X Congreso Internacional AEC: Clima, sociedad, riesgos y ordenación del territorio
DOI: http://dx.doi.org/10.14198/XCongresoAECAlicante2016-38

CONTROL DE CALIDAD DE SIETE VARIABLES DEL BANCO NACIONAL DE DATOS DE AEMET

Miquel TOMAS-BURGUERA[1], Azucena JIMÉNEZ CASTAÑEDA[2], María Yolanda LUNA RICO[3], Ana MORATA[3], Sergio VICENTE SERRANO[4], José Carlos GONZÁLEZ HIDALGO[2], Santiago BEGUERÍA[1]

[1]*Estación Experimental de Aula Dei, Consejo Superior de Investigaciones Científicas (EEAD-CSIC).*
[2]*Departamento de Geografía y Ordenación del Territorio. Universidad de Zaragoza.*
[3]*Agencia Estatal de Meteorología (AEMET).*
[4]*Instituto Pirenaico de Ecología, Consejo Superior de Investigaciones Científicas (IPE-CSIC).*
mtomas@eead.csic.es, geoazu.flysch@gmail.com, mlunar@aemet.es, amoratag@aemet.es, svicen@ipe.csic.es, jcgh@unizar.es, sbegueria@csic.es

RESUMEN

Se ha realizado un control de calidad de datos diarios de precipitación, temperatura máxima, temperatura mínima, humedad relativa, velocidad del viento, presión atmosférica e insolación del banco nacional de datos de AEMET, para el periodo 1950-2015. Los controles realizados pertenecen a tres grupos: i) control de errores de codificación; ii) control de valores propios; y iii) comparación con vecinos. Entre los errores de codificación destacan las series de días consecutivos con un mismo valor, la repetición de cadenas de valores entre estaciones distintas, y la aparición de cadenas idénticas dentro de una misma serie. El control de valores propios consiste en detectar los valores que están fuera del umbral de la variable o que suponen un extremo sospechoso. En el control de vecinos se realiza una comparación de los datos con los registrados en estaciones vecinas para marcar posibles anomalías. Mientras que algunos controles implicarían la eliminación directa del dato, otros únicamente marcan el dato como sospechoso y requieren de un control posterior para determinar si el dato tiene que ser eliminado. Al finalizar el proceso un dato puede estar en 4 estados: dato inexistente, dato original, dato sospechoso, dato eliminado. En la aplicación de estos controles se han detectado un número elevado de errores de codificación, con afectaciones ligeramente por encima del 0.5% de la serie en temperatura y viento, mientras que los porcentajes son menores en el resto. Para la comparación con vecinos se ha testado un control con percentiles móviles a 365 días, obteniendo resultados poco satisfactorios

Palabras clave: control de calidad, precipitación, temperatura, humedad relativa, velocidad del viento, presión atmosférica, insolación

ABSTRACT

A quality control of maximum temperature, minimum temperature, relative humidity, wind speed, atmospheric pressure and sunshine duration data from the national database of AEMET, for the period 1950-2015 has been done. Controls can be grouped into three groups: i) codification controls; ii) local values controls; and iii) neighbor comparison. Among codification controls we can find the detection of con-

secutive days with repeated values, the repetition of values between distinct weather stations. The local values controls are focused on the outliers detection. Neighbor control is focused on the comparison of values against nearby data in order to detect anomalous values . Some controls are programmed to suppress automatically detected errors, while others only mark the detected error as a suspicious value. At the end of the process, a value can be in 4 distinct states: Non-available value, original value, suspicious value or suppressed value

Key words: quality control, precipitation, temperature, relative humidity, wind speed, atmospheric pressure, sunshine duration

1. INTRODUCCIÓN

Tal y como aparece recogido en la guía de buenas prácticas climatológicas de la OMM (WMO, 2011), el control de calidad es una parte fundamental de cualquier red de observación meteorológica.

En un intento similar al realizado por Durre *et al*. (2010) con los datos meteorológicos de Global Historical Climatology Network (GHCN)-Daily se ha desarrollado una metodología que consta de diversos controles automáticos para aplicar a los datos diarios del banco nacional de datos de AEMET.

2. MÉTODOS

El control de calidad de los datos se divide en tres grandes bloques: i) control de errores de codificación; ii) control de valores propios; y iii) comparación con vecinos. Cada uno de estos bloques se compone de distintos controles, algunos de los cuales se aplican a todas las variables mientras otros son específicos de una variable. Además, los controles que se aplican a todas las variables pueden tener distintos parámetros en función de la variable de interés. Aunque de manera general estos controles trabajan con las variables de manera independiente, en unos pocos casos la temperatura máxima y la temperatura mínima pasan controles comunes.

El control de calidad se realiza sobre los datos diarios. Sin embargo, para las variables humedad relativa, velocidad del viento y presión atmosférica, se han obtenido valores sub-diarios con lecturas a las 00, 07, 13 y 18 h UTC. Dado que la lectura a las 00 está ausente en un gran número de casos, se ha decidido calcular los valores diarios para estas variables a partir de la media de los valores de las 07, 13 y 18 h. Algunos de los controles que se explican a continuación se pasan tanto a los valores diarios como a los valores sub-diarios, debido a i) el valor diario puede no saltar en ningún control, pero estar construido a partir de uno o varios datos sub-diarios erróneos y ii) la construcción de la media diaria puede provocar que se den ‘n’ días consecutivos con un mismo valor medio, pero con diferentes valores sub-diarios, lo que se trata en realidad de una casualidad estadística y no de un error de codificación

En cuanto a los datos de temperatura, se detectó durante el proceso de control de calidad que la resolución de los datos varía de manera espaciotemporal, pudiendo aparecer datos diarios con una resolución de 0.1 ºC, 0.5 ºC o 1.0 ºC. Dado que el funcionamiento de algunos controles puede verse alterado por este motivo, en primer lugar se obtiene la resolución de los datos de temperatura a partir de los valores presentes en cada mes. Si en un mes en concreto, todos los valores de temperatura son

enteros, se entiende que para dicho mes la resolución es de 1.0 ºC. De manera similar se obtienen los meses con resolución a 0.5 ºC y a 0.1 ºC

Una vez se han llevado a cabo estas tareas iniciales de preparación de los datos se aplican los controles de calidad, empezando por la detección de errores de codificación. Es importante remarcar que el sistema se ha diseñado de tal manera que para cada control se almacena un archivo de metadatos con los resultados del control y estructura similar al archivo de datos original. De esta manera se puede reconstruir todo el proceso, para cada dato.

2.1. Control de errores de codificación

Los controles de este apartado están encaminados a detectar y eliminar errores de codificación que puedan existir en la base de datos. Se consideran errores de codificación los valores que están en una estación meteorológica que no les corresponde; aquellos valores que estando en una estación meteorológica correcta, están en un mes erróneo; las cadenas de 'n' días con un valor repetido; los valores 0 que se corresponden en realidad con la ausencia de dato; y los valores que se encuentran en unidades incorrectas. Para la detección de todos estos casos se han programado los siguientes controles específicos.

Detección de meses duplicados. Se realiza una búsqueda de meses que tengan todos los valores idénticos, ya sea dentro de una misma serie o entre series diferentes. La existencia de meses completos idénticos se considera altamente improbable, y causada por errores de asignación durante el proceso de digitalización de los datos. Debido a la imposibilidad de determinar cuales son los datos correctos y cuales los repetidos, se procede a la eliminación de todos ellos y a su marcado en los metadatos. Únicamente se conservan los valores duplicados que se detectan entre estaciones meteorológicas que se encuentran muy próximas entre sí (menos de 1km), al considerarse que, aún no conociendo a qué estación pertenecen dichos datos, mantener el dato ofrece información climática válida para ese entorno geográfico. Este control se aplica a todas las variables por igual.

Detección de decenas duplicadas. Se trata una búsqueda similar a la anterior para períodos de 10 días (en lugar de por meses completos), y únicamente dentro de la serie de una misma estación meteorológica. El objetivo principal de este control es la detección de duplicidades que puedan aparecer entre decenas climáticas. Este control se aplica a todas las variables por igual.

Detección de 'n' días consecutivos con el mismo valor. Se realiza una búsqueda de secuencias de 'n' días consecutivos con mismo valor, es decir sin ninguna variación. Este control se aplica a todas las variables por igual, variando el valor de 'n' en función de la variabilidad natural de cada variable. Por defecto se utiliza un valor de n = 7, aunque en algunos casos se incrementa para evitar un exceso de falsos positivos. Por ejemplo, para series o trozos de series en las que la temperatura está codificada con una resolución de 0.5 o de 1 ºC se toma el valor n = 15. En el caso de la humedad relativa, y cuando esta toma valores por encima del 95%, se utiliza n = 20. Esto se debe a que humedades superiores al 95% pueden ser indicativas de presencia de nieblas, y un período de nieblas persistentes puede mantener fijado el valor de la humedad mientras persista la niebla. Dado que en España hay zonas con un elevado número de días de niebla al año (Cermak *et al* 2009), y después de probar con distintos valores de n, se ha decidido fijarlo a 20. En el caso de la insolación se ha utilizado un

valor de n = 10, excepto cuando ésta toma el valor de 0 en cuyo caso se utiliza n = 15. Además, si los valores de insolación que llevan a que salte el control están muy cerca del máximo teórico de horas de sol para esa localidad y momento del año se considera que el valor es correcto. Estos valores se resumen en la Tabla 1.

Variable	Valor[a]	n
Temperatura	0.1ºC	7
	0.5ºC o 1.0ºC	15
Humedad relativa	< 95%	7
	> 94%	20
Viento		7
Insolación	>0	10[b]
	0	15
Presión		7

Tabla 1: Valor de n en la detección de días consecutivos

a) Hace referencia al valor de la variable. En temperatura se refiere a la resolución de la medida
b) Se compara con el máximo teórico de horas de sol

Detección de 'n' días consecutivos con una variación del valor inferior a un umbral. Este control busca cadenas de 'n' días consecutivos en los que la diferencia entre el valor máximo y el valor mínimo esté por debajo de un umbral determinado. Es un control específico para humedad relativa, ya que durante el proceso se detectaron períodos de tiempo prolongados con una variación inusualmente baja del valor de la variable. Se consideran como erróneos aquellos periodos de 7 días o más, que presentan una variación inferior al 3% de humedad, siempre que esta esté por debajo del 95%, ya que de nuevo, las situaciones de niebla podrían estar forzando esta poca variabilidad del dato durante cortos períodos de tiempo.

Detección de 0's erróneos. En las bases de datos se codifican a veces 0's que se corresponden en realidad con la ausencia de dato. Este control se encarga de ello, aplicando una estrategia diferenciada para cada una de las variables en las que se aplica. En temperatura, tanto máxima como mínima, este control se lleva a cabo a partir de los datos de la propia estación, combinando la escala diaria con la mensual. Si eliminados todos los 0's que aparecen en un mes en concreto, la mínima de las temperaturas restantes es superior a 10ºC, se consideran falsos los 0's de ese mes. En insolación, el control de 0's se hace a partir de los vecinos más cercanos, a una distancia inferior a 50km, considerando que si la insolación es superior a 5h en 3 de los vecinos más cercanos, ese valor de 0 es erróneo.

Detección de meses mal codificados. Se detectan valores sospechosos de estar codificados en unidades incorrectas. Se aplica únicamente a temperatura e insolación. En el caso de la temperatura máxima el control salta si la temperatura máxima del mes es inferior a 4 ºC, y además la variable se mueve en un rango inferior a 3 ºC. En el caso de la temperatura mínima el control salta si la temperatura se encuentra entre –2 ºC y +2 ºC durante todos los días del mes, y el rango es inferior a 3 ºC. Para el caso de la insolación se buscan períodos de 15 días con una insolación máxima inferior a 2 horas de sol. Todos estos casos se consideran mal codificados y se eliminan.

2.2. Control de valores propios

Se lleva a cabo un control de los valores absolutos de la variable, día a día, que se realiza en dos pasos. En el primer paso los valores diarios se comparan con los registros extremos que se pueden llegar a considerar válidos en España (para detectar los casos que se descartan automáticamente), y en un segundo paso se compara con registros extremos que son posibles, pero que se considera que requieren una verificación posterior. En la Tabla 2 aparecen los valores que se han utilizado en ambos controles. Aunque en un principio los valores detectados no se eliminan (sólo se marcan como sospechosos), para el caso del viento hay un caso especial, que es la repetición sistemática de 3 valores específicos (108, 144 y 180), y que parece responder a un error. Cuando de los tres valores sub-diarios (07, 13, 18), al menos 2 contienen uno de estos tres valores se considera el dato erróneo y se elimina de la base de datos.

2.3. Comparación con vecinos

En un intento de poder aplicar el control de vecinos al máximo número de estaciones meteorológicas presentes en la base de datos, se ha programado un control basado en el cálculo móvil a 365 días del valor percentil y su comparación con el mismo valor registrado en las 5 estaciones más cercanas, siempre y cuando estas estaciones se encuentren a una distancia inferior a 50 km. Si la estación objetivo tiene un salto percentil superior a un umbral con al menos 3 de las 5 estaciones más cercanas, según este criterio se considera que el valor es erróneo. Un valor muy elevado de diferencia percentil garantiza que todos los casos que se detectan son erróneos, a costa de únicamente detectar los casos más extremos (p.ej: temperatura típica de invierno en un día de verano). A medida que decrece el umbral, la capacidad de detección se incrementa, pero a costa de aparecer falsos positivos.

Variable	**Aberrantes**	**Sospechosos**
Temperatura	> + 50ºC	Tmax > +45ºC
	< -35ºC	Tmax < -10ºC
	Tmax <= Tmin	Tmin < -25ºC
	Tmax – Tmin > 35ºC	Tmin > +30ºC
		Tmax – Tmin > +25ºC
Humedad relativa	< 0	<10 13,18 UTC
	> 100	<15 media diaria
		<20 07UTC
Viento	< 0	>150
	>220	= 108 ò =144 o =180
Insolación	<0	
	>máximo teórico	
Presión	>1055	
	<725	

Tabla 2. Límites utilizados en detección de aberrantes y valores sospechosos

3. RESULTADOS

En la tabla 3 se presenta el tamaño de la base de datos que se ha tratado, siendo las variables más abundantes las temperaturas máxima y mínima, que son además las que tienen un periodo de registro más amplio. El último año completo del que se tienen datos es 2014, apareciendo algunos meses de datos en 2015.

Variable	**Nº registros (meses)**	**Nº estaciones**	**Año inicio**
Tmax	1.172.684	5.111	1856
Tmin	1.172.684	5.111	1856
Humedad	160.617	1.132	1916
Viento	155.866	980	1916
Insolación	89.081	354	1917
Presión	88.641	393	1916

Tabla 3. Tamaño de la base de datos

Para los meses enteramente duplicados, la temperatura máxima, temperatura mínima y velocidad del viento tienen un porcentaje similar de la serie afectada, con un 0.53%. La variable que presenta un menor grado de afectación es la humedad relativa, con sólo un 0.056%. En cuanto a la evolución temporal de las duplicidades en temperatura máxima y temperatura mínima se observa un pico muy pronunciado que se corresponde con el año 1984 (Fig. 1). La duplicidad que más se repite en todos los casos es aquella que se produce entre dos estaciones distintas. En temperatura y en insolación existen algunos casos en los que hasta 3 estaciones aparecen implicadas en una duplicidad.

			Nº de casos en función del nº de estaciones afectadas		
	Duplicados	Porcentaje	1	2	3
Tmax	6.308	0,538%	960	5.342	6
Tmin	6.204	0,529%	953	5.235	16
Humedad	90	0,056%	26	64	0
Viento	830	0,532%	24	806	0
Insolación	269	0,301%	10	256	3
Presión	74	0,083%	4	70	0

Tabla 4. Número de registros (meses) que saltan con el control de meses duplicados

Un caso específico de duplicidad lo conforman aquellas estaciones que, teniendo las mismas coordenadas, presentan algunos meses duplicados entre sí. La hipótesis más plausible es que se trate de casos en los que coexisten, en un mismo jardín meteorológico, estaciones meteorológicas tradicionales, con estaciones meteorológicas automáticas, y que ante el fallo de una de las dos, se procede al relleno con el dato de la otra.

	Número de días	Número de casos	Estaciones afectadas
Tmax	2.615	191	115
Tmin	3.550	188	101
HR < 95%	141	11	10
HR > 94%	328	9	8
Viento	6.523	1.902	152
Insolación	276	26	14
Presión	159	15	6

Tabla 5. Estadísticas del control de días consecutivos con el mismo valor

Para las secuencias de valores idénticos (Tabla 5), la variable más afectada es el viento. Se debe al elevado número de casos en que hay más de 7 días consecutivos con viento a 0 km/h. En un principio estos casos se han considerado erróneos.

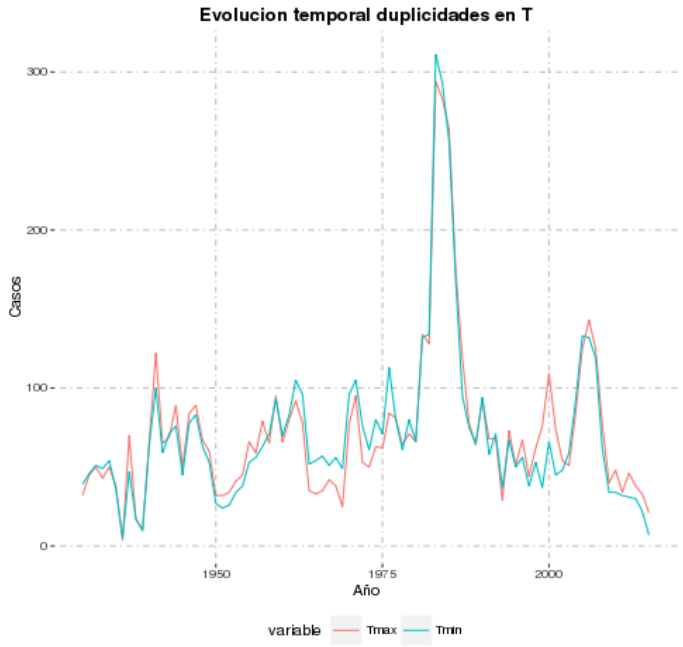

Fig. 1: Evolución temporal de las duplicidades en temperatura máxima y temperatura mínima.

El resto de controles aplicados tienen una detección menor. Como curiosidad, de los llamados controles aberrantes, únicamente se han detectado casos en temperatura.

En la tabla 6 se adjuntan el total de días eliminados para cada una de las variables y el número de estaciones afectadas. El porcentaje de afectación se encuentra siempre por debajo del 1%, siendo más elevado en temperatura máxima, con 0,85%, mientras que para la temperatura mínima el porcentaje es sólo del 0,57%. La humedad es la variable que presenta un menor porcentaje de afectación, con sólo un 0,10% del total. El elevado porcentaje de datos eliminados en insolación se debe a un error de codificación que afecta a diversas estaciones a principios de los años ’70, y del que

se presenta un caso particular en la figura 2, en la que se puede ver como entre los años 1969 y 1970 los valores son exageradamente inferiores a los del resto de la serie.

	Días con dato	Días eliminados	%	Estaciones afectadas
Tmax	35.146.529	298.614	0,85	3.129 (5.111)
Tmin	35.124.740	199.108	0,57	3.081 (5.111)
HR	4.606.384	4.626	0,10	92 (1.132)
Viento	4.460.299	19.929	0,44	190 (980)
Insolación	2.654.922	16.036	0,60	36 (354)
Presión	2.612.967	2.379	0,09	24 (393)

Tabla 6. Estadísticas de eliminación

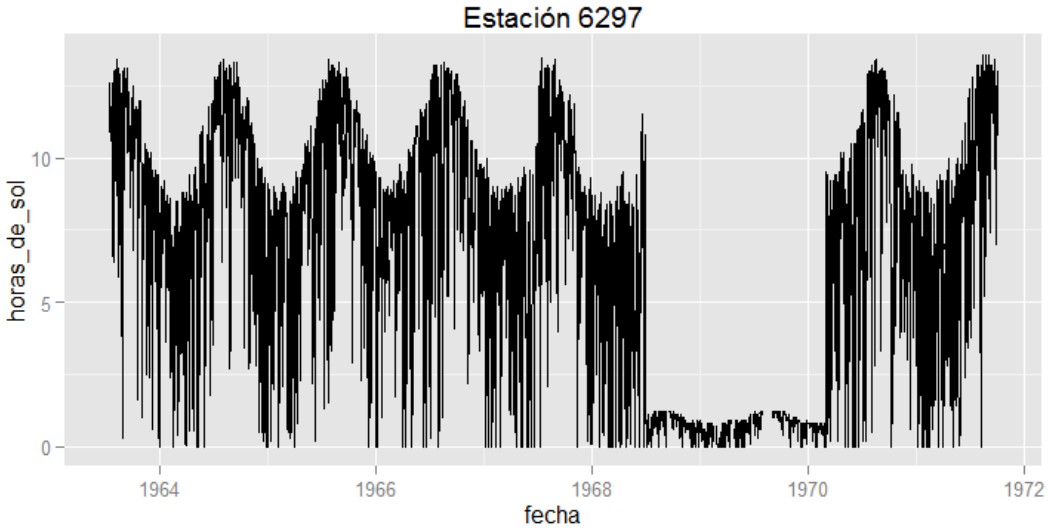

Fig. 2: Ejemplo de error de mala codificación en insolación en la estación 6297

4. DISCUSIÓN

Debido al tamaño de la base de datos se ha pretendido montar un sistema que, una vez implementado, permita llevar a cabo un control de calidad automático, del que, por una parte salen los datos meteorológicos y por otra parte salen los metadatos, que permiten conocer en todo momento que controles han saltado y cuáles no.

Aunque el control programado intenta detectar el máximo de errores que pueden aparecer en una base de datos de este estilo, hay controles, como el testado para detectar anomalías a través de los datos de los vecinos, que requieren ser reformulados para cumplir plenamente con su objetivo. Este control se pensó para ser aplicado al mayor número de series de datos posible, inclusive aquellas que tienen un período de datos muy corto, y que difícilmente permiten la aplicación de series de diferencias o series

de referencia. Sin embargo, la baja capacidad de detección de este control recomienda que, de cara al futuro, se evalúen otras opciones.

De hecho, se considera que la mayor contribución de este trabajo reside en la gran variedad de controles sobre errores de codificación que se han aplicado a los datos, y que permiten descartar un gran número de datos erróneos, existiendo todavía un margen de mejora, ya sea por la introducción de nuevos controles, por la modificación de los existentes o por la recuperación de parte de los datos que ahora se descartan.

Así mismo, resultaría interesante llevar a cabo trabajos dirigidos a detectar cuál es la afectación de los datos erróneos al cálculo de distintos índices climáticos.

AGRADECIMIENTOS

Este trabajo ha sido posible gracias a la financiación del proyecto CGL2014-52135-C03-01. El trabajo de Miquel Tomas-Burguera ha sido posible gracias a una beca predoctoral FPU del Ministerio de Educación, Cultura y deporte.

Los autores agradecen a AEMET la cesión de la base de datos para la realización de este estudio.

REFERENCIAS

Cermak, J., Eastman, R.M., Bendix, J. y Warren S.G. (2009). European climatology of fog and low stratus based on geostationary satellite observations. Q. J. R. Meteorol. Soc. (135). 2125-2130. DOI: 10.1002/qj.503

Durre, I., Menne, M.J., Gleason, B.E., Houston, T.G., Vose, R.S. (2010). Comprehensive Automated Quality Assurance of Daily Surface Observations. Journal of Applied Meteorology and Climatology. (49). 1615-1633

Eching, S.O. y Synder, R.L. (2004). Statistical control charts for quality control of weather data for reference evapotranspiration estimation. Acta Hortic. 664, 189-196. DOI: 10.17660/ActaHortic.2004.664.21

WMO. Guide to climatological practices. Ginebra. 2011. Recuperado de http://www.wmo.int/pages/prog/wcp/ccl/guide/documents/WMO_100_en.pdf