Araya López, José Luis.
Resultados de un control de calidad de datos de temperatura superficial
Tecnología en Marcha, *Vol. 24, N.° 1, Enero-Marzo 2011, P. 33-49*

Resultados de un control de calidad de datos de temperatura superficial del aire y humedad relativa

Fecha de recepción: 04/05/2010
Fecha de aceptación: 14/06/2010

José Luis Araya López [1]

Palabras clave

Control de calidad, datos meteorológicos, temperatura, humedad relativa, scilab, datos atípicos.

Key words

Quality control, meteorological observations, temperature, relative humidity, scilab, outliers.

Resumen

Se presenta un estudio básico sobre control de calidad de datos de temperatura y humedad relativa. Este trabajo fue efectuado en el Instituto Meteorológico Nacional (IMN) de Costa Rica. El objetivo fue determinar de una forma objetiva, cual es la calidad de los datos generados por la red de estaciones meteorológicas. Los programas de control de calidad marcaron datos sospechosos y erróneos de temperatura que pasaron desapercibidos en los niveles anteriores del control de calidad.

Los resultados muestran que existen valores atípicos infiltrados que escaparon a la detección del método de revisión manual tradicionalmente aplicado. De forma general, el porcentaje de valores atípicos de humedad relativa fue mayor que el encontrado para datos de temperatura. Dicho porcentaje nunca fue mayor al 20% del total de datos de la serie cronológica correspondiente. Los resultados de este trabajo también muestran cómo los programas desarrollados mejoran la eficiencia del protocolo actual de control de calidad.

Abstract

A recent study to quality-control hourly and daily data is presented. The meteorological parameters analyzed were temperature and relative humidity. The study took place at the National Meteorological Institute (NMI) from Costa Rica. The goal was to objectively determine the quality of the data generated by the data network. The quality control scripts flagged suspicious and wrong temperature values that went undetected in the previous steps of the current quality-control protocol. The results show that atypical data do exist in the database, in spite of being formerly checked using manual procedures.

1. Licenciado en Meteorología. Instituto Meteorológico Nacional. San José, Costa Rica. Teléfono: 2222-5616. Correo electrónico: *jlaraya@imn.ac.cr.*

As a whole, the percentage of atypical relative humidity data is higher than that detected for temperature data. In both cases, the percentage of atypical data for any of the datasets analyzed is never higher than 20%. The results of this research also suggest that the software tools developed improve the efficiency of the current quality-control procedure.

Introducción

El control de calidad de datos meteorológicos puede ser dividido en varias fases en función de la profundidad de la verificación y de los servicios disponibles para el usuario final. La Organización Meteorológica Mundial (OMM) sugiere que las etapas de control de calidad sean las siguientes: verificación de errores groseros, verificación de la coherencia interna, verificación de la coherencia temporal y verificación de la coherencia espacio temporal.

Parte de la nueva metodología que se desarrolla actualmente pretende dar un paso adelante en el uso de herramientas computacionales adecuadas que permitan analizar la información de un modo expedito, tanto analítica como visualmente, de modo que sea posible observar situaciones que eran difíciles de detectar sin la ayuda de los paquetes computacionales.

En la última fase puede realizarse un control combinado de estos métodos (OMM, 1992). Como parte de una nueva iniciativa desarrollada en el Instituto Meteorológico Nacional (IMN) de Costa Rica, se trabaja en el desarrollo de una forma alternativa de efectuar la revisión tradicional de los datos. El método tradicional se ha enfocado en la revisión de datos provenientes de estaciones meteorológicas convencionales, las cuales se han basado en una revisión visual.

Parte de la nueva metodología que se desarrolla actualmente pretende dar un paso adelante en el uso de herramientas computacionales adecuadas que permitan analizar la información de un modo expedito, tanto analítica como visualmente, de modo que sea posible observar situaciones que eran difíciles de detectar sin la ayuda de los paquetes computacionales. Más importante aún, las nuevas metodologías se están desarrollando de manera que sea posible hacer operativo un protocolo de control de calidad basado en su uso. Se llamará datos atípicos a aquellas observaciones que parecen haberse generado de manera distinta al resto de los datos (Peña, 2002). Por otro lado, la detección de valores atípicos en las series cronológicas es muy relevante, debido a las implicaciones que estos pueden tener en la matriz de covarianza y correlación, esto a la hora de analizar los datos para estudios varios. Otra razón por la que se llevó a cabo este control de calidad fue para recolectar información acerca de los puntos más vulnerables que pudieran existir en la generación actual del dato.

Método

Algoritmos y macros utilizados

La dificultad en la adquisición de herramientas que facilitaran la aplicación de metodologías de carácter más objetivo hacía difícil el desarrollo operacional de paquetes que se ajustaran a las necesidades particulares, específicamente la manipulación de series cronológicas en resolución horaria y diaria. Actualmente, mediante paquetes de libre distribución se está llevando a cabo la implementación operativa de dichas tecnologías que podrán ser de gran utilidad para una revisión y un tratamiento adecuado de series relativamente largas.

Para este proyecto, el paquete de programas Scilab se usó para el desarrollo de programas que permiten la revisión exploratoria de la información, junto con la aplicación de algoritmos de control de calidad sobre toda la serie. Este es un paquete de computación científica desarrollado por el Institut National de Recherche en Informatique et Automatique de Francia y cuenta con capacidades especiales para cálculo numérico. Utiliza un enfoque basado en operaciones matriciales que facilita y hace aplicables los más variados métodos estadísticos y matemáticos de uso en análisis de datos.

También es posible el despliegue visual de la información mediante gráficos y superficies diversas. Se utilizó este paquete debido a que es de libre distribución, es versátil y fácilmente extensible; su capacidad de incorporación de funciones permite hacer uso de contribuciones de otros desarrolladores, quienes libremente aportan de forma continua nuevas funciones, así como otras que puedan desarrollarse en el IMN para aplicaciones particulares.

Básicamente, la filosofía que suele adoptarse en el control de calidad de datos meteorológicos es analizar la validez de un dato desde varias perspectivas: la aplicación de Algoritmos de Control de Calidad (ACCD) facilita la evaluación de los datos generados desde diferentes puntos de vista, lo cual permite obtener conclusiones de una forma más objetiva. La labor operacional de control de calidad de datos a profundidad requiere que se acumulen diversos indicios en contra de la plausibilidad o no de los datos considerados atípicos. Los ACCD y las herramientas de visualización gráfica, así como una experiencia adecuada del analista, permiten encontrar situaciones anormales en los datos y clasificarlas convenientemente.

Los ACCD y las herramientas de visualización gráfica, así como una experiencia adecuada del analista, permiten encontrar situaciones anormales en los datos y clasificarlas convenientemente.

Meek y Hatfield (2001) enfatizan la importancia de la visualización gráfica de los datos para labores de control de calidad. En este sentido, el programa Scilab presenta diversos comandos para graficar que facilitan esta labor. Aparte de esto, la generación de reportes de listas de datos sospechosos permitió aislar rápidamente estos datos para su revisión detallada. Estas bondades hicieron del paquete de programas en cuestión una buena alternativa para la realización de este trabajo. En total se generaron tres diferentes programas para efectuar las siguientes tareas (véase cuadro 1).

a) Completar el eje de tiempo de los datos. La aplicación permitía tener una serie de datos con secuencia temporal, esto se efectúa antes del control de calidad. Esto es importante porque los datos almacenados en la base no toman en cuenta las brechas; de allí que estas fueron generadas de esta manera.

b) Control de calidad de datos en resolución horaria (temperatura superficial del aire y humedad relativa).

c) Control de calidad de datos diarios (temperatura superficial y humedad relativa).

Datos utilizados

El control de calidad descrito fue aplicado a los datos de Estaciones Meteorológicas Automáticas (EMA) existentes, hasta la fecha, en el Instituto Meteorológico Nacional (IMN). La razón por la que se escogió trabajar con estos datos se debe a que se prevé que el uso de las EMA en la red de estaciones será cada vez más común en comparación con las estaciones meteorológicas convencionales. El cuadro 1 muestra el conjunto de EMA utilizadas. Nótese que se trabajó tanto con estaciones cerradas como abiertas.

Se consideró correcto analizar las estaciones con periodos superiores a tres años. En todos los casos, los registros de información de las EMA inician después de 1995 y se extienden hasta el 2008, inclusive. Es importante mencionar que antes de que los datos fueran analizados con los criterios de control de calidad usados, se aplicó un programa para completar el eje de tiempo de las series meteorológicas analizadas. Esto era necesario, ya que algunas pruebas dependen de la relación entre dos lecturas consecutivas, además de que permiten tener una idea preliminar de la distribución de datos faltantes.

También se requirió revisar la información en los informes de gira de cada emplazamiento. Cuando se llevan a cabo giras de mantenimiento y recolección de datos en las diferentes EMA ubicadas a lo largo del país, se entrega un informe de gira donde se especifica el estado de la EMA en el momento en que se realiza la

Cuadro 1. Estaciones meteorológicas analizadas.

Nombre	Cuenca	Número	Activa	Cerrada	Elevación (m)	Latitud (Grados-Min.)	Longitud (Grados-Min.)
Boca Tapada	69	597		X	170	10 41	84 14
Comando los Chiles	69	633	X		40	11 02	84 43
Finca Brasilia	69	647	X		350	10 58	85 02
El Ensayo	69	649		X	550	10 57	85 24
Villa Blanca	69	651		X	1389	10 11	83 28
Los Lirios	69	655		X	100	10 45	84 35
Zurquí	69	659		X	1563	10 03	84 01
Ciudad Quesada	69	661	X		700	10 18	84 25
Laguna de Caño Negro	69	677	X		30	10 54	84 47
Upala	69	679	X		60	10 53	85 04
La Rebusca	69	681	X		26	10 29	84 01
Canta Gallo	71	15	X		40	10 03	83 04
La Rita	71	21		X	125	10 15	83 45
Santa Elena	72	141		X	285	10 55	85 36
San José Pinilla	72	149	X		50	10 15	85 05
Ferco Garza	72	155		X	10	09 54	85 36
Finca La Ceiba	72	157	X		20	10 06	85 19
Paquera	72	159	X		15	09 49	84 56
ITCR	73	123	X		1395	09 51	83 54
Recope Ochomogo	73	129	X		1546	09 53	83 05
Quebrada Grande	74	49		X	450	10 52	85 29
Liberia	74	51	X		144	10 35	85 35
Santa Cruz	74	53	X		54	10 17	85 25
Montezuma	76	51		X	519	10 04	85 04
Hacienda Mojica	76	55	X		40	10 27	85 09
Lagunilla de Miramar	78	23		X	960	10 08	84 42
Puntarenas	78	27	X		3	09 58	84 05
Caldera	80	1		X	2	09 55	84 43
Limón	81	5	X		7	09 57	83 01
Museo Nacional	84	0		X	1172	09 56	84 05
San Luis de Poás	84	124		X	2400	10 12	84 13
Finca Antolín	84	128		X	2400	10 01	84 01
Cigefi	84	139	X		1200	09 56	84 03
IMN	84	141	X		1172	09 56	84 05
Orotina	84	143		X	224	09 54	84 31
Escuela de Ganadería	84	145	X		450	09 56	84 22

Continúa...

Continuación

Nombre	Cuenca	Número	Activa	Cerrada	Elevación (m)	Latitud (Grados-Min.)	Longitud (Grados-Min.)
Barrio San José Atenas	84	167		X	700	10 00	84 24
El Coco	84	169	X		890	10 00	84 13
Istarú	84	181	X		1776	09 53	83 58
Fabio Baudrit	84	187	X		840	10 01	84 16
Fraijanes	84	189	X		1764	10 08	84 12
Recope la Garita	84	191	X		760	10 00	84 16
Pavas	84	195	X		997	09 58	84 08
Sta. Bárbara	84	197	X		1060	10 00	84 00
Hitoy Cerere	85	21	X		100	09 04	83 02
Manzanillo	85	23	X		8	09 38	82 39
Sixaola	87	13	X		10	09 32	82 37
Frailes	88	35	X		1600	09 46	84 04
La Lucha	88	41	X		1880	09 44	84 00
Quepos	90	11		X	6	09 25	84 10
Las Brisas de San Vito	98	75	X		900	08 51	82 57
La Linda	98	79		X	732	09 21	83 38
Pindeco	98	87	X		340	09 08	83 02
Chirripó	98	91		X	3630	09 27	83 03
Altamira	98	95		X	1360	09 01	83 00
Golfito	100	3	X		6	08 08	83 01
Guacaco	100	613		X	20	08 22	83 08
La Palma de Palmar Sur	100	615		X	16	08 57	83 28
Los Patios	100	625	X		22	08 06	83 04
Coto 47	100	631	X		8	08 36	82 59
Laurel	100	641	X		16	08 28	82 51

inspección. Muchas veces, las condiciones en las que la EMA es encontrada pueden ayudar a explicar valores sospechosos que se detectan en el control de calidad posterior. Para cada estación, se contó con el archivo histórico correspondiente y se consultó en aquellos casos en los que fueron detectados comportamientos sospechosos en los datos.

Algoritmos y macros utilizadas

Prueba de salto con resolución horaria

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

$$\left|\left(T_i - T_{(i-1)}\right)\right| > \delta \quad (1)$$

donde *ä* corresponde al valor umbral de la magnitud de la diferencia entre las horas consecutivas Ti y $T_{(i\text{-}1)}$ por encima del cual tales promedios horarios de temperatura serán abanderados (Reek *et ál*., 1992). El umbral se calcula basándose en un criterio de percentil a partir de la serie de datos analizada. Se decidió usar el percentil 99,9

con el fin de marcar un 0,01% de los saltos más altos.

Prueba de límites de tolerancia

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

$$T_{min} \leq T \leq T_{max} \quad (2)$$

donde T_{min} y T_{max} corresponden a 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. Para efectos de abanderar un porcentaje de valores altos, se utilizaron los percentiles 10 y 90, pues se consideró que era viable la revisión de un 10% de los datos para ambos límites. Dichos datos fueron aislados y analizados por medio de reportes generados a partir de los programas corridos sobre las series; además, fueron incluidos productos gráficos concretos, tales como gráficos de control estadístico (Montgomery, 2005) para la visualización expedita de dichos reportes. Esta prueba se aplicó a los valores diarios analizados.

Para efectos de aislar los datos que se alejaban del promedio histórico mensual u horario, se aplicó una prueba de anomalía. Esto permite la detección de datos que contrastan con la hora o el mes de su ocurrencia. La prueba incluyó una interfase gráfica para su revisión visual.

Prueba de anomalía estadística

En el vocabulario meteorológico, una anomalía se entiende como la resta de datos de un promedio relevante. El término *anomalía* no necesariamente denota un dato que indique un evento anormal o inusual (Wilks, 1995). Para efectos de aislar los datos que se alejaban del promedio histórico mensual u horario, se aplicó una prueba de anomalía. Esto permite la detección de datos que contrastan con la hora o el mes de su ocurrencia. La prueba incluyó una interfase gráfica para su revisión visual.

Generación de las estadísticas mensuales históricas

Con el fin de detectar situaciones atípicas serias en la información meteorológica, el reporte de los programas incluía una tabla de estadísticas básicas. Esas estadísticas eran los valores máximo, mínimo, media, moda, coeficiente de variación y desviación típica de cada parámetro a nivel mensual. Para curvas de frecuencia unimodales que sean ligeramente sesgadas o asimétricas, se conoce que se cumple la siguiente relación empírica (Spiegel y Stephens, 2001):

$$Media - Moda = 3 * (Media - Mediana) \quad (3)$$

Considerando que de forma general la distribución de frecuencias de la temperatura horaria se aproxima a una distribución de frecuencias sesgada a la izquierda y unimodal, en diversos casos, se revisó cuidadosamente la relación entre la media, la mediana y la moda. Para este tipo de distribución se verificó, en particular, que la relación entre estas medidas de tendencia central obedeciera la siguiente relación:

$$Moda < Mediana < Media \quad (4)$$

Se analizó la relación entre la media, la mediana y la moda basándose en la forma de la distribución de los datos, lo cual fue un criterio adicional para la detección de datos atípicos. Cuando la relación no se cumplía de forma satisfactoria después de analizar la forma de la distribución de frecuencias, se procedía a la revisión de los datos en busca de situaciones sospechosas con la información. Este tipo de reporte permite ver si las estadísticas mensuales son consistentes, por lo que en su análisis preliminar pueden encontrarse valores sospechosos.

Uso de gráficos de cajas y bigotes

En combinación con la generación de ciertos percentiles de la serie, también se aplicaron gráficos de cajas y bigotes, un método de detección de valores atípicos como el que se describe en Wilks (1995). Es importante destacar que este criterio asume que la distribución es normal, por lo que dependiendo de la distribución, tendía

a indicar como atípicos mayor o menor cantidad de datos dependiendo del sesgo de la distribución. Los valores atípicos eran abanderados según los siguientes criterios:

$$LIM\_SUP = Q_3 + 1.5 * (Q_3 - Q_1) \quad (5)$$

$$LIM\_INF = Q_1 - 1.5 * (Q_3 - Q_1) \quad (6)$$

Donde *LIM_SUP* y *LIM_ INF* son los límites superior e inferior, respectivamente. Los valores Q_3 y Q_1 son el tercer y primer cuartil de datos, respectivamente. Esta prueba fue aplicada y el significado de sus resultados evaluado dependiendo de si la distribución era normal o no. El método se aplicó de forma conjunta con la prueba de límites de tolerancia, esto con el fin de evaluar las ventajas y desventajas de ambos.

Pruebas de consistencia entre extremas

Aplicada para el caso de datos en resolución diaria. Se realizaron pruebas de consistencia entre máxima y mínima para aislar datos que puedan presentar dicho caso. Asimismo, se ha efectuado una prueba de rangos entre extremas, con el fin de detectar rangos que puedan generar algún tipo de duda. Otra prueba introducida es la comparación de valores extremos en días consecutivos, en la cual se buscaron secuencias de datos repetidos en los datos diarios (Feng *et. ál*., 2004). En el caso de la humedad relativa, se analizó cada caso por aparte, con el fin de determinar si la secuencia de valores repetidos era o no plausible.

Aplicaciones gráficas utilizadas

El primer paso en estas aplicaciones de control de calidad es la representación gráfica de los datos, ya que es muy útil para detectar asimetrías, heterogeneidad y datos atípicos (Peña, 2002). Se decidió hacer uso de la revisión gráfica de la información usando aplicaciones gráficas, como: histogramas de frecuencias relativas y series de tiempo de los saltos, de las anomalías y de los valores de temperatura y humedad relativa en general. Se insistió en la visualización redundante de valores mensuales, usando algunos criterios de la estadística descriptiva (Spiegel y Stephens, 2001). La figura 1 es un ejemplo de los reportes gráficos generados para una estación. Para este caso, nótese cómo visualmente algunos datos atípicos destacan en la serie de datos, esto al efectuar el gráfico de la serie de datos con los saltos y las anomalías horarias correspondientes. Una vez que el revisor observa estos reportes, cualquier dato atípico que destaque será fácilmente visible en el gráfico.

Reportes de valores atípicos generados

La aplicación generaba reportes de los valores atípicos encontrados. Se procedió a evaluar los resultados del reporte en conjunto con la evidencia gráfica obtenida. Tales reportes son muy importantes, pues permiten comparar la consistencia de los datos con la hora en la cual se generan. En muchos casos, esto permitió la identificación de datos erróneos que no fueron detectados por el protocolo tradicional.

Indicadores de alerta (IA) y marcación

Fue necesario definir ciertos IA, con el fin de otorgar un estatus al dato atípico detectado. Según Gandin (1988), el abanderamiento de datos es cuando las condiciones sospechosas señaladas por las pruebas de control de calidad son denotadas por dígitos especiales, pudiendo ser estos revisados durante las subsecuentes etapas (Araya, 2007). Durante el proceso se genera una lista de valores atípicos que es clasificada, según el caso, en datos erróneos o sospechosos.

Los IA que se han usado hasta el momento son tres: dato erróneo (-12), dato sospechoso (-4) y dato que contribuye a moda atípica (-16). El caso de moda atípica se dio en el análisis de los datos

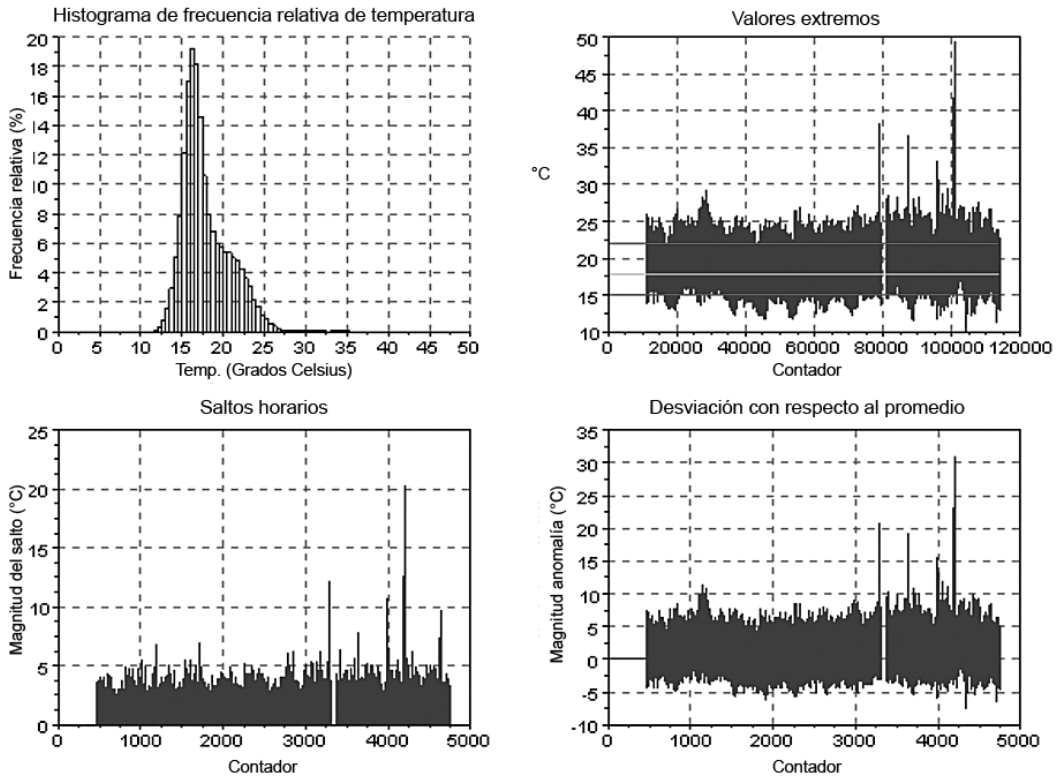

Figura 1. Ejemplo de valores atípicos de temperatura horaria que pueden visualizarse usando los programas de control de calidad desarrollados.

horarios, principalmente. El concepto de dato "sospechoso" en realidad requiere del criterio de experto e implica una potencialidad de error que no puede ser claramente establecida debido a falta de evidencia para catalogar el dato encontrado como "erróneo". Implica cierta subjetividad por parte del analista y está sujeto a los recursos y experiencia con que se cuente para poder descartar el dato. En todo caso, el dato original se retiene inalterado, así como cualquiera catalogado según estos criterios.

Sin pérdida de generalidad, el IA= -4 indica al usuario que el dato en cuestión generó al revisor cierto nivel de sospecha, pero que queda a criterio de este usuario tomar la decisión en cuanto a si lo considera erróneo. El caso IA=-12 declara una certeza por parte del revisor de que el dato no es físicamente posible, en tanto el IA=-16 indica al usuario que el dato en cuestión contribuye al cálculo de una moda atípica, es decir, el dato se repite más de lo que estadísticamente se espera. Este es un sistema de marcación para identificación y archivamiento de datos atípicos adaptado a las necesidades de este estudio y a la cantidad de personal disponible, el cual tiene como ventaja que permite evaluar la totalidad de la revisión efectuada a los datos (Araya, 2007).

Resultados

La figura 2 muestra la utilidad de las metodologías de despliegue gráfico para detectar rápidamente valores atípicos. En la figura 2 se muestra un gráfico de dispersión de los valores mínimos contra los valores máximos de temperatura diaria. Nótese cómo algunos puntos claramente se alejan de la nube de dispersión. Estos

datos fueron aislados y marcados como atípicos en la serie.

El cuadro 3 presenta un tipo de error que se encontró en las series cronológicas de temperatura horaria. Estos datos fueron detectados en la EMA Fabio Baudrit, ubicada en el Valle Central, San José, Costa Rica. Este error se caracteriza por la presencia de valores máximos en la madrugada y valores mínimos durante el día, en particular cerca del mediodía. Tales datos fueron detectados y abanderados con un indicador apropiado. Este tipo de errores es difícil de detectar visualmente y por medio de herramientas gráficas. Normalmente, salen a relucir en los informes de anomalía del control de calidad, donde se reporta el conjunto de datos con anomalías más altas de toda la serie. Posteriormente, se lleva a cabo una prueba de consistencia de esta anomalía con la hora, para determinar si es factible que

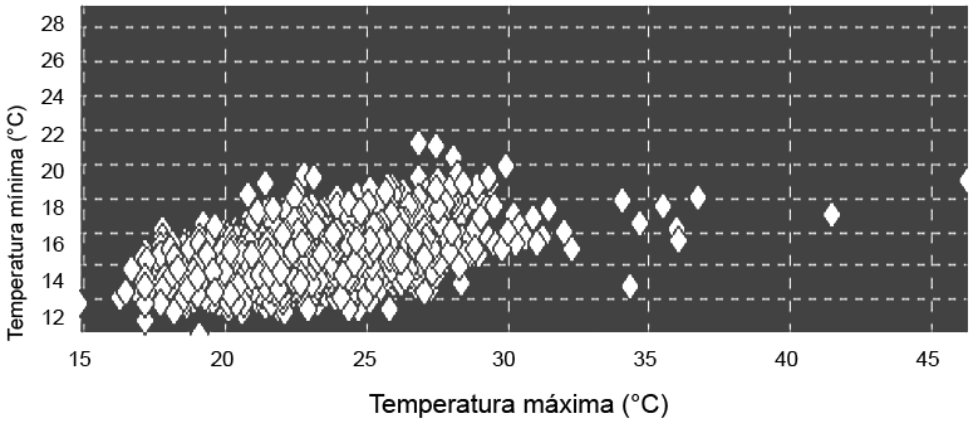

Figura 2. Datos atípicos detectados en Frailes, Cartago, Costa Rica.

Cuadro 2. Porcentaje de datos atípicos para temperatura y humedad relativa horaria.

Estaciones revisadas	Temperatura (°C)				Humedad relativa (%)			
	%IA (*)	IA (**)			%IA (*)	IA (**)		
Zona Norte		-4	-12	-16		*-4*	*-12*	*-16*
La Rebusca	*0*	X			*2.5*	X		
Laguna de Caño Negro	*0*				*0*			
Comando Los Chiles	*0.05*		X		*0*			
Upala	*0*				*1.4*			X
Los Lirios	*0*				*0*			
Boca Tapada	*0.1*	X	X		*0*			
Finca Brasilia	*0*				*0*			
El Ensayo	*0*				*0*			
Ciudad Quesada	*0*				*0*			
Villa Blanca	*0*		X		*0*			
Zurquí	*0*				*0*			

Continúa...

Continuación

Estaciones revisadas	Temperatura (°C)				Humedad relativa (%)			
Pacífico Norte	*%IA*	*-4*	*-12*	*-16*	*%IA*	*-4*	*-12*	*-16*
Caldera	*0*				*0*			
Puntarenas	*0.1*		X		*0*			
Ferco Garza	*6.2*	X	X		*0*			
Paquera	*0*				*0*			
Finca La Ceiba	*0*				*0*			
Hda. Mojica	*0*			X	*0*			
San Jose Pinilla	*0*	X			*0*			
Santa Cruz	*0*				*12.2*	X	X	X
Liberia	*0*				*0*			
Santa Elena	*0.2*	X	X		*0*			
Quebrada Grande	*0*				*0*			
Montezuma	*0*				*0*			
Lagunilla de Miramar	*0*				*0*			
Pacífico Sur	*%IA*	*-4*	*-12*	*-16*	*%IA*	*-4*	*-12*	*-16*
Golfito	*0*				*0*			
Coto 47	*0*				*0*			
La Palma de Palmar Sur	*0*				*10.1*			X
Laurel	*0*				*0*			
Guacaco	*0*				*0*			
Los Patios	*0*				*0*			
Pindeco	*0*		X		*0*			
La Londa de Pz	*0*				*0*			
Las Brisas De San Vito	*0*				*0*			
Altamira	*0*	X			*0*			
Chirripó	*0*				*0*			
Pacífico Central	*%IA*	*-4*	*-12*	*-16*	*%IA*	*-4*	*-12*	*-16*
Quepos	*1*	X			*0*			
Frailes	*0.5*	X			*0*			
La Lucha	*0*				*0*			
Valle Central	*%IA*	*-4*	*-12*	*-16*	*%IA*	*-4*	*-12*	*-16*
Orotina	*0*				*5.9*	X		X
Escuela de Ganadería	*0*				*0*			
Barrio San José Atenas	*0*				*19.0*		X	
Recope La Garita	*0*				*0*			
Fabio Baudrit	*0*		X		*0*			
El Coco	*0*				*0*			
Pavas	*0*				*0*			
Santa Bárbara	*0*			X	*0*		X	
Museo Nacional	*0*				*0*			
IMN	*0*				*0*			
Cigefi	*0.7*	X			*0*		X	
Fraijanes	*0*				*0*			

Continúa...

Continuación

Estaciones revisadas	Temperatura (°C)				Humedad relativa (%)			
Istarú	*0*				*0*			
San Luis de Poás	*0*				*0.3*			X
Finca Antolín	*0*				*0*			
Región Caribe	*%IA*	*-4*	*-12*	*-16*	*%IA*	*-4*	*-12*	*-16*
Limón	*0*		X		*0*			
Manzanillo	*0.1*			X	*0*			
Sixaola	*0*				*0*			
Canta Gallo	*0*	X	X		*0*			
Hitoy Cerere	*0*				*0*			
La Rita	*0*				*0*			
ITCR	*0*				*0*			
Recope Ochomogo	*0*				*0*			

(*) %IA: Porcentaje de indicadores de alertas en el total de datos.
(**) IA: Acrónimo para indicador de alerta.

Cuadro 3. Valores atípicos detectados en Fabio Baudrit.

Día	Mes	Año	Hora	Temperatura (°C)	Indicador de alerta
12	8	2002	13	15.95	-12
12	8	2002	17	29.48	-12
12	8	2002	21	28.41	-12
12	8	2002	22	28.87	-12
12	8	2002	23	30.24	-12
12	8	2002	24	30.95	-12
13	8	2002	1	31.39	-12
13	8	2002	2	31.49	-12
13	8	2002	3	31.83	-12
13	8	2002	4	34.31	-12
13	8	2002	5	37.08	-12
13	8	2002	6	39.28	-12
13	8	2002	7	34.00	-12
13	8	2002	11	16.70	-12
13	8	2002	10	17.84	-12
13	8	2002	11	16.70	-12
13	8	2002	12	17.66	-12
13	8	2002	13	17.05	-12

se presente a la hora señalada. Obsérvese que estos algoritmos son exitosos en el aislamiento de situaciones como estas en series de tiempo, considerando, desde luego, que se cuenta con un técnico experimentado capaz de aislar la situación.

Otro tipo de error que se encontró en las series de tiempo analizadas fue la combinación de series cronológicas, lo cual se dio en tres estaciones meteorológicas diferentes. Este tipo de error obedece al hecho de que hubo un cambio de emplazamiento en la estación meteorológica y los datos del nuevo emplazamiento se siguieron almacenando junto con los datos generados en el emplazamiento anterior.

En la figura 2 se muestra el caso de una serie de tiempo contaminada con los datos de dos emplazamientos diferentes. Nótese la disminución de la varianza de la serie. Este ejemplo señala la importancia de contar con un adecuado sistema de metadatos que permita dar seguimiento a cambios de emplazamiento. Lo recomendable es, siempre que se haga un cambio de emplazamiento, documentarlo adecuadamente y diferenciar la serie anterior y posterior de una forma clara, ya que el desconocimiento de este hecho podría llevar a conclusiones equivocadas con respecto a los datos; esto por el riesgo de que un usuario no enterado de la situación puede hacer inferencias equivocadas.

La figura 4 muestra el caso de la EMA de Boca Tapada, ubicada en la zona norte de Costa Rica. Las etiquetas en el margen

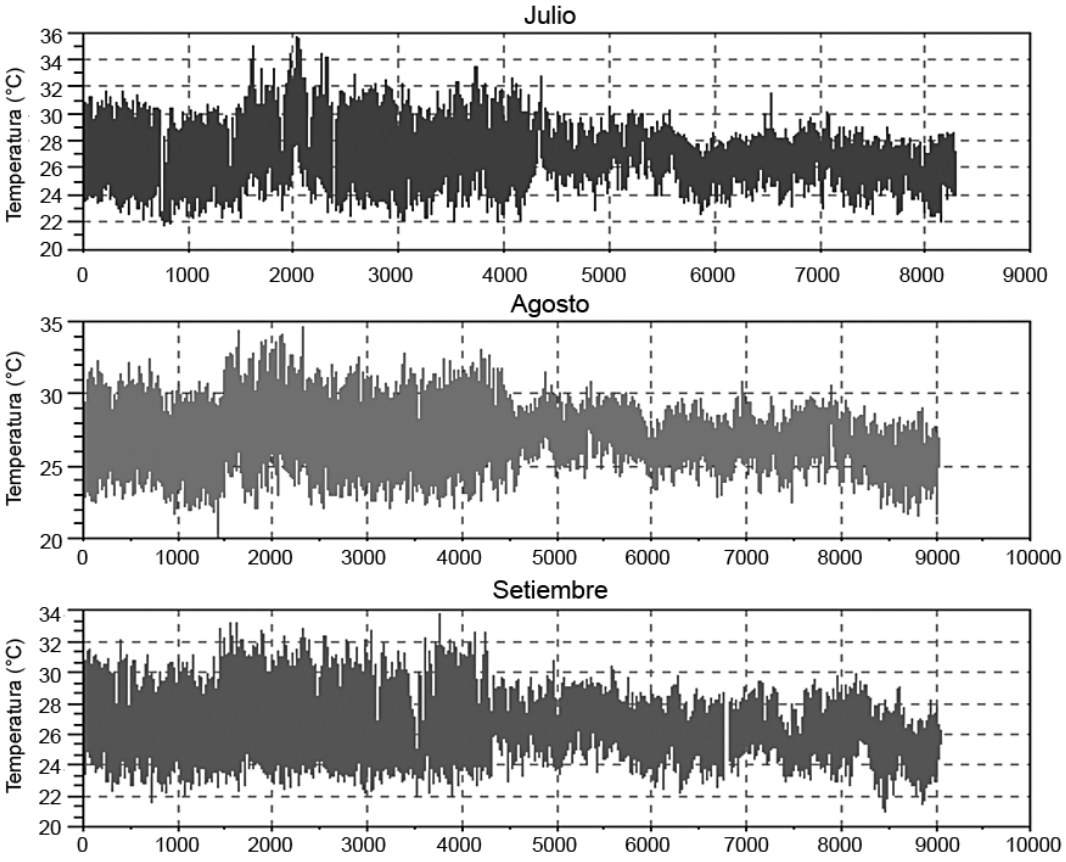

Figura 3. Ejemplo de series de datos con diferente variabilidad.

Cuadro 4. Algunos de los datos atípicos en Boca Tapada

Día	Mes	Año	Hora	Temperatura (°C)	IA
31.0	10.0	1997	23.0	29.6	-12
31.0	10.0	1997	24.0	30.4	-12
1.0	11.0	1997	3.0	30.3	-12
1.0	11.0	1997	23.0	30.0	-12
1.0	11.0	1997	24.0	31.7	-12
2.0	11.0	1997	1.0	32.5	-12
2.0	11.0	1997	2.0	33.2	-12
2.0	11.0	1997	3.0	31.6	-12
2.0	11.0	1997	21.0	30.1	-12
2.0	11.0	1997	22.0	30.0	-12
2.0	11.0	1997	23.0	29.9	-12
3.0	11.0	1997	2.0	31.1	-12
3.0	11.0	1997	3.0	30.5	-12
3.0	11.0	1997	22.0	30.1	-12
3.0	11.0	1997	23.0	31.8	-12

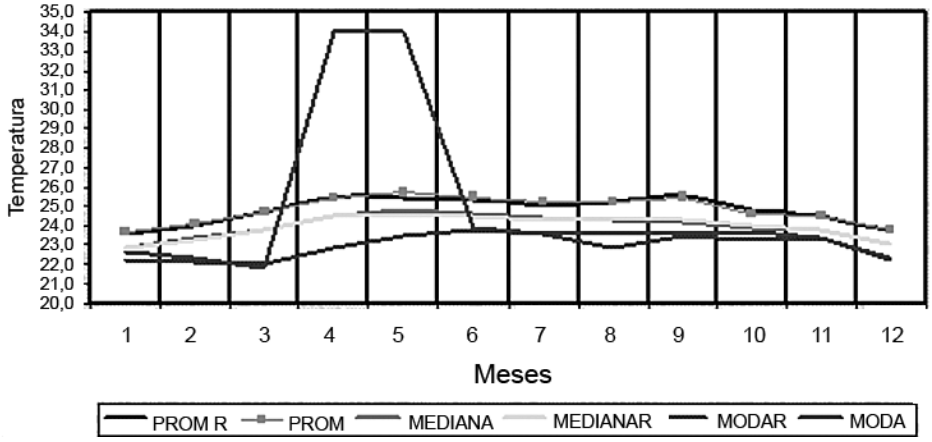

Figura 4. Ejemplo de problemas con la moda mensual de temperatura superficial del aire en la estación Boca Tapada.

inferior terminadas en "R" indican que los datos en cuestión recibieron un tratamiento de control de calidad y de relleno. Los datos sin "R" muestran los datos con los valores atípicos originales. Este ejemplo muestra cómo la metodología de control de calidad permite encontrar situaciones que se salen de lo esperado en la información meteorológica. Una vez que los datos atípicos fueron aislados, se procedió a hacer un análisis de los estadísticos de la serie sin control de calidad y con él, y rellenado de brechas posterior. Nótese cómo en general la moda de la serie es inferior a las estadísticas media y mediana, como corresponde en una distribución de frecuencias sesgada a la izquierda.

Para la mayoría de los meses, la moda tiene el comportamiento regular esperado, excepto para abril y mayo. Al efectuar la

revisión de los datos atípicos generados en los reportes de los programas de control de calidad, se encuentran situaciones de valores de temperatura anormalmente altos en horas de la madrugada y noche, como puede observarse en el cuadro 3. Estos valores son claramente incorrectos, dado que no puede esperarse máximos horarios en horas de la madrugada y noche, de allí que se procedió a indicarlos como erróneos. Este ejemplo muestra la importancia de una revisión progresiva de los datos basándose en diferentes criterios, de modo que pueda acumularse suficiente evidencia de la plausibilidad de los datos.

Los cuadros 2, 5 y 6 muestran el total de valores atípicos detectados en todos los datos de temperatura humedad relativa en resoluciones horaria y diaria, respectivamente. Con el fin de contabilizar la cantidad de datos atípicos mostrados, se definió un índice que corresponde al total de valores atípicos encontrados (erróneos y sospechosos), dividido entre el total de datos de la serie y multiplicado por cien. Nótese cómo en el caso de los datos de temperatura horaria el porcentaje de datos atípicos siempre se presenta en el orden de las décimas de este, con excepción de la estación de Garza-Ferco, en la región del Pacífico Norte, que presentó un porcentaje del 6% de datos atípicos.

Para la humedad relativa horaria, la cantidad de valores atípicos encontrados tiende a ser mayor, particularmente en las regiones Pacífico Sur y Pacífico Norte. Con respecto a lo anterior, está demostrado que los instrumentos electrónicos de sensor higroscópico usados en las EMA son delicados; en los trópicos sus circuitos se deterioran al término de meses, de allí que requieren una calibración constante después del primer año de uso (Castro e Hidalgo, 1997). Hubo estaciones como San José de Atenas donde la cantidad de valores sospechosos encontrados llega a ser incluso del orden del 20% del total de la serie. Upala, Santa Cruz y Palmar Sur fueron estaciones que mostraron incongruencias en los datos, particularmente en lo que se refiere a modas atípicas que distorsionaban los estadísticos generados.

Cuadro 5. Porcentaje de datos atípicos de temperatura diaria.

Estación	Máximas % Atípicos	4	-12	Mínimas % Atípicos	-4	-12
Villa Blanca	7.5	188		17.8	454	
La Rebusca	0			0.1	2	
Canta Gallo	0		1	0		1
Santa Elena	0		1	0		
Finca La Ceiba	0	1		0.2	1	5
ITCR	0	2		0	2	
Recope Ochomogo	0.4	14	1	0		
Puntarenas	0.1		2	0.1		2
Cigefi	0.1	4		0		
Fabio Baudrit	0			0		1
Recope La Garita	0		1	0		
Santa Bárbara	0.7	8	10	0.7	8	10
Frailes	1.2	36	23	0		

Cuadro 6. Porcentaje de datos atípicos de humedad relativa diaria.

Estación	Máximas % Atípicos	4	-12	Mínimas % Atípicos	-4	-12
Villa Blanca	0			0.3	6	1
Santa Elena	0			8.7	222	
Recope Ochomogo	0			0	1	
Santa Cruz	0			0		1
Limón	0			0	1	
Barrio San José Atenas	0.1		2	0		
Fabio Baudrit	0			0.3		17
La Linda	0	33		0.8	33	
Chirripó	4.3	156		4.3	156	

El uso de estos programas permitió la manipulación de los archivos de los datos de una forma eficiente, pero no exime al analista de su interpretación de los reportes y gráficos generados. Esto es importante, ya que sin experiencia climatológica adecuada el usuario podría encontrar difícil tomar decisiones sobre la plausibilidad de los valores detectados.

En el caso de los datos de temperatura diaria, se observa que la estación de Frailes fue la que presentó una mayor cantidad de valores erróneos no detectados para las máximas extremas, en tanto que en la estación de Villa Blanca fueron detectados más valores sospechosos, subiendo a más del 7% la cantidad de datos abanderados, según este criterio.

En total, para los datos de cada emplazamiento se encontró algún tipo de valor atípico en el 18% de las series de máximas, así como en un 13% para los datos de temperatura mínima. En el caso de los datos extremos diarios de humedad relativa, se observa que las estaciones de Chirripó y Santa Elena fueron las que contribuyeron con mayor cantidad de valores atípicos, para los valores máximo y mínimo, respectivamente. En ambos casos, la contribución fue primordialmente de datos sospechosos y el porcentaje de datos atípicos no sobrepasó el 8,7%. En el caso de los datos de máximas de humedad relativa, se encontró algún tipo de valor atípico en un 5% de las series y un 13% en las series de humedad relativa mínima.

Conclusiones

De la experiencia con estos métodos de control de calidad se concluye que es muy importante tener experiencia en el manejo de datos para que dichas herramientas sean usadas e interpretadas adecuadamente. Las aplicaciones propuestas requieren de personal con conocimiento de los diferentes niveles de generación del dato, así como conocimiento de los sistemas de medición que lo generan, esto con el fin de estar en capacidad de explicar situaciones anómalas en los datos observados. También se observó que la aplicación de paquetes de computación científica es fundamental para visualizar situaciones atípicas que no pueden ser detectadas usando paquetes convencionales, por lo que un control de calidad efectivo debería usar este tipo de tecnologías al máximo.

El uso de estos programas permitió la manipulación de los archivos de los datos de una forma eficiente, pero no exime al analista de su interpretación de los reportes y gráficos generados. Esto es importante, ya que sin experiencia climatológica adecuada el usuario podría encontrar difícil tomar decisiones sobre la plausibilidad de los valores detectados. Asimismo, se observó que después de

aplicar los ACCD y las herramientas de visualización indicadas, en el caso datos horarios de temperatura generados por EMA, el porcentaje de datos atípicos encontrados es bajo. En el caso de los datos diarios de temperatura, la cantidad de datos atípicos es similar, aunque en una estación este porcentaje llegó, incluso, al 17% del total de datos de la serie.

Se han aislado diferentes circunstancias de generación de datos atípicos para estos datos: los máximos diarios en horas de la madrugada normalmente fueron asociados, y en ocasiones influyeron en el cálculo de la moda mensual, alterando los estadísticos, la serie. En el caso de los saltos atípicos, estos eran comunes y no necesariamente un salto entre horas consecutivas correspondía a un dato erróneo. El criterio usado fue considerar la hora en la que se generaba dicho salto, ya que los saltos consecutivos de mayor magnitud tienden a darse en horas en la tarde o de la mañana, dependiendo de la estación que se esté considerando. Algunas series mostraban modas inconsistentes que señalaban la presencia de datos atípicos enmascarados, los cuales debían detectarse para ser indicados.

El control de calidad en etapas posteriores al almacenamiento final es un excelente control cruzado que permite identificar deficiencias en el protocolo de procesamiento y control de calidad existente. Esto permitirá caracterizar mejor las condiciones bajo las cuales se pueden introducir inhomogeneidades y valores atípicos en las series.

En lo referente a las causas de generación de valores atípicos, pueden listarse las siguientes: aislamiento del detector térmico resistivo de temperatura, debido a la presencia de panales. (Suárez, 1996), voltaje bajo de la batería de la EMA y caducidad del tiempo de vida útil del instrumento. En el caso del higristor capacitivo usado para la medición de humedad relativa, se observó que hubo una tendencia a la generación de mayor cantidad de valores atípicos usando este instrumento.

Un posible tratamiento del problema para disminuir el porcentaje de valores atípicos generado es sugerido por Castro e Hidalgo (1997) y tiene que ver con el empleo de termistores, además de usar el método psicométrico para lidiar con el problema de estabilidad en la calibración de los higristores capacitivos. También es fundamental el reemplazo y revisión rigurosa, así como comparaciones o calibraciones regulares de los diversos instrumentos, tanto en el emplazamiento como antes de ser instalados. Otro problema detectado en algunas estaciones fue la combinación de series debido a cambios no documentados del emplazamiento. Esta situación fue corregida a raíz de los hallazgos de este estudio.

En este trabajo quedó claro que la inclusión de diferentes ACCD en conjunto con herramientas visuales permite detectar situaciones peculiares en los datos. Los procedimientos de control de calidad manuales pueden permitir que datos atípicos pasen sin ser detectados; de hecho, los datos atípicos mostrados en este estudio no fueron detectados al aplicar el método de revisión manual. Esta experiencia reitera que estudios de control de calidad como este permiten encontrar vulnerabilidades en los protocolos de revisión de datos. Revisiones periódicas de información almacenada en las bases de datos meteorológicos deberían ser parte de un protocolo efectivo de control de calidad a profundidad; esto, en las instituciones preocupadas por asegurar cierta calidad en la información de la que disponen.

Además, el control de calidad en etapas posteriores al almacenamiento final es un excelente control cruzado que permite identificar deficiencias en el protocolo de procesamiento y control de calidad existente. Esto permitirá caracterizar mejor las condiciones bajo las cuales se pueden introducir inhomogeneidades y valores atípicos en las series. Esto es también un punto importante por considerar para las compañías o empresas estatales que cuenten con bancos de datos que hayan basado sus controles de calidad en la inspección manual de los datos, ya que, según se ha visto, datos erróneos podrían haber pasado desapercibidos; esto, por las vulnerabilidades propias del método

de control manual y su dependencia de la habilidad del analista; de ahí, la importancia de aplicar un enfoque de control de calidad del dato basado en la revisión visual de la información, así como en la aplicación de ACCD sobre la serie.

Resulta de gran importancia que un control de calidad como este enfatiza la importancia del desarrollo de programas para tal fin, lo cual resulta ser una experiencia de desarrollo de capacidades; institucionalmente, se pueden desarrollar programas tomando en cuenta la experiencia del personal y las peculiaridades de la red y del equipo implicado, de modo que la experiencia de analistas experimentados puede ser incluida en dichos programas. También pueden ser aplicados de forma reiterativa por personal técnico entrenado para mejorar su trabajo de revisión de datos.

Agradecimientos

Se agradece al Instituto Meteorológico Nacional por el apoyo brindado en esta iniciativa. Especial gratitud a Rafael Pacheco, funcionario del Departamento de Redes y Procesamiento de Datos, por su empeño y dedicación para hacer que este trabajo sea posible. También se agradece al Ingeniero José Retana, por sus importantes sugerencias y comentarios.

Bibliografía

Araya, J. L., (2007). *Algoritmos de Control de Calidad de Datos en Estaciones Meteorológicas Automáticas.* Tesis de Licenciatura. Escuela de Física, Universidad de Costa Rica. 172 pp.

Castro, V.; Hidalgo, H., (1997). *Mediciones climáticas de humedad del aire en los trópicos, con termistores y la ecuación psicométrica*. *Top. Meteoro. Oceanogr*., 4(1): 91-94.

Feng, S.; Hu, Q.; Qian, W. (2004). *Quality Control of Daily Meteorological Data in China, 1951-2000: A new dataset*. *Int. J. Climatol*., 24: 853-870.

Gandin, L., S. (1988). *Complex Quality Control of Meteorological Data*. *Mon. Wea. Rev*., 116, 1137-1156.

Meek, D. W. & Hatfield, J. L. (2001). *Single Station Quality Control Procedures*. Automated Weather Stations for Applications in Agriculture and Water Resources Management.WMO/TD N.º 1074.

Montgomery, D. C. (2005). *Introduction to Statistical Quality Control.* 5th ed. J. Wiley & Sons.

OMM, 1992: Manual del Sistema Mundial de Proceso de Datos. Vol. II. Nº 485.

Peña, D. (2002). *Análisis de Datos Multivariantes*. 1 edición. Mc Graw Hill. 539pp.

Reek, T.; Doty R. & Owen T.W. (1992). *A deterministic approach to the validation of historical daily temperature and precipitation data. Bull Amer. Meteor. Soc*., 73, 753-762.

Spiegel, M. R. & Stephens, L. J. (2002). *Estadística*. 3ª edición. Mc Graw Hill.

Suárez, M., E. (1996). *Comparación de los datos generados por una estación meteorológica automática y una manual*. *Top. Meteoro. Oceanogr*. 3(2): 153-170.

Wilks, D.S. (1995). *Statistical Methods in Atmospheric Sciences*. Academic Press, Inc.