

Marfil
Una Herramienta de Software para la
Validación y Caracterización de
Datos Instantáneos

Marcelo F. Cagliolo - Martín Domínguez
Coordinador R. D. Sbarato
Municipalidad de la ciudad de Córdoba
Subsecretaría de Ambiente de la ciudad de Córdoba
Observatorio Ambiental
Laprida 854 - Barrio Observatorio - (5000) CORDOBA
fidel@obsambi.oac.uncor.edu

Resumen

Uno de los principales objetivos del Observatorio Ambiental de Córdoba es determinar la calidad del medio ambiente de la Ciudad de Córdoba. Esta tarea se realiza en tres etapas: Adquisición de datos, su validación y caracterización, y su posterior análisis para elaborar conclusiones con los mismos. *Marfil* es una herramienta de software que posibilita la validación y caracterización de los datos adquiridos, teniendo gran capacidad de almacenamiento de datos y brindando comodidades gráficas para el análisis de los mismos.

Introducción

Uno de los principales objetivos del Observatorio Ambiental de Córdoba es determinar la calidad del medio ambiente de la Ciudad de Córdoba. Para ello es necesario medir el nivel de cada contaminante en distintos puntos de la ciudad, y realizar el posterior análisis de estas mediciones. Pero antes de comenzar con su análisis, es necesario asegurar que las mediciones hayan sido válidas, es decir, que no hayan habido fallas en los equipos de medición y/o almacenamiento - transmisión. Por ésto, la determinación de la calidad del medio ambiente puede dividirse en tres etapas:

Adquisición de Datos

Para esto, el Observatorio Ambiental cuenta con dos *estaciones móviles de monitoreo automático* de los contaminantes criterio. Las estaciones adquieren los datos de cada equipo de medición situado en las mismas, para su posterior transmisión al Observatorio. Los datos son adquiridos cada un segundo por cada equipo de medición, por este motivo, son llamados *datos instantáneos*.

Validación y Caracterización de Datos

Una vez que los datos adquiridos por las estaciones móviles llegan al Observatorio, se debe verificar la *validez* de los mismos de acuerdo a criterios previamente establecidos, dejando como resultado la anulación de valores detectados como incorrectos. Además, si fuera necesario, los datos pueden ser *caracterizados*. En esta caracterización se agregan a los datos instantáneos nueva información, la cual especifica un suceso a destacar.

Análisis y Conclusiones

En esta etapa se trabaja con los *promedios horarios* (ver definición más adelante) de los datos instantáneos ya validados y caracterizados en la etapa anterior, los cuales se consideran los más representativos de los contaminantes. El análisis consiste en seguir el comportamiento temporal y espacial de cada contaminante, y la interferencia que hay entre ellos, como así también la incidencia de las variables meteorológicas y del factor humano debido a sus actividades en la sociedad. También se efectúa el pronóstico de contaminación en la ciudad para las próximas 24 hs.

En este trabajo específico nos concentraremos en la validación de datos.

Inicialmente, los datos que llegaban al Observatorio eran los *promedios horarios* de los datos instantáneos, los cuales se computaban en las estaciones móviles de monitoreo, ya que el análisis directo sobre los datos instantáneos era impracticable debido al gran volumen de datos adquiridos y a lo ineficientes que eran las pocas herramientas con que se contaba para tareas de este tipo. Entonces, la validación y caracterización se realizaba observando los promedios horarios. Si se

detectaba una irregularidad en algún promedio horario, entonces se pasaba al análisis de los datos instantáneos correspondientes a la hora en cuestión. Dicho procedimiento es evidentemente impreciso, ya que una falla en los datos instantáneos se refleja sobre su promedio horario cuando ésta es demasiado grave.

Toda corrección se realizaba entonces sobre los promedios horarios, adjuntando en cada una la especificación de qué tipo de corrección y criterio fue utilizado. Con esta manera de proceder no se deja registro alguno sobre los datos instantáneos de las irregularidades encontradas así como de las modificaciones y los criterios aplicados en el análisis de los promedios horarios, por lo tanto, el análisis posterior de los datos instantáneos era imposible sin el chequeo de los datos promediados, lo cual es una gran molestia.

Objetivo

Para solucionar las falencias anteriormente mencionadas, se incorporó en el Observatorio un sistema de transmisión de datos, el cual posibilita la recepción de los datos instantáneos diariamente. El objetivo de nuestro trabajo es entonces realizar una herramienta de software que pueda trabajar con esta gran cantidad de datos (instantáneos) de manera directa y diaria, permitiendo validarlos y caracterizarlos, para luego realizar los promedios horarios sobre los datos instantáneos ya tratados.

Desarrollo

El proceso de validación y caracterización de datos instantáneos se efectúa sobre los datos adquiridos durante todo un día. Por lo tanto, la herramienta a desarrollar deberá tener la capacidad de trabajar con todos los datos que involucra este período de tiempo. El volumen de datos que se maneja es entonces de 86.400 valores diarios para cada uno de los 22 contaminantes que se miden.

Ya que el Observatorio cuenta con computadoras que en su mayoría tienen sistema operativo Microsoft Windows, y ya que el personal está adecuado a este entorno, y además porque el Observatorio cuenta con el software Microsoft Access 97, decidimos realizar nuestro trabajo sobre este último, el cual soporta perfectamente la cantidad de datos requerida y brinda además un entorno gráfico muy conveniente para dicha tarea.

Desarrollamos entonces una herramienta de software llamada *Marfil*, la cual permite realizar el proceso de validación de datos instantáneos, como así también permite la caracterización de los mismos y el cómputo de los promedios horarios. Sus principales características son las siguientes:

Carga de los Datos Instantáneos

Los datos instantáneos que llegan al Observatorio se almacenan en archivos de texto. Marfil reconoce este formato y permite la carga de los mismos al sistema, hasta un intervalo total de 26 horas. Además, permite la carga de archivos de datos instantáneos con formato de planilla de cálculo y con formato propio, es decir, el formato con que Marfil almacena los datos. Una vez que se han cargado los datos instantáneos, comienza el proceso de análisis de los mismos, para lo cual, Marfil brinda un entorno gráfico en el cual se representan los valores de cada contaminante como una función sobre el tiempo. Se puede ver gráficamente a más de un contaminante a la vez, lo cual permite comparaciones entre tales. Cada contaminante se visualiza en una *ventana gráfica*.

Información sobre un Contaminante

Una ventana gráfica muestra el gráfico en ejes cartesianos de un determinado contaminante. El eje X (abscisa) representa el tiempo y el eje Y (ordenada) representa el valor del componente. La cantidad de valores para un determinado contaminante será típicamente los que comprenden un día entero, es decir, 84.600 valores (uno por cada segundo). Debe ser posible visualizar el intervalo total de tiempo en la pantalla, la cual tiene una resolución horizontal cuantitativamente menor (por ejemplo, de 1.024 píxeles). Por lo tanto, los valores de un contaminante deberán estar agrupados por cada elemento de resolución mínima, por ejemplo, si tomamos como resolución del eje X 1.024 puntos, entonces cada punto deberá representar a un conjunto de $84.600/1.024$ valores, es decir, cada punto sobre el eje X representará un intervalo de tiempo de $84.600/1.024$ segundos.

Este agrupamiento tiene la desventaja de que no se pueden distinguir los máximos y mínimos locales contenidos en cada grupo de valores, es decir, no se puede visualizar el comportamiento del contaminante dentro de un punto o elemento de resolución mínima.

Los datos instantáneos que son adquiridos por las estaciones de monitoreo pueden verse afectados por distintos factores técnicos, y como consecuencia de esto, es posible que haya intervalos de tiempo en donde los datos no sean tomados o sean perdidos durante su transmisión al observatorio. Esta *falta de datos* debe ser reflejada en el gráfico de un contaminante. Entonces, tiene que haber una forma de decir cuándo el intervalo de tiempo que representa un punto tiene falta de datos.

Además de los valores de los contaminantes, junto con los datos instantáneos hay información de sucesos preestablecidos que son propios de las estaciones de monitoreo, como por ejemplo, cuando se abre la puerta de la estación o cuando hay fallas en el suministro de energía. Estos sucesos son llamados *alarmas*. Las alarmas no dependen de un contaminante, son generales a todos, pero es necesario visualizar las alarmas junto con el gráfico de cada contaminante, ya que con esta información adicional se puede determinar o prever el comportamiento irregular de tales. Por lo tanto, tiene que haber una forma de ver las alarmas en el gráfico de un contaminante.

En Marfil resolvimos estos inconvenientes de la siguiente manera:

El gráfico de un contaminante tiene su eje X dividido en 300 elementos de resolución mínima, los cuales son barras, por lo tanto, cada barra representa a un intervalo de tiempo que abarca una trescientasava parte de lo comprendido por el intervalo de tiempo total graficado. La altura de una barra y su ubicación vertical respecto al eje Y nos dice el valor máximo y mínimo que alcanza el componente en el intervalo de tiempo que representa dicha barra. Además, las barras nos brindan mucha más información a través de otra dimensión, su color. Una barra tiene dos zonas en donde se colorea: el *borde* o contorno y el *relleno* o superficie. Cada color tiene un significado particular, dependiendo de los siguientes factores:

Cambios

Gama Amarillo - Rojo (color de relleno)

Si bien (como vimos anteriormente) es imposible visualizar los máximos y mínimos locales de un contaminante en una barra o elemento de resolución mínima, podemos sin embargo brindar cierta información que ayuda a conocer de alguna manera dicho comportamiento. Dentro del intervalo de tiempo que está comprendido por una barra pueden existir variaciones en el cambio de signo de la derivada de la función que representa al contaminante graficado. En Marfil estas variaciones serán llamadas *cambios*. Cada barra posee una cantidad determinada de cambios, por lo tanto, podemos clasificarlas según este factor. De este modo, en el gráfico de un contaminante, las barras que poseen la menor cantidad de cambios son pintadas con color amarillo puro en su relleno, mientras que las que poseen la mayor cantidad de cambios están pintadas con color rojo puro en su relleno. Los colores entre el amarillo y rojo son utilizados por consiguiente para pintar los rellenos de las barras cuyos cambios estén entre el mínimo y máximo total de cambios, dependiendo el tono de cuán cerca estén de éstos (mientras más cerca del mínimo más amarilla y mientras más cerca del máximo más roja).

El color rojo nos dice de cierta manera en qué zona del gráfico hay mayor ruido en los datos. Esto es relativo al intervalo de tiempo graficado ya que el color rojo no significa siempre que los datos son muy ruidosos ni el color amarillo significa siempre que los datos no tienen ruido,

simplemente nos indican en qué zonas del intervalo actualmente graficado hay mayor y menor cantidad de cambios en los valores del componente.

Falta de Datos

Azul (color de relleno/borde)

Como vimos anteriormente, es posible que dentro del intervalo de tiempo que se quiere graficar existan faltas de datos. El color azul en el borde o en el relleno de las barras nos alertan de la falta de datos. Cuando todo el intervalo horario que contiene una barra no posee datos para el contaminante que se está graficando dicha barra es pintada de azul en su relleno y abarca todo el eje Y graficado. Cuando una barra está pintada con azul en su borde significa que en el intervalo de tiempo que contiene esa barra existe al menos un punto (segundo) el cual no tiene datos para el contaminante graficado, pero que esta falta no es total, es decir, que también existe al menos un punto el cual sí posee datos.

Alarmas

Verde (color de borde)

Las alarmas son representadas en el gráfico de un contaminante con el color verde en el borde de las barras. Cuando una barra tiene su borde de color verde significa que dentro del intervalo de tiempo contenido por esa barra existe al menos un punto (segundo) en el cual hay una alarma.

Falta de Datos - Alarmas

Celeste (color de borde)

Existe un caso en el que los colores de las barras están en conflicto: cuando hay falta de datos no total en una barra (es decir, que la barra debe llevar borde azul), y a su vez dentro de esa barra hay alguna alarma (es decir, que la barra debe llevar borde verde). En este caso, se utiliza el color celeste para pintar el borde de la barra.

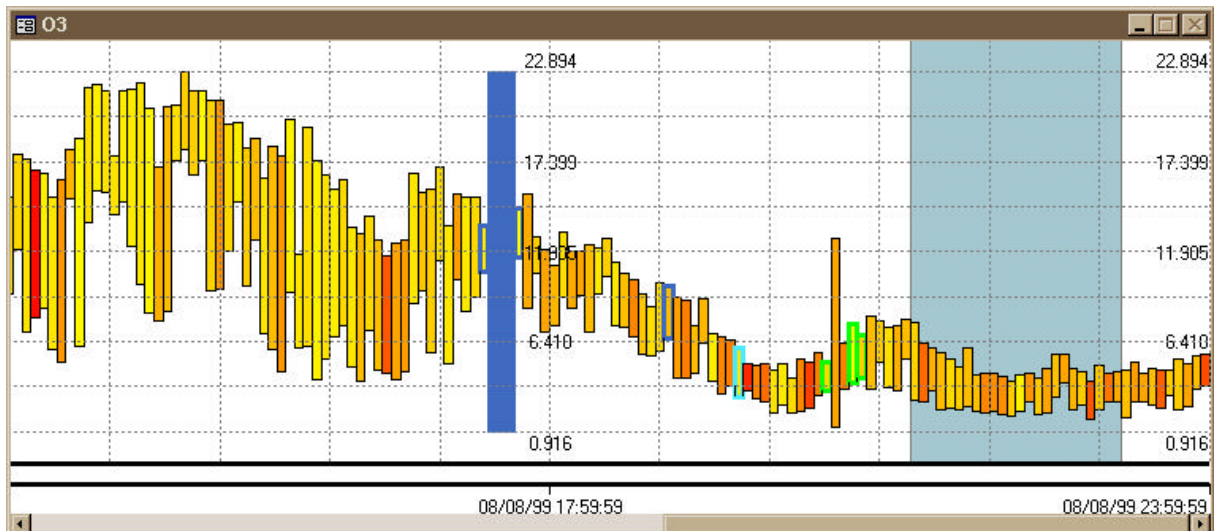


Figura 1. Ventana gráfica

En la figura 1 se muestra una ventana gráfica correspondiente al contaminante O_3 donde se pueden observar faltas de datos de ambos tipos, alarmas, la variedad en la cantidad de cambios de cada barra, la superposición de alarmas con falta de datos y la selección de un rango de tiempo.

Aparte de toda la información sobre un contaminante que nos ofrece una ventana gráfica, Marfil ofrece más información sobre un contaminante a través de la selección de un *rango de tiempo*. Esta selección se realiza directamente sobre el gráfico del contaminante (mediante el mouse), y se reconoce por el cambio de color (gris) del fondo del gráfico en la zona abarcada por dicho rango. Al seleccionar un rango de tiempo en un gráfico, el *Menú Principal* muestra información (en forma numérica) relativa al comportamiento del contaminante dentro de ese período de tiempo seleccionado. La información que se muestra es la siguiente:

- Fecha/hora inicial y final del rango de tiempo seleccionado.
- Valores máximo y mínimo que alcanza el contaminante en el rango de tiempo seleccionado.
- Cantidad total de cambios que se producen dentro del rango de tiempo seleccionado.
- Cantidad total de segundos que contiene el rango de tiempo seleccionado.
- Alarmas (en caso de haber) que se producen dentro del rango de tiempo seleccionado.

También es posible regraficar un contaminante de manera que el intervalo total graficado coincida con el rango de tiempo seleccionado, ésto se denomina aplicar *zoom*, y su efecto es el de ver con más precisión una zona determinada del contaminante. Toda la información que estaba disponible para el rango de tiempo total seguirá estando disponible ahora para el nuevo rango de tiempo graficado. Esto puede ser aplicado sucesivamente.

Validando y Caracterizando un Contaminante

El análisis de los datos se realiza utilizando toda la información anteriormente mencionada. La esencia de este proceso es detectar irregularidades en los valores de los contaminantes, para lo cual es necesario distinguirlos de los demás valores, y puede ser útil también dejar remarcado algún período de tiempo en donde se presume la influencia de un agente externo fuera de lo normal sobre los valores de un contaminante. Ambas cosas son posibles realizar en Marfil, mediante la inserción *propiedades*. Entonces, el proceso de validación y caracterización se realiza en Marfil adjuntando una lista de propiedades a cada uno de los contaminantes.

Para un contaminante dado, una propiedad es la especificación de un acontecimiento o una corrección que debería hacerse a los valores de dicho contaminante dentro de un determinado intervalo de tiempo. Por ejemplo, si consideramos el contaminante *Polvo en Suspensión*, podemos indicar con una propiedad que a una determinada hora del día los valores de dicho contaminante podrían haber sido afectados por una quema de neumáticos realizada en las proximidades al lugar de medición; también podemos indicar que en un determinado intervalo de tiempo los datos medidos no son válidos por lo cual habría que eliminarlos, debido a una falla en el suministro de energía del instrumento de medición. Existen cuatro tipos de propiedades:

- **Eliminar:** Se indica que todos los valores comprendidos en el intervalo de tiempo de la propiedad deberían ser eliminados.
- **Destacar:** Se indica algún acontecimiento que podría haber afectado los valores normales de muestreo dentro del intervalo de tiempo de la propiedad.
- **Desplazar - Escalar:** Se indica que debería aplicarse un desplazamiento y un reescalado de los valores comprendidos en el intervalo de tiempo de la propiedad.

- **Eliminar Selectivamente:** Se indica que deberían ser eliminados todos valores comprendidos en un determinado rango de valores, dentro del intervalo de tiempo de la propiedad.

Para agregar una propiedad a un contaminante, se debe seleccionar en su correspondiente gráfico el rango de tiempo al cual se le desea adjuntar dicha especificación, y luego desde el *Menú Principal* decir qué tipo de operación se quiere. Las propiedades que se le agregan a un contaminante no efectúan modificación alguna sobre los datos instantáneos, solamente conforman información adicional para cada contaminante. De esta manera, se mantiene siempre intacta la fuente original de datos, pudiendo entonces dejar documentado mediante las propiedades qué corrección es necesaria hacer y qué criterio se utilizó para tal determinación.

Promedios Horarios

Marfil permite también la realización de *promedios horarios* con los datos instantáneos. Los promedios horarios son una tabla de valores que muestra el valor promedio de cada contaminante sobre cada hora de datos instantáneos. Los promedios horarios pueden estar influidos si se desea por el conjunto de propiedades que se agregaron durante el proceso de validación y caracterización. En este caso, los promedios horarios se realizan aplicando primero todas las modificaciones sobre los datos que indican las propiedades, y luego efectuando los cálculos sobre el nuevo conjunto de datos instantáneos "filtrados", además, a cada valor de promedio se le asociará un identificador que especificará el tipo de propiedad que dicha hora contiene.

Resultados

Actualmente, en nuestro Observatorio Ambiental, se utiliza Marfil para realizar el proceso de validación y caracterización de los datos instantáneos adquiridos diariamente por las estaciones móviles de monitoreo. Se comprobó en la práctica su superioridad frente a otras posibilidades para el análisis de datos como por ejemplo, el uso de planillas de cálculo, ya que estas últimas no pueden brindar la información indispensable (falta de datos, alarmas y cambios) para el análisis de forma práctica, además de no permitir la realización de manera directa sobre el gráfico de un componente de acciones tales como agregar una propiedad, cálculo de máximos y mínimos, cambios y zoom sobre determinadas áreas del gráfico.

Conclusiones

Podemos concluir que Marfil es una muy buena herramienta para el análisis y posterior validación, caracterización y promedio de datos instantáneos ya que, gracias a la abundante información que ofrece sobre un contaminante, es posible realizar el análisis de los datos de manera muy completa y práctica.