Estimación de parámetros meteorológicos secundarios en la zona de la Cujae utilizando técnicas de minería de datos

Estimation of secondary meteorological parameters in CUJAE's zone using data mining techniques

Gil Cruz Lemus
Telf. 266 3277, gil@tesla.cujae.edu.cu

Rosabel Zerquera Díaz,
Telf. 266 3277, rzerquera@tesla.cujae.edu.cu

Ayleen Morales Montejo
Telf. 266 3578, ayleen@dcrhmail.cujae.edu.cu

Alejandro Rosete Suárez
Teléf. 266 3787 rosete@ceis.cujae.edu.cu

Instituto Superior Politécnico "José Antonio Echeverría"
Calle 114 no. 11901 entre Rotonda y Ciclovía, Marianao, La Habana, Cuba, CP 19390

Recibido: noviembre 6, 2011; aceptado: marzo 22, 2012.

Resumen

El presente trabajo desarrolla un proceso de descubrir conocimiento en bases de datos en el grupo de Medio Ambiente del Instituto Superior Politécnico José Antonio Echeverría (CUJAE), en colaboración con el Centro de Gestión de la Información y Desarrollo de la Energía (Cubaenergía), con el objetivo de obtener un modelo de datos para estimar el comportamiento de parámetros meteorológicos secundarios a partir de datos de superficie. Se detallan algunos aspectos relacionados con la minería de datos y su aplicación en el entorno meteorológico, así como se selecciona y se describe la metodología CRISP-DM y la herramienta de análisis de datos WEKA. Asimismo, se utilizan las tareas de selección de atributos y de regresión, la técnica de redes neuronales de tipo perceptrón multicapas y los algoritmos CfsSubsetEval, BestFirst y MultilayerPerceptron. Se obtienen modelos de estimación para los parámetros meteorológicos secundarios: *altura de la capa de mezcla convectiva*, *altura de la capa de mezcla mecánica*, *velocidad de fricción*, *flujo de calor superficial* y *velocidad convectiva de escala*, necesarios para el estudio de los modelos de dispersión de contaminantes en la zona de la CUJAE. Los resultados obtenidos constituyen un precedente para futuras investigaciones, así como para la continuidad de esta.

Palabras Clave: Meteorología, estimación, parámetros meteorológicos secundarios, minería de datos, redes neuronales.

Abstract

This work develops a process of knowledge discovery in databases for the group of Environmental Research at the Higher Polytechnic Institute José Antonio Echeverría in collaboration with the Center of Information Management and Energy Development (Cubaenergía) in order to obtain a data model to estimate the behavior of secondary weather parameters from surface data. There are described some aspects of data mining and its application in the meteorological environment, also there are selected and described the CRISP-DM methodology and data analysis tool WEKA. Tasks used: attribute selection and regression, technique: neural network of multilayer perceptron type and algorithms: CfsSubsetEval, BestFirst and MultilayerPerceptron. Estimation models are obtained for secondary meteorological parame-

ters: *height of convective mixed layer, height of mechanical mixed layer, friction velocity, surface heat flux* and *convective velocity scale*, necessaries for the study of patterns of dispersion of pollutants in CUJAE's zone. The results set a precedent for future research and for the continuity of this.

KEYWORDS: Meteorology, estimation, secondary meteorological parameters, data mining, neural networks.

Introducción

En la actualidad, la automatización de las actividades de los negocios produce un flujo creciente de datos, puesto que incluso la información referente a acciones tan simples como una llamada telefónica o un test médico es almacenada en una computadora. Las empresas e instituciones se encuentran abrumadas por este crecimiento acelerado del tamaño y la cantidad de datos. Es imprescindible convertir los grandes volúmenes de datos existentes en experiencia, conocimiento y sabiduría, formas que son útiles para la toma de decisiones y el desarrollo económico-social contemporáneo.

Tal es el caso del grupo de Medio Ambiente del Instituto Superior Politécnico José Antonio Echeverría (CUJAE), el cual cuenta con una estación meteorológica automática que recoge datos de superficie desde el 15 de abril de 2008, a los que no se les da ningún tratamiento o explotación, entre estos están: la *dirección* y la *velocidad del viento*, la *temperatura*, la *humedad relativa* y la *presión*. La figura 1 muestra una estación meteorológica automática como la instalada en la CUJAE, con los sensores que miden cada una de las variables meteorológicas.

Estos datos son almacenados en soporte digital (con restricciones de capacidad) y procesados por la consola Vantage Pro 2 de Davis, que solo los muestra en forma numérica y, además, predice el comportamiento de algunos de estos para las próximas 12 ho-

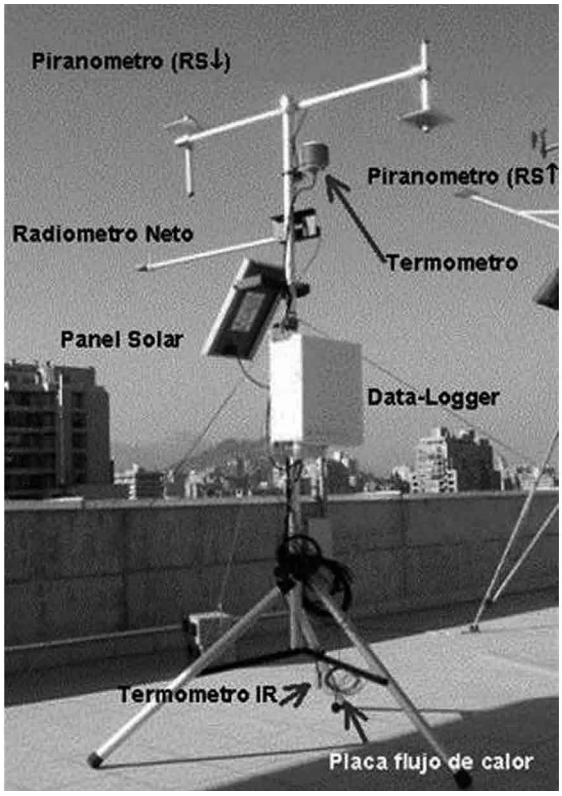

Figura 1. Ejemplo de estación meteorológica automática

Figura 2. Consola Vantage Pro2

ras (Fig. 2). Existe también el *software* WeatherLink (en su versión 5.7 para Windows, de 2006), que conecta la estación meteorológica a la computadora, lo cual posibilita el intercambio de datos y su almacenamiento total, así como permite plotear, analizar, exportar e imprimir los datos meteorológicos, y configurar la estación y monitorear las alarmas.

Actualmente, en Cuba existe una red de monitoreo automático de parámetros meteorológicos de superficie, con la que se han obtenido series temporales para calcular variables meteorológicas necesarias con vistas a estudiar los modelos de dispersión de contaminantes en una zona.

El grupo está interesado en el análisis de la dispersión local de contaminantes gaseosos y de partículas en la zona de la CUJAE; para ello, es necesario poseer los valores de parámetros meteorológicos secundarios como: la *altura de la capa de mezcla convectiva*, la *altura de la capa de mezcla mecánica*, la *velocidad convectiva de escala*, el *flujo de calor superficial*, la *velocidad de fricción*, entre otros. Hoy día, no se cuenta con estos valores, pero pueden obtenerse mediante cálculos a partir de los parámetros meteorológicos primarios siguientes: la *dirección* y la *velocidad del viento*, la *temperatura exterior*, la *humedad relativa*, las *precipitaciones*, la *presión barométrica* y la *radiación solar*; estos últimos son los datos de superficie que se recogen en la estación meteorológica automática.

Para la obtención de los parámetros meteorológicos secundarios se propone la utilización del preprocesador meteorológico Aermet, del sistema de modelos Aermod, establecido por la Agencia de Protección Ambiental de los Estados Unidos (EPA). La empresa Cubaenergía dispone del *software* Lakes Enviromental, una versión mejorada del Aermet de la EPA, que brinda entre otros, un potente entorno visual. (Lakes Enviromental es propiedad de una empresa canadiense que limita su distribución y uso). El empleo de cualquiera de estas dos versiones de Aermet requiere de datos de superficie y sondeo que, actualmente, en Cuba no se realizan de forma sistemática, por lo cual Cubaenergía desarrolló la versión Aermet+.

La versión Aermet+ de Cubaenergía, consume un tiempo de procesamiento considerable, una parte del cual está destinado a la preparación de los datos y la creación de ficheros de entrada para el sistema. Además, esta versión solo permite trabajar con datos horarios; por ello, si se poseen varias mediciones por hora, es necesario hacer un promedio vectorial que puede implicar la pérdida de grados de exactitud en las mediciones.

Por todo lo expuesto, se considera engorroso el cálculo de los parámetros meteorológicos secundarios; además, no resulta factible el procesamiento de los datos, lo cual imposibilita el análisis de la dispersión local de contaminantes. Con los valores que se obtienen mediante Aermet+, no se pueden obtener patrones de comportamiento de los datos. Además, no se cuenta con los métodos y herramientas de procesamiento y análisis que den sentido y utilidad a la información existente.

Como objetivo del trabajo se propone obtener modelos que permitan analizar las dependencias entre los parámetros meteorológicos y estimar los parámetros meteorológicos secundarios respecto a los datos de superficie con las tareas de selección de atributos y regresión mediante redes neuronales y el empleo de la metodología CRISP-DM y la herramienta de análisis de datos WEKA.

Materiales y métodos

Sistema de modelos Aermod

Muchas actividades industriales pueden afectar de manera directa la calidad del aire y, en consecuencia, la salud humana. Conocer en qué proporciones se afecta la calidad del aire es importante cuando se trazan las estrategias de desarrollo energético, entre otras, para lo cual es imprescindible el uso de modelos de calidad de aire. Estos modelos simulan los procesos

físico-químicos que afectan y transforman los contaminantes en el aire, y estiman los incrementos de las concentraciones de los contaminantes primarios que son emitidos directamente a la atmósfera y, en algunos casos, los contaminantes secundarios resultantes de reacciones químicas que tienen lugar en esta, apoyando la búsqueda de soluciones para mitigar la contaminación atmosférica.

Dentro de los modelos de calidad de aire están los clasificados como de dispersión, los cuales son usados para estimar la concentración de contaminantes en receptores que rodean las fuentes a determinado nivel sobre la tierra; estos, a su vez, pueden ser divididos en: modelos de dispersión local o regional, según la distancia de la fuente en la que se dispersan los contaminantes. Los modelos de dispersión se construyen a partir de los parámetros secundarios meteorológicos: *la velocidad de fricción, la altura de la capa de mezcla convectiva y el flujo de calor superficial, entre otros.*

El sistema de modelos Aermod, desarrollado por la EPA en conjunto con la Sociedad Meteorológica Americana, despliega modelos de dispersión de la calidad del aire. El Aermod incluye tres preprocesadores de datos de entrada que son componentes regulatorios del sistema: Aermet, preprocesador de datos meteorológico, Aermap, preprocesador de los datos del terreno y Aersurface, un preprocesador de datos de uso de suelos.

El Aermet se encarga de preparar los datos meteorológicos disponibles, según el formato apropiado, con miras a su posterior utilización por Aermod; está diseñado para operar sobre los tipos de datos siguientes:

1. Datos de superficie (horarios).
2. Datos de sondeo, de aire superior, tomados dos veces al día.
3. Colección de datos provenientes de un programa de mediciones *in situ*, como torres instrumentadas.

Para utilizar el procesador meteorológico Aermet son obligatorios los datos referentes a: *la dirección y la velocidad del viento y la temperatura ambiente;* así como la tasa de precipitación, imprescindible para modelar la deposición húmeda. De no contar con los valores de la nubosidad y la altura de la base de nubes, se recomienda conocer la radiación solar; se necesitan, además, los siguientes valores apropiados para las características del terreno: el coeficiente de rugosidad de la superficie (Z0), Albedo (r) y la tasa de Bowen (B0). Aermet se nutre de las tres etapas siguientes:

1. Extracción y verificación de los datos.
2. Unión de todos los datos meteorológicos.
3. Cálculo de variables secundarias y creación de los archivos meteorológicos que Aermod necesita.

Datos de superficie

Estación meteorológica automática

Una estación meteorológica automática es una herramienta mediante la cual se obtienen datos de los parámetros meteorológicos, leídos por medio de sensores eléctricos. Las lecturas son acondicionadas para luego ser procesadas mediante la tecnología de microcontroladores o microprocesadores, y transmitidas a través de un sistema de comunicación (radio, satélites, teléfono, etc.) en forma automática. La estación automática funciona en forma autónoma, las 24 horas, con un sistema de alimentación a base de energía solar (paneles solares) o mediante el uso de la energía eólica.

Consola vantage pro 2

La consola Vantage Pro 2 cuenta con un conjunto de sensores integrados: pluviómetro de cazoletas, sensores de temperatura, humedad y anemómetro en un solo kit, entre otros, y está disponibles en versiones inalámbricas y cableadas. La consola Vantage Pro 2, diseñada para proporcionar lecturas extremadamente precisas, transmite y recibe datos a una

distancia de hasta 300 m en línea visual. Registra y visualiza los datos de la estación y muestra su propio pronóstico en la pantalla gráfica, sin necesitad de una computadora; no obstante, proporciona funciones de gráficos y alarmas, e interfaces hacia la computadora.

Weatherlink

WeatherLink constituye un *software* para el seguimiento de las condiciones meteorológicas. Su cargador de datos encaja perfectamente en el interior de la consola Vantage Pro 2 y almacena los datos meteorológicos incluso con la PC apagada. Posibilita al usuario elegir el período de archivo con el que desea almacenar los datos (1 min, 5 min, 10 min, 15 min, 30 min, 60 min o 120 min), y puede almacenar hasta seis meses de datos en dependencia del intervalo del archivo. Este *software* permite generar gráficos, sumarios, analizar dependencias, entre otras, así como comprobar las condiciones meteorológicas en tiempo real en el boletín instantáneo o en gráficos de base diaria, semanal, mensual o anual. Tiene funciones especiales de seguimiento energético para la medida de *grados-día* de frío y de calor, y de radiación solar; asimismo, permite estimar el riesgo de quemadura solar para cada tipo de piel a partir del índice de UV.

Datos meteorológicos primarios

Velocidad del Viento: *Distancia recorrida por el viento en la unidad de tiempo. Tiene como unidades de medida: kilómetro por hora (km/h), metro por segundo (m/s), nudo.*

Dirección del viento: Punto cardinal del cual procede el viento, según la Rosa de los Vientos.

Temperatura del Aire: Magnitud proporcional a la energía cinética media de las moléculas de aire. Tiene como unidades: *grado Celsius (˚C), grado Kelvin (˚K), grado Fahrenheit (˚F).*

Humedad Relativa del Aire: Humedad que contiene una masa de aire, en relación con la máxima humedad absoluta que podría admitir sin producirse la condensación, conservando las mismas condiciones de temperatura y presión atmosférica. Esta es la forma más habitual de expresar la humedad ambiental. Se expresa en *tanto por ciento (%).*

Insolación: Tiempo durante el cual los rayos solares inciden directamente (radiación directa) sobre la parte sensible del instrumento, sobrepasando un cierto umbral de radiación (aproximadamente, 200 W/m^2).

Presión Atmosférica: Presión ejercida por el aire en cualquier punto de la atmósfera; peso de la columna de aire por unidad de superficie. La medida de presión del Sistema Internacional de Unidades (SI) es *el newton por metro cuadrado (N/m^2), pascal (Pa) o milibar (mbar).*

Precipitación: Volumen de agua caída (líquida), por metro cuadrado de superficie, en el lugar de observación. La unidad de medida es el *milímetro (mm).*

Datos meteorológicos secundarios

Velocidad de Fricción: Velocidad de referencia del viento que suele aplicarse a los movimientos cerca del suelo donde, a menudo, se supone que la cizalladura es independiente de la altura y aproximadamente proporcional al cuadrado de la velocidad media del viento.

Flujo de Calor Superficial: Parámetro físico muy relacionado con los procesos atmosféricos. La transferencia de calor es el proceso mediante el cual se intercambia energía en forma de calor entre distintos cuerpos o entre diferentes partes de un mismo cuerpo que están a distintas temperaturas. El calor se transfiere mediante la convección, la radiación o la conducción, y aunque estos tres procesos pueden tener lugar simultáneamente, podría ocurrir que uno de los mecanismos predomine sobre los otros dos. Para el caso de la tierra con respecto al sol, ocurre el fenómeno de transferencia por radicación. El flujo de calor sensible muestra una clara dependencia de la insolación diurna. Los valores máximos se presentan durante el intervalo cuando se registra la temperatura máxima. El flujo de calor sensible tiene valores

positivos mínimos durante la salida y la puesta del sol, y es negativo durante la noche.

Altura de la Capa de Mezcla. Altura de la capa límite atmosférica o altura de la capa de mezcla (Z); es un parámetro fundamental que caracteriza la estructura de la troposfera baja, zona inferior de la atmósfera donde ocurre fundamentalmente el transporte turbulento de masa y energía, y donde los contaminantes se trasladan e interaccionan. Las sustancias emitidas en la capa se dispersan gradualmente (horizontal y verticalmente) por la acción de la turbulencia y, por último, se mezclan completamente en esta capa si permanecen el tiempo suficiente y no hay ningún sumidero significativo. Por consiguiente, en la meteorología de la contaminación atmosférica se usa, a menudo, el término capa de mezcla o capa mezclada. La altura de la capa de mezcla es de gran importancia para los modelos de contaminación atmosférica, por cuanto determina la profundidad vertical de la atmósfera donde se produce el mezclado y la dispersión de los contaminantes, y está envuelta en muchos métodos y(o) modelos predictivos y de diagnóstico para evaluar las concentraciones de los contaminantes; es también un parámetro importante en los modelos de flujo atmosférico.

Velocidad Convectiva de Escala. También denominada velocidad vertical de escala (w*), caracteriza la porción convectiva de la turbulencia, y es necesaria para estimarla en la capa límite planetaria. Es un parámetro esencial que, físicamente, representa la velocidad típica de los remolinos grandes, como son las térmicas, en la capa de mezcla dominada por efectos convectivos intensos.

Minería de datos

Entre las múltiples definiciones que identifican a la *minería de datos* se encuentra:

> *...es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos.*

Áreas de aplicación de la minería de datos

La integración de las técnicas de minería de datos en las actividades cotidianas se está convirtiendo en algo habitual; para la mayoría de las empresas y organizaciones que la aplican, su empleo ha significado un aumento de sus ofertas, así como una reducción de sus costos o una replanificación de su estrategia de trabajo.

Es posible encontrar ejemplos de su aplicación, entre otras, en las áreas siguientes: financieras, seguros, científicas (medicina, farmacia, astronomía, sicología), políticas, económicas, sanitarias o demográficas, educación, policiales, procesos industriales.

En la actualidad, la aplicación de técnicas de minería de datos en el campo de la meteorología se ha incrementado considerablemente. Esta es un área en la que es abundante el volumen de datos que se genera, y en la que se han desarrollado diferentes conjuntos de datos con numerosas variables meteorológicas. Una de las ramas en la que se han empleado estas técnicas es el pronóstico del tiempo, que constituye un problema real donde el análisis de datos cobra una alta significación, dada la complejidad y la variabilidad de este proceso a lo largo de los años. Como ejemplos pueden citarse lo siguientes:

1. Análisis del viento en el valle de Río Negro, en Argentina, utilizando procedimientos de redes neuronales a partir de la integración de mapas autoorganizados y algoritmos de inducción.
2. Se desarrolló, en 2003, el sistema CubaForecast, en el centro meteorológico de Cienfuegos, con el cual es posible realizar, para las diferentes provincias del país, el pronóstico de variables meteorológicas.
3. Estudio de series temporales de contaminación ambiental mediante técnicas de redes neuronales artificiales en Chile.
4. Empleo de las redes neuronales artificiales para identificar la fuente de los contaminantes observados.

Tareas de minería de datos

Una tarea de minería de datos es un problema de minería de datos. Las tareas pueden ser predictivas o descriptivas. Entre las tareas predictivas se encuentran, fundamentalmente, la clasificación y la regresión; entre las descriptivas, el agrupamiento, la asociación y la correlación. Las tareas predictivas tratan de problemas y tareas en los que hay que predecir uno o más valores para uno o más ejemplos. Las tareas descriptivas se encargan de describir los datos existentes, puesto que los ejemplos se presentan como un conjunto. No obstante, también es posible obtener modelos descriptivos a partir de tareas predictivas, y viceversa.

Técnicas de minería de datos

Existen muchas técnicas empleadas en la minería de datos para extraer la información y obtener modelos o patrones a partir de los datos; entre estas: los árboles de decisión, las redes neuronales artificiales, los algoritmos genéticos y la lógica difusa.

Las redes neuronales artificiales han recibido un interés particular, puesto que ofrecen los medios necesarios para revelar categorías comunes en los datos, dado que son capaces de detectar y aprender complejos patrones, y características dentro de los datos; así como para modelar de manera efectiva y eficiente problemas grandes y complicados, de forma individual o en combinación con otros métodos.

Una de las características de las redes neuronales artificiales es que son capaces de trabajar con datos incompletos e, incluso, paradójicos, lo cual, según el problema, puede resultar una ventaja o un inconveniente. Entre su desventajas suelen nombrarse su sensibilidad a valores anómalos; el hecho que necesitan muchos ejemplos para el aprendizaje y que son relativamente lentas; y, fundamentalmente, su incomprensibilidad.

Metodologías de minería de datos

El empleo de una metodología bien estructurada y organizada en la realización de un proyecto de minería de datos facilita la planificación y la dirección del proyecto, permite realizar un mejor seguimiento de este y facilita la realización de nuevos proyectos con características similares.

Encuestas realizadas demuestran qué tan diversa puede ser la elección de la metodología para la realización de un proyecto de minería de datos; entre las más difundidas se encuentran CRISP-DM y SEMMA.

La metodología CRISP-DM estructura el ciclo de vida de un proyecto de minería de datos en seis fases que interactúan entre sí de forma iterativa durante el desarrollo del proyecto. CRISP-DM ha sido diseñada como una metodología neutra respecto a la herramienta de desarrollo que se utilice. CRISP-DM 1.0, que, además de ser de libre distribución y estar en constante perfeccionamiento por la comunidad internacional, se considera más completa que SEMMA, puesto que posee una fase de desarrollo dedicada, íntegramente, al entendimiento del negocio, con un estándar que muestra una precisa y sólida distribución de tareas de carácter general con sus salidas más significativas, así como una guía para su desarrollo.

Herramientas de minería de datos

Las herramientas de minería de datos utilizadas para resolver problemas del mundo real en la ingeniería, la ciencia y los negocios, implementan las técnicas de minería antes expuestas por medio de algoritmos (casi todos de gran complejidad y costo computacional) en diferentes lenguajes de programación.

El número de herramientas para realizar análisis de minería de datos aumenta con el vertiginoso desarrollo de la tecnología en la actualidad. Algunas encuestas realizadas en mayo de 2007 en el portal de minería de datos y gestión de conocimientos Kdnuggets, evidencian la variedad en el uso de estaspor organizaciones y personas. Clementine se presenta como líder en el mercado, seguida de las de libre distribución Yale y WEKA.

WEKA (Waikato Enviroment for Knowledge Analysis), desarrollada por un equipo de investigadores de la

Universidad de Waikato (Nueva Zelanda), en sus inicios, era solo una librería; en cambio, hoy día es un paquete integrado. Constituye un entorno de experimentación de análisis de información, con diferentes técnicas de preprocesado, clasificación, asociación y visualización de datos. Posee una interfaz gráfica de usuario compuesta de cuatro entornos que permiten diferentes funcionalidades y formas de análisis, además de tres soportes para cargar las fuentes de datos. Su desarrollo sobre el lenguaje Java la hace multiplataforma; asimismo, el hecho de ser de código abierto, unido a su prestigio, hace que se encuentre en evolución constante por parte de la comunidad internacional.

Resultados

La CUJAE cuenta con una estación meteorológica automática que almacena datos en formato digital, desde abril de 2008, hasta la fecha. El grupo de Medio Ambiente del Instituto necesita obtener parámetros meteorológicos secundarios a partir de esta información, los cuales permitan la obtención de modelos de dispersión local de contaminantes gaseosos y de partículas en la zona. Estos parámetros se obtienen mediante algunas de las variables medidas por la estación utilizando el *software* Aermet, que consume un tiempo de procesamiento elevado, una parte destinado a la preparación de los datos y la creación de los ficheros de entrada al sistema. Esta situación motiva el uso técnicas y herramientas que posibiliten la extracción del conocimiento oculto en los datos y permitan una mejor compresión del comportamiento de los parámetros meteorológicos primarios y secundarios, así como la obtención de estos últimos mediante la estimación.

Con las herramientas de análisis de datos Microsoft Excel 2007 y WEKA 3.5.8 se realiza una exploración del comportamiento de los datos meteorológicos; su frecuencia, en caso de ser nominales; y su distribución, si son numéricos (valor máximo, mínimo, media, desviación estándar y cantidad de valores distintos). Los resultados más significativos del reporte exploratorio de los datos son siguientes:

La *temperatura* se mueve dentro del rango de valores entre 10,2 ˚C y 34,9 ˚C, con una media de 22,55˚C. La *humedad relativa* presenta un valor mínimo de 38 % y un valor máximo de 98 %, con una media de 78,44 %. La *dirección del viento*: ENE es el valor predominante con 6 014 instancias; al igual que ESE, con 4 171; WSW es el valor menos frecuente, con solo 158 instancias.

Para el cálculo de los parámetros meteorológicos secundarios mediante Aermet+, se ha dividido la fecha en *año*, *mes* y *día*, y del tiempo se ha tomado solo la *hora*; la *velocidad del viento* está expresada en *kilómetro por hora* (km/h), por lo que se utiliza la ecuación 1, para llevarla a *metro por segundo* (m/s):

$$Km/h * h/3600s * 1000m/Km \qquad (1)$$

La *presión barométrica* es multiplicada por *10*; las *precipitaciones* son multiplicadas por *1000*; los valores de la *dirección del viento*, según la Rosa de los Vientos, se sustituyen por sus respectivos valores numéricos (en grados) desde el norte.

Como resultado del proceso de selección, transformación y construcción de datos, se obtiene un conjunto de datos compuesto por 17 campos y un total de 20 854 instancias. Como parte de la limpieza se eliminan los valores perdidos, para no introducir errores en los modelos de estimación; además, puesto que los atributos pueden moverse en un rango tan amplio de valores, se eliminan registros con valores extremos en sus atributos, los que presentan una baja frecuencia de aparición y lejanía de la media numérica del grupo. Se emplean los datos normalizados, con vistas a evitar trabajar con rangos tan amplios de valores y, de este modo, disminuir los errores en la modelación.

Como técnicas de modelado se escogen los algoritmos *CfsSubsetEval* y *BestFirst* (métodos de evaluación y búsqueda, respectivamente) para la selección de los atributos que más influyen en los parámetros meteorológicos se-

cundarios, y *MultilayerPerceptron* para estimar los valores de los parámetros secundarios a partir de los primarios.

Para entrenar y probar los modelos se emplean conjuntos distintos en aras de no sobrestimar su precisión. En este sentido, se utiliza una validación cruzada (*cross-validation*) de diez pliegues, la cual, de forma aleatoria, divide el conjunto de datos en diez subconjuntos y realiza diez iteraciones, donde en cada una se reserva un grupo diferente para el conjunto de prueba y los nueve restantes para entrenar el modelo.

Con miras a la selección de atributos se realizan diez experimentos (cinco de estos con los atributos normalizados y cinco sin normalizar), con los cuales se determinaron las variables fundamentales que influyen en los parámetros meteorológicos secundarios y su porcentaje de incidencia.

En este experimento se comentan los atributos que contribuyen con la estimación de cada uno de los parámetros meteorológicos secundarios, entre 70 % y 100 %, por considerarse este valor significativo.

Para la *altura de la capa de mezcla convectiva* se destacan los atributos: *año*, *velocidad del viento* y *humedad*, con 100 % de influencia. Para la *altura de la capa de mezcla mecánica* se destacan los atributos: *velocidad del viento* y *radiación*, con 100 % de influencia. Para la *velocidad de fricción* se destacan los atributos: *año*, *velocidad del viento* y *radiación*, con 100 % de influencia. Para el *flujo de calor superficial* se destacan los atributos: *mes*, *hora* y *humedad* (con 90 %) y *año*, *dirección del viento* y *radiación*, con 100 % de influencia. Para la *velocidad convectiva de escala* se destacan los atributos: *año*, *velocidad del viento*, *radiación* y *humedad*, con 100 % de influencia.

Los modelos anteriores están enfocados a la estimación de los parámetros meteorológicos secundarios; en este sentido, el conocimiento que aportan favorece considerablemente los resultados de próximos experimentos.

En el caso de la estimación numérica, se realizan 20 experimentos, cuatro por cada variable secundaria (con los atributos normalizados y sin normalizar), y tomando en cuenta (y no) la selección de atributos del proceso anterior, con la finalidad de comparar los resultados (Tablas 1, 2, 3, 4 y 5).

Tabla 1
Resultados de la estimación para la altura de la capa de mezcla convectiva

Datos	Todos los atributos	Con los atributos seleccionados
Sin normalizar	Coeficiente de correlación: 0,974 Error medio absoluto: 83,307	Coeficiente de correlación: 0,8567 Error medio absoluto: 184,963
Normalizados	Coeficiente de correlación: 0,9722 Error medio absoluto: 0,0414	Coeficiente de correlación: 0,716 Error medio absoluto: 0,1238

Tabla 2
Resultados de la estimación para la altura de la capa de mezcla mecánica

Datos	Todos los atributos	Con los atributos seleccionados
Sin normalizar	Coeficiente de correlación: 0,8418 Error medio absoluto: 138,5182	Coeficiente de correlación: 0,8104 Error medio absoluto: 137,573
Normalizados	Coeficiente de correlación: 0,8485 Error medio absoluto: 0,0607	Coeficiente de correlación: 0,8096 Error medio absoluto: 0,0667

Tabla 3
Resultados de la estimación para el flujo de calor de superficie

Datos	Todos los atributos	Con los atributos seleccionados
Sin normalizar	Coeficiente de correlación: 0,97 Error medio absoluto: 8,25	Coeficiente de correlación: 0,95 Error medio absoluto: 11,47
Normalizados	Coeficiente de correlación: 0,97 Error medio absoluto: 0,02	Coeficiente de correlación: 0,95 Error medio absoluto: 0,0282

Tabla 4
Resultados de la estimación para la velocidad de fricción

Datos	Todos los atributos	Con los atributos seleccionados
Sin normalizar	Coeficiente de correlación: 0,9274 Error medio absoluto: 0,0492	Coeficiente de correlación: 0,8686 Error medio absoluto: 0,0737
Normalizados	Coeficiente de correlación: 0,9274 Error medio absoluto: 0,0465	Coeficiente de correlación: 0,8686 Error medio absoluto: 0,0697

Tabla 5
Resultados de la estimación para la velocidad convectiva de escala

Datos	Todos los atributos	Con los atributos seleccionados
Sin normalizar	Coeficiente de correlación: 0,9884 Error medio absoluto: 0,065	Coeficiente de correlación: 0,746 Error medio absoluto: 0,3491
Normalizados	Coeficiente de correlación: 0,9888 Error medio absoluto: 0,0255	Coeficiente de correlación: 0,7644 Error medio absoluto: 0,1301

Como resultado del análisis comparativo de los modelos de estimación para cada variable meteorológica secundaria, aplicando la técnica de normalización, se decide seleccionar aquellos modelos en los cuales esta se aplica, puesto que se obtienen valores elevados del coeficiente de correlación y valores más bajos del error

absoluto; de este modo, los resultados son más confiables. Por tanto, se decide —por opinión de los expertos— que ambos modelos (normalizado con todos los atributos y normalizado con selección de atributos) pueden utilizarse en la práctica para estimar los parámetros meteorológicos secundarios.

Los modelos de estimación de las variables meteorológicas secundarias en estudio (normalizado con selección de atributos y normalizado con todos los atributos) proyectaron resultados favorables, donde los errores medios absolutos oscilan entre 0,02 y 0,34 para todas las variables, y los coeficientes de correlación son mayores de 76 %.

Como propuesta de los expertos de Cubaenergía para la validación de los modelos obtenidos, se utiliza la desviación fraccional, la cual plantea que los valores estimados son aceptables si tienen un valor de desviación fraccional acotado entre (-0,67; +0.67); esta se calcula mediante la ecuación 2, donde V_1 es el valor estimado y V_2 es el valor real. Este análisis se realiza para los valores reales, predichos durante la fase de prueba de cada modelo, y los resultados se muestran en la tabla 6 para las cinco variables (con los datos normalizados y con todos los atributos).

$$V1 - V2/V1+V2 \qquad (2)$$

El porcentaje de las instancias que está fuera del rango de la desviación fraccional para cada variable es relativamente bajo en todos los casos, por lo cual los modelos pueden tomarse como aceptables y válidos para la estimación.

Tabla 6. Resultados de los cálculos de la desviación fraccional

Parámetro meteorológico secundario	Porcentaje de instancias que están fuera de rango
Altura de la Capa de Mezcla Convectiva	2,07% de 10322 Instancias
Altura de la capa de mezcla mecánica	10,35% 20835 Instancias
Velocidad de fricción	5,40% de 18492 Instancias
Velocidad convectiva de escala	3,141% de 10322 Instancias
Flujo de calor de superficie	2,67% de 16667 Instancias

Conclusiones

Al término del presente trabajo de investigación, teniendo en cuenta los resultados obtenidos, se concluye lo siguiente:

1. Se determinaron los atributos o parámetros meteorológicos primarios que más influyen en la estimación de cada uno de los parámetros meteorológicos secundarios.
2. Se presentan varios modelos de regresión basados en redes neuronales artificiales, que permiten estimar los valores de los parámetros meteorológicos secundarios: la *altura de la capa de mezcla convectiva*, la *altura de la capa de mezcla mecánica*, la *velocidad de fricción*, el *flujo de calor superficial* y la *velocidad convectiva de escala*, a partir de los datos meteorológicos primarios.
3. Los modelos normalizados obtenidos con todos los atributos y con selección de atributos normalizados pueden ser utilizados, puesto que presentan un coeficiente de correlación alto, un error medio cuadrático pequeño y un bajo porciento de instancias fuera de rango.

Bibliografía

Addendum (2006): *User's guide for the Aermod Meteorological Preprocessor (Aermet)*, Estados Unidos, Carolina del Norte, EPA, Environmental Protection Agency.

Chapman, P., Clinton, J. (2000): *CRISP-DM 1.0: Step-by-step data mining guide*, Estados Unidos, SPSS Inc. CRISP-DM Consortium.

Cogliati, M. G., Britos, P., García-Martínez, R. (2006): *Análisis del viento en el valle del Río Negro mediante mapas auto organizados y algoritmos de inducción*, Argentina.

Díaz, Y., Fernández, A. (2003): *CubaForecast para el pronóstico de variables meteorológicas*, Cuba, Cien-

fuegos, Centro Meteorológico.

Environmental Protection Agency, EPA (2004): *User's guide for the Aermet meteorological preprocessor (Aermet)*, Estados Unidos, Carolina del Norte.

Environmental Protection Agency, EPA (2008): *AERSURFACE: User's Guide*, Estados Unidos, Carolina del Norte.

Gondar Nores, J. E. (2004): *Metodologías para la realización de proyectos de data mining*, España, Madrid, Data Mining Institute, Consultado: 15 de febrero 2010, http://www.estadistico.com/index.html.

Hernández Orallo, J, Ramírez Quintana, M. J.; Ferri Ramírez, C. (2004): *Introducción a la minería de datos*, Madrid, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Pearson Education S.A.

Kdnuggets (2007): *Polls: Data Mining Methodology*. Consultado: 15 de febrero 2010,
http://kdnuggets.com/polls/2007/data_mining_methodology.htm

_____ (2007): *Polls: Data Mining / Analytic Software Tools*, consultado: 15 de febrero 2010, http://kdnuggets.com/polls/2007/data_mining_software_tools.htm.

Núñez Crespí, S. (2002): "*Altura de la capa de mezcla: Caracterización experimental y aplicación de un modelo meteorológico para el estudio de su evolución diurna*, Doctora en Ciencias Físicas", Universidad Complutense de Madrid.

Reich, S. L., D. R. Gómez and L. E. Dawidowski (1999): "Artificial neural network for the identification of unknown air pollution sources", *Atmospheric Environment*.

Rodríguez Valdés, D., L. Echevarría Pérez y O. Cuesta Santos (2009): *Automatización de métodos para estimar emisiones de contaminantes y la altura de la capa límite atmosférica*, Cuba, Pinar del Río. Disponible en http://www.monografias.com/estimar-emisiones-altura-capa-limite2.html. [Consulta en línea 17 febrero 2010].

Turtós Carbonell, L. (2007): *Proyecto Programa Ramal Nuclear. Sistema de modelos Aermod para dispersión local de contaminantes atmosféricos. Salida 1/2007: Ampliación de la propuesta de Guía de modelación de la dispersión local de contaminantes gaseosos y partículas con el sistema de modelos Aermod*. Cuba, La Habana.

Witten, I. H. y E. Frank (2000): *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Estados Unidos, San Diego, Morgan Kaufmann Publishers.

Zerquera Díaz, R. (2009): *Predicción de parámetros meteorológicos secundarios: Altura de la capa de mezcla convectiva, altura de la capa de mezcla mecánica y velocidad convectiva de escala, en la zona de la Cujae, utilizando técnicas de Minaría de Datos*. Tesis (Ing.), Cuba, La Habana, Centro de Estudios de Ingeniería y Sistemas (CEIS), Instituto Superior Politécnico José Antonio Echeverría (CUJAE).