Comunicación B-6

SISTEMA EXPERTO EN INTERPRETACIÓN DE SALIDAS DE MODELOS NUMERICOS

Rafael Cano Trueba

SED del CMT de Cantabria y Asturias (INM)

RESUMEN

Se trata de ofrecer una interpretación objetiva (de alta resolución espacio-ten?-poral), de los fenómenos meteorológicos que cabría esperar de producirse las condiciones previstas por un Modelo «X». Para ello se emplea una base de datos con información diaria de completas, automáticas, radiosondeos y estaciones termo y pluviométricas de Asturias, Cantabrin, País Vasco y su entorno (esta BD contiene las series diarias 01/01/86-31/12/94 con 4 551 550 datos interrelacionados). La metodología empleada es puramente estadística, basada en criterios de ***Analogías y Regresiones Múltiples.*** *Ti-as 5 años de pruebas, según va creciendo la BD, los resultados mejoran, esperándose operatividad con garantías para fin el de 1997. Así el predictor ejercerá una labor más crítica.*

1. Introducción

Un Sistema Experto, es la primera aplicación práctica de la Inteligencia Artificial, y está considerado en el nivel más bajo de la escala de sistemas inteligentes. Se trata de programas de ordenador capaces de simular la experiencia de un experto en una disciplina dada. Técnicamente se podría definir como un **programa interactivo entre una BD y un usuario, que sirve para tomar decisiones.** Algunos hitos son: Jugador de Damas *(Arthur Samuel, 1946)*, Diagnosis Clínica *(ELIZA, 1964)*, Estructuras Moleculares *(DENDRAL, 1965)*, Conversaciones *(SHRDLU, 1971)*, Traductores *(SYSTRAN, 1981)*, Ingeniería Genética *(MOLGEN, 1975)*, Resolución de Ecuaciones, etc. Por otra parte hay que citar otros campos alternativos de la IA, bastante bien desarrollados como son: Redes Neurales (simulando el comportamiento de las neuronas cerebrales), Lógica Difusa (donde Verdadero y Falso se difunden de manera que un enunciado puede ser 20% Verdadero y 80% Falso) y la Realidad Virtual (capaz de simular casi cualquier cosa).

2. Base de Datos

Es la materia prima sobre la que trabaja el Programa Principal, su contenido se halla estructurado de la siguiente manera:

2.1. Predictores

Son los Campos Meteorológicos Fundamentales (z, T, h, u & v) a diferentes niveles (1 000,850,700, 500 & 300) en puntos donde se realiza diariamente Radiosondeo-RS- (La Coruña-CR-, Santander-ST-, Burdeos-BX- y Zaragoza-ZG-). Hay por tanto, 25 predictores por estación RS (total 100). Estos datos se encuentran organizados en 8 ficheros, 2 por estación RS (uno para 12 Z y otro para 00 Z). Cada fichero contiene la serie 01/01/86-31/12/94, sumando un total de 730 000 datos entre los 8 ficheros. Estos datos serán utilizados inicialmente como filtros con el fin de identificar a qué familia pertenece el tipo de día que predice el Modelo *«X»* a través de los Campos Previstos —interpolados para los puntos con estación RS por la *Macro* YLEGD (facilitada por el STAP del INM)— y posteriormente serán empleados como predictores en las ecuaciones de regresión.

2.2. Predictandos

Hay 3 tipos de ficheros:

2.2.1. Ficheros de estaciones completas

. . Hay un fichero para cada estación completa de la zona; las estaciones empleadas son: Avilés, Oviedo, Gijón, Santander, Sondica, Foronda y Fuenterrabía. En cada fichero hay un resumen de datos diarios que contiene las siguientes variables: **Tx** (temperatura máxima), Tn (temperatura mínima), R07 (precipitación acumulada de 00 Z a 07 Z), R713 (precipitación acumulada de 07 Z a 13 Z), **R1318** (precipitación acumulada de 13 Z a 18 Z), **R1824** (precipitación acumulada de 18 Z a 24 Z), **I07** (insolación entre 00 Z y 07 Z), 1713 (insolación entre 07 Z y 13 Z), 11318 (insolación entre 13 Z y 18 Z), 11824 (insolación entre 18 Z y 24 Z), VO (velocidad a las 00 Z), V7 (velocidad a las 07 Z), **V13** (velocidad a las 13 Z), **V18** (velocidad a las 18 Z), DO (dirección a las 00 Z), D7 (dirección a las 07 Z), **D13** (dirección a las 13 Z), **D18** (dirección a las 18 Z), VX (racha máxima), DX (dirección de la racha máxima) y VI7 (visibilidad a las 07 Z). Total 21 variables por estación. Las series cubren 01/01/86-31/12/94, resultando en total 536 550 datos en el apartado de completas.

2.2.2. Ficheros de estaciones termo-pluviométricas

Hay 6 ficheros, uno por variable, con las 60 estaciones más fiables y representativas de la región (según un estudio realizado por la Sección de Climatología del CMT de CAS). Cada fichero contiene para las 60 estaciones una de las siguientes variables: **Tx** (temperatura máxima), Tn (temperatura mínima), R77 (precipitación acumulada de 07 Z a 07 Z del día siguiente), NV (si hubo o no nieve), TR (si hubo o no tormenta) y NB (si hubo o no niebla). El contenido total de estos 6 ficheros es de 1 752 000 datos.

2.2.3. Ficheros de estaciones automáticas

Hay 20 estaciones automáticas seleccionadas, cuya estructura es exactamente la misma que las Completas, resultando 20 ficheros con un total de 1 533 000 datos.

3. Esquema de funcionamiento

HIPÓTESIS FUNDAMENTAL: Campos meteorológicos similares producen —salvo casos aislados de Inestabilidad Estiuctural (Saunders, 1983; Woodcock, 1986; «Teoría de Catástrofes») —fenómenos meteorológicos similases.

DEFINICIÓN: Dos días se dicen meteorológicamente similares cuando lo son sus campos meteorológicos fundamentales (Z, T, h, u & v).

3.1. Similitud estructural. Búsqueda y Selección

Asumiendo esta perspectiva, el primer problema a resolver es encontrar entre los 3 285 días que de momento contiene la BD, los 30 más parecidos al día problema. Si se es demasiado exigente, puede ocurrir que ningun día de la BD supere los filtros establecidos, y si se es demasiado tolerante, no seremos capaces de conseguir un grupo con características bien definidas. En lenguaje técnico hay que conseguir un grupo de días que cumplan simultáneamente la pertenencia al *cluster* cuyo centro está definido por el día problema, y que además cubra al máximo el espectro de variación dentro del *cluster* —de esta manera se podrá extraer información acerca del comportamiento de dicha familia o *cluster* bajo diferentes condiciones— éste es el motivo principal de elegir 30 como número de días similares.

Fig. 1. *Diagramas de los programas de filtrado*

3.1.1. Filtros y Algoritmos empleados

Este apartado resume todo el contenido físico del Sistema Experto (que es muy escueto) y el resto es pura Estadística Empírica. Trata de la elección subjetiva de cuáles son las variables predictoras —o combinaciones de ellas— que serán empleadas como filtros para seleccionar los 30 días estructuralmente similares al día problema.

Hay dos categorías de filtros, los Excluyentes (E), que descartan los días que claramente pertenecen a otros *cluster* —es un mecanismo conocido como PODA, que consiste en eliminar, p. ej., días con flujo de sur en una situación de norte—; y los Selectivos (S), que seleccionan del *cluster* los 30 días que más cerca se hallan del centro del *cluster* —es decir aquellos cuya Distancia Generalizada N-dim al día problema es menor—.

La lista de filtros empleados es:

1. $\Delta(\nabla\phi\ (1\ 000)) < 150$ mgp/ud. Entre Bx y Cr
2. $\Delta(\nabla T\ (1\ 000)) < 3$ °C/ud. Entre Bx y Cr.
3. $\Delta(\nabla D\ (500)) < 25$ grados/ud. Entre St y Cr.
4. $\Delta(\nabla T\ (500 - 1\ 000)) < 3$ °C/ud. Para St.
5. AD (850) < 25 grados. Para todos.
6. AD (500) < 25 grados. Para St.
7. A$ (1 000) < 150 mgplud. Para St.
8. AT (1 000) < 3 °C. Para St.
9. ΔHr (700) < 30%. Para St.
10. AU (850) < 6 + ABS (U850-mod). Para St.
11. AV (850) < 6 + ABS (V850-mod). Para St.

El símbolo «delta» colocado delante de todas las ecuaciones indica diferencias entre el día problema y el día de la BD. El símbolo «grad» que aparece a continuación en las 4 primeras significa «gradiente horizontal entre dos puntos)) (tomando como unidad la distancia entre dichos puntos) en las tres primeras y «gradiente vertical entre 500 y 1 000 mb» en la 4.ª (tomando como unidad la distancia media 500-1 000). A modo de ejemplo, la interpretación literal de la ecuación 1 sería: sólo se admiten días cuyo gradiente horizontal de geopotencial en 1 000 mb entre Bx y Cr difiera en menos de 150 mgp con el mismo gradiente del día problema. Los días aceptados como estructuralmente similares (ES), han de superar simultáneamente todos los filtros. Evidentemente a medida que la BD vaya creciendo se podrán exigir umbrales más pequeños consiguiendo así familias de días mucho más parecidas al día problema.

Aunque la elección de estos filtros y umbrales es subjetiva no es ni mucho menos arbitraria, ya que se han tratado de combinar aspectos dinámicos y térmicos de manera que el perfil de la atmósfera quede caracterizado al máximo mediante estos filtros. Se obliga a que la onda tenga parecida forma, parecida **curvatura**, parecida advección de temperatura y **vorticidad,** parecida intensidad de viento y contenido de **humedad**, parecida inestabilidad; todo esto haciendo más hincapié en los aspectos térmicos en niveles bajos y en los dinámicos en niveles medios y altos.

De entre los días aceptados como ES hay que seleccionar los 30 mejores, aquí es donde intervienen los filtros selectivos, que son los 7 restantes, y que mediante una combinación lineal generan un número que indica el grado de proximidad al día problema. La c. 1. actual es:

$$\text{SIMIL} = 50 \cdot A\,[\nabla T\,(1\,000)] + 50 \cdot A\,[\nabla\,(T\,(500 - 1\,000)] + \\ + \Delta\phi\,(1\,000) + 50 \cdot AT\,(1\,000) + 5 \cdot \Delta Hr\,(700) + 25 \cdot \Delta U\,(850) + 25 \cdot AV\,(850)$$

Los 30 días cuyo valor de SIMIL sea menor, se ordenan de menor a mayor, con el objeto de realizar una regresión múltiple ponderada, en la cual el día más parecido pesará más que el menos parecido del grupo de 30. Así, al día más parecido se le asigna un peso de 30, al segundo 29, ..., y al trigésimo se le asigna un peso de 1. Las fechas y los pesos asociados se guardan en un archivo intermedio llamado **«Simil»** que actúa como *input* de un subprograma llamado **«Cepp1»** que es el encargado de generar las tablas de predictores y predictandos seleccionadas para cada PINI del modelo. La BD sería perfecta si no tuviese lagunas ni datos erróneos *o* al menos sospechosos, como no es así hay que realizar un estricto control de calidad.

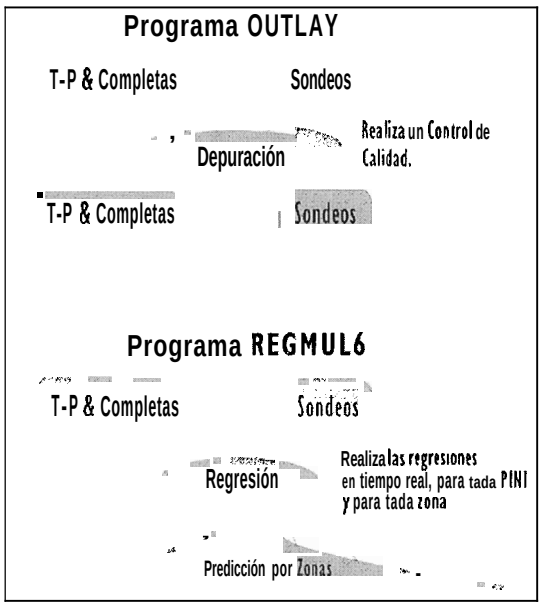

***Fig.** 2. Diagrama de los programas OUTLAY y REGMUL6*

3.2. Control de calidad. Eliminación de *outlayers*

Dentro de cada *cluster* de 30 días que contiene 100 predictores y 947 predictandos, es obligado proceder a la eliminación de datos que puedan ser erróneos. Para ello empleamos una técnica muy sencilla que consiste en calcular la media y la varianza de cada una de las 1 047 variables de 30 elementos cada una, eliminando aquellos elementos que se salgan del rango

$$\mu \pm 6 \cdot \sigma$$

hecho esto se itera el proceso de cálculo de medias y varianzas sin los eliminados, hasta que media y varianza sean constantes.

3.3. Regresiones

Dentro de cada *cluster* (muestra de 30 elementos) se calculan todos los coeficientes de corre-

lación simple entre predictores y predictandos, seleccionando el que mejor coeficiente de correlación presente y calculando a continuación los residuos, sobre los que se repite el proceso hasta que el predictor seleccionado no supere un coeficiente de correlación mínimo o se llegue a 5 términos en la regresión.

A continuación, a partir de la ecuación de regresión obtenida, se genera la serie de predicciones y los correspondientes residuos, rechazando aquellos elementos de la muestra cuyo residuo se salga del rango

$$\mu \pm 6 \cdot \sigma$$

recalculándose nuevamente la regresión hasta que media y varianza permanezcan constantes.

Una vez obtenida la ecuación de regresión con su coeficiente de correlación y su error estándar, no queda más que introducir los predictores predichos por el modelo para obtener la predicción en la estación, variable y PINI especificados.

Este proceso se repetirá tantas veces como *Pinis* × Variables × Estaciones, resultando 1 047 regresiones por *Pini.* Como es lógico, tal volumen de datos ha de ser informatizado y presentado en forma de gráficos y mapas para su consulta. Otro subprograma se encarga de traducir las predicciones a lenguaje técnico por zonas, de acuerdo con los criterios establecidos en el Manual **de** Términos Meteorológicos.

3.4. Traducción a lenguaje técnico

El subprograma **«Lope»** realiza esta tarea que consiste en generar automáticamente un boletín con referencia a la fecha, período de predicción y área de aplicación, así como el modelo numérico empleado, la pasada y el PINI.

Este programa se encarga de seleccionar aquellas variables alfanuméricas que se corresponden con los umbrales establecidos. Por ejemplo:

IF prec≤30 THEN prec$=«Precipitaciones fuertes))
IF prec≤15 THEN prec$=«Precipitaciones moderadas))
IF prec≤2 THEN prec$=«Precipitaciones débiles))
IF prec=0 THEN prec$=«No se esperan precipitaciones))

De igual modo se opera con los términos: posible, probable, intervalos, variable, ocasional, persistente, frecuente, intermitente, continuo, disminución, aumento, generalizado, disperso, arreciar, amainar, subir, bajar, sin cambios, etc.

Solamente queda encadenar las variables alfanuméricas para cada zona y ya está generado el boletín para que el predictor pueda hacer la consulta pertinente.

4. Evaluación de prestaciones (provisional)

En su estado actual el S.E. ha sido probado para una muestra aleatoria de 30 salidas D+1 del Modelo CEPPM, con los siguientes resultados:

Lugar de predicción: Santander-Centro.
Muestra aleatoria de 30 días de Oct-95 a Feb-96.
Período de predicción D+1.
Variables analizadas: Tª Máx, Tª Min y Prec acumulada cada 12 h.
Método de evaluación: coeficiente de correlación lineal simple entre los valores previstos y los reales —téngase en cuenta que aquí se acumula el error propio del modelo empleado junto con el propio del S. E.—.

Variable	Núm. Eltos.	Coef. Corr.
Tª Máxima	30	,89
Tª Mínima	30	,51
Prec. Acu. 12 h	60	,68

Para los mismos días, el Modelo CEPPM dio un coeficiente de correlación lineal simple de 0,57 en la predicción de precipitación acumulada cada 12 h para Santander-Centro. Por supuesto esta evaluación es provisional, ya que hay que probar en diferentes puntos y con series más largas. Por otro lado el S. E. no está todavía al máximo rendimiento por lo que cabe esperar alguna ligera mejoría en sus prestaciones según vaya creciendo su base de datos y se vayan mejorando los filtros y algoritmos que emplea.

5. Presentación de resultados

Como ya se ha dicho anteriormente, para poder asimilar la gran cantidad de datos de salida que proporciona el Sistema Experto, es necesario recurrir a representaciones en forma de mapas, gráficos y tablas. Este cometido se realizará empleando programas comerciales como WINSURFER para la obtención de mapas a partir de ficheros en formato «.GRD» que son generados directamente por el S. E.; los gráficos y tablas se realizarán mediante LOTUS a partir de ficheros en formato «.TXT» tambien generados directamente por el S. E. Quedan por hacer las *«macros»* que automaticen completamente estas tareas.

A continuación se da un ejemplo de salida gráfica del S. E.:

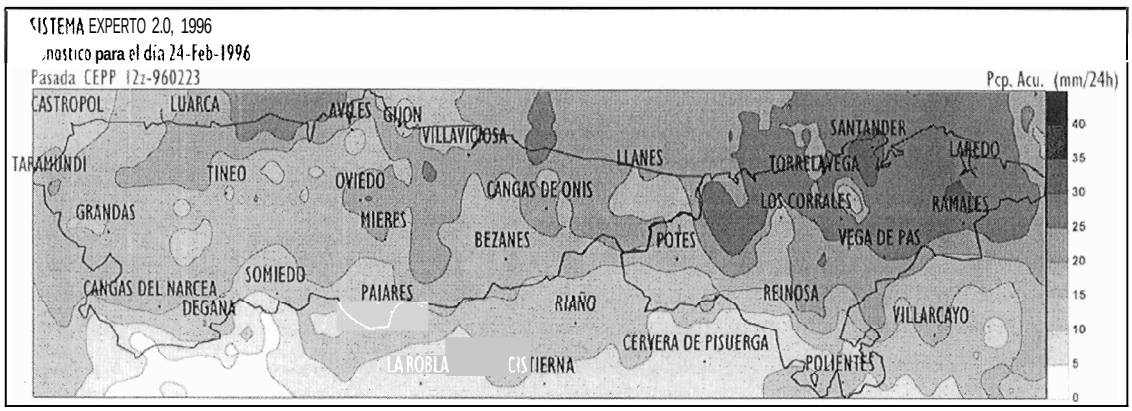

Fig. 3. *Pronóstico general* ***deprecipitacion acumulada en 24 h. Todos lospredictandos de estaciones T-P se representan en mapas de este tipo. Para las zonas*** *Oviedo-Gijón-Avilés y* ***Santander-Torrelavega se realiza*** *una* ***anzpliación*** *que incluye* ***nzayor*** *resolución* ***espacio-temporal*** *y mayor* ***número de variables***

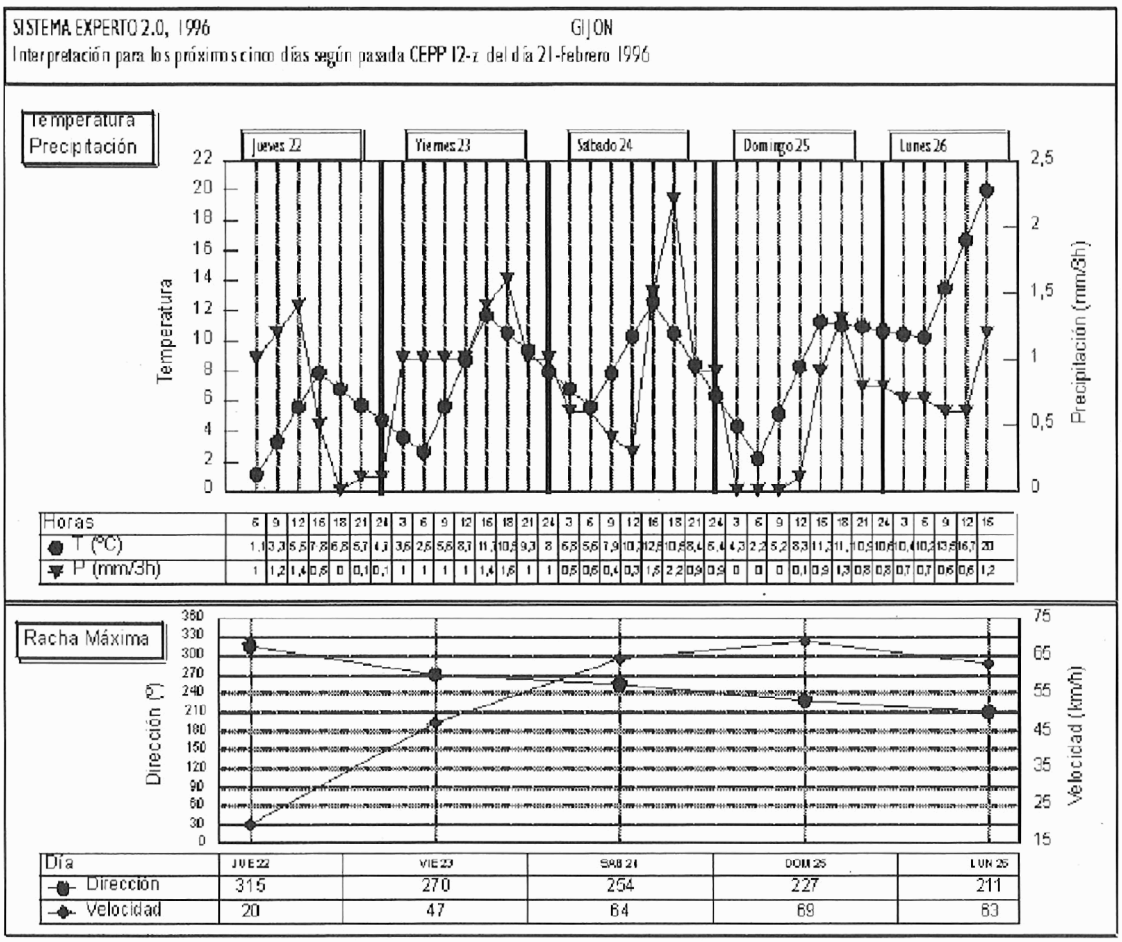

Fig. 4. *Pronóstico hasta D+5 de una estación completa, incluye* ***Temperatura,*** *Precipitación y* ***Racha Máxima,*** ***aunque*** *se puede dar también: Insolación, Viento y Visibilidad cada tres horas*

6. Conclusiones

Sólo gracias al enorme potencial de cálculo que tienen hoy los nuevos ordenadores personales es posible realizar Estadísticas Empíricas en Tiempo Real (EETR). En ellas se pueden recalcular las ecuaciones de regresión para cada caso particular, por lo que su nivel de especialización y capacidad de adaptación es mucho mayor.

Por sus características, se adapta automática e instantáneamente a cualquier tipo de modelo numérico. Cuanto más perfecto sea el modelo numérico más perfecta será la interpretación. Por supuesto tambien se verá arrastrado por cualquier fallo de predicción del modelo numérico (recuérdese que el Sistema Experto sólo se limita a interpretar y matizar los campos predichos por el modelo).

La mejor virtud que de momento se le puede atribuir es la de ofrecer la posibilidad de emitir pronósticos locales, lo que tiene la ventaja de no necesitar conocer perfectamente el comportamiento del clima de cada valle o subregión.

Aunque ofrece datos numéricos y probabilidades, debe ser interpretado cualitativamente en el sentido de dónde lloverá más, dónde menos, dónde hará más calor, dónde más viento, cuándo culminará determinada situación o cuándo se extinguirá.

Ha sido empleado con gran éxito en el episodio de nevadas de Feb-1996, como apoyo a la predicción.

Otras utilidades que admite el S. E. son las de Simulador de Casos Extremos, sin más que introducir manualmente las condiciones límite supuestas como si de una predicción numérica se tratase. También es un

Detector de Leyes, para lo que sólo hay que emplear su propia base de datos y generar un fichero particionado de ecuaciones de regresión.

En los 5 años de pruebas y diferentes métodos de operación, la efectividad del S. E. ha ido creciendo fundamentalmente por dos razones: primero porque ha crecido la base de datos que le sustenta y segundo porque se ha ido refinando el método de selección y regresión.

Referencias

Ayuso, J., Predicción Estadistica Operativa en el INM. B-34. INM.

Barghava, M. & M. Danard, 1994: Application of optimum interpolation to the analisys ofprecipitation in complex terrain. Journal of Applied Met., vol. 33, pp. 508-518.

Borrajo, D. y otros: Inteligencia artificial, Métodos y Técnicas. Ed. Ramón Areces.

Box & Hunter, 1988: Estadística para investigadores. Ed. Reverté.

Cano Trueba, R., 1992: Atlas Climático de la Región de la Cordillera Cantábrica. Biblioteca CMT CAS.

Cano Trueba, R., 1991: Realce orográfico de Normales Climatológicas. Biblioteca INM.

Hartnell, T., 1985: Inteligencia Artificial, conceptos y programas. Ed. Anaya.

INM, 1992: ESTILO-Manual de términos meteorológicos.

Lenat, D. B.: Inteligencia Artificial. Rev. Investigación y Ciencia núm. 230.

López Cachero, M., 1987: Fundamentos y Métodos de Estadística. Ed. Pirámide.

Luthe y otros: Métodos numéricos. Ed. Limusa.

Méndez, A. y F. Elizaga, 1993: Procedimiento de Análisis a Mesoscala Interactivo en SAIDAS (PAMIS). STAP. Nota Técnica Núm. 11. INM.

Peña Sánchez, D., 1993: Estadistica. Modelos y Métodos. Vols. I y II. AUT Núms. 109 y 110.

Press y otros: Numerical recipe-The art of scientific computing. Ed. Cambridge Un. Press.

Rich, E. y K. Knight: Inteligencia Artificial. McGraw Hill.

Saunders, P., 1983: Introducción a la Teoría de Catástrofes. Ed. Siglo XXI.

Sokal y Rohlf, 1984: Introducción a la Bioestadistica. Ed. Reverté.

Thom, R., 1972: Stabilité structurelle et morphogenese.

Woodcock, A., 1986: Teoría de las Catástrofes. Ed. Cátedra

Agradecimientos

Muchas son las personas que Izan colaborado y lo siguen haciendo, cada tina de ellas ha supuesto una vital ayuda para superar tino o varios de los numerosos obstáculos que han surgido y surgirán hasta la realización de este proyecto. Aunque intentaré citar a todos, puede que se me olvide alguno:
A todos los Predictores del GPV del CMT CAS y a su Jefe.
A Gonzalo Moreno, Mari-Sol De Andrés, Toño Fdez.-Cañadas y Ramón Celis de la Sección de Climatología del CMT de CAS.
A Eduardo Arasti y J. Salvador Martín de la SE&D del CMT de CAS.
A Antonio Gª Méndez, Benito Elvira, Paco M. León y Ricardo Riosalido del STAP.
A J. Antonio Guijarro del CMT de Baleares.
A Luis Hernando de Explotación.
A Météo-France
A Ernesto Rodriguez y J. Antonio Garcia-Moya de Predicción Numérica.
A J. J. Ayuso de Predicción Estadística.
Y a Pilar Fernández de Satélites y Radares.