EFECTO DE LA VALIDACIÓN DE DATOS DE PRECIPITACIÓN EN EL ANÁLISIS REGIONAL EN LA PROVINCIA DE MÁLAGA

Javier Estévez

Amanda P. García-Marín

María T. Medina

José L. Ayuso

Departamento de Ingeniería Rural. Área de Proyectos de Ingeniería. Universidad de Córdoba.

Abstract

The province of Malaga, located in Southern Spain, is characterized by irregular rainfall Mediterranean climate. In this work, two regional frequency analyses of maximum daily annual precipitation data have been carried. One of them was conducted using the raw data of daily precipitation provided by the Spanish 'Confederación hidrográfica de la Cuenca Sur', and the other was conducted after applying various quality control tests to this dataset. As it is known, all sources of weather information have an associated uncertainty that must be considered. These data validation procedures are used to eliminate possible errors in the climatic series and if it is possible, correct them. In our case, we have identified potentially erroneous data not to be taken into account in the second regional frequency analysis. Finally, the results obtained using initial precipitation series (raw data) were compared to those obtained with validated data series.

Keywords*: all separated by semicolons; use italics; use lower–case; minimum 3 words, maximum 6 words; without period at the end*

Resumen

El presente trabajo se enmarca en una zona caracterizada pluviométricamente por la irregularidad del clima Mediterráneo, como es la provincia de Málaga (Sur de España). Se ha llevado a cabo un análisis regional de frecuencias de los datos de precipitación máxima diaria anual antes y después de aplicar diversos tests de control de calidad a los datos brutos de precipitación diaria suministrados por la Confederación hidrográfica de la Cuenca Sur. Como es sabido, toda fuente de información meteorológica lleva una incertidumbre asociada que debe ser considerada. Estos procedimientos de validación de datos se emplean para eliminar posibles errores en las series climáticas y en su caso, corregirlos. En nuestro caso, se han identificado datos potencialmente erróneos para no ser tenidos en cuenta en el análisis regional. Finalmente, se han comparado los resultados obtenidos utilizando la serie de precipitación inicial (datos brutos) y la serie de precipitación validada.

Palabras clave: *validation; precipitation; regional frequency analysis*

1. Introducción

Toda fuente de información meteorológica lleva una incertidumbre asociada que debe ser considerada en cualquier serie climática (Cuadrat et al., 2002). Por ello, como requisito previo a la utilización de los registros de lluvia, es necesario un depurado previo de las series de datos, así como un control de calidad de los mismos (Estévez et al., 2011), para lo que se aplicarán un conjunto de test de validación (Feng et al., 2004; Hubbard et al., 2005).

El control de calidad es un concepto muy amplio que comienza con la elección adecuada de la ubicación de cada estación meteorológica (WMO, 1993). Posteriormente resulta esencial un correcto mantenimiento de la misma y la calibración periódica de los sensores que se encuentren instalados en ella. Finalmente la validación de los datos generados engloba un conjunto de técnicas, procedimientos, algoritmos y tests que sirven como herramientas para la identificación y detección de errores, asegurando la fiabilidad de dichas observaciones (Estévez et al., 2011). Existen sistemas de validación potentes donde se inserta un dato informativo que se conoce con el nombre de “flag” o bandera. Generalmente son dígitos que equivalen a descripciones del tipo “dato bueno”, “dato sospechoso”, “dato erróneo”, “sin dato”, etc. en función de las alertas que generen cada uno de los test aplicados (Shafer et al., 2000). Se trata por tanto de generar una información añadida al registro meteorológico que sirva para describir el nivel o grado de confianza de ese valor. En el proceso de validación siempre hay una etapa final de monitorización o análisis manual que es efectuado por personal cualificado, con capacidad para decidir si ciertos valores o registros que son potencialmente erróneos están asociados a fenómenos meteorológicos poco habituales como tormentas, olas de calor, etc. (Graybeal et al., 2004). Por esta razón, aunque los procedimientos de validación se automaticen, siempre existirá una toma de decisiones del equipo investigador, como paso previo a la utilización de los datos.

La mayoría de las estaciones meteorológicas españolas cuentan con registros de series de datos demasiado cortas. El inconveniente de utilizar o disponer de series no muy extensas de datos puede soslayarse empleando técnicas relativamente recientes, como la del análisis regional de frecuencias (Hosking & Wallis, 1997). Este enfoque permite paliar el problema de la carencia de datos en el tiempo con la abundancia de datos en el espacio, siendo actualmente la tendencia generalizada en el análisis de frecuencias de eventos extremos (Álvarez et al., 1999). Diferentes trabajos avalan la regionalización como técnica que mejora las estimaciones de los cuantiles al trabajar con lluvia o con caudal (Sáenz de Ormijama et al., 1991; Hosking & Wallis, 1997; Parida et al., 1998; Yun & Chen, 1998; Ferrer & Mateos, 1999; Chiang et al., 2002a,b; Garcia-Marin et al., 2011). Dentro de la regionalización, la determinación de regiones homogéneas es el paso más complejo y de su resultado dependerá la continuidad del análisis. Por ello, el objetivo de este trabajo es la comprobación de la influencia de la validación de datos de lluvia en los resultados del test de homogeneidad dentro del AR.

2. Metodología

2.1. Tests de Validación

Los principios básicos de la mayoría de los tests de validación que se aplican utilizando datos registrados en una sola estación provienen de las tres reglas introducidas por Meek & Hatfield (1994) y que están basadas en O'Brien & Keefer (1985). Estas reglas son: límites fijos o dinámicos para cada variable (lo que se conoce habitualmente como test de rango), límites fijos o dinámicos para los cambios entre observaciones sucesivas (se conoce con el nombre de Step test o test de consistencia temporal) y por último, límites para detectar medidas consecutivas que son iguales o su variabilidad es baja (se conoce con el nombre de test de persistencia).

Para el caso de nuestros datos de precipitación, se han aplicado un test de rango fijo (Shafer et al., 2000; Feng et al., 2004; Estévez et al., 2011) y la comparación con efemérides (AEMET, 2011) y se ha seguido la metodología desarrollada por Hubbard et al. (2005), utilizando los test de rango dinámico y test de persistencia y que son tests que están basados en decisiones estadísticas.

2.1.1. Test de Rango dinámico

Este test consiste en llevar a cabo una verificación donde se compruebe que cada dato de precipitación registrado se encuentra dentro de un rango específico para cada mes del año. Siendo x la precipitación diaria que queremos validar, la Eq.1 muestra la formulación del test.

$$\overline{x} - f\sigma_x \leq x \leq \overline{x} + f\sigma_x \tag{1}$$

donde $\overline{x}$ es el promedio diario, σ_x es la desviación típica de los valores diarios para cada mes y f un factor que va desde 0.2 a 5.0 (Hubbard et al., 2005). Este procedimiento indica que para valores grandes de f (factor), el número de outliers potenciales disminuye.

2.1.2. Test de Persistencia

Este test se basa en la hipótesis de que cuando el sensor falla siempre registra un valor constante, de manera que la desviación típica será menor. En otros casos, cuando el sensor trabaja intermitentemente, se registrarán valores razonables y valores cercanos a cero, obteniendo una desviación típica muy elevada. Esto quiere decir que cuando la variabilidad se salga fuera de unos límites determinados según Eq. 2 los valores serán etiquetados como sospechosos y por tanto, se considerarán potencialmente erróneos.

$$\overline{\sigma}_j - f\sigma_{\sigma_j} \leq \sigma_j \leq \overline{\sigma}_j + f\sigma_{\sigma_j} \tag{2}$$

donde σ_j es la desviación típica de los valores diarios para cada mes (j) y año, y σ_{σ_j} es la desviación típica de σ_j de cada mes en cuestión.

2.2. Análisis regional

El objetivo fundamental del análisis de frecuencias es la estimación de sucesos extremos correspondientes a diferentes períodos de retorno mediante el uso de funciones de distribución de probabilidad. La regionalización se utiliza normalmente en hidrología para facilitar la extrapolación desde lugares en los que existen registros a lugares donde se requieren pero no están disponibles, o lo están en menor cantidad. Dividiendo la zona de estudio en subregiones homogéneas de similar comportamiento hidrológico, los registros

pueden extrapolarse con más precisión, y las ecuaciones que se deduzcan a partir de las características de la cuenca podrán utilizarse con una mayor confianza a la hora de predecir determinadas variables hidrológicas (Nathan & McMahon, 1990).

El análisis regional de frecuencias permite calcular datos para un determinado sitio de interés utilizando datos de otros lugares diferentes a los del sitio en cuestión. Si se cuenta con *N* sitios o estaciones cada uno de ellos con n años de registros de eventos máximos, puede suponerse que *N x n* datos de la región darán estimaciones más precisas de cuantiles tan extremos como Q_{Nn}.

Para la estudiar la homogeneidad de una debe analizarse el valor de la discordancia de las estaciones que a priori pueden formar parte de ella. La medida de la discordancia Di permite identificar estaciones inusuales en comparación con el resto de las que componen la región de estudio, para lo que se calculan los momentos lineales de variación (LC_v), sesgo (LC_s) y curtosis (LC_k) de las series de datos disponibles en cada lugar considerado. Se considera que el vector de momentos lineales de una estación i constituye un punto en un espacio tridimensional. Un grupo de estaciones producirá una nube de puntos en este espacio de forma que cualquier punto que se ubique lejos del centro de gravedad del conjunto de estos, deberá ser considerado como discordante. Numéricamente la medida de la discordancia viene dada por:

$$D_i = \frac{1}{3} N (u_i - \bar{u})^T A^{-1} (u_i - \bar{u}) \tag{3}$$

Siendo $A = \sum_{i=1}^{N} (u_i - \bar{u})(u_i - \bar{u})$, $\bar{u} = N^{-1} \sum_{i=1}^{N} u_i$ y $, u_i = \left[LC_v^i, LC_s^i, LC_k^i \right]$.

Una vez identificadas y, en su caso, eliminadas las estaciones discordantes, el valor de la homogeneidad de la posible región vendrá dado por

$$H = \frac{(V - \mu_v)}{\sigma_v} = \frac{\left(\left\{ \sum_{i=1}^{N} n_i (t^{(i)} - t^R)^2 / \sum_{i=1}^{N} n_i \right\}^{1/2} - \mu_v \right)}{\sigma_v} \tag{4}$$

siendo $t^R = \sum_{i=1}^{N} n_i t^{(i)} / \sum_{i=1}^{N} n_i$, N el número de estaciones, n_i el de registros y $t^{(i)}$ los momentos lineales muestrales.

Una región podrá considerarse homogénea si $H < 1$, posiblemente heterogénea si $1 < H < 2$, y heterogénea para valores de *H* superiores a 2 (Hosking & Wallis, 1997).

3. Datos

Para el desarrollo de este trabajo se han utilizado los datos existentes en 72 estaciones de la provincia de Málaga (Figura 1), suministrados por la Confederación Hidrográfica de la Cuenca Sur. El número de años para los que existen datos varía de unas estaciones a otras (tabla 1), considerándose sólo aquellos lugares que disponen de series de datos superiores a diecisiete años consecutivos.

Figura 1. Situación de las estaciones utilizadas en el análisis

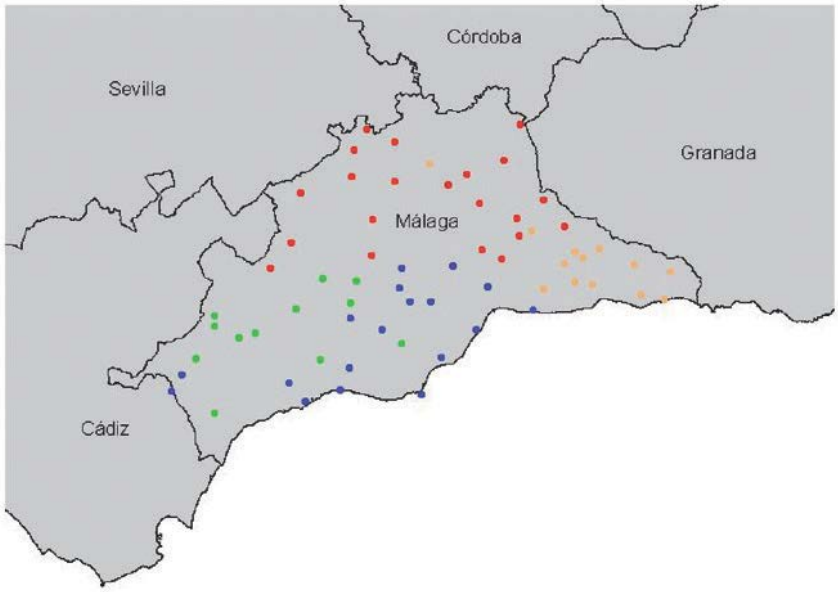

Tabla 1. Estaciones utilizadas

Agujero	Benahavís	Casaber. VP	Fte. P. Herr.	Parauta	Viñuela
Alcaucín Cjo.	Benalmádena	Casapalma	Govantes	Parchite	La Yedra
Alcaucín For.	Benamargosa	Casarabonela	Humilladero	Peña	
Alfarnate	Benamocarra	Casares	Hundidero	Periana	
Alhaurín	Benaoján	Chorro	Istán	Pizarra	
Aljaima	Bobadilla	Coín	Jimena	Rincón V.	
Almargen	Borregos	Colmenar	Las Mellizas	Riogordo	
Almogia	Buitreras CE	Cómpeta	Málaga	Ronda CE	
Álora	Buit. Presa	Conde	Marbella	SP Alcántara	
Alozaina	Campillos	Contaderas	Mijas	Tolox	
Alpandeire	Canillas	Corchado	Moclinejo	Torrox	
Antequera	Cartajima	Cuevas	Montejaque	Vegueta	
Archidona	Cártama	El Burgo	Nerja	Vélez	
Arriate	Casabermeja	Fte. Piedra	Ojén	Villanueva	

4. Resultados

4.1. Tests de validación

El primer test aplicado ha sido el test de rango fijo que consta de dos comprobaciones, una primera donde se verifica que no existen valores inferiores a 0 ni superiores a 508 mm de precipitación en la serie de datos inicial (no se ha encontrado ningún dato fuera de este rango), y otra donde se han utilizado los valores de efemérides obtenidos de la Agencia Estatal de Meteorología (AEMET), con el fin de seleccionar aquellos datos de la serie que superen estos valores extremos, que son los máximos de cada mes en las series históricas de registros de datos de precipitación de la provincia de Málaga.

Posteriormente se ha aplicado el test de rango dinámico, en el que se han usado distintos valores de *f* siguiendo la Eq.1, haciéndolo variar desde 0.2 hasta 5.0 en cada una de las estaciones meteorológicas del área de estudio. De esta forma se ha establecido una relación entre el número de errores potenciales y el factor *f* empleado. Como ejemplo, se muestra en la figura 2, la aplicación de este test a los datos de la estación Agujero, donde se puede comprobar que utilizando un f de 4.2 el test marca el 1.5% de los datos como potencialmente erróneos.

Figura 2. Fracción de datos marcada (log) en función de los valores de f para el test de rango dinámico en la estación Agujero

Por último se ha aplicado el test de persistencia, siguiendo la Eq.2. La aplicación de este test relaciona un conjunto de datos marcados como potencialmente erróneos, en función de los distintos valores de f aplicados, que oscilarán entre 0,2 y 5, igual que en el test de rango dinámico explicado anteriormente.

Ambos tests, basados en decisiones estadísticas, han marcado como sospechosos un conjunto de datos de precipitación en base a un grado de restricción (factor f utilizado).Como toma de decisiones dentro de la validación de la serie de precipitación inicial se ha seleccionado, en primer lugar, un valor de f que marque como potencialmente erróneos el 1,5% de los datos (Estévez et al, 2011), tanto para el test de rango dinámico como para el test de persistencia. Este proceso ha sido llevado a cabo en las 72 estaciones del área de estudio, obteniéndose un conjunto de datos marcados por ambos tests para cada una de las localizaciones.

Posteriormente se ha realizado la comprobación de que los datos marcados en ambos tests, es decir, los que se encuentran dentro de ese 1,5% de datos potencialmente erróneos, coinciden o no con los datos que no han superado satisfactoriamente el test de rango fijo basado en las efemérides obtenidas de la AEMET (Málaga Aeropuerto).

Continuando con el proceso de validación y en base a García Marín et al. (2011), se ha tenido en cuenta la diferencia entre aquellas estaciones que se encuentran en la misma región homogénea que Málaga Aeropuerto, tomada como estación de referencia por la presencia de datos en la AEMET, de aquellas que se encuentran fuera de ella, por tanto, situadas a más distancia de Málaga, en las que las condiciones meteorológicas pueden oscilar notablemente.

Las estaciones incluidas en la región homogénea son Agujero, Aljaima, Almogia los Llanes, Alora, Benahavis, Benalmádena, Buitreras CE, Cártama, Coín, Corchado Central, Málaga Oficina y Marbella Instituto Laboral.

Para la depuración final de la serie de precipitación inicial se ha tenido en cuenta, por tanto, el criterio de que en las estaciones cercanas a Málaga Aeropuerto y, por tanto, situadas en dicha región, se descartarán los datos marcados como potencialmente erróneos por el test de rango dinámico, el test de persistencia y el test de rango fijo empleando las efemérides de la AEMET, obteniéndose un rango de datos marcados entre 1 y 5 valores para cada una de las estaciones. Para el caso de aquellas estaciones situadas fuera de la citada región los datos que no se tendrán en cuenta serán los marcados por el test de rango dinámico y el test de persistencia únicamente, en cuyo caso se han obtenido un rango de 15 a 20 datos, como máximo, para cada una de las estaciones.

4.2. Análisis regional

La región de estudio sobre la que se quiere analizar la posible homogeneidad desde el punto de vista de las precipitaciones máximas diarias es la provincia de Málaga y dentro de ella 72 estaciones se han tenido en cuenta (Tabla 1).

El primer paso consiste en la búsqueda de estaciones discordantes dentro del conjunto considerado. Para ello ha de obtenerse el valor de Di y compararse con un valor crítico que depende del número de estaciones consideradas en el análisis. Para más de 15 estaciones ese valor es de 3 (Hosking & Wallis, 1997). Todas las estaciones para las que se obtenga un valor de Di superior a este límite, deberán eliminarse del análisis y por lo tanto, de la posible región.

Al trabajar con las series de datos máximos diarios sin validar el conjunto de datos de partida, ninguna de las estaciones aparece como discordante. Sin embargo, Alozaina, Alpandeire y Rincón de la Victoria son estaciones discordantes al trabajar con series de datos validadas, resultando unos valores de Di de 3.52, 3.30 y 5.81 respectivamente. Por lo tanto, estas tres estaciones deben eliminarse del análisis para obtener el valor del estadístico H de la región de Málaga.

La tabla 2 muestra los valores de H para las dos opciones consideradas. Como puede observarse, en ningún caso la región de Málaga al completo puede considerarse homogénea desde el punto de vista de las precipitaciones máximas diarias. Tanto con los datos sin validar como con los datos validados, los valores obtenidos del estadístico H superan el valor de 2, indicando la heterogeneidad de la región de estudio.

Tabla 2. Valores del estadístico *H*

Datos sin validar	4.45

Datos validados	4.44

Para conseguir valores inferiores de este estadístico con ambos conjuntos de datos, la región de partida habría de dividirse en subregiones (e.g. Garcia-Marin et al., 2011) y volver a repetir tanto el análisis de discordancia como el de homogeneidad.

5. Conclusiones

La validación de datos de precipitación empleando los tests comentados anteriormente ha permitido garantizar la calidad de los registros de lluvia como paso previo al análisis regional de frecuencias llevado a cabo en la provincia de Málaga. En el presente trabajo, estos tests han marcado como potencialmente erróneos un conjunto de datos y mediante diversos criterios se han eliminado un subconjunto de ellos. Como primera aproximación para comprobar el efecto de la validación de datos en el análisis regional de frecuencias se ha seleccionado, para cada estación, un máximo del 1.5% de datos de la serie inicial de precipitación como posibles errores, y por tanto, descartables para su posterior análisis.

A la vista de los resultados, no existen grandes diferencias entre los valores obtenidos con las series de datos validadas respecto a los obtenidos con las series sin validar. En este sentido, cabe destacar que los criterios para eliminar datos en los procedimientos de validación han poco restrictivos. Sin embargo, al haber eliminado los datos erróneos la caracterización estadística de las series de datos máximas es más correcta y por lo tanto, se mejorarán los resultados que puedan obtenerse al aplicar técnicas formación de subregiones.

6. Bibliografía

AEMET, 2011. Resumen de extremos climatológicos en España. <http://www.aemet.es/documentos/es/divulgacion/resumen_efemerides/Resumen_extremos.pdf> (02.12.11)

Álvarez, M., Puertas, J., Soto, B. and Díaz, F. (1999). Análisis regional de las precipitaciones máximas en Galicia mediante el método del índice de avenida. Ingeniería del Agua, 6 (4), 379-386.

Chiang S-M, Tsay T-K and Nix YS. (2002a). Hydrologic regionalization of watersheds. I: Methodology Development. Journal of Water Resources Planning and Management 128 (1), 3-11.

Chiang S-M, Tsay T-K and Nix YS. (2002b). Hydrologic regionalization of watersheds. II: Applications. Journal of Water Resources Planning and Management 128 (1), 3-11.

Cuadrat, J.M., Vicente, S.M., Saz, M.A., 2002. Fuentes de información en climatología: Incertidumbres de las series de datos climáticos en España. VII Reunión Nacional de Climatología, 27–29 June 2002, Albarracín (Zaragoza), Spain.

Estévez, J., Gavilán, P., Giráldez, J.V., 2011. Guidelines on validation procedures for meteorological data from automatic weather stations. Journal of Hydrology, 402,pp. 144 – 154

Feng, S., Hu, Q., and Qian, Q., 2004. Quality control of daily meteorological data in China, 1951-2000: a new dataset. Int. J.Climatol., 24, 853-870.

Ferrer, F. J. and Mateos, C. (1999). Análisis de máximas lluvias diarias. Un nuevo método regional de estimación de parámetros de la función de distribución SQRT-ET máx. Ingeniería Civil 115, 109-118.

Garcia-Marin, AP, Ayuso-Muñoz, JL, Taguas-Ruíz, EV, Estévez, J. 2011. Regional Analysis of the annual maximum daily rainfall in the province of Malaga (Southern Spain) using the principal component analysis. Water and Environment Journal, 25: 522-531.

Graybeal, D. Y., A. T. DeGaetano, and K.L. Eggleston, 2004: Complex quality assurance of historical hourly surface airways meteorological data. J. Atmos. Oceanic Technol., 21, 1156 - 1169.

Hosking, J. R. M. and Wallis, J. R. (1997). Regional Frecuency Analysis. Cambridge University Press. New York.

Hubbard, K. G., Goddard, S., Sorensen, W. D., Wells, N. and Osugi, T.T., 2005. Performance of quality assurance procedures for an applied climate information system. J. Atmos. Oceanic Technol., 22, 105-112.

Meek, D. W., and J. L. Hatfield, 1994. Data quality checking for single station meteorological databases. Agric. For. Meteor., 69, 85-109.

Nathan, R.J. and McMahon, T.A. (1990). Identification of homogeneous regions for the purposes of regionalization. J. Of Hydrology 121: 217-238.

O'Brien, K.J., and T. N. Keefer, 1985. Real-time data verification. Proc. ASCE Special Conf., Buffalo, NY, American Society of Civil Engineers, 764-770.

Pariba B.P., Kachroo, R.K., and Shrestha, D.B. (1998). Regional flood analysis of Mahi-Sabarmati Basin (Subzone 3-1) using index flood procedure with L-moments. Water Resources Management 12, 1-12.

Saenz De Ormijana, F., Hidalgo, F.J., and Santa, A. (1991). Estimación de precipitaciones máximas mediante el método regional del índice de avenida. Revista de Obras públicas 138, 9-22.

Shafer, M.A., C.A. Fiebrich, D. S. Arndt, S. E. Fredrickson, and T. W. Hughes, 2000. Quality assurance procedures in the Oklahoma Mesonet. J. Atmos. Oceanic Technol., 17, 474-494.

World Meteorological Organization, 1993. Guide on the Global Data-Processing System. WMO-No. 305, Geneva, Switzerland.

Yun, P., and Chen, C. (1998). Incorporating uncertainty analysis into a regional IDF formula. Hydrological Processes 12, 713-726.