# Técnicas de minería de datos para el análisis de periodos de sequía en España

**F. Belda, M.C. Penadés y F.J. García-Haro**

## Introducción

La sequía es un fenómeno recurrente del clima Europeo, de especial influencia en la regiones mediterráneas. Este fenómeno necesita la definición de un marco adecuado para poder describirlo. La sequía afecta a una amplia variedad de sectores, su diversidad geográfica y distribución temporal, y la demanda de agua para uso humano hacen difícil establecer una definición única. Es posible definir la sequía en términos de las condiciones meteorológicas, hidrológicas, agronómicas y/o socio-económicas dominantes, razón por la cual existen un gran número de índices y parámetros asociados a ella (WMO, 1975).

El presente estudio está referido al concepto de sequía meteorológica, a saber, condiciones meteorológicas que provocan ausencia o reducción de la precipitación durante un período prolongado de tiempo (semanas, meses, años). Desde el punto de vista meteorológico es necesario el estudio de las sequías cortas (importantes para la agricultura) o muy prolongadas (relevantes para evaluar la disponibilidad de agua subterránea, la escorrentía y los niveles de reservas de agua).

La precipitación y la evapotranspiración son los principales factores que controlan la aparición y persistencia de las condiciones de sequía. Dificultades históricas para la cuantificación de la evapotranspiración han sugerido la definición de esquemas de clasificación utilizando solamente la precipitación. En este sentido, índices basados solamente en la precipitación han sido comparados con índices meteorológicos-climatológicos más complejos (Oladipio, 1985). En el presente trabajo, utilizamos el índice SPI (McKee et al., 1993) que ha sido contrastado frente a índices de cálculo más complejo (Lloyd-Hughes and Saunders, 2002).

Hayes et al. (1999) analizaron las ventajas y desventajas del SPI. Por una parte el SPI es simple de calcular, está basado solamente en la precipitación, comparado, por ejemplo, con los 68 términos necesarios para describir el PDSI (*Palmer Drought Severity Index*), otro índice ampliamente utilizado en la bibliografía. Es aplicable al ámbito de la meteorología, agricultura o hidrología teniendo en cuenta que puede ser calculado para escalas de tiempo variables. Ésta versatilidad temporal es también útil para el análisis de la dinámica de la sequía, especialmente la determinación del comienzo y el fin. Como su nombre indica, es un índice estandarizado, lo que asegura que la frecuencia de los eventos extremos en cualquier localidad y en cualquier escala de tiempo es consistente. Sin embargo, por otra parte, el SPI depende solamente de la calidad de los datos de precipitación utilizados. Las sequías extremas (o cualquier otro tipo de sequías) tienen la misma probabilidad de ocurrencia en cualquier lugar, por lo tanto, éste índice no es capaz de identificar regiones que son más propensas que otras a la ocurrencia de sequías, y finalmente el SPI es un índice bastante estacional. En estas situaciones, un completo conocimiento de la climatología de estas regiones mejora la interpretación del SPI.

Para la realización del estudio se han utilizado técnicas de búsqueda y extracción de conocimiento, entendido como información relevante, a partir de grandes cantidades de datos almacenados. En la Fig. 1 se muestran los pasos generales del proceso de descubrimiento del conocimiento utilizados en el presente trabajo (Penadés, 2005).


**Fig. 1** - Pasos del proceso de descubrimiento de conocimiento.

A partir de los repositorios de datos (información disponible) que puede estar almacenada en cualquier formato y soporte, se realiza un proceso de limpieza e integración de la misma, seleccionándose y transformándose los datos si fuera necesario. Posteriormente se construye el almacén de datos como una colección de datos orientados a temas, integrados, historiados y no volátiles que sirven de apoyo al proceso de toma de decisiones (Inmon, 1996). A partir de aquí empieza la evaluación de patrones y presentación del conocimiento, para lo que aplicamos la tecnología OLAP-Mining.. En Han (1997) se propone OLAP-Minig como un mecanismo que integra técnicas propias de la tecnología OLAP (Codd, 1993) con las de minería de datos (Fay, 1996). Esta integración facilita la búsqueda de patrones o conocimiento interesante de forma multidimensional y a varios niveles de abstracción, puesto que las herramientas de análisis trabajan directamente sobre un cubo de datos construido a partir del almacén de datos.

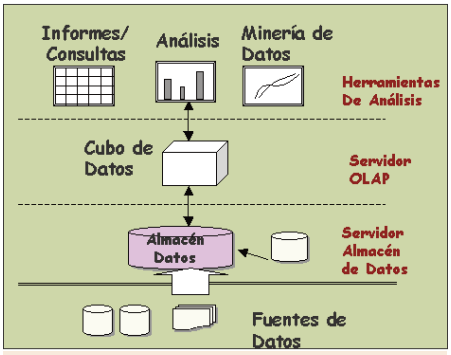


**Figura 2** - Arquitectura de 3-niveles del almacén de datos.

La Fig. 2 muestra la arquitectura de 3-niveles del almacén de datos que tomamos como referencia para nuestro análisis. Como puede observarse las herramientas de consulta e informes, de análisis y/o minería de datos para la exploración y visualización del los datos del almacén se encuentran en el tercer nivel.

## Metodología

El objetivo fundamental es monitorizar y cuantificar la sequía en diferentes áreas de la Península. En la Fig. 3 se muestran los actores del modelo a desarrollar y el flujo de información.

Los datos climatológicos utilizados proceden de la red termopluviométrica de la AEMET (Agencia Estatal de Meteorología). A partir de los datos de precipitación se calcula el índice SPI siguiendo el método definido por McKee et al. (1993). Se consideran los patrones sinópticos climatológicos definidos por Riosalido (1999) utilizándose en los reanálisis de los campos de 500hPa y 850hpa del NCEP/NCAR (Kistler et al., 2001).

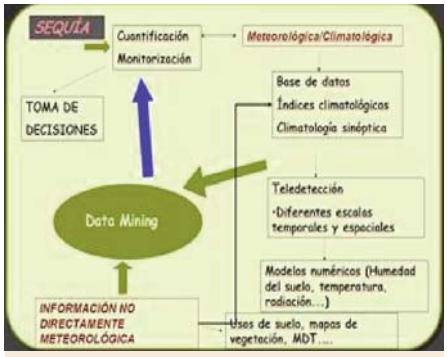


**Figura 3** - Modelo de procesos.

Técnicamente, el SPI es calculado ajustando la distribución de frecuencia de la precipitación de un lugar dado, en la escala de tiempo de interés, con una función teórica de densidad de probabilidad. De acuerdo a varios autores como Thorn (1966), Young (1992) y Lloyd-Hughes (2002) entre otros, la función más apropiada para este ajuste es la Gamma, si bien ésta ofrece algunas dificultades en las zonas de muy poca precipitación, debido a que no se encuentra definida para valores de la variable iguales a 0. La función de densidad es luego transformada a una distribución normal estandarizada (con media igual a 0 y varianza igual a 1), siendo el SPI el valor resultante de esta transformación. Este índice representa el número de desviaciones estándar en que el valor transformado de la precipitación se desvía del promedio histórico (el cual queda representado por 0). Los valores negativos del SPI representan déficit de precipitación y, contrariamente, los valores positivos indican que la precipitación ocurrida ha sido superior al promedio histórico. Con los valores obtenidos se establece una categorización de los eventos tal como muestra la Tabla 1. Para el análisis de la sequía tendremos en cuenta las categorías de sequía moderada, severa y extrema a partir del SPI3.

A partir de los datos climatológicos se generan grids mensuales de precipitación, temperatura y SPI (Fig. 4) a diferentes escalas. Se utilizan entre 2000 y 5000 (dependiendo del período) estaciones distribuidas por todo el territorio. Hay casos en los cuales dos estaciones diferentes presentan similares coordenadas. Nosotros promediamos cuando la distancia que las separa es menor de 0.02º, para evitar un gran número de matrices de covarianzas.

**Tabla 1 –** Clasificación de la sequía según el valor del SPI.

| SPI | Categoría |
|---|---|
| >= 2.00 | Extremadamente húmedo |
| 1.50 a 1.99 | Muy húmedo |
| 1.00 a 1.49 | Moderadamente húmedo |
| 0 a 0.99 | Ligeramente húmedo |
| 0 a –0.99 | Ligeramente seco |
| -1.00 a -1.49 | Moderadamente seco (sequía moderada) |
| -1.50 a -1.99 | Muy seco (sequía severa) |
| <=-2.00 | Extremadamente seco (sequía extrema) |

Para la interpolación de los datos utilizamos el kriging ordinario, eligiendo el modelo que más se ajustaba. Después de un primer análisis, fueron elegidos cuatro diferentes, a saber, exponencial, esférico, gaussiano y pentaesférico. En general el mejor modelo fue el esférico. (García-Haro et al. 2008).

Para el análisis de los condiciones sinópticas dominantes en cada período de sequía, teniendo en cuenta los patrones sinópticos definidos en Riosalido (1999), se consideran tres patrones diferentes: Tipo subsidencia generalizada (Tipo S). Esta situación favorece los descensos de aire sobre toda la península. Viene caracterizada por una potente dorsal en 500hPa y de un anticiclón que afecta a toda la península. Este tipo se presenta en cualquier época del año. Tipo baja dinámica (BD) caracterizadas por vaguadas en 500hPa al oeste peninsular. Dependiendo de la profundidad de la vaguada y de su colocación latitudinal afecta a diferentes áreas del oeste peninsular. En superficie existe un predominio de borrascas que no llegan a afectar al SE peninsular. El tipo BD presenta un máximo invernal y un mínimo estival. Finalmente, el tipo baja térmica (tipo BT), caracterizada por una pequeña vaguada con el eje al oeste de la Península Ibérica y en superficie caracterizada por una baja térmica como extensión de la baja térmica africana. Utilizamos los reanálisis de 1000hPa y 500hPa del NCEP/NCAR(Fig. 5).

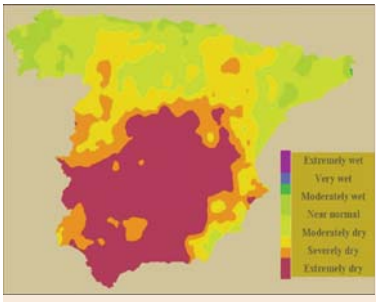


**Figura 5** - SPI-12 correspondiente a mayo de 1995

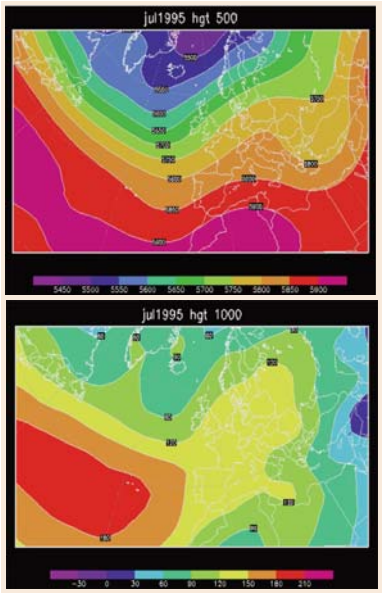


**Figura 5** - Reanálisis NCEP/NCAR de 1000 y 500 hgt correspondientes a julio de 1995.

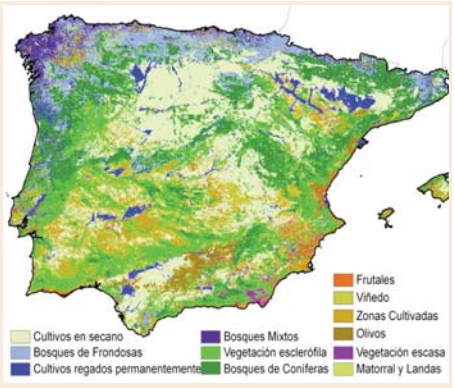


**Figura 6** - Tipos de vegetación

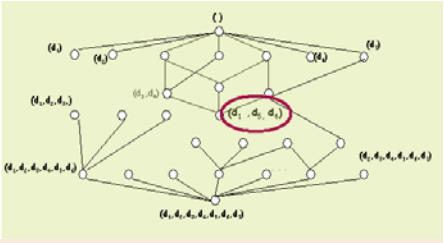

**Figura 6** - Vista de un cubo de datos de 7 dimensiones.

Se van incorporando al cubo de datos información procedente de diferentes instrumentos de teledetección (MERIS, MODIS, SEVIRI,…). Se considera fundamentalmente series de tiempo de FVC (*Fraction Vegetation Cover*) y LAI (*Leaf Area Index*) desde el 2000 hasta el 2008 (1km, 8 días). Finalmente se incorpora la imagen de tipos de vegetación (Fig. 6).

## Caso práctico

A partir de los datos climatológicos y del valor del SPI calculado, aplicamos la técnica de minería de datos, utilizando como herramienta de análisis la proporcinada por Han (2001) y que se denomina DBMiner. En primer lugar determinamos los elementos de interés a analizar (precipitación, temperatura, viento, presión y patrones sinópticos), a continuación definimos el modelo de datos multidimensional. Para ello debemos identificar las medidas de interés, en nuestro caso la sequía (M1) y las dimensiones o perspectivas respecto a las cuales se desea almacenar información, pudiendo definirse jerarquías de agregación (relaciones de orden total o parcial dentro de cada dimensión). En nuestro caso, las dimensiones seleccionadas son: precipitación (D1), temperatura media mensual (D2), recorrido del viento (D3), presión media al nivel del mar (D4) y patrones sinópticos (D5). Consideramos también como dimensión el tiempo (D6) con jerarquía de agregación (Mes < Trimestre < Año) y la zona (D7) con jerarquía de agregación (Punto local < Provincia < Comunidad Autónoma < Delegación Territorial).

En la Fig. 7 se muestra el cubo de datos construido. Potencialmente puesto que hay 7 dimensiones, hay un total de $2^7$ vistas posibles. En el estudio aquí presentado siempre dispondremos como mínimo de tres dimensiones, a saber, una variable meteorológica, el tiempo y la zona. Utilizaremos el SPI3 como el valor de la medida (M1). Una vez construido el almacén de datos y el consiguiente cubo de datos se aplican las técnicas de OLAP-Mining para la búsqueda de información con las siguientes reglas de asociación y clasificación. Los algoritmos de Minería de Datos organizan los datos como una secuencia de eventos (SPI <-1 ó <-2) teniendo cada tiempo un período de ocurrencia. Definimos una secuencia de eventos por la terna, (tc, tf, S), siendo tc comienzo y final del evento respectivamente y S es la serie de SPI3 desde 1971 a 2005. Definimos como episodio a una combinación de elementos en un orden de tiempo específico. Buscamos el patrón sinóptico correspondiente a cada uno de los eventos del episodio. Definimos el par (C, clase), donde C es la colección de eventos que cumplen una condición predeterminada y clase es un vector de tipos de tiempo que nos determinará si existe un orden fijo o no en el episodio. Aplicamos la técnica al Sureste peninsular.

## Resultados

En primer lugar escogemos un punto de grid representativo de la Vega Media del Rio Segura. Analizamos los meses en los que el SPI ha sido inferior a –2 (Tabla 2), y para los que el SPI ha sido inferior a –1 (Tabla 3).

En el algoritmo aplicado nos aparecen casos en los que le SPI es inferior a –1 pero no tenemos una secuencia de más de tres meses seguidos. Ocurre en los años 1981 y 1994. A partir de este análisis construimos una tabla (tabla 4) indicándose relacionándose el episodio (C) y el tipo de tiempo asociado (clase). Aplicamos una sencilla regla de ocurrencias con las siguiente condiciones, que el SPI es inferior a –1 y un número seguido de meses superior a 3, o que el SPI es inferior a –2 y un número seguido de meses superior a 2.

**Tabla 2 –** Períodos en los el SPI ha sido inferior a –2 en el punto de grid escogido.

| tc | tf | S | N° meses |
|---|---|---|---|
| Ago1971 | Ago1971 | 1971-2005 | 1 |
| Ene1975 | Ene1975 | 1971-2005 | 1 |
| Ago78 | Oct78 | 1971-2005 | 3 |
| Sep80 | Oct1980 | 1971-2005 | 2 |
| May1983 | Jun1983 | 1971-2005 | 2 |
| Feb1994 | Mar1994 | 1971-2005 | 2 |
| Ago1994 | Ago1994 | 1971-2005 | 1 |
| Ago2001 | Ago2001 | 1971-2005 | 1 |
| Ago2003 | Ago2003 | 1971-2005 | 1 |

**Tabla 3 –** Períodos en los el SPI ha sido inferior a –2 en el punto de grid escogido.

| tc | tf | S | N° meses |
|---|---|---|---|
| Jul78 | Oct78 | 1971-2005 | 4 |
| Jul81 | Sep81 | 1971-2005 | 3 |
| Ene1983 | Jul1983 | 1971-2005 | 7 |
| Ene1995 | Jul1995 | 1971-2005 | 7 |
| Jun1999 | Ago1999 | 1971-2005 | 3 |

**Tabla 4 –** Secuencia de tipos sinópticos para períodos en los el SPI ha sido inferior a –1 y un número seguido de meses superior a 3, o que el SPI es inferior a –2 y un número seguido de meses superior a 2.

| SPI<-2,N>2; SPI<-1,N>3. | Secuencia tipo sinóptico | N° meses |
|---|---|---|
| Ago-Oct1978 | BT – BT - BT | 3 |
| Ene-Jul1983 | S – BT – BD – BT – BD –BT - BD | 7 |
| Ene-Jul1995 | S – BT – BT – BT – BT –BT - BT | 7 |

## Conclusiones

Estamos definiendo un modelo de datos multidimensional y su gestión automatizada en una arquitectura de tres niveles. Establecemos unos requisitos, definimos un modelo, utilizamos la tecnología para su explotación y finalmente sacamos conclusiones.

El método presentado es una alternativa a los existentes estando abierto a la aparición de nuevos sistemas de información y mejora de la tecnología existente. La aplicación en el campo de la meteorología es incalculable, se pueden incorporar cualquier tipo de información directa o indirecta (teleconexiones). El meteorólogo puede definir las condiciones y reglas de asociación según las características del estudio que se esté realizando. Debido a la incorporación de gran cantidad de información, éste método debe ser introducido gradualmente con mínimos cambios.

Al aplicar estos procedimientos, con ésta técnica se detectan los episodios y sus características "fácilmente". En concreto, se observa un predominio de situaciones de BT (61 %), BD (28 %) y S(11%). Una sequía extrema se da con 4 meses seguidos con situación BT. Períodos más largos y no intensos dan una tendencia BT-BD.

Es de vital importancia la correcta y óptima parametrización de la base de datos. La técnica será mucho más eficiente si los datos son de una alta fiabilidad y de una máxima precisión. Necesitamos mejorar la clasificación sinóptica para establecer secuencias de tipos de tiempo más exigentes. Éste modelo de datos multidimensional nos permitirá de una forma eficiente y sencilla introducir parámetros oceánicos, más estacionarios, que afecten a la circulación general de la atmósfera, así como incorporar reanálisis del ECMWF. De esta forma podremos encontrar, por ejemplo, períodos de sequía precedidos por determinados valores del SOI, MEI, PNA; NAO.

## Referencias


**Belda, F. (1997)** *Climatología y teledetección en zonas forestales de la provincia de alicante. Aplicación a zonas incendiadas*. Tesis Doctoral. Serv. de Pub. de la Univ. de Valencia. ISBN:84-370-3206-7.

**Belda F.** and M.C Penades. (2010**)** *Applying Data-Minig techniques to study drought periods in Spain*. 8th ECAC. Vol. 7. EMS2010-444.

**Codd, E.F.**, Codd, S.B., Salley, C.T. (1993) *Beyond Decision Support*, Computer World, 27.

**Fayyand, U.M.,** Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (1996) *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press.

**García-Haro,** F.J. Belda, F. and Poquet, D. (2008). *Estimation of climatological variables in Spain during 1950-2008 period using geostatistical techniques*, 7th ECAC. Abs. A-003**19.**

**García-Haro, F. J.,** Belda, F., Gilabert Navarro, M.A, Meliá, J., Moreno, A., Poquet, D., Pérez-Hoyos, A., Segarra, S. (2008b), *Monitoring drought conditions in the Iberian Peninsula using moderate and coarse resolution satellite data*, In Proc. of the '2nd MERIS/(A)ATSR User Workshop', ESA SP-666, Noordwijk, The Netherlands, ISBN 978-92-9221-230-8, 7 pp.

**Han, J.** (1997) *OLAP-Mining: An Integration of OLAP with Data Mining*, In Proc. IFIP Conferenceon Data Semantics, Leysin, Switzerland, 1-11

**Han, J.** (2001) *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers.

**Hayes, M.,** Svoboda, M., Wilhite, D.,A. and Vanyarkho (1999): *Monitoring the 1996 drought usuing SPI*. Bulletin of American Meteorology Society, 80, 429-438.

**Inmon, W.H.** (1996) *Building the Data WareHouse*, John Wiley & Sons.

**Lloyd-Hughes, B.** and Saunders, M.A. (2002): *A drought climatology for Europe*. Int. Jour.of Climatology, 22, 1571-1592.

**Kistler R., Kalanay,** E., Collins, W., Saha, S., White, G., Woollen, J.,Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V., Van den Dool, H., Jenne, and Fiorino, M. (2001). *The NCEP-NCAR 50 year reanalysis: Monthly means CD-Rom and documentation*. Bulletin of AMS, 82, 247-267.

**McKee, TB.,** Doesken, NJ. and Kliest, J. (1993): *The relationship of driught frecuency and duration to time scales*. Proceedings of the 8th Conf. of App. Climat., 17-22 January, Anaheim, CA. AMS: Boston, MA; 179-184.

**Oladipio, EO.** (1985): *A comparative performance analysis of threee meteorological drought indices*. Int. Jou. of Climat., 5, 655-664.

**Penadés, M.C.** (2002) *Una Aproximación Metodológica al Desarrollo de Flujos de Trabajo*. Tesis Doctoral. Univ. Polit. de Valencia. Ed.l: ProQuest.I.S.B.N.: 0-493-82722-6, 264 pp.

**Penadés, M.C.** (2005) *Workflow Mining. Minería de Datos: Técnicas y Aplicaciones*. Ediciones de la UCLM, 187-212.



**Fernando Belda Esplugues**
**Agencia Estatal de Meteorología (AEMET)**
**Delegación Territorial en Murcia. fbeldae@aemet.es.**

**María del Carmen Penadés Gramaje**
**Departamento de Sistemas Informáticos y Computación**
**Univ. Politécnica de Valencia. mpenades@dsic.upv.es**

**Francisco Javier García-Haro**
**Departamento de Física de la Tierra y Termodinámica.**
**Univ. de Valencia (Estudi General). j.garcia.haro@uv.es**