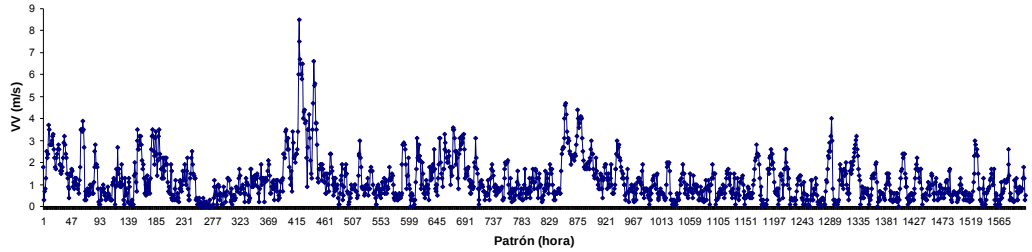
c) Ciclo 2003-2004



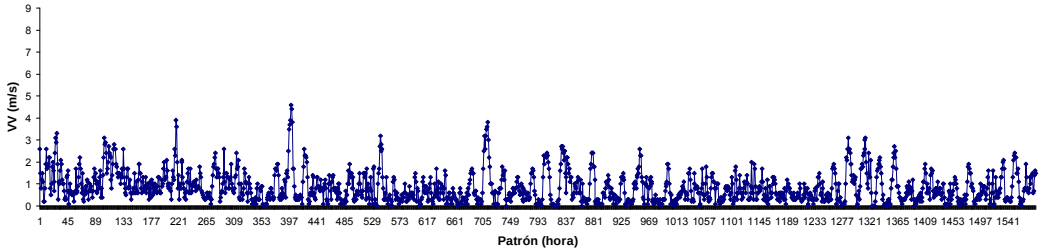
d) Ciclo 2004-2005

Fuente: Elaboración propia.

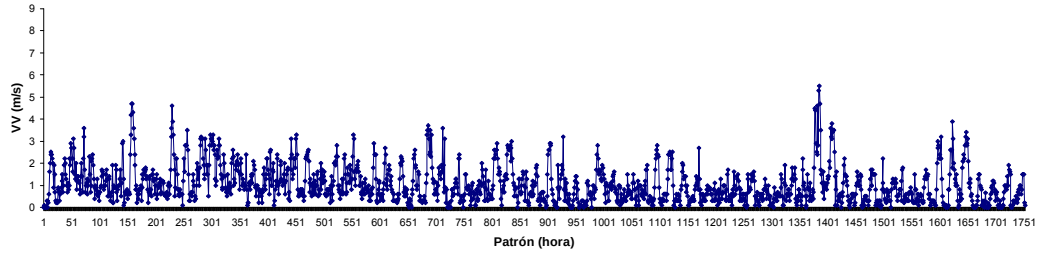
Figura A.5 Comportamiento de la velocidad del viento durante el periodo de estudio



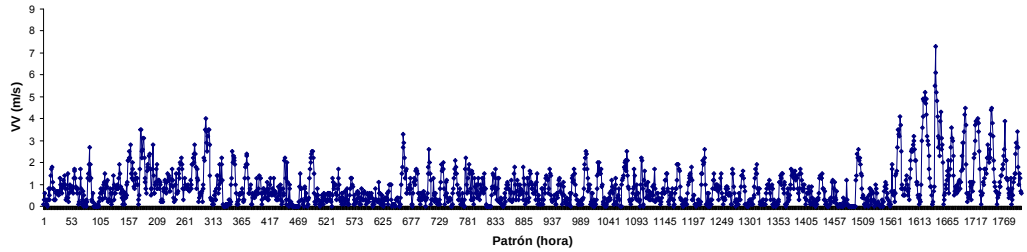
d) Ciclo 2001-2002



e) Ciclo 2002-2003



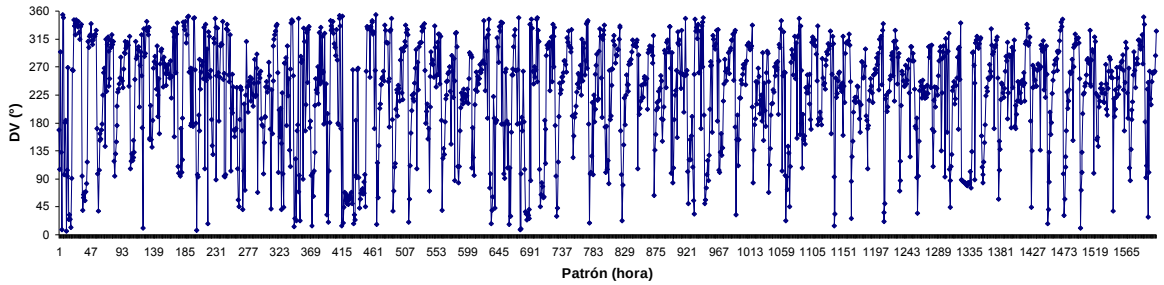
f) Ciclo 2003-2004



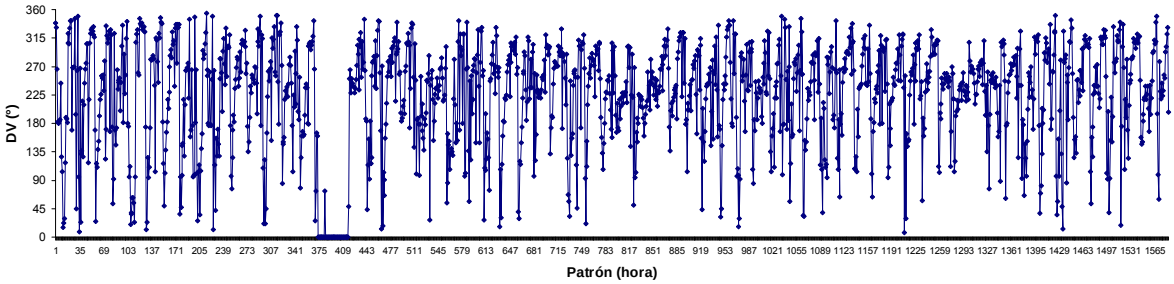
d) Ciclo 2004-2005

Fuente: Elaboración propia.

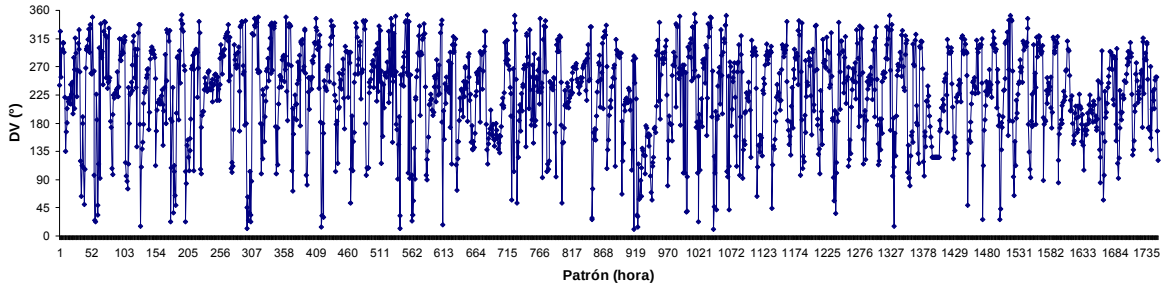
Figura A.6 Comportamiento de la dirección del viento durante el periodo de estudio



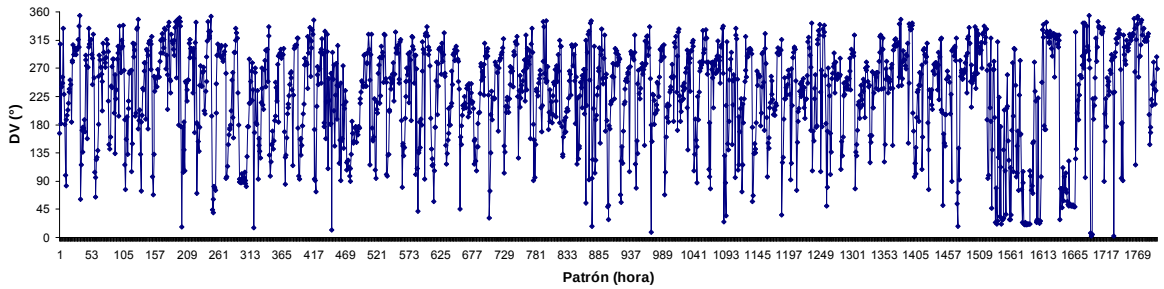
a) Ciclo 2001-2002



b) Ciclo 2002-2003



c) Ciclo 2003-2004



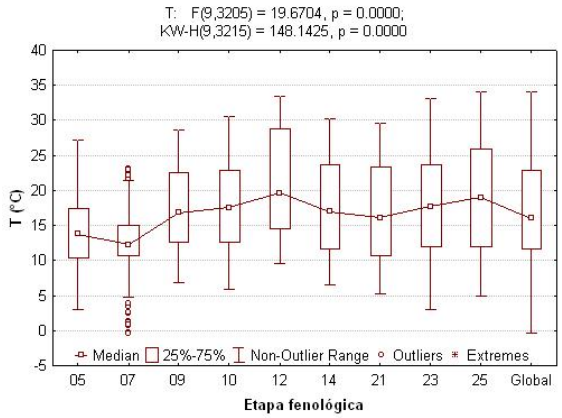
d) Ciclo 2004-2005

Fuente: Elaboración propia.

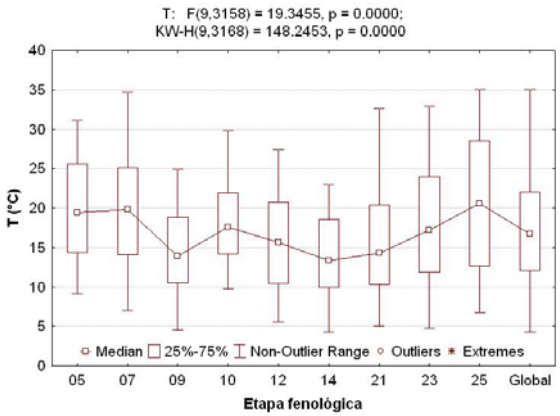
Anexo B

Dispersión de los valores de las variables climáticas distribuidos por etapa fenológica para los cuatro ciclos productivos

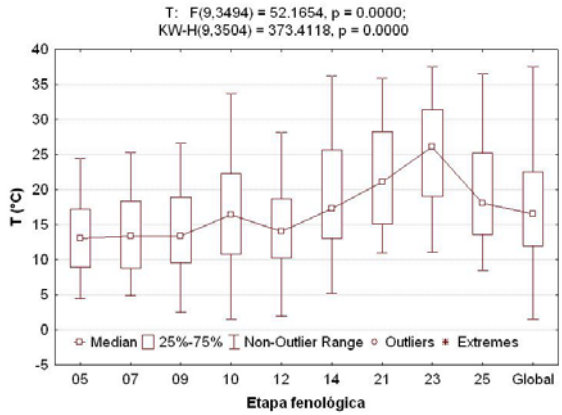
Figura B.1 Dispersión de los valores de la temperatura distribuidos por etapa fenológica para los cuatro ciclos productivos.



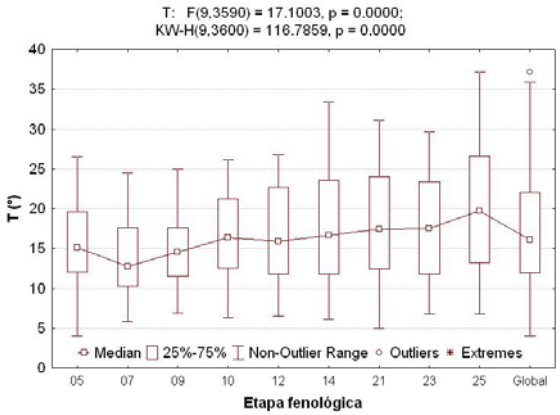
a) Ciclo 2001-2002



b) Ciclo 2002-2003



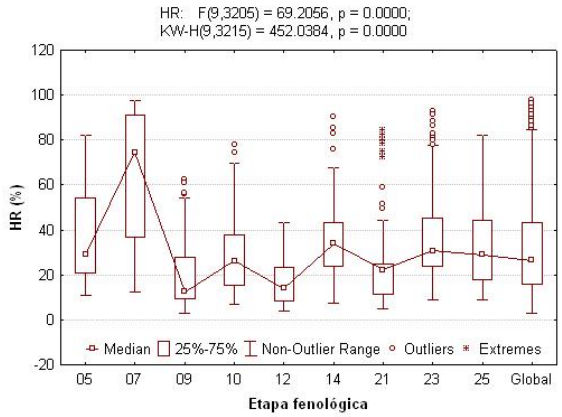
c) Ciclo 2003-2004



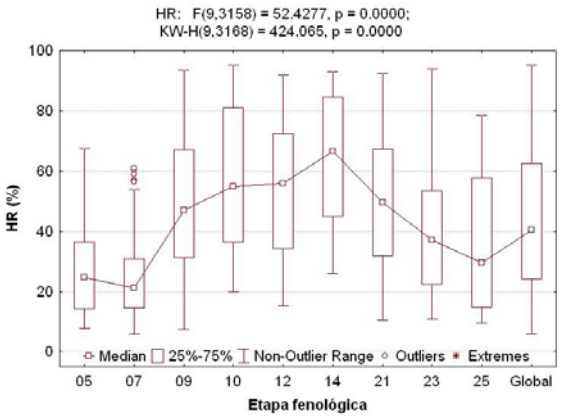
d) Ciclo 2004-2005

Fuente: Elaboración propia.

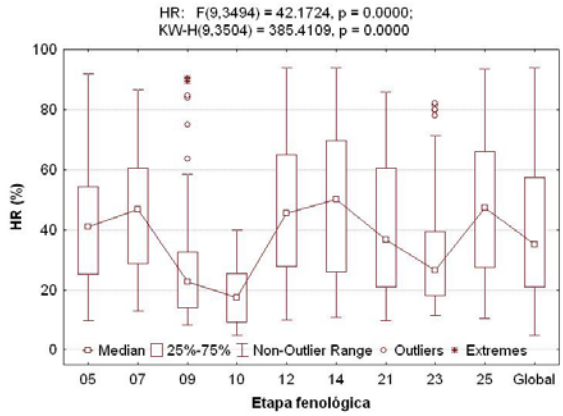
Figura B.2 Dispersión de los valores de la humedad relativa distribuidos por etapa fenológica para los cuatro ciclos productivos.



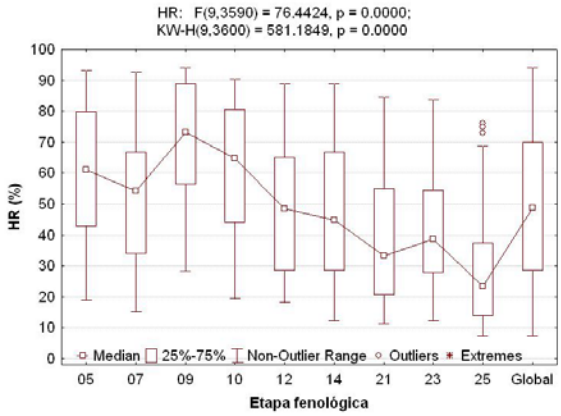
a) Ciclo 2001-2002



b) Ciclo 2002-2003



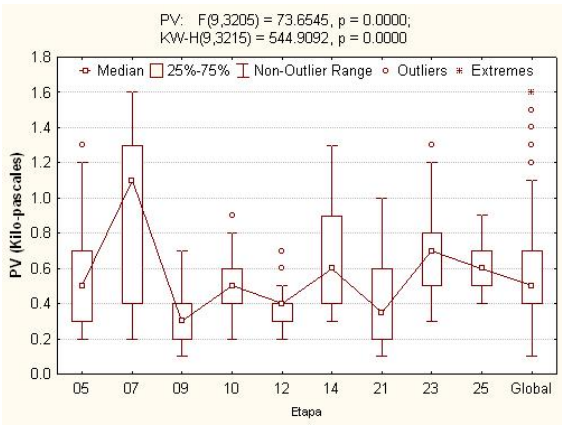
c) Ciclo 2003-2004



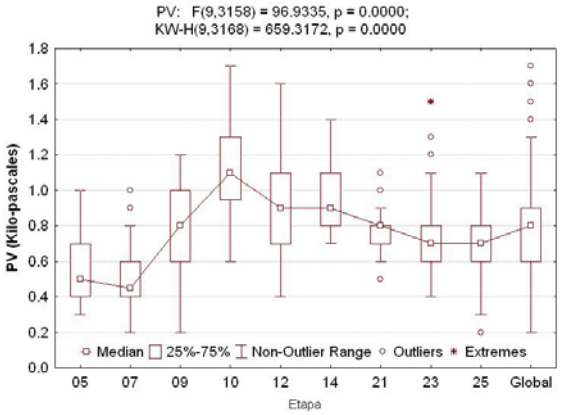
d) Ciclo 2004-2005

Fuente: Elaboración propia.

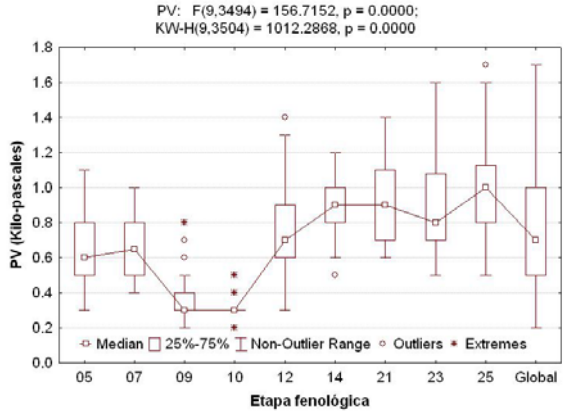
Figura B.3 Dispersión de los valores de la presión de vapor distribuidos por etapa fenológica para los cuatro ciclos productivos.



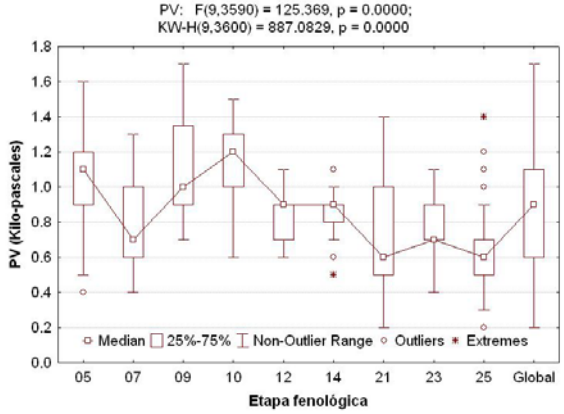
a) Ciclo 2001-2002



b) Ciclo 2002-2003



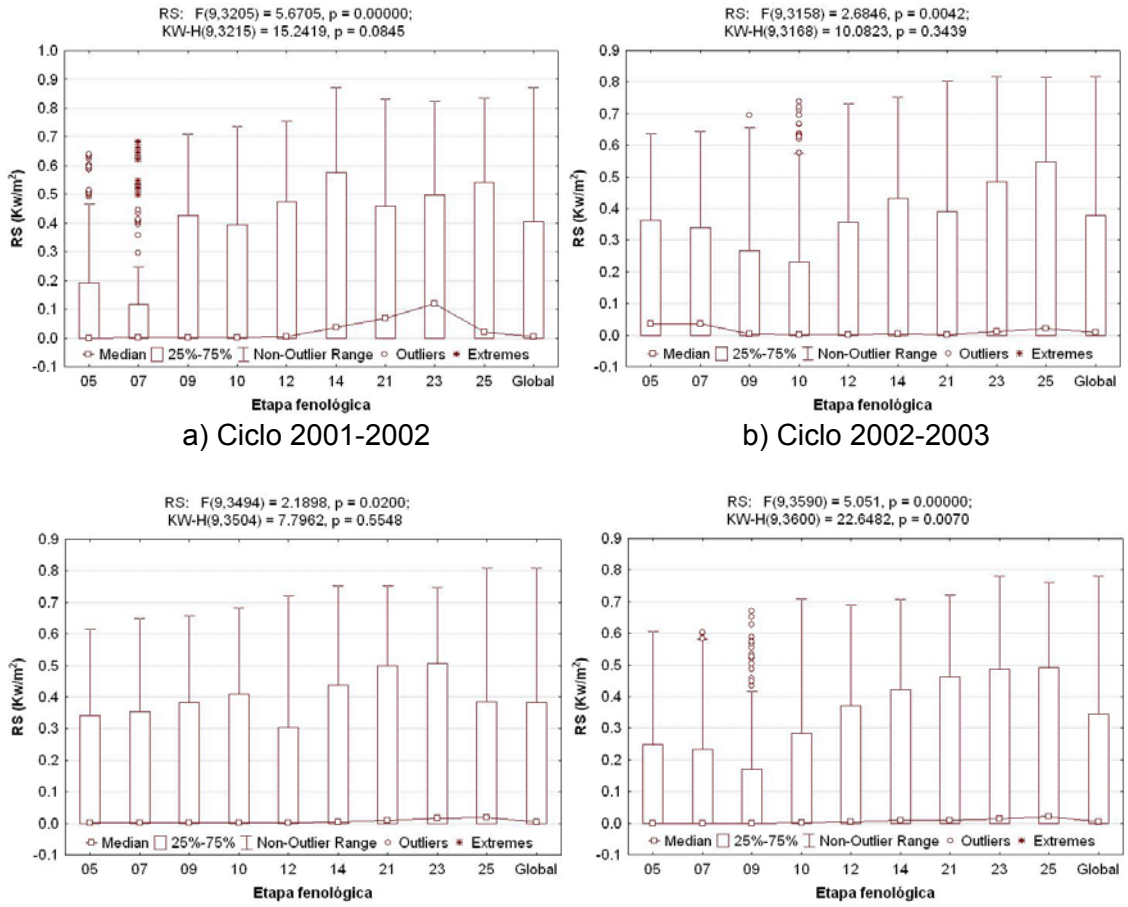
c) Ciclo 2003-2004



d) Ciclo 2004-2005

Fuente: Elaboración propia.

Figura B.4 Dispersión de los valores de la radiación solar distribuidos por etapa fenológica para los cuatro ciclos productivos.



a) Ciclo 2001-2002

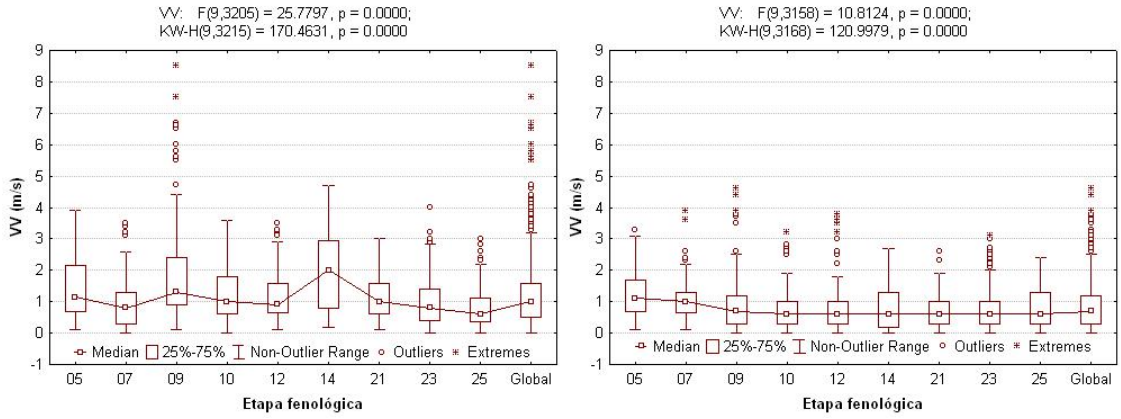
b) Ciclo 2002-2003

c) Ciclo 2003-2004

d) Ciclo 2004-2005

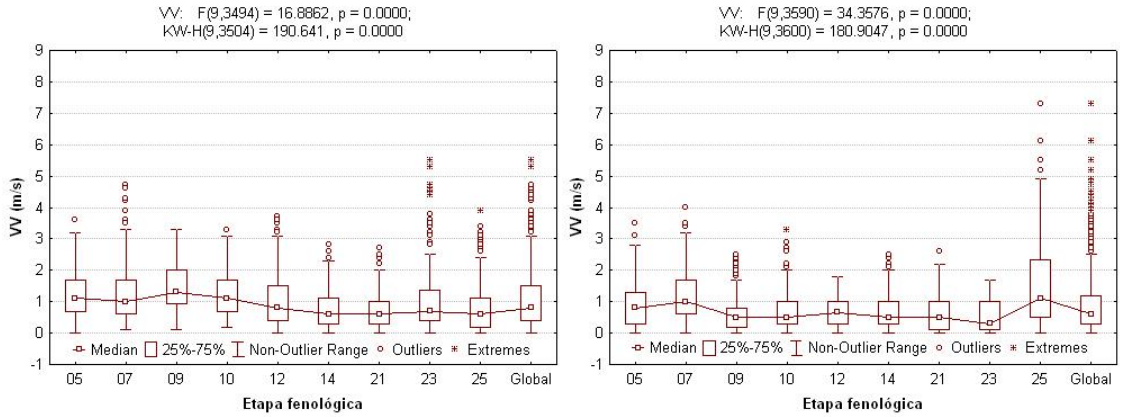
Fuente: Elaboración propia.

Figura B.5 Dispersión de los valores de la velocidad del viento distribuidos por etapa fenológica para los cuatro ciclos productivos.



a) Ciclo 2001-2002

b) Ciclo 2002-2003

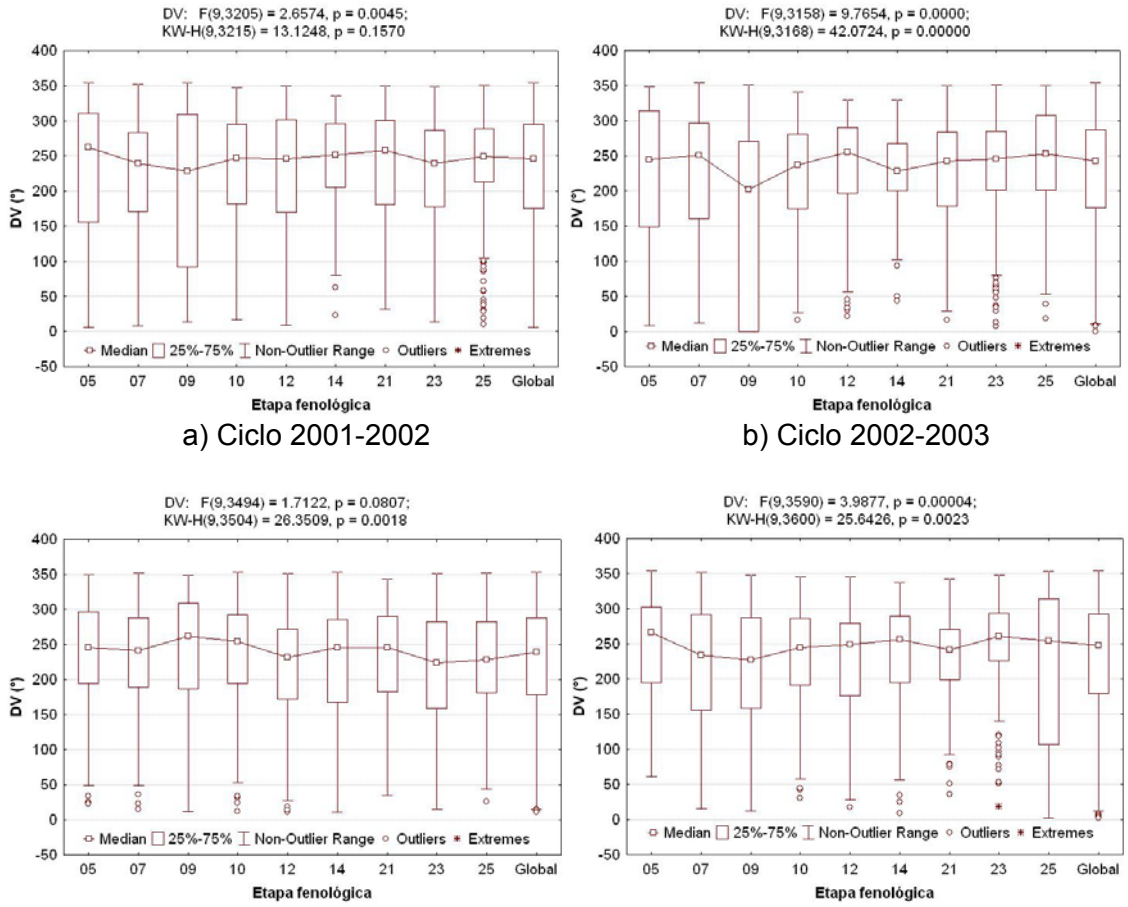


c) Ciclo 2003-2004

d) Ciclo 2004-2005

Fuente: Elaboración propia.

Figura B.6 Dispersión de los valores de la radiación solar distribuidos por etapa fenológica para los cuatro ciclos productivos.



a) Ciclo 2001-2002

b) Ciclo 2002-2003

c) Ciclo 2003-2004

d) Ciclo 2004-2005

Fuente: Elaboración propia.

Anexo C

Escala de Beaufort del Viento

Fuerza Beaufort	Velocidad del Viento (KmPH)	Velocidad del Viento (MPH)	Indicadores	Términos Usados en las Predicciones del NWS
0	0-2	0-1	Calma; el humo sube verticalmente.	Calma
1	2-5	1-3	La dirección se puede apreciar por la dirección del humo, pero no por medio de veletas.	Ventolina
2	6-12	4-7	El viento se siente en el rostro, las hojas se mueven ligeramente; las veletas ordinarias se mueven con el viento.	Ligero
3	13-20	8-12	Las hojas y las ramas delgadas se mueven constantemente; el viento extiende las banderas ligeras.	Suave
4	21-29	13-18	Levanta polvo y papeles sueltos; las ramas pequeñas se mueven.	Moderado
5	30-39	19-24	Los árboles pequeños empiezan a balancearse; en los lagos pequeños se observan olas con crestas.	Fresco
6	40-50	25-31	Se mueven las ramas grandes; los cables telefónicos silban; es difícil usar sombrillas.	Fuerte
7	51-61	32-38	Los árboles enteros se mueven; es incómodo caminar contra el viento.	Muy fuerte
8	62-74	39-46	Se rompen las ramas de los árboles; generalmente no se puede avanzar.	Ventarrón
9	75-87	47-54	Daños estructurales ligeros.	Ventarrón Fuerte
10	88-101	55-63	Pocas veces se siente en tierra firme; los árboles son arrancados de raíz; ocurren daños estructurales considerables.	Temporal
11	102-116	64-72	Casi nunca sucede en tierra firme; acompañado de daños graves generalizados.	Borrasca
12	117 o más	73 o más	Casi nunca sucede; acompañado de devastación.	Huracán