Araya, José L., Alfaro, Eric J.
Prueba e implementación de algoritmos de control de calidad de datos de temperatura superficial del aire en un contexto operativo
Tecnología en Marcha, *Vol. 21, N.° 3, Julio-Setiembre 2008, P. 47-61*

Prueba e implementación de algoritmos de control de calidad de datos de temperatura superficial del aire en un contexto operativo

Fecha de recepción: 30/01/2008
Fecha de aceptación: 05/02/2008

José L. Araya[1]
Eric J. Alfaro[2]

La investigación realizada muestra que estos algoritmos son efectivos para la detección de valores atípicos que de otra manera podrían ser detectados tardíamente y pasar inadvertidos.

Palabras clave

Control de calidad, estación meteorológica automática, programación en grabador de datos, bases de datos, temperatura superficial del aire.

Key words

Quality control, ranges, automatic weather stations, Data Logger programming, databases, air surface temperature.

Resumen

Se presenta una metodología para el cálculo de rangos de temperatura, así como algoritmos de programación simple para la detección de errores obvios en datos meteorológicos con el fin de mostrar cómo un sistema de control de calidad sencillo, en tiempo real, puede implementarse de forma exitosa. Estos algoritmos fueron probados a través de su programación en un grabador de datos, el cual es el núcleo procesador en una estación meteorológica: primero, bajo condiciones controladas; y, luego, en dos estaciones meteorológicas automáticas con capacidad de transmisión en tiempo real. La investigación realizada muestra que estos algoritmos son efectivos para la detección de valores atípicos que de otra manera podrían ser detectados tardíamente y pasar inadvertidos.

Abstract

A methodology for the calculation of temperature ranges and easily programmable algorithms for detecting gross errors in meteorological measurements are suggested to show how a basic real-time quality control system can be successfully established. As a first approximation, these algorithms

1. Meteorólogo. Instituto Meteorológico Nacional. Tel. 2222-5616. Correo electrónico: *jlaraya@imn.ac.cr.*
2. Profesor e investigador. Escuela de Física, Centro de Investigaciones Geofísicas (CIGEFI) y Centro de Investigación en Ciencias del Mar y Limnología (CIMAR), Universidad de Costa Rica. Tel. 2207-5096. Correo electrónico: *ealfaro@cosmos.ucr.ac.cr.*

were implemented through Data Logger programming and tested under controlled conditions. Determination of ranges was made analysing data sets of two real-time automatic weather stations. The results of this research show that these algorithms can detect suspicious values in a straightforward manner, which avoids making such checks at a later stage and preventing outliers from going undetected.

Introducción

El Control de Calidad de Datos (CCD) es fundamental para la adecuada realización de investigaciones en meteorología. Además, el área de CCD es, por sí misma, un área de investigación consolidada y activa en meteorología y ha adquirido un auge sobresaliente en las últimas décadas debido al proceso de automatización, el cual ha implicado un incremento significativo del volumen de datos obtenidos (Araya, 2007). Incluso, con la ayuda de computadoras y paquetes de graficación es bastante difícil para el usuario examinar todo el conjunto de datos generados por una estación meteorológica automática y, por lo tanto, los errores más groseros podrían pasar inadvertidos, al menos que se definan procesos de monitoreo automatizados de datos para el Proceso de Control de Calidad de Datos (PCCD), la automatización de la toma de datos perfectamente puede causar una disminución en la calidad de los datos generados por la red (Brock y Richardson, 2001). Sin información sobre la calidad de los datos será imposible llegar a conclusiones serias en cualquiera de las áreas propias del quehacer meteorológico tales como climatología, meteorología sinóptica, aeronáutica, etc. Por ejemplo, una gran parte del quehacer científico en las investigaciones de cambio climático tiene que ver con la calidad de los datos ya que las tendencias climáticas son muy sensibles a valores erróneos o extremos generados por una variedad de circunstancias (Eischeid et al, 1995).

Por ejemplo, una gran parte del quehacer científico en las investigaciones de cambio climático tiene que ver con la calidad de los datos ya que las tendencias climáticas son muy sensibles a valores erróneos o extremos generados por una variedad de circunstancias (Eischeid et al, 1995).

Aunque la calidad del dato puede obedecer a más o menos exigencias de calidad dependiendo del usuario, lo óptimo es siempre perseguir la mejor confiabilidad posible.

Justificación de la presente investigación

En esta investigación se pretende, como objetivo general, demostrar la posibilidad técnica de implementar un PCCD en tiempo real que permita detectar valores sospechosos o anómalos que pueden ser generados por una estación meteorológica automática.

Particularmente, se pretende probar diferentes ACCD, además de generar metadatos "in situ" que sirvan como referencia al usuario para verificar la fiabilidad de los datos generados. En este estudio, la variable que será objeto de análisis es la temperatura superficial del aire, no obstante, las conclusiones de esta investigación arrojarán luz en cuanto a la implementación de este tipo de pruebas en otras variables de interés, tales como humedad relativa, radiación, presión atmosférica, velocidad y dirección del viento.

Diseño de experimentos y metodología de validación de los algoritmos

Los emplazamientos elegidos para las pruebas de los métodos de control de calidad fueron las localidades de Manzanillo y Pavas, Costa Rica. Se decidió probar los algoritmos del PCCD a estas estaciones debido a que ofrecen la posibilidad de detectar cualquier inconsistencia en tiempo real o casi real.

Para la programación de los ACCD se utilizó el lenguaje propio de los GD Campbell Sci., con el cual se establecen diferentes especificaciones referentes a la forma, regularidad y cálculos para la generación

de los datos; es decir, características tales como intervalos de ejecución, cálculos de valores extremos, promedios, tablas de datos, entre otras (Campbell Sci, 2003a). El paquete de cómputo utilizado para estos GD, proporcionado por el fabricante, es LoggerNet, el cual incluye el compilador Edlog para el desarrollo de tales programas. Entre las más importantes bondades de Edlog se incluyen (Campbell Sci., 2005):

- Visualización de datos u otros archivos en ASCII.
- Edición y búsqueda de errores en el programa.
- Edición de localidades de almacenamiento de datos y de localidades de entrada.
- Detección y ubicación de errores de compilación.
- Tres tablas de programación, dos para aquellos casos en los que se requiera programar uno o varios sensores con un intervalo de ejecución diferente y una tercera para la inclusión de subrutinas.

Un programa desarrollado en Edlog contiene instrucciones de programación que se clasifican de la siguiente manera (Campbell Sci, 2003b):

- **Instrucciones de entrada /salida:** Son instrucciones especiales para configuración de los sensores utilizados, en las cuales se incluyen las constantes de calibración del instrumento, voltajes de entrada o de excitación aplicados.
- **Instrucciones de proceso:** estas instrucciones permiten la programación de algoritmos y la conversión de los voltajes o pulsos registrados por los sensores a unidades físicas.
- **Instrucciones de control:** Habilita las salidas de datos en cierto orden, una vez cumplida alguna condición lógica, así como el inicio de telecomunicaciones en el GD.
- **Instrucciones de salida:** Procesa los valores medidos colectados en tiempo y espacio para su almacenaje final en la memoria del GD en forma de tablas de datos.

El paquete de cómputo también incluye otras ventanas para configuración, diseño de pantallas de monitoreo en tiempo real, panel de datos que permiten generar gráficos rápidos del archivo de datos generado por la estación automática y un programa que permite generar reportes y manipular estos archivos de datos (Campbell Sci., 2004c).

Por ejemplo, para la visualización de los indicadores de alerta generados por la EMA se procedió a elaborar una pantalla de visualización de los datos en tiempo real. La Figura 1 muestra el diseño de dicha pantalla, la cual fue realizada de tal manera que, en el momento de activarse uno de los IA, la computadora genera una alarma sonora y visual (Araya, 2007). Esta pantalla de visualización de datos en tiempo real fue diseñada, específicamente, para esta investigación y muestra como las bondades del paquete de cómputo LoggerNet pueden usarse eficazmente en la labor operativa del PCCD. En general, el lenguaje de programación Edlog también

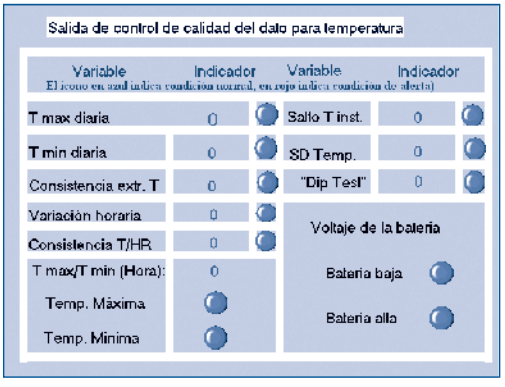

Figura 1. pantalla de visualización de datos en tiempo real para vigilancia de los indicadores de alerta incorporados en el programa de la Estación Meteorológica Automátic.

mostró ser lo suficientemente versátil como para la programación de los ACCD.

Para la programación y prueba de los algoritmos de temperatura empleados se utilizaron los laboratorios de la Gestión del Dato del Instituto Meteorológico Nacional. Las etapas de validación de los algoritmos fueron:

- **Pruebas heurísticas de consistencia**: se realizaron pruebas de laboratorio en las cuales se inducían situaciones para que los ACCD las detectaran. Esto se efectuó de diversas formas. En el caso de los algoritmos para temperatura en superficie se varió la temperatura ambiente, los parámetros de los diferentes ACCD y la hora en el Grabador de Datos (GD). Es importante recalcar que este tipo de pruebas son parte del quehacer operativo de la Gestión del Dato del IMN ya que es necesario probar la consistencia de los diversos algoritmos antes de hacerlos operativos. Con estas pruebas de laboratorio se generaron archivos de datos que fueron, posteriormente, analizados para determinar si los IA operaban correctamente. Esta fase permitió, además, la determinación de los requerimientos de memoria de tales algoritmos dentro del GD.
- **Prueba de los algoritmos en condiciones reales**: Instalación de un programa con formato de EMA climática en la estación de Manzanillo con el fin de probar los algoritmos de temperatura en condiciones reales. Instalación de un programa con formato de EMA sinóptica en Pavas con el fin de hacer operativos los algoritmos de temperatura. En esta etapa algunos de los parámetros se definieron de manera muy general y fueron sometidos a modificación conforme se acumulaba nueva experiencia en torno al sistema. El Cuadro 1 muestra los detalles de ubicación de las estaciones utilizadas, así como la longitud de las series de datos con las cuales se trabajó.

Dependiendo del algoritmo analizado y de los rangos utilizados, en ocasiones, los únicos datos con los que se pudo contar en esta investigación corresponden a datos de laboratorio. No necesariamente en todos los casos se presentó una situación real de detección de valores sospechosos durante el periodo en que se hicieron operativos los ACCD.

Dependiendo del algoritmo analizado y de los rangos utilizados, en ocasiones, los únicos datos con los que se pudo contar en esta investigación corresponden a datos de laboratorio. No necesariamente en todos los casos se presentó una situación real de detección de valores sospechosos durante el periodo en que se hicieron operativos los ACCD. Algunos casos reales de abanderamiento serán mostrados más adelante como casos de estudio. Para mayor detalle sobre los ACCD implementados, consúltese Araya (2007).

ACCD implementados

Prueba de salto para Valores Instantáneos (PSVI)

Un salto se define como el valor absoluto de la diferencia entre lecturas consecutivas que sobrepasa cierto valor límite previamente definido. En este caso, se consideran dos mediciones instantáneas

Cuadro 1. Localización de estaciones meteorológicas utilizadas y longitud del registro de datos utilizado.

Emplazamiento	Latitud	Longitud	Altitud (m)	Longitud de la serie de datos	
				Inicio	Final
Manzanillo	09º38’	82º39’	5	27/02/2005	12/12/2005
Pavas	09º51’	84º08’	997	15/08/1995	14/08/2005

consecutivas (a intervalos de 10s). Después de cada medición de una señal en el Detector Térmico Resistivo (DTR) la muestra actual debe ser comparada con la precedente. Si la diferencia entre estas dos muestras sobrepasa el valor límite especificado, entonces la muestra actual es identificada como sospechosa (Zahumenski, 2005). Dado que la menor resolución del archivo de datos reportado por las EMA del IMN es en resolución horaria, se programó el algoritmo de modo que dé información concerniente a los saltos instantáneos en esta resolución, a pesar de que el tiempo de muestreo entre mediciones consecutivas es de diez segundos, cada sesenta minutos el IA reporta si los valores instantáneos medidos por el GD, en esa hora pasaron con éxito la prueba de control. Esto se logró estableciendo un contador a nivel de programa que determina el número de saltos que son mayores a 0.1ºC, una vez que este contador excede la tercera parte del número total posible de saltos en una hora (360 en total) entonces un IA mostrará un valor no nulo para señalar que dicha condición se presentó. La Figura 2 muestra un conjunto de datos generado en pruebas de laboratorio en 10s de resolución donde se aprecia cómo el IA abandera una secuencia de saltos elevada inducida de forma artificial sobre el DTR.

La Figura 2 muestra un conjunto de datos generado en pruebas de laboratorio en 10s de resolución donde se aprecia cómo el IA abandera una secuencia de saltos elevada inducida de forma artificial sobre el DTR.

Pruebas de Rango en resolución Horaria (PRH)

Este algoritmo se implementó con el objeto de señalar como sospechosos los promedios horarios de temperatura fuera de rango según la siguiente expresión:

$$T_{\min} \leq T \leq T_{\max} \quad (1)$$

donde T es el promedio horario de temperatura, $T_{\min}$ y $T_{\max}$ son los límites inferior y superior previamente determinados. Lo que se hace es calcular el promedio horario de temperatura y compararlo con los rangos previamente calculados. Si el promedio está sobre el límite máximo de temperatura un IA mostrará una constante positiva; si el promedio se encuentra dentro de un rango aceptable el IA reportará un valor nulo; en tanto que si el promedio está debajo del límite inferior entonces el IA reportará una constante negativa.

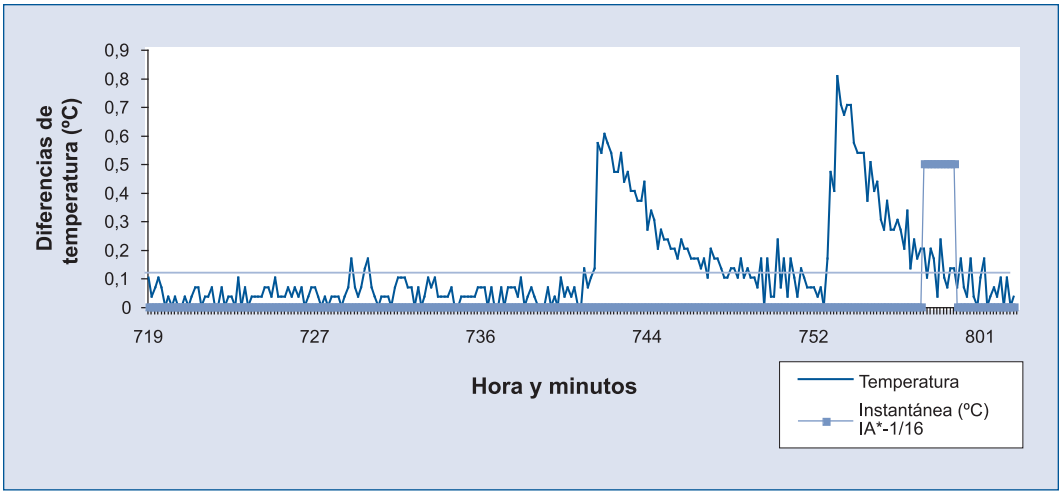

Figura 2. Serie de tiempo para diferencias absolutas de valores instantáneos de temperatura.

Prueba de detección de picos (PDP)

El objetivo de este algoritmo es encontrar valores atípicos para cualquier parámetro geofísico que presente una distribución continua, tal como la temperatura, humedad relativa y la velocidad del viento (Vejen et al., 2002) y ya ha sido utilizado en el Instituto Nacional de Meteorología de Dinamarca para analizar datos con resolución horaria. Otras pruebas similares a estas han sido analizadas, también, por Graybeal et al. (2004).

Sea un $\delta \in \Re^+$ que depende del parámetro *x(t)* cuyos valores atípicos se desean abanderar. El algoritmo procede a abanderar una observación $x_i = x_i\,(t)$ tal que cumpla con la relación:

$$(x_{i-1} - x_i)(x_{i+1} - x_i) \geq \delta^2 \quad . \quad (2)$$

Desde el punto de vista de programación se calculó el promedio de temperatura de dos horas consecutivas y, con ello, el valor absoluto de la diferencia entre ellas, la cual se compara con un valor previamente calculado con base en la serie de datos del emplazamiento.

Esta prueba está relacionada con la curva implícita, la cual corresponde a una hipérbole.

Nótese que si se considera una prueba *T* particular del tipo donde la observación x_i se rechaza debido a que es mayor que alguna constante δ que no depende de x_i , entonces el área :

$$\Omega = \{(x, y): T(x, y) > \delta\} \quad (3)$$

se define como el área de rechazo de la prueba.

Obsérvese que para que la ecuación (2) sea válida ambos términos de la parte izquierda de la igualdad deben ser positivos o negativos. Es así como este algoritmo lo que hace es detectar puntos de subida o de bajada anormalmente altos comparando el promedio horario de temperatura central con el de la hora anterior y la hora posterior. Según Ogland (1993) la PDP se comporta de varias formas posibles: en la primera de ellas, si cada salto es de magnitud moderada, mayor que δ^2, y en dirección opuesta, el valor es indicado como sospechoso; en el segundo caso, se encuentra que si los saltos son en direcciones opuestas, y uno de ellos es mucho mayor que el otro de modo que el producto es mayor que δ^2, entonces el valor es indicado como sospechoso. Esto implica que si uno de los saltos es muy grande en magnitud y el otro suficientemente pequeño, entonces el valor podría no ser indicado como sospechoso.

La PDP depende de series de datos completos, por lo que presentarán problemas cuando se tenga datos faltantes, por lo que puede requerirse para su implementación interpolar valores con el fin de hacerlo aplicable.

Prueba de Salto con resolución Horaria (PSH)

Esta prueba tiene que ver con la diferencia absoluta entre dos promedios de horarios de temperatura consecutivos. La PSH viene dada por la expresión:

$$|T_i - T_{i-1}| \geq \zeta \quad (4)$$

donde ζ corresponde al valor umbral de la magnitud de la diferencia entre las horas consecutivas T_i y T_{i-1} por encima del cual tales promedios horarios de temperatura serán abanderados (Reek et al., 1992).

Desde el punto de vista de programación se calculó el promedio de temperatura de dos horas consecutivas y, con ello, el valor absoluto de la diferencia entre ellas, la cual se compara con un valor previamente calculado con base en la serie de datos del emplazamiento. Si la diferencia absoluta entre promedios horarios consecutivos sobrepasa el valor límite, entonces el IA del algoritmo generará un valor no nulo para indicar qué situación se presentó; de lo contrario, permanecerá nulo.

Prueba de Consistencia entre Promedios Horarios de

Temperatura y Humedad Relativa (PCTHR)

La idea de implementar este tipo de prueba está relacionada con el hecho de que, como parte del control manual de datos en el IMN, se acostumbra revisar la variación de la temperatura con la humedad relativa en el transcurso del día con el fin de detectar posibles errores evidentes en los datos. El algoritmo aquí propuesto efectúa dicha prueba de forma automática, de forma que no es necesario inspeccionar visualmente cuando se dé una razón de cambio positiva de la temperatura con la humedad relativa.

Se utilizó el registro climatológico de los valores extremos de temperatura y sus horas y se procedió a agrupar por clases las horas a las cuales tienden a darse los valores extremos.

Si bien este algoritmo supone que la tensión de vapor se mantiene constante (lo que no es siempre cierto) ofrece información relevante con respecto a la frecuencia con que se dan razones de cambio positivas entre humedad relativa y temperatura al nivel de medición de ambos parámetros, lo que permite identificar algunos patrones generales en su recurrencia de forma automática, como por ejemplo, las horas más frecuentes a las que este tipo de casos se presentan. Esto puede ser útil para quien revisa los datos ya que una frecuencia muy grande de estos IA podría brindar información interesante con respecto a factores locales, de pequeña escala, de mesoescala o escala sinóptica que, quizás, incidieron en el emplazamiento de la EMA. Sea la medición $T_i = T_i\,(t)$ el promedio de la temperatura media horaria y $U_i = U_i\,(t)$ la observación de la humedad relativa horaria a la misma hora t en la que se efectuó la observación de la temperatura. Los datos que cumplan con la relación:

$$\frac{T_i - T_{i-1}}{U_i - U_{i-1}} \geq \varepsilon \quad (5)$$

son abanderados. Aquí $\varepsilon \geq 0$ es el valor umbral de abanderamiento del dato, no se consideran los casos con denominador nulo. Para efectos de verificación del algoritmo se asumirá $\varepsilon = 0$, suponiendo que la tensión de vapor es constante.

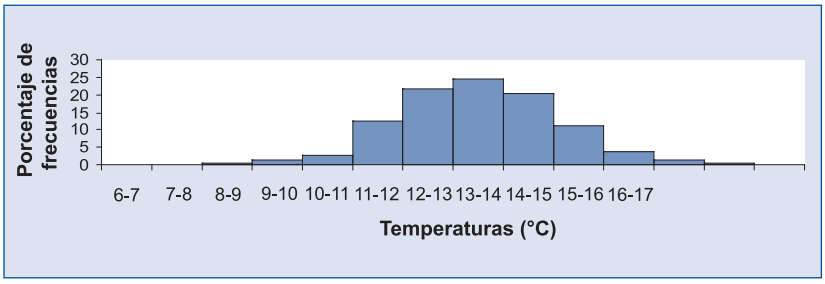

Figura 3. Histograma de frecuencias relativas para las horas de las máximas diarias de temperatura en Pavas del 14 de setiembre de 1995, al 14 de marzo del 2003.

Prueba de Rango de las Horas de temperaturas Extremas diarias (PRHE)

Esta prueba se efectuó sobre los valores absolutos extremos diarios de temperatura, y se basa en determinar las horas en las cuales tiene sentido que se den la máxima y la mínima absolutas de temperatura diaria. Para definir los rangos, se utilizó el registro climatológico de los valores extremos de temperatura y sus horas y se procedió a agrupar por clases las horas a las cuales tienden a darse los valores extremos (Araya, 2007).

Se utilizó el registro climatológico de los valores extremos de temperatura y sus horas y se procedió a agrupar por clases las horas a las cuales tienden a darse los valores extremos. Las Figuras de la 3 a la 6 muestran los histogramas con un intervalo de clase de un grado centígrado para el caso de las estaciones de Pavas y Manzanillo. A partir del registro de datos de ambas estaciones se eligió el rango de horas que incluyera, aproximadamente, el 95% de los casos. Este es un ejemplo de cómo se tomó una decisión con respecto a los rangos utilizados en los ACCD implementados.

Prueba de Consistencia entre la temperatura máxima diaria y la temperatura mínima diaria (PCE)

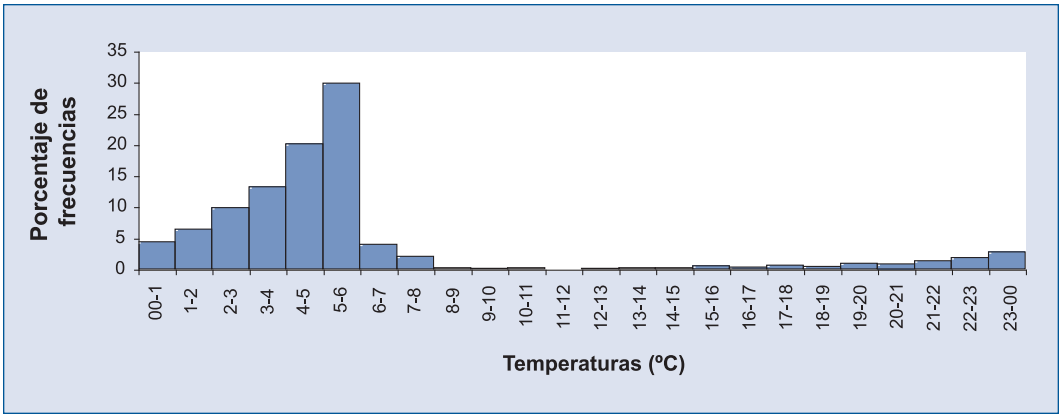

Figura 4. Histograma de frecuencias relativas para las horas de las minimas diarias de temperatura en Pavas del 14 de setiembre de 1995, al 14 de marzo del 2003.

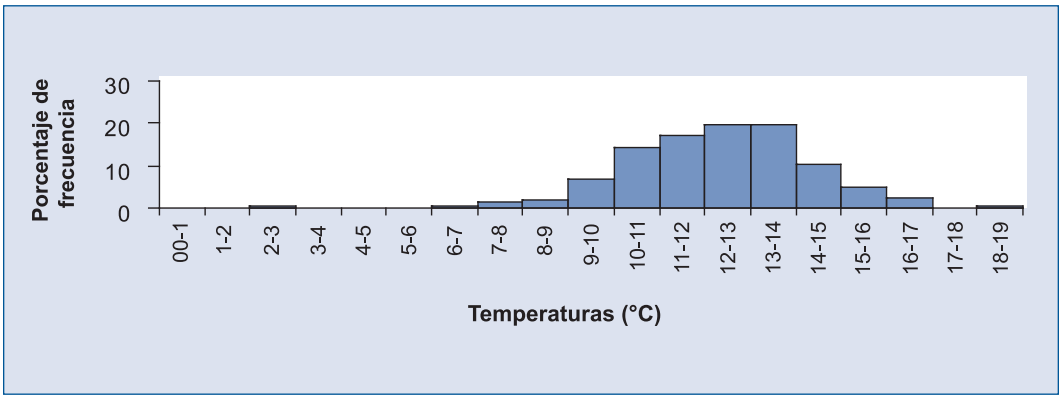

Figura 5. Histograma de frecuencias relativas para las horas de las temperaturas diarias máximas en Manzanillo del 02 de febrero, al 12 de diciembre 2005.

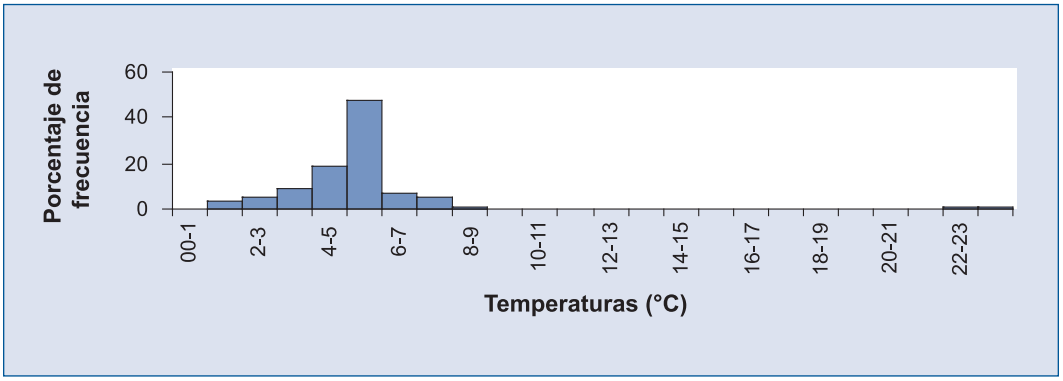

Figura 6. Histograma de frecuencias relativas para las horas de las temperaturas diarias mínimas en Manzanillo del 02 de febrero, al 12 de diciembre 2005.

La PCE tiene como objeto alertar acerca de la existencia de inconsistencias entre las extremas diarias de temperatura. Se decidió indicar aquellos valores atípicos en los que la temperatura máxima diaria pudiera ser igual o inferior a la temperatura mínima diaria, por lo que se acepta que las extremas diarias son consistentes si se cumple la relación:

Parte de la problemática con un sistema automatizado en el PCCD es determinar cuántos datos se desean revisar ya que puede ser difícil desde el punto de vista de disponibilidad de recursos humanos y económicos explicar una cantidad muy grande de valores individuales.

$$T_{max} \geq T_{min} \text{ , (6)}$$

donde T_{max} corresponde a la temperatura máxima absoluta diaria y T_{min} es la temperatura mínima absoluta diaria. Nótese que (6) indica que en caso en que las extremas de temperatura diaria sean iguales, la PCE debe indicar la situación como atípica.

Prueba de Desviación Estándar de la temperatura promedio horaria (PDE)

El cálculo de la desviación estándar para temperatura es usado como un criterio importante de decisión para la aceptación o rechazo del promedio horario de temperatura ya que, si la desviación estándar del parámetro se encuentra debajo de cierto mínimo aceptable, los correspondientes datos deberían ser indicados como sospechosos o bien como erróneos, a partir del momento en que se da el primer valor por debajo del umbral mínimo aceptado (Shafer et al., 2000). Este valor es de gran utilidad para la detección de un sensor bloqueado así como para la detección de histéresis a largo plazo (Zahumensky, 2004). Siguiendo estos lineamientos se desarrolló un algoritmo para la detección de valores atípicos en la desviación estándar horaria. Sea σ_{min} el umbral de desviación estándar tolerable para una variable en particular, mientras que σ_{max} es el límite superior de aceptación para la desviación estándar. Se aceptará un promedio horario para la temperatura en los casos en que se cumpla la siguiente relación:

$$\sigma_{min} < \sigma \leq \sigma_{max} \quad (7)$$

En el caso de la PDE se decidió explicar los valores de desviación estándar que se salieran de rango, con particular atención a los valores nocturnos muy altos ya que los gradientes de temperatura son menores durante la noche con respecto a los de las horas del día.

Resultados: Decisiones con respecto a los valores atípicos obtenidos

El objetivo de los ACCD aquí sugeridos es proponer criterios de decisión con el fin de identificar valores sospechosos o condiciones especiales en los datos para su posterior validación, por lo que el abanderamiento de un dato por parte de estos algoritmos no significa necesariamente que el dato es erróneo, sino que señala situaciones que, dependiendo del caso de análisis pueden ser posibles o no. Por esta razón, tales valores atípicos deben ser cuidadosamente estudiados en la OPD utilizando experiencia y criterio experto con el fin de determinar si el dato puede considerarse correcto, aceptable, probablemente correcto, probablemente erróneo o inconsistente del todo, dependiendo de cuán sofisticado sea el sistema de control de calidad del dato aplicado. Parte de la problemática con un sistema automatizado en el PCCD es determinar cuántos datos se desean revisar ya que puede ser difícil desde el punto de vista de disponibilidad de recursos humanos y económicos explicar una cantidad muy grande de valores individuales. Por ello se determinó que lo mejor era explicar solo un porcentaje operativamente manejable. Los criterios de elección de la cantidad de abanderamientos atípicos a explicar según los ACCD fueron los siguientes:

- Para la PSVI se decidió explicar todas las posibles alertas generadas. En total, el IA correspondiente a este algoritmo indicó dos datos sospechosos. Esta prueba mostró que había condiciones de inestabilidad entre lecturas consecutivas instantáneas de temperatura.
- Para la PRH se decidió explicar todas las alertas que el ACCD pudiera generar.
- Para la PDP, la PSH y la PDE se analizó el 0,1% de los datos atípicos detectados (aunque la PSH abanderaba el 5% de los saltos anormalmente altos, debido a que se determinó que, operativamente, el tiempo para la explicación de ese porcentaje es alto). Los criterios de decisión para el rechazo o aceptación de los datos abanderados por la PSH fueron los

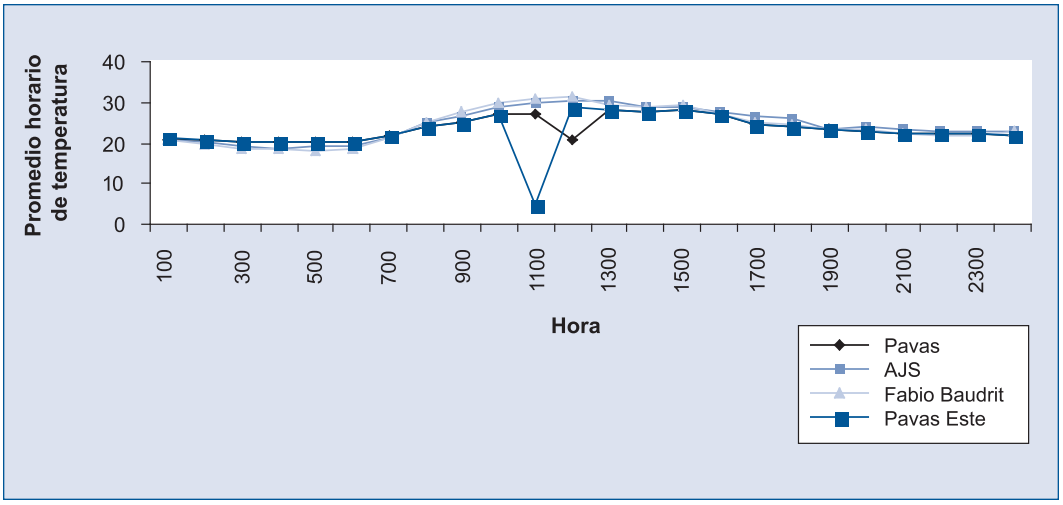

Figura 7. Comportamiento diurno de los promedios horarios de temperatura para estaciones cercanas a Pavas, el 21 de abril del 2006.

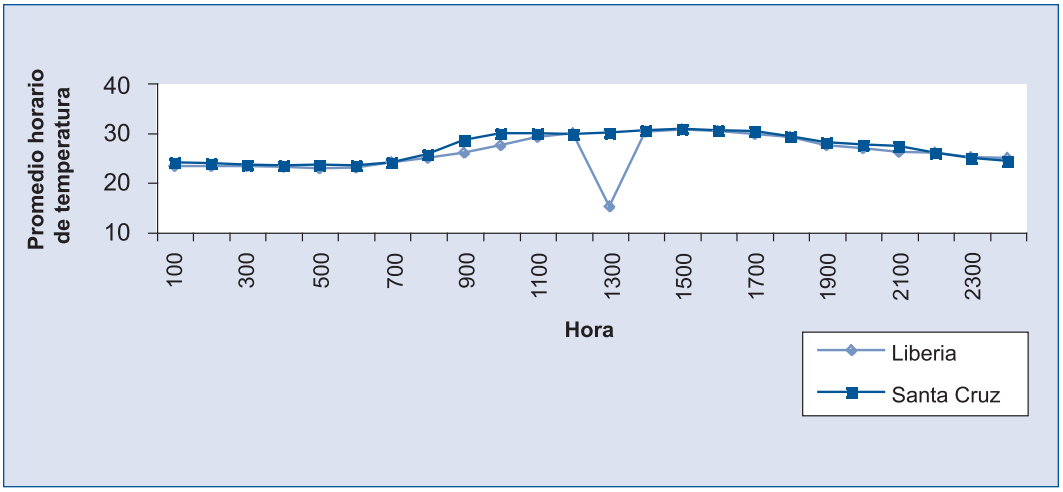

Figura 8. Valor atípico detectado en la serie de datos de Liberia, el 3 de junio de 1999.

Cuadro 2. Valores atípicos erróneos para las horas de las extremas diarias de temperatura detectadas en San José.

Fecha	16/07/2005				17/07/2006			
EMA	San José	Pavas	Santamaría	Cigefi	San José	Pavas	Santamaría	Cigefi
Temperaturas máxima (ºC)	**26,7**	26.69	29.93	25,6	**28,3**	26	27,8	27,7
Horas temperaturas máximas	**21:34h**	9:23h	10.48h	12:16h	**21:41h**	9:32h	12:53h	10:33h

mismos que los utilizados en la PDP. Debido a que el algoritmo abanderó el 5% de los saltos de mayor magnitud la cantidad de alertas generadas fue alta si se compara con el número de valores indicados con la PDP. Esto demuestra que en la labor operativa se debe llegar a un equilibrio entre el número de datos atípicos que se deseen explicar y el número de horas que el personal está en capacidad dedicar a la explicación de los mismos, de allí que se consideró revisar solo las diferencias superiores a 4,7 ºC para Manzanillo y 4,6 ºC para el caso de Pavas ya que, por encima de estos valores, solo se encuentran el 0,1% de los saltos.

- El procedimiento seguido con los datos marcados por la PCTHR fue explicar solo las secuencias consecutivas de razones de cambio positivas de temperatura y humedad relativa de al menos tres horas que se dieran durante el día ya que el IA asociado a esta prueba genera un número alto de alertas. En total, se encontraron ocho secuencias diurnas de, al menos, tres horas consecutivas, de las cuales solo una mostró ser un valor erróneo y fue detectado por otras pruebas. Quedó claro que una alerta generada por este algoritmo no tiene el grado de severidad que se le puede dar a las otras pruebas por lo que una alerta debe interpretarse en el contexto de la frecuencia de estas y las horas del día en las que se presentaron (véase Araya, 2007 para más detalles).
- En el caso de la PDE se decidió explicar los valores de desviación estándar que se salieran de rango, con particular atención a los valores nocturnos muy altos ya que los gradientes de temperatura son menores durante la noche con respecto a los de las horas del día.
- Para las pruebas con resolución diaria (PRHE y PCE) se determinó analizar todos los posibles datos atípicos generados. Solo se detectó un valor sospechoso que fue aceptado como correcto. Para la PCE no se detectó ningún valor sospechoso.
- Se consideró a los ACCD de temperatura con resolución horaria como una metodología unificada de análisis; en caso de que un dato fuera indicado como sospechoso también se consideraba si los otros ACCD

Si bien el objetivo de los ACCD analizados aquí es ser implementados en tiempo real, también pueden ser aplicados para analizar secuencias largas de datos en busca de valores atípicos.

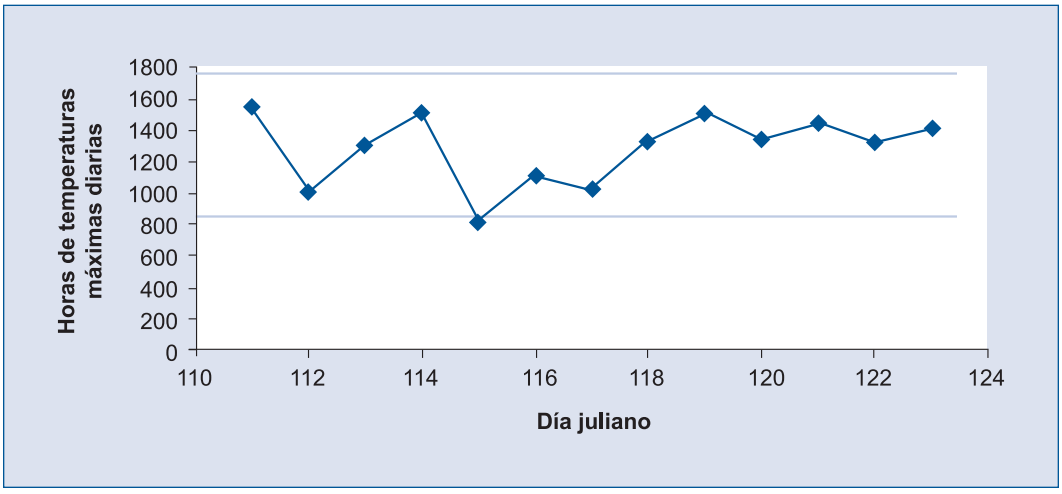

Figura 9. Hora de temperatura máxima atípica detectada en Manzanillo, 25 de abril del 2006.

también abanderaban el dato. Puede consultarse Araya (2007) para más detalles en cuanto a ejemplos concretos en los cuales esta metodología de control de calidad tuvo éxito en la detección de valores atípicos. En este artículo solo se expondrán dos de ellos a manera de ejemplos. Las Figuras 7, 8 y 9, así como el Cuadro 2, son ejemplos de casos detectados por esta metodología de control de calidad en Pavas (en este caso a través de algoritmos programados en tiempo real) y Liberia.

Los diversos algoritmos aplicados en servicios meteorológicos y afines de otras latitudes pueden implementarse en Costa Rica, siempre y cuando se definan adecuadamente rangos nuevos, especialmente para aquellos que son calculados tomando como referencia la climatología del emplazamiento.

La Figura 7 muestra la secuencia del promedio horario de temperatura para otras estaciones cercanas, incluida la estación meteorológica ubicada en Pavas, así como otra estación conocida como "Pavas Este", la cual se encuentra ubicada cerca del mismo emplazamiento de la EMA Pavas, usada en esta investigación. Dado que el dato en cuestión no pasó prácticamente ninguna de las pruebas se decidió investigar su validez.

Nótese en la Figura 7 que el valor atípico fue el de las 12:00h, si bien los IA de los correspondientes algoritmos para este dato fueron generados hasta las 13.00h, esto debido a la interdependencia entre datos horarios consecutivos que los diferentes ACCD presentan. Obsérvese cómo este descenso abrupto en el promedio horario de temperatura en Pavas es inexplicable en términos del comportamiento en las estaciones vecinas Juan Santamaría (latitud 9º 59', longitud 84º 9') y Fabio Baudrit (latitud 10º 0', longitud 84º 16'), por lo que se trata de un dato dudoso. Comparando con Pavas Este se concluye que el pico atípico esta relacionado con labores de reemplazo de los DTR en ambas estaciones, ya que a las horas en que se registró estos picos se reportó labores de mantenimiento en emplazamiento. El valor tan bajo detectado en temperatura se debe al hecho de que cuando los canales a los que el DTR está conectado quedan libres y se generan valores espurios que son interpretados y promediados como datos de temperatura, y que provocan los picos observados. Nótese que las dos EMA en Pavas detectó un valor atípico con una diferencia de una hora. En este caso se llegó a la conclusión que el dato es erróneo.

Los IA generados fueron posteriormente investigados con el objetivo de validar los datos. Es importante recalcar que los ACCD solo señalan situaciones que pueden ser errores o que pueden ser situaciones meteorológicas muy especiales; es decir, raramente vistas o bien eventos extremos.

Si bien el objetivo de los ACCD analizados aquí es ser implementados en tiempo real, también pueden ser aplicados para analizar secuencias largas de datos en busca de valores atípicos. Para ampliar lo anterior, la Figura 8 muestra el caso de un valor atípico erróneo detectado en una serie de datos obtenida en Liberia (Latitud 10º 35', longitud 85º 32'). Nótese cómo después de aplicar la PDP y la PSH el valor atípico destaca en la secuencia de datos. Nótese en la Figura 8 la variación grande del salto (15 ºC). Tras comparar este dato con el de una estación cercana quedó claro que el campo de temperaturas de la región del Pacífico Norte no experimentó variaciones de temperaturas como la indicada en Liberia a las 13:00 h, por lo que se trataba de un dato erróneo. En este caso, en particular, se encontró información muy valiosa en el expediente de informes de campo para la EMA de Liberia. Es interesante, en este caso, que el técnico estuviera efectuando trabajos en emplazamiento entre las 11:00 am y las 13:35 pm, por lo que este valor atípico está relacionado con las labores de mantenimiento que se efectuaban en ese periodo del día, y que indujeron condiciones especiales en la lectura del DTR que se manifestaron en forma del valor atípico aquí detectado.

La Figura 9 muestra el caso de un dato atípico detectado en la EMA de Manzanillo

usando la PRHE. El único valor que sobrepasa el límite inferior establecido para las horas de temperatura máxima diaria corresponde al valor señalado como sospechoso y se observa que el umbral para las horas de temperaturas máximas fue sobrepasado en este caso. En principio una temperatura máxima a las 8:13h es posible, ya que la PCHE fue programada dentro del GD de modo que se marcaran alrededor del 5% de las horas atípicas. Se comparó este valor con el registrado por la EMA de Limón y la hora a la que se dio la máxima fue exactamente la misma, de modo que, al menos, una estación cercana registró la misma hora para este caso. También se comparó con el registro de promedios horarios de temperatura para las mismas estaciones y se observó que el promedio horario máximo se presentó alrededor de la misma hora, por lo que se aceptó este dato como correcto.

La meta información producida por los IA sería de gran utilidad para cierto tipo de usuarios especializados para diferentes estudios, y les indicará con un criterio objetivo que tan buenos son los datos que pretenden utilizar para sus fines.

Aunque en el caso anterior la hora a la que se dio el valor extremo es creíble, es posible encontrar casos en los que este algoritmo señala horas erróneas. Con el fin mostrar otro caso en el que este ACCD puede ser muy útil, el Cuadro 2 presenta las horas de las máximas absolutas de temperatura en San José, Costa Rica, los días 16 y 17 de julio de 2005 y 2006 respectivamente, así como la hora a la que estaciones cercanas ubicadas en reportaron dicho valor. Nótese que es poco probable encontrar una máxima absoluta en temperatura entre las 21:00h y las 22:00h, además en las otras estaciones cercanas el rango de la hora para el reporte de la máxima fue entre 9:00h y 13:00h, lo que acumula evidencia en contra de la plausibilidad de dicho valor en San José. En este caso se decide que los datos son, probablemente, incorrectos, por lo que tal información debería registrarse de forma adecuada a través de un IA final para conocimiento de los eventuales usuarios de estos datos.

Conclusiones

Analizando los resultados obtenidos se concluye que se cumplió el objetivo geneya que todos los ACCD fueron debidamente probados en el laboratorio. Además, los ACCD para temperatura fueron implementados en dos de las EMA de la red de estaciones automáticas del IMN exitosamente, indicaron valores atípicos en tiempo real y demostraron ser una herramienta sistemática y objetiva de evaluación de los datos generados por una EMA, detectando anomalías en los datos desde una perspectiva física y estadística.

Específicamente se concluye que:

- Los algoritmos para el PCCD pueden implementarse en tiempo real.
- Los diversos algoritmos aplicados en servicios meteorológicos y afines de otras latitudes pueden implementarse en Costa Rica, siempre y cuando se definan adecuadamente rangos nuevos, especialmente para aquellos que son calculados tomando como referencia la climatología del emplazamiento.
- La experiencia de los analistas de la OPD y de los especialistas de la red de estaciones meteorológicas puede ser incorporada al proceso, esto con el fin de definir cuál es la forma más directa en la que puede abordarse el cálculo de tales rangos de modo que se disminuyan los costos y el tiempo, el nivel en donde deben establecerse los debidos controles de calidad y el análisis de la posible redundancia de algoritmos.
- La cantidad de datos que deseen abanderarse dependerá de la capacidad de recursos y material humano para dedicar tiempo en la explicación de estos.
- Las EMA pueden enviar la información codificada de los resultados de los ACCD junto con el resto de la información meteorológica desde el

emplazamiento. Desde el punto de vista de programación, los GDs CR10X muestran suficiente flexibilidad para la inclusión de ACCD. No obstante, esto no puede llevarse a cabo sin disminuir el espacio de memoria para el almacenamiento de programas.

- Los ACCD aquí expuestos son capaces de detectar valores atípicos y alertar al personal acerca de posibles anomalías o problemas de carácter técnico que pudieran estarse presentando.
- Las pruebas de control de calidad aquí efectuadas son aplicables tanto en tiempo real, como después de que los datos son traídos del campo por el personal técnico, por lo que es posible incorporarlos a un protocolo de control de calidad de datos más estructurado en el que son analizados con pruebas similares antes de ser asimilados por la base de datos, ya sea de forma automatizada o manual .
- La implementación de un PCCD de este tipo no garantiza por sí mismo un mejoramiento de la calidad de los datos generados por la red de EMA del IMN, pero si ayudará en la determinación de qué tan buena es. Para poder garantizar el mejoramiento de la calidad de los datos es necesario un mejoramiento integral de todos los niveles involucrados en la producción y procesamiento de datos.
- La meta información producida por los IA sería de gran utilidad para cierto tipo de usuarios especializados para diferentes estudios, y les indicará con un criterio objetivo que tan buenos son los datos que pretenden utilizar para sus fines.