Aportaciones de las Redes Bayesianas en Meteorología. Predicción Probabilística de Precipitación.

Tesis Doctoral

Presentada por
D. Rafael Ancell Trueba

bajo la dirección de
Dr. José M. Gutiérrez

Dpto. de Matemática Aplicada y Ciencias de la Computación.
E.T.S.I. Caminos, Canales y Puertos
Universidad de Cantabria

Santander, mayo de 2009

Bayesian Networks for Probabilistic Weather Forecast

Rafael Ancell
rct@inm.es, Agencia Estatal de Meteorología (AEMET)
Departamento Matemática Aplicada y C.C. University of Cantabria
39005 Santander, Spain

Prefacio

"La sabiduría inútil,
se diferencia de la tontería
en que da mucho más trabajo"
Proverbio Sueco

Llevo casi 10 años con la Tesis, y muy posiblemente, ahora esté escribiendo las últimas líneas; ha sido más trabajo de lo que pensaba. Nos ha sobrado ilusión, motivación, mal rollo, buen rollo, trabajo, desilusión, parón, más trabajo, falta de prisa, juergas y agonías. Con todo, puedo decir que ha sido una etapa muy intensa que ha cumplido nuestras expectativas con creces. Haciendo una breve historia cronológica, primero presentaré la trayectoria profesional que me planteó el problema y después presentaré el propio problema del que trata esta Tesis desde su punto de vista más reciente, pues he comprobado que se da mejor respuesta cuando se hace una buena pregunta.

En 1986 obtuve la Licenciatura en Ciencias Físicas, año en el cual también ingresé en el Cuerpo de Observadores (CO) del entonces Instituto Nacional de Meteorología (INM). En el año 1991, ingresé en el Cuerpo Superior de Meteorólogos del Estado con la firme convicción de sacarle partido a la enorme cantidad de conocimientos en estadística básica que me habían obligado a estudiar en las pruebas de acceso (la estadística que estudié en la licenciatura de Ciencias Físicas me sirvió, pero necesitaba más, ... bastante más). Dado que procedía del CO (eramos los que hacíamos las puntuales y periódicas anotaciones del estado del tiempo) tenía un buen conocimiento y fácil acceso a las bases de datos imprescindibles para poner manos a la obra. Y así lo hice, conseguí (y muchas veces tuve que transcribir) series de precipitación, manchas solares, efemérides, vientos, presiones y temperaturas, a diferentes horas y en diferentes localidades, con errores, con lagunas, etc; por otra parte, también conseguí acceder a salidas de modelos numéricos de predicción, por supuesto más simples y menos precisos que los actuales (estoy hablando de 1991). Como dato anecdótico, el único ordenador accesible era un Olivetti M-24

compartido y cuya capacidad de cálculo era tan pequeña que se pasaba los fines de semana enteros calculando lo que hoy se hace en segundos; en resumidas cuentas: más información que conocimiento y menos aún, recursos.

Por supuesto comencé con la estadística lineal, regresión múltiple, componentes principales y poco más. En tres años ya le había dado demasiadas vueltas al problema de la predicción local de precipitación, desarrollando un modelo de regresión múltiple por etapas que funcionaba como cabía esperar: bien en las predicciones corrientes y mal en las difíciles. Llegué a la conclusión de que la idea sólo podría prosperar con la ayuda de otras personas con más conocimientos de estadística que me sacaran del callejón sin salida en el que me hallaba, pues las técnicas lineales eran lo más que yo podía desarrollar entonces dada mi formación y los limitados recursos computacionales disponibles. Una verdadera pena ya que la base de datos era fastuosa y, estadísticamente hablando, virgen.

Un amigo de juventud, Jaime Gutiérrez, de la Universidad de Cantabria, me pareció la persona perfecta para contarle todo esto y pedirle consejo. Así tuve la suerte de contactar con José Manuel Gutiérrez, en aquél entonces del Departamento de Matemática Aplicada y Ciencas de la Computación y con Miguel Angel Rodríguez, del Instituto de Física de Cantabria. Después de mi exposición del asunto, no hubo ninguna duda, el tema merecía la pena y además ya teníamos todo: el conocimiento, los datos y la ilusión (creo firmemente que un gramo de ilusión equivale a un kilogramo de conocimiento). Desde entonces hemos trabajado juntos en numerosos proyectos y personalmente me ha supuesto un impagable beneficio tanto académico como personal.

Para brindar al lector una visión de conjunto más apropiada, tras esta descripción cronológica, se ofrece una perspectiva actual del planteamiento del problema abordado en esta Tesis.

Posiblemente la biología, la socio-economía y el medio ambiente sean los mayores retos a los que se enfrenta la civilización actual y aunque básicamente son problemas independientes, los tres tienen muchas cosas en común, desde el gran número de variables que interactúan y regulan el sistema, hasta el comportamiento caótico que se deriva de su naturaleza no lineal. Para analizar, diagnosticar y predecir estos sistemas disponemos de potentes ordenadores que nos sirven para predecir su estado futuro con un grado de detalle espacial y alcance temporal, que han crecido enormemente gracias al reciente desarrollo tecnológico. Esto ha favorecido una rápida evolución en la predicción meteorológica de forma que actualmente, los modelos de circulación atmosférica son capaces de pronosticar la fenomenología meteorológica hasta escalas del orden de decenas de kilómetros. Sin embargo, no todo se resuelve a base de musculatura tecnológica: por un lado, el crecimiento exponencial de las incertidumbres iniciales con el tiempo, impone límites mas allá de los cuales la solución determinista pierde sentido, dando razón de ser a la concepción probabilística cuando la anticipación en los pronósticos supera los 3 días, y por otro, para obtener pronósticos a escala local es necesario realizar un post-proceso, utilizando una técnica conocida como *downscaling* (disminución de escala, o aumento de resolución), que

consiste en estimar valores en localidades a partir de las salidas de baja resolución de los modelos numéricos y que será el tema principal de esta Tesis.

El trabajo presentado en esta Tesis ha sido desarrollado dentro del grupo *Meteorology & Data Mining*, de la Universidad de Cantabria `http://meteo.unican.es`. Tengo que agradecer el apoyo y la colaboración de la Agencia Estatal de Meteorología (AEMET), de la Universidad de Cantabria y de la Dirección General de Ciencia y Tecnología en forma de becas y proyectos de investigación. También al European Center for Medium-Range Weather Forecast (ECMWF) y a la AEMET por la disponibilidad de datos, que ha permitido llevar a cabo este trabajo.

Todo esto ha sido posible gracias a mis amigos Jose, Antonio, Miguel, Carmen, Dani y Jaime en la Universidad, junto con la confianza de Charo, Bartolomé, Salva y Mariona en la AEMET, aunque por otra parte, hubiese sido imposible sin el necesario apoyo de Mari en casa, lidiando, además de conmigo, con otros tres proyectos llamados Luis, Marina y Violeta. Con vuestro permiso, este trabajo va dedicado a Paquita y Piqui, que son mis padres.

Rafael Ancell Trueba
Santander, 27 de Abril de 2009

Índice general

Prefacio III

Índice general 0

1. Objetivos y Resumen de la Tesis 5
- 1.1. Motivación . . . 6
- 1.2. Objetivos . . . 7
  - 1.2.1. Objetivo general . . . 7
  - 1.2.2. Objetivos particulares . . . 8
- 1.3. Resumen de la Tesis . . . 9
  - 1.3.1. Estado del Conocimiento . . . 10
  - 1.3.2. Aportaciones de la Tesis . . . 11
- 1.4. Algunos Acrónimos y Terminología Utilizados . . . 13

I Estado del Conocimiento 19

2. Predicción Meteorológica 21
- 2.1. La Atmósfera . . . 21
  - 2.1.1. Descripción Física . . . 22
  - 2.1.2. Escalas de la Atmósfera . . . 26
  - 2.1.3. No Linealidad y Caos . . . 27
- 2.2. Predicción Numérica . . . 29
  - 2.2.1. Asimilación de Datos . . . 31
  - 2.2.2. Tratamiento de la incertidumbre . . . 32

2.2.3. Tipos de Predicciones 33
2.3. Series de Observaciones y Simulaciones 44
2.3.1. Redes Globales de Observación 45
2.3.2. La Red de Observación de la AEMET 46
2.4. Simulaciones de Modelos Numéricos: Reanálisis 49

3. Minería de Datos. Redes Bayesianas 51
3.1. Sistemas Inteligentes 52
3.2. Dependencia e Independencia de Variables 55
3.3. Redes Probabilísticas 55
3.3.1. Codificación de Dependencias. Criterios de Separación 57
3.3.2. Factorización de la Probabilidad 59
3.4. Redes Bayesianas. Definición y Tipos 61
3.4.1. Tipos de Redes Bayesianas 61
3.4.2. Razonamiento Probabilístico. Inferencia 64
3.5. Algoritmos de Aprendizaje Automático 65
3.5.1. Algoritmos Basados en Test de Independencia 67
3.5.2. Algoritmos de Evaluación y Búsqueda 70
3.6. Aprendizaje de Clasificadores 76
3.7. Sobreajuste 79

4. Medidas de Verificación de Predicciones Probabilísticas 81
4.1. Aspectos de la Calidad de una Predicción 83
4.2. Medidas Estándar de Validación 86
4.2.1. Medidas de Validación de Predicciones Categóricas 88

II Aportaciones 95

5. Dominios de Estudio y Datos Utilizados 97
5.1. Áreas de Estudio y Observaciones 97
5.2. Predictandos: Observaciones en Superficie 99
5.3. Predictores: Patrones Atmosféricos 102

6. Aplicaciones de las Redes Bayesianas en Meteorología 109
6.1. Estructura Espacial de la Precipitación 110
6.1.1. Ejemplo de Aplicación de Redes Bayesianas 112
6.2. Generadores de Tiempo (Weather Generators) 115

6.2.1. Cadenas de Markov . . . 116
6.2.2. Modelos de Markov Ocultos (HMM) . . . 118
6.2.3. Ejemplo de Aplicación de Redes Bayesianas . . . 119
6.3. Predicción Local . . . 123
6.3.1. Ejemplo de Aplicación de Redes Bayesianas . . . 125

7. Paradigmas de Predicción Probabilística Local 129
7.1. Paradigmas de Predicción Local . . . 130
7.2. Downscaling Estadístico . . . 131
7.2.1. Métodos Basados en Funciones de Transferencia . . . 133
7.2.2. Métodos basados en Análogos . . . 135
7.3. Downscaling con Redes de Estructura Ingenua . . . 138
7.3.1. Dependencia Espacial . . . 141
7.3.2. Predicciones Marginales y Conjuntas . . . 143
7.3.3. Observaciones Disponibles y Sobreajuste . . . 146

8. Algoritmos de Aprendizaje para la Predicción Local 149
8.1. Algoritmos de Búsqueda en la Predicción Local . . . 150
8.1.1. Deficiencias de los Algoritmos con Pre-ordenación . . . 150
8.1.2. Algoritmos de Búsqueda Local . . . 152
8.2. Medidas Específicas de Calidad . . . 158
8.2.1. Efectos del Término de Penalización . . . 160
8.2.2. Medidas Específicas para Clasificación y Predicción . . . 161
8.2.3. Capacidad Predictiva Marginal: Medida *RSA* . . . 162

9. Predicción Local con Redes Bayesianas 169
9.1. Comparativa de Algoritmos . . . 170
9.1.1. Modo Downscaling . . . 171
9.1.2. Modo Predicción . . . 172
9.1.3. Modo Interpolación . . . 173
9.2. Resultados . . . 173
9.2.1. Modo Downscaling . . . 177
9.2.2. Modo Predicción . . . 178
9.2.3. Modo Interpolación . . . 179

10. Downscaling en Redes Heterogéneas 189
10.1. Redes Heterogéneas. El Algoritmo *H2* . . . 190
10.2. Ejemplo de Aplicación . . . 192

10.2.1. Resultados 193

11. Conclusiones y Trabajo Futuro **201**

11.1. Conclusiones 202

11.2. Trabajo Futuro 204

III Bibliografía **207**

Bibliografía **209**

CAPÍTULO 1

Objetivos y Resumen de la Tesis

Esta Tesis está dirigida principalmente a investigadores interesados en la aplicación de técnicas de *minería de datos* en Meteorología y otras ciencias medioambientales afines, aunque algunas de las innovaciones también son susceptibles de ser aplicadas en otras áreas científicas, siempre que traten con sistemas definidos por muchas variables, cuyas relaciones de dependencia no sean conocidas a priori y tengan que ser inferidas a partir de un conjunto de datos que describan el sistema. Buena parte de los logros conseguidos son fruto del intercambio de conocimiento (y desconocimiento) entre un meteorólogo y un matemático, en el afán de aplicar los recientes desarrollos producidos en el área de la minería de datos a problemas prácticos relacionados con el diagnóstico y la predicción probabilística local en Meteorología, considerando el problema de la coherencia espacial. En concreto, el eje central de esta Tesis ha sido el desarrollo de modelos gráficos probabilísticos, en particular Redes Bayesianas (RBs), para su aplicación en la predicción probabilística local, aunque como se verá posteriormente también se han realizado otras aportaciones interesantes. Hay que decir también que esta Tesis supone un punto de partida en este campo y, por ello, en su redacción se comienza con una descripción genérica de algunas de las posibilidades que ofrece esta herramienta en distintos problemas considerados estándar en la Meteorología actual, para terminar con una aplicación específica más desarrollada, relativa a la predicción probabilística local de eventos de precipitación.

1.1. Motivación

Son muchos y diversos los motivos que han impulsado la realización de esta Tesis; por ello dedicamos este breve apartado a describir algunos de los que consideramos más relevantes.

Se entiende por *predicción meteorológica local* aquella que está calibrada de forma fiable con la climatología propia de una localidad y, por tanto, puede considerarse un producto final. En la práctica, la predicción meteorológica operativa es un proceso complejo, y tiene como primer elemento de la cadena las salidas de los modelos numéricos de simulación de la atmósfera que se integran diariamente, resolviendo las ecuaciones diferenciales de la física que rigen este problema. Cuando la predicción local se obtiene como postproceso de estas salidas, el problema se denomina *downscaling* (descenso de escala). Este problema es hoy en día una actividad prioritaria en los planes estratégicos de numerosos servicios meteorológicos debido, entre otras, a las siguientes razones:

- En primer lugar, cada vez es mayor la necesidad de proyectar a escala regional y local las predicciones obtenidas a diferentes alcances por los modelos numéricos del clima (globales y regionales), ya que la resolución demandada por los usuarios es generalmente mayor que la ofrecida. La resolución actual de los modelos globales es de unos 25-50km (para la predicción a plazo medio: 3-15 días), 100km (para la predicción estacional) y 200km (para la predicciones interdecadales de cambio climático).

- En segundo lugar, existe una creciente demanda de este tipo de predicciones para variables de superficie (principalmente precipitación y temperatura) por numerosos sectores, entre los que se pueden citar el ambiental, biológico, hidrológico, sanitario, agrícola, industrial, energético, transportes, seguros y turístico. Estas variables son muy sensibles a procesos físicos que habitualmente se hallan parametrizados en los modelos numéricos y, por tanto, requieren un post-proceso antes de ser utilizadas en estos ámbitos.

Estas razones, de índole computacional (en el primer caso) y físico (en el segundo), ponen de manifiesto la importancia de las técnicas de post-proceso para la generación de predicciones locales, ya que son el vínculo de conexión entre los productos meteorológicos generados directamente por los modelos numéricos y los requerimientos específicos de los sectores usuarios, mejorando la potencial utilidad de estos productos. El creciente interés por estas técnicas también se refleja en los

proyectos internacionales, que recientemente vienen incluyendo esta tarea como elemento final en la cadena de predicción. Por ejemplo, los siguientes proyectos llevados a cabo en el V y VI Programa Marco de la UE han contemplado paquetes especiales para el desarrollo y la implementación de técnicas de *downscaling*:

- DEMETER (Development of a European Multimodel Ensemble system for seasonal to inTERannual prediction) que ha desarrollado un experimento de predicción estacional multi-modelo que abarca un período de 40 años, `www.ecmwf.int/research/demeter`
- STARDEX (Statistical and Regional dynamical downscaling of Extremes for European regions), que ha analizado el problema de la proyección regional del cambio climático utilizando técnicas de *downscaling*, `www.cru.uea.ac.uk/cru/projects/stardex`
- ENSEMBLES (ENSEMBLE-based Predictions of Climate Chages and their Impacts), del VI Programa Marco de la UE (2004-2009), que cuenta con grupo de trabajo de *downscaling estadístico* del que forman parte el grupo de trabajo de la Universidad de Cantabria en el que se ha desarrollado esta Tesis, dentro de la tarea RT2B: *Production of Regional Climate Scenarios for Impact Assessment*), `www.ensembles-eu.org`

Por tanto, se puede concluir que el presente trabajo se desarrolla en un momento oportuno, pues el reto a corto y medio plazo en este ámbito será la adaptación e implementación de métodos multi-dimensionales en proyectos de distintos alcances temporales. Por todo ello, tenemos la esperanza de que la investigación desarrollada en esta Tesis pueda ser de utilidad, tanto para la comunidad científica, como para la Meteorología operativa y sus usuarios, y además pueda tener una continuación futura en el marco de proyectos nacionales y europeos.

1.2. Objetivos

Una vez descritos los principales motivos que han impulsado la realización de esta Tesis, a continuación se describen los principales objetivos propuestos para la misma, que se derivan naturalmente de aquellos.

1.2.1. Objetivo general

El objetivo general que se plantea, es desarrollar un método global e integrado capaz de proporcionar predicciones locales de forma colectiva, considerando la

coherencia espacial para un conjunto de localidades en un dominio específico. En la mayoría de los trabajos publicados en la literatura, y de los proyectos anteriormente citados, este problema se aborda desde un punto de vista unidimensional, considerando cada localidad de forma independiente. En comparación, son escasos los trabajos que abordan el problema desde el punto de vista multi-dimensional. En esta Tesis se propone abordar este problema para el caso discreto o categórico (por ejemplo, "lluvia" y "no lluvia"), considerando un modelo conjunto que permita predecir de forma coherente todas las localidades simultáneamente. Para ello, se adaptan y aplican las redes probabilísticas (en particular las redes Bayesianas), que son herramientas de minería de datos populares en otras disciplinas, pero que han sido aplicadas en contadas ocasiones en Meteorología; por ejemplo, en predicción local de granizo (Abramson et al., 1996) y en predicción de brisa (Kennett et al., 2001).

Por tanto, un primer objetivo de la Tesis consistió en realizar una primera aproximación a distintos problemas meteorológicos (diagnosis, predicción, clasificación e interpolación) utilizando esta metodología, con diversas publicaciones resultantes (Cofiño et al., 2002; Cano et al., 2004; Gutiérrez et al., 2004b).

El segundo objetivo de esta Tesis fue analizar en detalle la aplicación de estos modelos para la predicción local multivariada, aprovechando las ventajas que ofrece esta metodología y adaptando aquellos aspectos que han de ser optimizados para su correcto funcionamiento en el problema analizado (en particular el problema de aprendizaje). En este caso también se han realizado diversas publicaciones que están en proceso de revisión (Ancell et al., 2009; Ancell and Gutiérrez, 2009). Todo esto se ha llevado a cabo manteniendo una continua exploración del estado del conocimiento, que nos ha ayudado a componer el marco de referencia para desarrollar esta metodología.

1.2.2. Objetivos particulares

Los objetivos generales se ramifican de forma natural en diferentes objetivos particulares, sin que en esta Tesis se hayan agotado, ni mucho menos, todas las posibles aplicaciones.

- En primer lugar, se identificarán aquellos problemas de Meteorología que mejor se adecúan al tratamiento probabilístico con esta metodología, y se estudiarán algunas aplicaciones prácticas "piloto" basadas en RBs para ilustrar su potencial (ver Cap. 6). El procedimiento seguido para describir cada aplicación es el mismo en todos los casos; en primer lugar se plantea el problema, después se

establecen las limitaciones de la metodología al uso y, finalmente, se proponen soluciones y se contrastan las ventajas y desventajas de las RBs cuando se aplican en dichos problemas. En particular, el objetivo en este caso ha sido describir tres aplicaciones particulares, como son el tratamiento colectivo de la dependencia espacial en una red de observaciones, los generadores de tiempo y la predicción local.

- En segundo lugar, y casi como una consecuencia derivada del objetivo anterior, se ha tenido como objetivo específico abstraer distintos problemas de Meteorología relacionados con la predicción local para que puedan ser expresados como un único problema (con distintas fuentes de evidencia) en el marco de las redes Bayesianas. Así se han definido diferentes paradigmas de estimación probabilística local como la predicción, el *downscaling*, la clasificación y la interpolación (ver Cap. 7). El objetivo es poder resolver todos estos problemas como variantes de una única metodología de análisis, a través de un modelo global único para operar en los diferentes paradigmas, con las ventajas de integración, coherencia y robustez que ello conlleva, sobre todo cuando se considera la gran heterogeneidad, complejidad e incoherencia (no solo espacial) que actualmente presenta la predicción estadística operativa en los servicios meteorológicos nacionales.

- Finalmente, se trata de analizar a fondo una de estas aplicaciones (la predicción probabilística local de precipitación), desarrollando algoritmos específicos de aprendizaje, junto con medidas de calidad orientadas específicamente a contrastar la capacidad predictiva de los modelos resultantes, para los diferentes modos de operación propios de la Meteorología. El objetivo es tratar de evaluar la aportación operativa de esta nueva metodología, comparando con los métodos existentes.

1.3. Resumen de la Tesis

La Tesis está dividida en dos partes: la primera describe el estado del conocimiento y la segunda las aportaciones desarrolladas. Un último capítulo recoge las principales conclusiones y enumera las líneas futuras previstas de investigación.

1.3.1. Estado del Conocimiento

En esta primera parte se describe el estado del conocimiento, tanto de la *Predicción Meteorológica* a diferentes alcances temporales, como de la *Minería de Datos*, desde el punto de vista de su aplicación en la predicción probabilística. Esta parte está dividida en tres capítulos.

En el Cap. 2, bajo el título de *Predicción Meteorológica*, se muestra una panorámica actual de los modelos numéricos de circulación atmosférica y su aplicación a la predicción operativa a las distintas escalas de tiempo. En particular, se describen los importantes avances tanto en las técnicas de observación (estaciones automáticas, satélites, radares, etc.), como en las de cálculo, que permiten asimilar y manejar esta cantidad ingente de información para integrar los modelos que simulan la evolución de la atmósfera para predecir su estado futuro. Sin embargo, a pesar de estos avances, los modelos numéricos de simulación sufren limitaciones de origen matemático (incertidumbre en las condiciones iniciales), físico (dinámica de pequeña escala mal resuelta) y computacional (requerimientos de cálculo excesivos), que afectan a la resolución espacial a la que pueden ser resueltos. Para tener estas limitaciones e incertidumbres en cuenta, recientemente se ha abandonado el paradigma determinista y se ha recurrido a la predicción por conjuntos (Ensemble Prediction Systems, o EPS), realizando distintas integraciones a partir de condiciones iniciales y/o modelos ligeramente perturbados. Así, se obtienen un conjunto de predicciones que se resumen de forma probabilística para cuantificar la incertidumbre asociada.

A continuación, en el Cap. 3, titulado *Minería de Datos. Redes Bayesianas*, se describen brevemente el problema de la minería de datos desde la perspectiva de los modelos de redes probabilísticas (las redes Bayesianas en particular). En primer lugar se muestra la capacidad de estos modelos para codificar en un grafo todas las asociaciones/dependencias relevantes entre un conjunto de variables, y también para transferir de forma automática esta estructura de dependencia a una Función de Probabilidad Conjunta (FPC), definida a través de una factorización que tiene en cuenta esas dependencias y requiere un número reducido de parámetros. De esta manera, se puede calcular de forma eficiente cualquier probabilidad a posteriori y razonar sobre distintos aspectos de los problemas tratados de forma probabilística. Una vez introducida esta metodología, se dedica especial atención a los algoritmos de aprendizaje automático, que permiten obtener a partir de un conjunto de datos tanto el grafo como los parámetros que definen la FPC. Finalmente, dada la similitud con el problema que se pretende resolver en esta Tesis, se presta especial atención al aprendizaje de clasificadores.

Finalmente, en el Cap. 4 (*Medidas de Verificación de Predicciones Probabilísticas*) se analiza el problema de la validación o verificación de predicciones probabilísticas, puesto que serán utilizados profusamente a lo largo de la Tesis. En primer lugar se ilustra la complejidad del problema y se analizan aquellos aspectos de la calidad de una predicción que más nos interesa evaluar en el contexto de esta Tesis. A continuación, se introducen las distintas medidas de validación estándar para predicciones binarias y se describen las ventajas y desventajas de cada una de ellas.

1.3.2. Aportaciones de la Tesis

En esta segunda parte de la Tesis se describen las aportaciones realizadas tanto en el campo del aprendizaje estructural de los modelos de redes Bayesianas, como en las aplicaciones en el campo de la Meteorología y, en especial, la predicción probabilística local de fenómenos meteorológicos. Esta parte está organizada en seis capítulos, mas uno final de conclusiones.

En primer lugar, en el Cap. 5, titulado *Dominios de Estudio y Datos Utilizados*, se ofrece una descripción detallada del área geográfica de interés (la cornisa Cantábrica), y de los datos que se utilizan en los diferentes ejemplos prácticos desarrollados en esta Tesis. En primer lugar, se describen los datos observacionales y simulados (salidas de modelos numéricos) disponibles; en particular, se han considerado 42 estaciones con observación diaria de precipitación en una red de estaciones de la Agencia Estatal de Meteorología (AEMET) y se han considerado simulaciones procedentes de reanálisis y modelos operativos del ECMWF. En este capítulo también se analiza el problema de la discretización previa de las variables, ya que la metodología de las redes Bayesianas multinomiales utilizada es propia de variables discretas. A continuación, se describen en detalle los patrones atmosféricos utilizados para caracterizar el estado de la atmósfera en el área de estudio: un patrón de baja resolución, abarcando toda la península Ibérica y Baleares, y otro de alta resolución, sobre la región del Cantábrico.

Una vez descritas las bases de datos y sus dominios de aplicación, el Cap. 6, titulado *Aplicaciones de las Redes Bayesianas en Meteorología*, describe las aplicaciones más inmediatas de las redes Bayesianas en diferentes problemas de interés meteorológico, obtenidas sin realizar ningún tipo de adaptación de los algoritmos existentes, y que constituyen un análisis exploratorio de esta metodología en el ámbito de la Meteorología. En este caso, las aplicaciones han sido desarrolladas sobre casos sencillos, tanto de diagnosis como de predicción, manteniendo un espíritu más descriptivo que operativo. En particular se analizan los problemas de la distribución

espacial de la precipitación, los generadores de tiempo y, finalmente, la predicción local.

En el Cap. 7, titulado *Paradigmas de Predicción Probabilística Local*, se analiza la versatilidad de las RBs en todos los paradigmas de estimación probabilística local: diagnósis, predicción, *downscaling*, clasificación e interpolación, mostrando que todos estos problemas se pueden resolver con una misma metodología, y que responden a problemas de inferencia con distintos conjuntos de evidencia. A continuación se analiza el caso del *downscaling* estadístico, objeto de esta Tesis, comenzando con una descripción de los distintos métodos estándar que se aplican para resolver este problema, incidiendo en algunas virtudes y defectos, y destacando la capacidad de las RBs para generalizar algunos de estos métodos. En particular se presentan los *modelos ingenuos de Bayes* como una metodología estándar de predicción probabilística que generaliza la idea del *downscaling* basado en tipos de tiempo. Asimismo se describen las deficiencias de estos modelos, en cuanto a la independencia espacial de las predicciones y se describen vías de mejora utilizando redes Bayesianas que aprendan de forma automática la estructura óptima para llevar a cabo este tipo de predicción.

En el Cap. 8, titulado *Algoritmos de Aprendizaje para la Predicción Local*, se entra de lleno en el problema del aprendizaje estructural, teniendo encuenta que cuando se plantean modelos basados en RBs para representar sistemas de predicción local, el coste computacional del aprendizaje estructural por un lado, y la escasa longitud del conjunto de entrenamiento disponible por otro, hacen que solo sean prácticos modelos con muy pocos parámetros. En esta Tesis se considera un predictor multinomial y un conjunto de predictandos binarios, que se corresponden con la ocurrencia de precipitación sobre distintos umbrales en distintas estaciones distribuídas espacialmente. Los métodos de aprendizaje existentes no permiten optimizar la estructura de la red para el problema concreto de la predicción, obteniedo modelos óptimos para el diagnóstico pero que no reproducen todo el potencial de las relaciones de dependencia existentes de cara a la predicción. Por tanto, en este capítulo se proponen técnicas de aprendizaje específicas para este problema, basadas en algoritmos de búsqueda de estructuras óptimas con una medida especial de calidad que cuantifica la capacidad predictiva de la red. Esta técnicas permiten obtener topologías "ingenuas extendidas" para abordar la predicción local.

Una vez desarrollados los algoritmos específicos, en el Cap. 9, titulado *Predicción Local con Redes Bayesianas*, se utilizan para comparar experimentalmente las diferentes opciones de RB que son de aplicación en problemas meteorológicos probabilísticos y discretos: *ingenua*, *extendida* y *genérica*, sobre los paradigmas: *down-*

scaling, *predicción* e *interpolación*, desde el punto de vista de la predicción local, es decir evaluando únicamente las probabilidades marginales. En este caso, para simplificar los cálculos, se consideran redes formadas por nodos que representan series sin lagunas y de la misma longitud, y que por tanto, realizan su aprendizaje en condiciones homogéneas. El capítulo se divide en dos partes, en la primera se describe el experimento, mientras que en la segunda, se analizan los resultados tanto en su conjunto, como de forma separada para cada uno de los tres paradigmas.

Para terminar, en el Cap. 10, titulado *Downscaling en Redes Heterogéneas*, nos centramos en la aplicación práctica del problema del *downscaling* en condiciones realistas, en las que se dispone de redes de observación heterogéneas, con lagunas y con series de distinta longitud (series históricas largas junto a series cortas procedentes de nuevas estaciones). Esta heterogeneidad es inevitable para abordar estudios de alta resolución, que utilicen las nuevas redes que se han implantado durante la última década en distintas zonas geográficas, de procedencia diversa. Para tratar este problema, se introduce un nuevo algoritmo de tipo jerárquico, denominado *h2*, diseñado específicamente para utilizar conjuntos de entrenamiento diferentes (largos y cortos) que aprendan la estructura predictor-predictando o predictando-predictando, según sea más conveniente, para eliminar problemas de sobreajuste e integrando en un mismo modelo varias redes de observación.

Finalmente, el Cap. 11, titulado *Conclusiones y Trabajo Futuro* se ofrece un resumen de las principales conclusiones y posibilidades de trabajo futuro que se derivan de esta Tesis.

1.4. Algunos Acrónimos y Terminología Utilizados

- ACM: Modelo de Circulación Atmosférica (*Atmospheric Circulation Model*).
- AEMET: Agencia Estatal de Meteorología (antes INM).
- Atributo: En esta Tesis este concepto se utiliza para designar a las variables, casi siempre binarias, que representan a los predictandos.
- AUC: Area bajo la curva ROC (*Area Under Curve*).
- Augmented: Tipo de red Bayesiana extendida que se obtiene agregando enlaces a una red *ingenua*.
- BD: Base de Datos.

- BS: Brier Score, error cuadrático medio de la probabilidad prevista (ver Cap. 4).

- BSS: *Brier Skill Score*, índice de pericia que se obtiene como $1 - BSP/BSR$, donde BSP es el BS de la predicción y BSR es el BS de un sistema de referencia, normalmente la climatología (ver Cap. 4).

- CCA: Análisis de Correlación Canónica (*Canonical Correlation Analysis*).

- Clase: En esta Tesis, por analogía con los problemas de clasificación, se utiliza este concepto para designar al escalar que representa al estado de la atmósfera (variable clase).

- Clasificador Ingenuo: Red Bayesiana en la que todos los atributos están enlazados exclusivamente a la clase.

- CP: Componente principal (ver Sec. 3.7).

- CPD: Distribución de Probabilidad Condicional (*Conditional Probability Distribution*, ver Cap. 3).

- CSI: Independencia específica (*Context-Specific Independence*, ver Sec. 3.5.1).

- DAG: Grafo Dirigido Acíclico (*Directed Acyclic Graph*, ver Sec.3.4).

- Diagnosis: Paradigma de predicción en el que la evidencia es algún atributo o predictando.

- Downscaling: Mejora de resolución de una predicción de un modelo numérico. Paradigma de predicción en el que la evidencia es el estado de la atmósfera.

- ECMWF: Centro europeo para predicción a plazo medio (*European Center for Medium-Range Weather Forecast*, `www.ecmwf.org`).

- ENSO: Patrón de oscilación de El Niño-Pacífico Sur (*El Niño-Southern Oscillation*).

- EOF: Función empírica ortogonal (*Empirical Ortogonal Function*).

- EPS: Sistema de predicción por conjuntos (*Ensemble Prediction System*).

- FAR: Probabilidad de predecir un evento, dado que no se ha observado (*False Alarm Rate*, ver Cap. 4). No confundir con *False Alarm Ratio* que corresponde a la probabilidad de no observar un evento, dado que se ha previsto, que también se suele utilizar bajo el mismo acrónimo.
- FPC: Función de Probabilidad Conjunta.
- HIR: Probabilidad de predecir un evento, dado que se ha observado (*Hit Rate*, ver Cap. 4).
- HIRLAM: Modelo atmosférico regional (*High Resolution Limited Area Model*, `hirlam.knmi.nl`).
- IA: Inteligencia Artificial.
- INM: Instituto Nacional de Meteorología, `www.INM.es`.
- IPCC: *Intergubernamental Panel on Climate Change*.
- *h2*: Acrónimo utilizado para el algoritmo *jerárquico*.
- MD: Minería de Datos.
- MDL: *Minimum Description Length*, ver Sec. 3.5.2.
- MI: Información Mutua (*Mutual Information*, ver ec. 3.6).
- MM5: Modelo atmosférico regional (*Penn State/NCAR Mesoscale Model*, `box.mmm.ucar.edu/mm5`).
- Modelo Conceptual: En el ámbito de la Meteorología, descripción de los procesos característicos de las diferentes escalas y dinámicas de la atmósfera (una tormenta, un frente, etc.).
- Modelo Discriminativo: Modelo que estima directamente la CPD de la clase dados los atributos para un problema de clasificación (ver Sec. 7.1).
- Modelo Generativo: Modelo que estima previamente la FPC y después deduce de ella la CPD para un problema de clasificación, ver Sec. 7.1.
- Naive: Ingenuo. Se utiliza para designar la red bayesiana en la que todos los atributos están enlazados exclusivamente al clasificador.
- NAO: Oscilación del Atlántico Norte (*North Atlantic Oscillation*).

- NCEP: *National Centers for Environmental Prediction*: `www.ncep.noaa.gov`.
- Nowcasting: Paradigma de predicción, a muy corto plazo, en el que además de conocer los campos previstos por un modelo numérico, se pueden conocer eventualmente datos de observación en tiempo real.
- Patrón: Estructura numérica con una cierta organización de sus elementos (por ejemplo, una estructura matricial o tensorial relativa a una rejilla sobre una region geográfica concreta). Para su tratamiento matemático los patrones serán considerados vectores, descomponiendo su estructura de forma apropiada, y componiéndola de nuevo cuando sea necesario.
- Predicción Local: Predicción adaptada a un punto geográfico concreto.
- PDF: *Probability Distribution Function*.
- RB o BNet: Red Bayesiana (*Bayesian network*, ver Sec.3.4).
- Re-análisis: Experimento de simulación numérica de la circulación atmosférica para un largo período de tiempo con un mismo modelo numérico, asimilando toda la información disponible (ver Sec. 2.3).
- RMSE: Error cuadrático medio (*Root Mean Square Error*).
- ROC, Curva obtenida a partir de los aciertos y las falsas alarmas de un sistema de predicción binario (*Relative Operating Characteristic*, ver 4).
- RSA, Area comprendida entre la diagonal y la curva ROC a partir de los aciertos y las falsas alarmas de un sistema de predicción binario (*Roc Skill Area*, ver 4).
- SAF: *Satellite Application Facilities*.
- SLP: Presión a nivel del mar (*Sea Level Pressure*).
- SOM: Redes auto-organizativas (*Self-Organizing Maps*) para agrupamiento y visualización de datos.
- SP: Algoritmo *Super-Parent* de Keogh-Pazzani.
- SST: Temperatura de la superficie del agua del mar (Sea Surface Temperature).

- TAC: Turbulencia en aire claro.
- TAN: Red Bayesiana extendida (*tree-augmented naive*).
- UKMO: Servicio meteorológico británico. *United Kingdom Meteorological Office*. También designa al modelo numérico de dicho servicio.
- UTC: Coordenadas universales de tiempo (*Universal Time Coordinates*).
- WG: Generador de tiempo (*Weather Generator*).
- WMO: *World Meteorological Organization*, `www.wmo.ch`.

Parte I

Estado del Conocimiento

CAPÍTULO 2

Predicción Meteorológica

Este es el primer capítulo de los dedicados al estado del conocimiento, donde se trata de ofrecer al lector una perspectiva actual del problema de la predicción meteorológica, y está dividido en cuatro secciones mas una breve introducción. En la Sec. 2.1 se explican las principales ecuaciones que gobiernan el sistema atmosférico; posteriormente, la Sec. 2.2 describe brevemente el problema de la predicción numérica, comentando aspectos como la asimilación de datos (para inicializar el modelo), los diferentes alcances de la predicción y los distintos modelos que son actualmente utilizados en la predicción operativa a distintas escalas temporales, desde la predicción inmediata (con horas de anticipación), hasta la climática (décadas). En la Sec. 2.3 se muestran diferentes sistemas de observación de tipo global y local, describiendo en particular la red de observación de la AEMET, puesto que será la utilizada en esta Tesis. Finalmente, en la Sec. 2.4, se describen las series de simulaciones obtenidas a partir de los modelos numéricos, con particular interés en la serie de reanálisis *ERA40* generada por el modelo de predicción numérica ECMWF (European Center for Medium-Range Weather Forecast), ya que también será la utilizada en esta Tesis.

2.1. La Atmósfera

Desde un punto de vista planetario, la atmósfera es uno de los componentes más activos del *sistema climático*, formado además por la hidrosfera, criosfera, litosfera y biosfera (ver Fig. 2.1). Sin embargo, la atmósfera se puede estudiar de forma aislada considerando escalas de tiempo meteorológicas (del orden de una semana), pues los

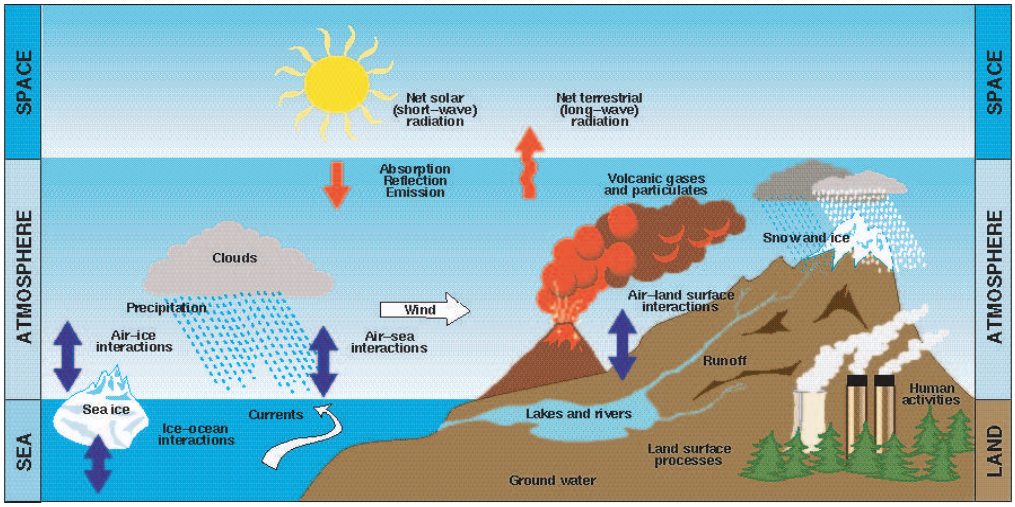

Figura 2.1: Vista esquemática de los componentes del sistema climático.
FUENTE: http://www.dsostenible.com.ar/.

restantes componentes del sistema tienen muy poca variabilidad a esta escala (la atmósfera es el componente que tiene la dinámica más rápida). En esta Tesis nos centramos en estas escalas de tiempo, en concreto en la predicción a corto y medio plazo.

2.1.1. Descripción Física

En este apartado se ofrece una breve descripción de las ecuaciones que controlan la dinámica de la atmósfera, con la única intención de ilustrar la complejidad de los modelos involucrados en la predicción numérica del tiempo a corto plazo. Para un estudio detallado, existen numerosos libros de Meteorología dinámica con excelentes explicaciones y ejemplos de las deducciones y descripciones físicas de las ecuaciones; ver, por ejemplo Holton (1992), Barry and Chorley (1998) o Kalnay (2003).

Desde un punto de vista físico, la atmósfera es una mezcla de gases y agua en sus distintos estados, moviéndose en el campo gravitatorio de una esfera heterogénea en rotación, cuyo eje se encuentra ligeramente inclinado respecto al plano de la eclíptica y, lo fundamental, calentada por el Sol. Como se muestra en la Fig. 2.2, estos forzamientos y restricciones confieren a la atmósfera una dinámica particular, que se manifiesta en un cierto comportamiento promedio estadístico que muestra una organización a gran escala en células de circulación para redistribuir la energía y lograr un equilibrio.

En este sistema se deben cumplir las leyes de conservación: energía, masa y

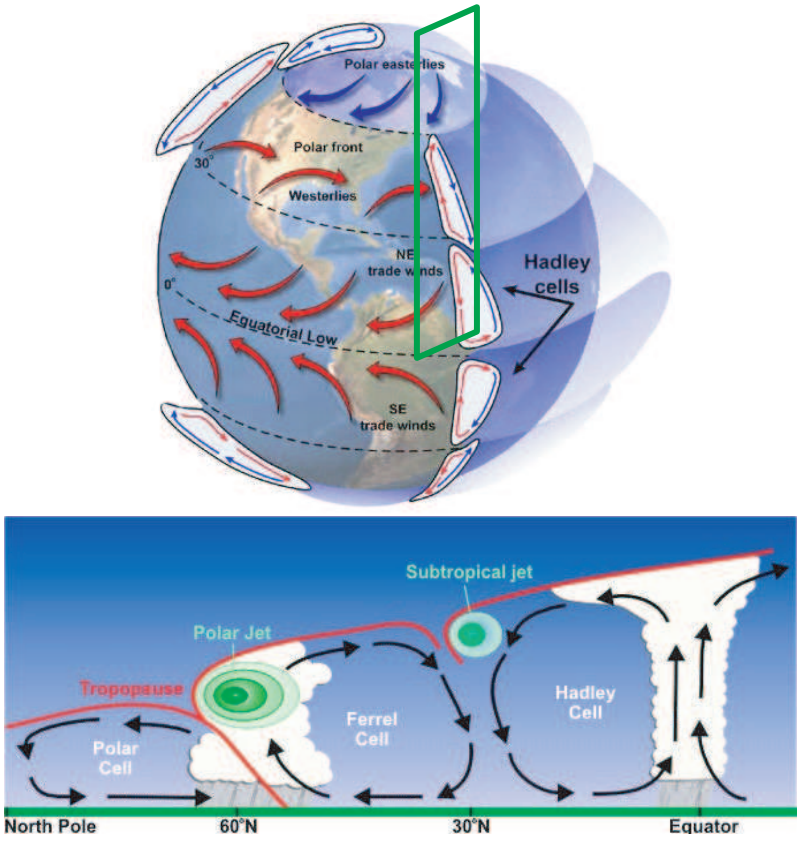

Figura 2.2: Esquema sencillo de la circulación atmosférica global (arriba) y un corte vertical de las tres células básicas de la circulación (abajo). *FUENTE: http://earthobservatory.nasa.gov/* .

momento, y la ecuación de estado de los gases. La mayoría de estas leyes se formulan con ecuaciones que expresan las derivadas totales de ciertas magnitudes físicas en el tiempo (Lagrangiana o individual), descompuestas en sus términos local (Euleriana o parcial) y advectivo.

- Según la *ecuación de conservación de masa*, la variación total en el tiempo de la masa de una parcela de aire es nula ($dM/dt = 0$). La masa de aire contenida en un volumen $\Delta x \Delta y \Delta z$ es:

$$M = \rho \Delta x \Delta y \Delta z \tag{2.1}$$

donde ρ es la densidad del aire. Considerando la ecuación

$$\frac{1}{M}\frac{dM}{dt} = 0, \tag{2.2}$$

sustituyendo (2.1) y descomponiendo la derivada total, se tiene la ecuación de conservación de masa

$$\frac{1}{\rho}\frac{d\rho}{dt} + \boldsymbol{\nabla}\cdot\boldsymbol{v} = 0 \Leftrightarrow \frac{\partial\rho}{\partial t} = -\boldsymbol{\nabla}\cdot(\rho\,\boldsymbol{v}), \tag{2.3}$$

ya que $\partial u/\partial x = \Delta x^{-1} d\Delta x/dt$.

- La *ecuación de conservación de vapor de agua* determina que la variación total de vapor de agua en una parcela de aire se debe exclusivamente a procesos internos de evaporación E (fuente) y/o condensación C (sumidero):

$$\frac{dq}{dt} = E - C \tag{2.4}$$

donde q es la proporción en masa de vapor de agua en la parcela de aire (g/kg). Si esta ecuación se multiplica por ρ, se descompone la derivada total, y se suma a la ecuación de conservación de masa (2.3) multiplicada por q, queda

$$\frac{\partial\rho\,q}{\partial t} = -\boldsymbol{\nabla}\cdot(\rho\,\boldsymbol{v}\,q) + \rho\,(E - C). \tag{2.5}$$

De la misma forma que se ha incluido una ecuación de conservación para el vapor de agua, se podría incluir cualquier otra ecuación de conservación para otros elementos como agua líquida, ozono, etc., mientras también se incluyan sus correspondientes *fuentes* y *sumideros*.

- La *ecuación de estado de los gases ideales* aplicada a la atmósfera impone la siguiente relación entre las variables de estado termodinámicas:

$$p\,\alpha = R\,T, \tag{2.6}$$

donde p es la presión (mb ó Hpa), T la temperatura (^oC ó K), R es la constante de los gases ideales, y α es el volumen específico (m^3/kg), inverso de la densidad ρ (kg/m^3).

- *Conservación de energía:* La atmósfera es transparente a la radiación solar directa, de manera que su principal fuente de calor es la superficie terrestre (calentada por el Sol), entendiendo como tal tierra y océano. Este calor, cuando es absorbido por la atmósfera se emplea en variar su temperatura, su volumen o ambas cosas a la vez; es decir, si se aplica una tasa de calor Q por unidad de masa a una parcela de aire $(cal/s\,g)$, esta energía es empleada en aumentar la energía interna $C_v\,T$ y(ó) producir un trabajo de expansión

$$Q = C_v \frac{dT}{dt} + p\frac{d\alpha}{dt}, \tag{2.7}$$

donde los coeficientes de calor específico a volumen constante (C_v) y a presión constante (C_p) se relacionan mediante $C_p = C_v + R$. Haciendo uso de la ecuación de estado se puede obtener una forma alternativa de la ecuación de conservación de energía:

$$Q = C_p \frac{dT}{dt} - \alpha\frac{dp}{dt}. \tag{2.8}$$

- *Conservación del Momento:*

$$\frac{d\boldsymbol{v}}{dt} = -\alpha\boldsymbol{\nabla}p - \boldsymbol{\nabla}\phi + \boldsymbol{F} - 2\boldsymbol{\Omega}\times\boldsymbol{v}; \tag{2.9}$$

La aceleración sobre la unidad de masa es debida a cuatro fuerzas: gradiente de presión ($-\alpha\boldsymbol{\nabla}p$), gravedad aparente ($-\boldsymbol{\nabla}\phi$), rozamiento ($\boldsymbol{F}$) y Coriolis ($-2\boldsymbol{\Omega}\times\mathbf{v}$).

Por tanto, resumiendo lo anterior, se tienen siete ecuaciones y siete incógnitas: $\boldsymbol{v} = (u, v, w)$, T, p, $\rho = 1/\alpha$ y q, que gobiernan la dinámica a gran escala de la atmósfera:

$$\begin{aligned}
\frac{d\boldsymbol{v}}{dt} &= -\alpha\boldsymbol{\nabla}p - \boldsymbol{\nabla}\phi + \boldsymbol{F} - 2\boldsymbol{\Omega}\times\boldsymbol{v} && (2.10)\\
\frac{\partial\rho}{\partial t} &= -\boldsymbol{\nabla}\cdot(\rho\,\boldsymbol{v}) && (2.11)\\
p\,\alpha &= R\,T && (2.12)\\
Q &= C_p\frac{dT}{dt} - \alpha\frac{dp}{dt} && (2.13)\\
\frac{\partial\rho\,q}{\partial t} &= -\boldsymbol{\nabla}\cdot(\rho\,\boldsymbol{v}\,q) + \rho\,(E - C) && (2.14)
\end{aligned}$$

Considerando que el objetivo de la predicción numérica del tiempo es obtener el estado de la circulación atmosférica en un tiempo futuro a partir de la condición inicial, es necesario disponer de un modelo numérico capaz de integrar las ecuaciones y que incluya los intercambios energéticos (principalmente en la capa límite) más importantes (radiación, turbulencia, calor latente, etc.); dada la no linealidad y complejidad de las ecuaciones, en la práctica se suele recurrir a distintas aproximaciones simplificadoras. Las simplificaciones más elementales de las ecuaciones anteriores ayudan a aislar y entender las características de los diferentes tipos de ondas básicos presentes en la atmósfera que, por supuesto, también aparecen en las soluciones más generales. Por ejemplo, considerando un *flujo inicial zonal* ($v = w = 0$) y un campo

uniforme de densidad, cualquier perturbación al flujo zonal supone un cambio de latitud y por tanto un cambio en la fuerza de Coriolis, que siempre actúa tratando de restaurar el flujo zonal produciendo las llamadas *ondas de Rossby* que se propagan alrededor del globo con un periodo de desarrollo de varios días a una semana y con una longitud de onda del orden de 3000 km para latitudes medias ($\varphi = 45^{o}$).

Por tanto, los modelos operativos utilizados en la predicción numérica del tiempo son aproximaciones más o menos completas de las ecuaciones de la atmósfera.

2.1.2. Escalas de la Atmósfera

Los fenómenos de la atmósfera ocurren a distintas escalas espacio-temporales, dependiendo de la homogeneidad e intensidad de las fuerzas participantes. Por ejemplo, cuando las fuerzas del gradiente de presión son homogéneas en el espacio y el tiempo, la fuerza de Coriolis produce un equilibrio originando los grandes sistemas de circulación de escala sinóptica, como las ondas de Rossby, anticiclones y borrascas (estos dos últimos contando también con la fuerza de rozamiento), con longitudes de onda en torno a $3000km$ y periodos entorno a 5 días. Cuando las fuerzas del gradiente de presión no son homogéneas en el espacio y el tiempo (debido principalmente a la interacción entre calentamientos diferenciales, al ciclo diurno y a las inhomogeneidades en superficie) no pueden ser estabilizadas por la fuerza de Coriolis, provocando fluctuaciones mesoescalares en la circulación sinóptica, con longitudes de onda típicas de $200km$, para una descripción detallada ver Orlanski (1975) o Jansá (1990). En la figura 2.3 se muestra la clasificación de Orlanski de las diferentes escalas meteorológicas.

Estas distintas escalas espaciales y temporales que componen la dinámica atmosférica han de ser tenidas en cuenta en el proceso de resolución numérica para considerar sólo aquellos factores relevantes a la escala que se desee analizar. En general, interesa que las soluciones obtenidas para la circulación de la atmósfera en un problema operativo muestren el comportamiento de interés meteorológico, por lo que es deseable eliminar algunas de las ondas básicas que no intervienen en esta dinámica (por ser de menor magnitud y/o distinta escala temporal), pero que pueden producir inestabilidades en el modelo numérico al ser propagadas (esto ocurre, por ejemplo con las ondas sonoras y gravitatorias). Este procedimiento de selección de tipos de ondas se denomina *filtrado* y consiste en eliminar parcialmente las fuerzas restauradoras que originan la onda que se desea filtrar. Por ejemplo, la *aproximación hidrostática* impide la propagación vertical de las ondas sonoras.

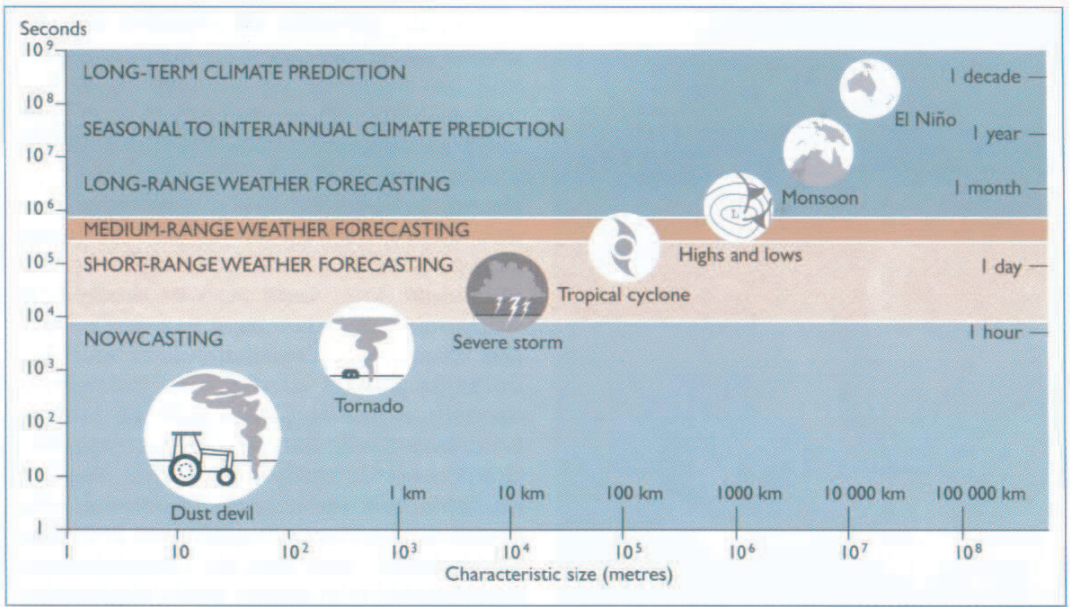

Figura 2.3: Escalas meteorológicas; adaptado de Orlanski (1975).

2.1.3. No Linealidad y Caos

El comportamiento colectivo de la atmósfera tiene como consecuencia que todas las escalas son dependientes entre sí, siendo el estado de las escalas grandes (lentas) dependiente del estado de las escalas pequeñas (rápidas) y viceversa. Esto supone un reto añadido para los modelos de circulación atmosférica (ACMs, según sus siglas en Inglés) pues no sólo deben ser capaces de reproducir los patrones y la variabilidad característicos de cada escala sino que además han de ser capaces de tener en cuenta la interacción entre las escalas.

Este aspecto está contemplado en los términos no lineales que contienen los modelos de circulación atmosférica y, lejos de suponer una ventaja, introducen una gran sensibilidad a pequeñas perturbaciones en el sistema, es decir, cualquier error crecerá exponencialmente en el tiempo acotando el horizonte de predecibilidad de los ACMs. Este comportamiento es conocido como caos determinista, e impone una limitación teórica al horizonte de predecibilidad de la circulación atmosférica, según el tipo de aproximaciones físicas y escalas que se consideren en el modelo; para una introducción a los sistemas no lineales, ver Strogatz (2001) o Kalnay (2003). Otro hecho destacable es que la dinámica caótica se desarrolla sobre un soporte fractal en el espacio de fases, llamado *atractor* (Grassberger and Procaccia, 1983). Así, geometría fractal y dinámica caótica están íntimamente ligados (Mandelbrot,

2004).

El caos fue analizado por primera vez en el siguientes sistema de ecuaciones:

$$\begin{cases} \dot{x} = \sigma(y - x) \\ \dot{y} = r\,x - y - x\,z \\ \dot{z} = x\,y - b\,z \end{cases} , \tag{2.15}$$

propuesto por Lorenz (1963) mientras estudiaba la posibilidad de predecir estructuras convectivas, tan frecuentes en la naturaleza y en tan diversas escalas. Para ello tomó como punto de partida las formaciones convectivas hidrodinámicas de Rayleigh-Bénard, por ser fáciles de generar y reproducir en el laboratorio. Basta con calentar por debajo una capa muy fina de un fluido viscoso, como la glicerina. El fluido caliente formado en la parte baja tiende a elevarse creando corrientes convectivas en un proceso parecido a lo que sucede en la atmósfera. Para un calentamiento suficientemente intenso la evolución temporal de estas corrientes convectivas tiene un comportamiento caótico; es decir, pequeñas perturbaciones producen cambios impredecibles en su dinámica tras un tiempo suficiente. Las ecuaciones que describen este experimento tienen en cuenta los tres efectos más importantes que aparecen: La fuerza debida al gradiente térmico, la viscosidad y la difusión térmica. Lorenz propuso una simplificación de este problema, suponiendo que el campo de velocidades era el gradiente de una función Ψ y que existían simetrías espaciales. El modelo resultante (2.15) presenta dinámicas caóticas para ciertos valores de los parámetros (para otros valores su comportamiento es periódico). Por ejemplo, para los valores de los parámetros $\sigma = 10$, $b = 8/3$, y $r = 28$, el sistema presenta una dinámica caótica. Partiendo de la condición inicial $(x(0), y(0), z(0)) = (-10, -5, 35)$ y utilizando un método de integración de Runge-Kutta de cuarto orden y paso de integración $\Delta t = 10^{-2}$ se obtiene la evolución del sistema en el espacio de fases que se encuentra dibujado en la Fig. 2.4(a); las Fig. 2.4(b)-(d) muestran la evolución en el tiempo de las tres variables. Por otra parte, la Fig. 2.5 ilustra la sensibilidad de este sistema a las perturbaciones, que se manifiesta de manera distinta según la posición en el atractor del mismo: en la figura izquierda, las condiciones iniciales dentro de una bola de error se mantienen agrupadas a lo largo de la evolución del sistema, mientras que en la parte derecha vemos que otras condiciones iniciales en una bola de error similar presentan un comportamiento notablemente divergente para el mismo sistema.

El principal problema de este simple y pedagógico sistema de Lorenz es que no tiene en cuenta la componente espacial y, por tanto, ilustra sólo parcialmente el problema real que ocurre en la circulación atmosférica y en otros sistemas espacio-

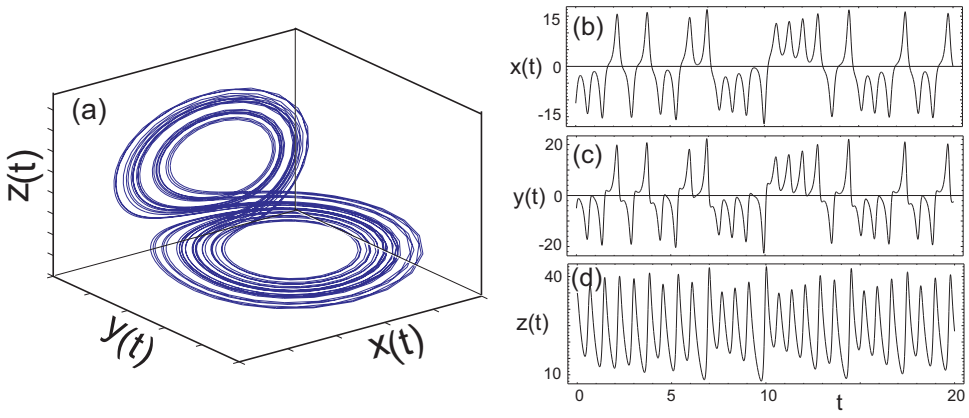

Figura 2.4: (a) Atractor del sistema de Lorenz. (b)-(d) Orbitas descritas por cada una de las tres variables del sistema.

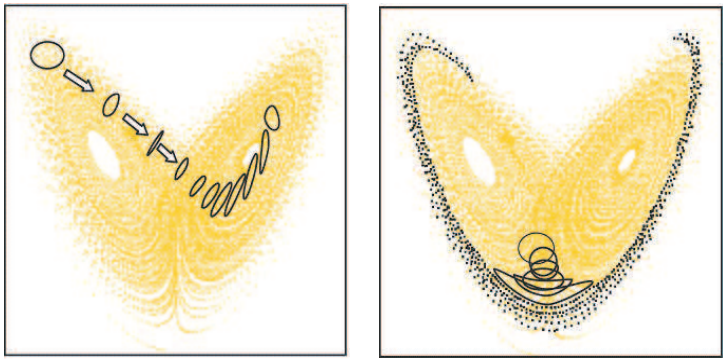

Figura 2.5: Ejemplo de sensibilidad a las condiciones iniciales. *Fuente ECMWF.*

temporales, y que responde a leyes de crecimiento más complejas debidas a la interacción de las componentes espacial y temporal de la dinámica (ver Gutiérrez et al., 2008; Primo et al., 2007, para una descripción más detallada de este problema).

2.2. Predicción Numérica

A nivel operativo, las ecuaciones mostradas en la Sec. 2.1.1 se integran numéricamente discretizando el espacio y el tiempo, partiendo de una condición inicial. Por ejemplo, en coordenadas cartesianas se considera una rejilla 4D sobre la que se aplican técnicas de diferencias y elementos finitos, mientras que en coordenadas esféricas se considera un número finito de armónicos esféricos para la descripción del sistema (para una introducción a estos métodos ver Kalnay, 2003). La Fig. 2.6 ilustra esquemáticamente este proceso.

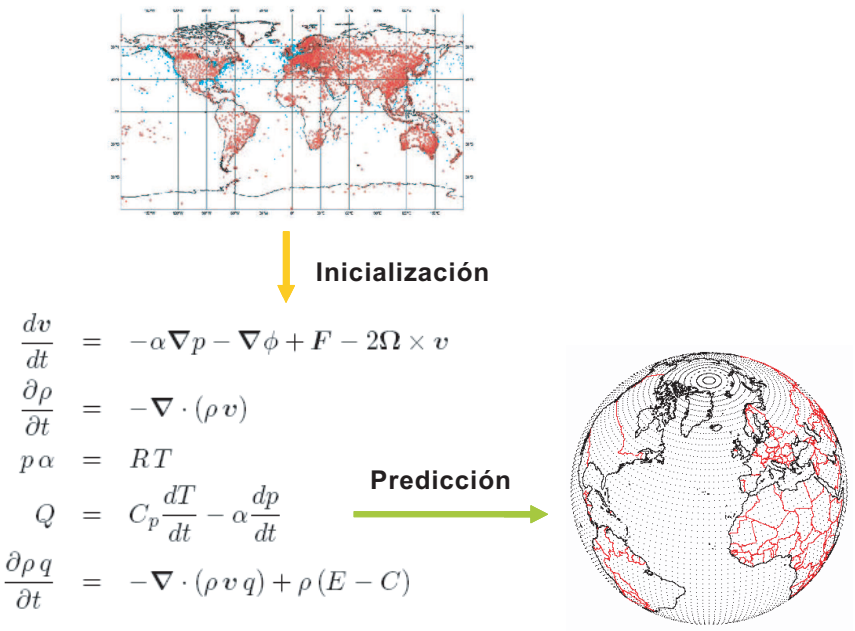

Figura 2.6: Esquema básico de un modelo global de predicción numérica. Se muestran únicamente los procesos fundamentales de la inicialización, a partir de un conjunto de observaciones, y la predicción resultante de la integración.

Dependiendo de su resolución y alcance, una predicción está afectada en mayor o menor medida por la no linealidad de la atmósfera, que impone un horizonte a la predicción determinista y hace necesario plantear la predicción en términos probabilísticos. La no linealidad también limita la resolución espacial con que puede obtenerse una predicción para un alcance dado (a medida que aumenta el alcance, decrece la resolución espacial de los fenómenos que se pueden predecir). La Figura 2.7 ilustra estas relaciones y detalla el tipo de predicción (determinista o probabilística) adecuada para cada uno de los alcances (ver Sec. 2.2.3 para más detalles). La predicción en términos probabilísticos también puede ser una elección, por ejemplo en el corto plazo, para cuantificar la incertidumbre de las predicciones.

Tradicionalmente, las predicciones se han obtenido integrando un único modelo numérico a partir de una condición inicial dada. Sin embargo, en la década de los noventa, tras el descubrimiento del caos determinista, las técnicas de predicción por conjuntos han permitido considerar los efectos causados por las distintas fuentes de incertidumbre en la evolución del modelo, y producir predicciones probabilísticas. Hoy día se utilizan en la predicción operativa distintos esquemas de *predicción por conjuntos* que perturban las condiciones iniciales, o combinan las salidas de distintos modelos (*multi-modelo*), para obtener un conjunto de predicciones con el que

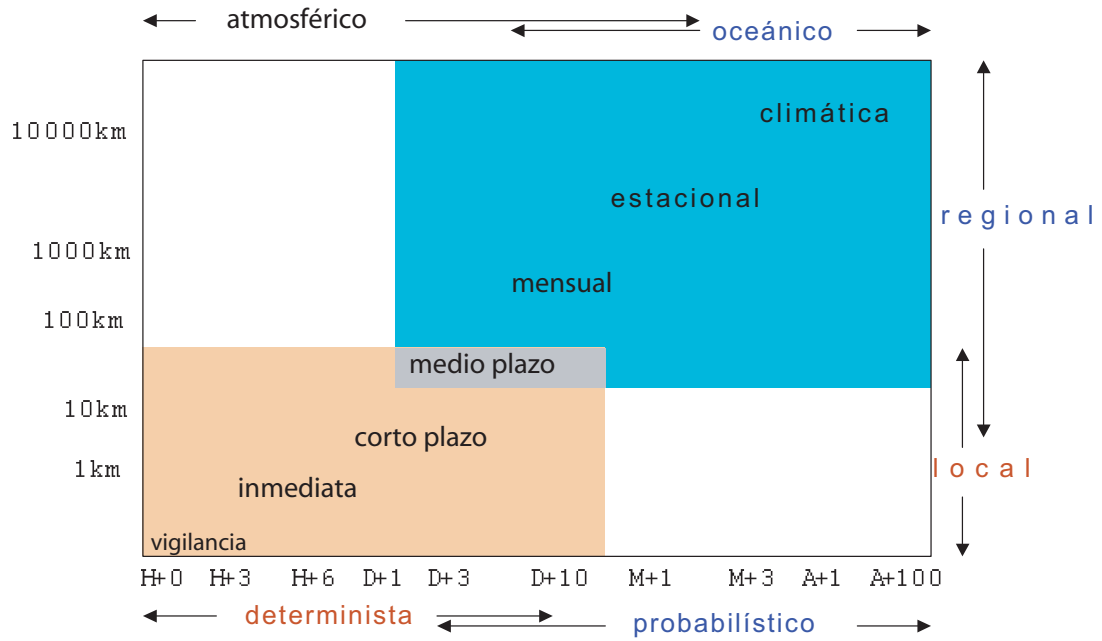

Figura 2.7: Relación entre el alcance de las predicciones, la escala espacial considerada y algunas de sus características fundamentales.

anticipar la evolución de la atmósfera (ver más detalles en Sec. 2.2.2).

Las predicciones operativas que se realizan en la Agencia Estatal de Meteorología (AEMET, `www.aemet.es`) y en otros centros meteorológicos europeos (MeteoFrance, UK MetOffice, etc.) se basan en este esquema de modelos numéricos y técnicas de predicción por conjuntos, que constituyen actualmente las herramientas fundamentales para la predicción meteorológica.

2.2.1. Asimilación de Datos

La predicción numérica es un problema de condicionas iniciales: dada una estimación actual del estado de la atmósfera, un modelo numérico simula su evolución para obtener una predicción. Esta condición inicial se establece a partir de la interpolación y adaptación dinámica de las observaciones disponibles a los puntos de rejilla del modelo, y el proceso de obtención es denominado *asimilación*. Hay que tener en cuenta que la distribución espacial y temporal de las observaciones no es uniforme, y que la cantidad de observaciones disponibles no es suficiente para definir una condición inicial del modelo en todos sus grados de libertad.

Para compatibilizar dinámicamente las observaciones con las ecuaciones del ACM, se suele usar información a priori (*first-guess*) proporcionada por una predicción a corto plazo realizada en los ciclos de asimilación. Estos ciclos de asimilación son

interpolaciones, cada vez más refinadas, donde la matriz de pesos se determina minimizando los errores en cada punto de rejilla. Otros métodos modernos de asimilación utilizan técnicas variacionales 4D, que tienen en cuenta la dinámica espacial y temporal del sistema para lograr dinámicamente un estado asimilado que pase lo más cerca posible de las observaciones (ver Kalnay, 2003, para una introducción moderna al problema de asimilación). Este problema es tan complejo que supone una gran parte del tiempo computacional requerido en la predicción operativa.

2.2.2. Tratamiento de la incertidumbre

Como se ha visto en la Sec. 2.1.3, la no linealidad de los modelos numéricos de circulación atmosférica plantea un problema de predecibilidad, ya que cualquier incertidumbre o error introducido en el sistema crece exponencialmente en el tiempo. Por lo tanto, para caracterizar correctamente el espacio de posibles estados futuros, la predicción debe considerar la incertidumbre real inicial.

Analizado de forma general, el origen de las fluctuaciones en estos modelos puede provenir de errores en la condición inicial o de errores en la modelización del sistema. La dispersión producida por los errores en la condición inicial se suele analizar realizando conjuntos de predicciones a partir de diferentes condiciones iniciales perturbadas. Esta técnica se denomina *predicción por conjuntos* o EPS (Ensemble Prediction System). Estos conjuntos se pueden formar de muy diversas formas, siendo el método de *vectores singulares* (Buizza and Palmer, 1995) o el método de *breeding* (Pu et al., 1997) los más utilizados para la generación de las diferentes predicciones, denominadas miembros, a partir de perturbaciones específicas de las condiciones iniciales (ver Gutiérrez et al., 2004b, para más detalles). El método de vectores singulares ha sido adoptado por el ECMWF y el método de breeding por el NCEP. Por ejemplo, la Fig. 2.8 representa un esquema de la evolución de predicciones obtenidas a partir de condiciones iniciales perturbadas y la correspondiente trayectoria de control obtenida de las condiciones iniciales sin perturbar.

Para tener en cuenta los errores en la modelización del sistema, también se pueden aplicar técnicas de predicción por conjuntos; en este caso se consideran distintos modelos (o perturbaciones de un mismo modelo) para generar cada uno de los miembros, y por ello es conocida como *ensemble multimodelo*. Estas técnicas han mostrado su eficacia en distintos proyectos, como el proyecto DEMETER para predicción estacional (Palmer et al., 2004). La Figura 2.9, muestra el caso en el que la incertidumbre la aporta el modelo y, por tanto, cada uno de los miembros del conjunto de predicciones está obtenido utilizando un modelo de circulación diferente

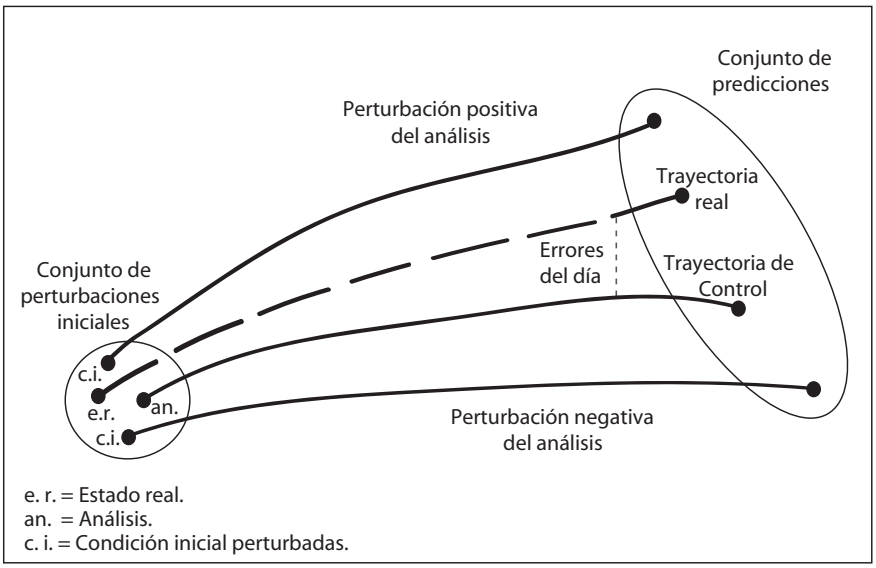

Figura 2.8: Evolución de un conjunto de dos predicciones obtenidas a partir de condiciones iniciales perturbadas.

(alternativa a considerar pequeñas perturbaciones de un mismo modelo).

Las distintas técnicas de predicción por conjuntos son complementarias, por lo que pueden ser combinadas entre sí dando lugar a toda una batería de posibilidades aún por explorar.

2.2.3. Tipos de Predicciones

La estabilidad numérica de los métodos de integración obliga a mantener un compromiso entre la resolución espacial y temporal. Esto es debido a que los procesos locales (nubes, precipitación, transferencia de calor en régimen turbulento, humedad, radiación, etc.) ocurren en escalas de tiempo más pequeñas (es decir, su ciclo de vida es más corto) que los globales. Por tanto, resolución espacial y temporal están íntimamente ligadas, de manera que para duplicar la resolución en el espacio tridimensional, también hay que reducir a la mitad el paso de tiempo para lograr un esquema numérico estable; así, el coste computacional total de duplicar la resolución crece con un factor de $2^4 = 16$. Las integraciones con gran resolución espacial requieren un paso de integración pequeño limitando el alcance operativo de las mismas a unos pocos días, mientras que las integraciones de modelos de baja resolución pueden prolongarse en el tiempo hasta escalas estacionales o climáticas. Así, cada modelo debe caracterizar los fenómenos propios de su escala de operación, desde fenómenos sinópticos de miles de kilómetros y períodos de días, hasta fenómenos

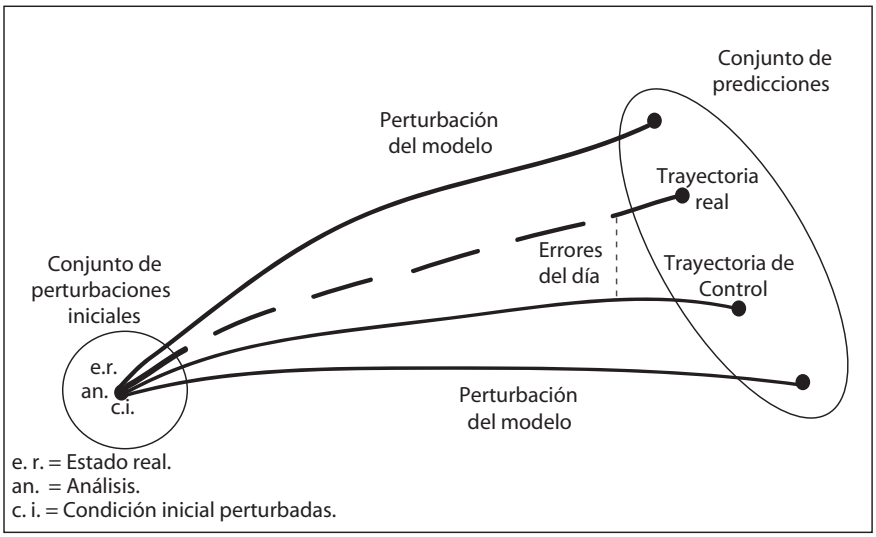

Figura 2.9: Evolución de un conjunto de predicciones obtenidas con distintos modelos aplicados a la misma condición inicial.

mesoescalares con tamaños de kilómetros y períodos de horas (ver Fig. 2.3). Este hecho ha motivado que operativamente se consideren distintas configuraciones o tipos de modelos numéricos según el alcance y resolución de la predicción deseada.

Vigilancia y Predicción Inmediata: Nowcasting

El *nowcasting* es la predicción de plazo más corto (minutos/horas) en el cual las labores de vigilancia sobre las estaciones automáticas, el radar, las imágenes de satélite, etc., llevadas a cabo por observadores y predictores humanos desempeñan el papel principal, relegando a los modelos numéricos a un segundo plano. El objetivo principal de este tipo de predicción es actualizar las predicciones con la información que se recibe continuamente, e incluso estimar la intensidad de eventos extremos (fuertes tormentas, etc.), así como la localización geográfica de los mismos. Aunque este tipo de predicciones es relativamente reciente, ya existen algunos modelos sencillos, que se basan en datos de teledetección, para predicción de fenómenos severos, ver `http://www.eumetsat.de/en/area4/saf/internet/index.html`.

Cuando a principios de los años noventa EUMETSAT, la Organización que gestiona los satélites meteorológicos europeos, estaba preparando el programa de satélites Meteosat de Segunda Generación (MSG), se acordó crear centros de aplicaciones de los datos producidos en los satélites para su utilización en diferentes campos de Meteorología y climatología. Estos centros serían adjudicados a diversos

Servicios Meteorológicos Nacionales de Europa con el nombre de SAF (Satellite Application Facilities). La AEMET dispone actualmente de un SAF para *nowcasting*, que ayuda a los meteorólogos a predecir los fenómenos atmosféricos inusuales y de aparición repentina con más precisión y mayor anticipación que antes. El denominado Centro de Aplicaciones de Satélite para el apoyo a la Predicción inmediata y a Muy Corto Plazo, ha sido concebido gracias al último satélite meteorológico europeo, el Meteosat-8. Este satélite es el primero de una serie de satélites meteorológicos muy avanzados dentro del Programa Meteosat de Segunda Generación (MSG). El satélite comenzó sus operaciones en enero de 2004 y está operado por EUMETSAT, la Organización Europea para la Explotación de los Satélites Meteorológicos, con sede en Darmstadt, Alemania.

Entre los usuarios de la predicción inmediata y a muy corto plazo se incluyen la agricultura, la construcción, la energía, el transporte y diversos sectores públicos. La predicción, por ejemplo, de la velocidad y dirección del viento ayuda a los agricultores a realizar de manera efectiva la siembra desde aviones. Los trabajadores de la construcción usan la información para evitar daños a elementos de obras sin terminar como el cemento. Los usuarios del sector energético pueden optimizar la distribución de electricidad y gas natural. En el transporte la información sobre factores como la formación de hielo, el techo de nubes y la visibilidad pueden usarse para minimizar riesgos e incrementar la eficiencia de decisiones previas al vuelo o en ruta.

El software y los productos del SAF permitirán a los meteorólogos, independientemente del lugar de Europa donde trabajen, disponer de los siguientes datos y productos:

- Máscara nubosa y cuantificación de nubes (incluyendo niebla).
- Temperatura y altitud del tope de las nubes.
- Tasa de lluvia.
- Humedad atmosférica.
- Contenido de agua en las nubes.
- Estabilidad termodinámica.
- Dirección y velocidad del viento.

En la figura 2.10 se muestra, a modo de ejemplo, el producto *Cloud Type* recientemente introducido en el SAF de predicción de la AEMET.

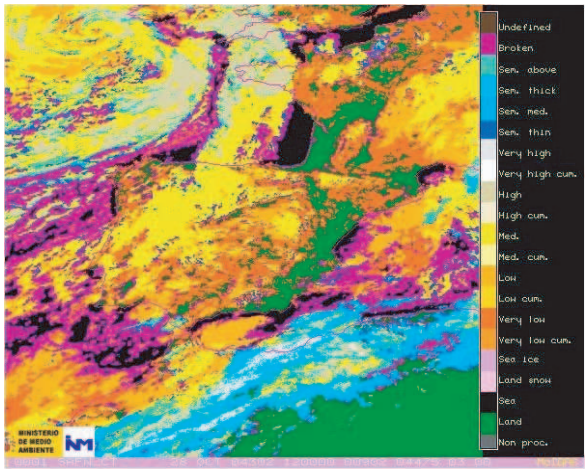

Figura 2.10: Producto de tipo de nubes (Cloud Type) generado en el AEMET dentro del SAFNWC. Las nubes observadas por el satélite con sus diferentes canales son clasificadas según la escala y realce de la derecha. *FUENTE: AEMET*.

Predicción a Corto Plazo

Es la más popular y antigua de todas las predicciones y se refiere al rango entre 1 y 3 días de alcance. Aunque la resolución de los modelos globales que se utilizan en este tipo de predicción aumenta de año en año, actualmente la resolución horizontal media de los modelos operativos está en 0.5º (aproximadamente 50 kilómetros). Por ejemplo, el ECMWF proporciona salidas cada 6 horas sobre todo el globo con 0.6º a partir de un modelo de circulación general T511. Las salidas de los modelos del ECMWF se utilizan en distintos servicios meteorológicos europeos como condiciones de contorno para modelos regionales de mayor resolución. Por ejemplo, la actual predicción operativa a corto plazo, generada por el AEMET para España, se obtiene aplicando el modelo regional HIRLAM, con una resolución espacial de 0.2º, a las salidas del ECMWF. De forma análoga, en EEUU el NCEP (National Center for Environmental Prediction, `www.ncep.noaa.gov`) realiza una predicción global que distribuye libremente a través de Internet, y que es luego utilizada por distintos centros regionales en diversas aplicaciones. Por ejemplo la Fig. 2.11 muestra el campo de precipitación estimado por el modelo del NCEP y los campos estimados por el modelo regional MM5 utilizado en *Meteo Galicia* `meteo.usc.es` con 30 y 10 km de resolución.

En España, son numerosos los grupos de investigación que llevan a cabo integraciones a corto plazo sobre regiones concretas de nuestra geografía. Por ejemplo la red Ibérica MM5 `redibericamm5.uib.es` aglutina a los grupos de investigación

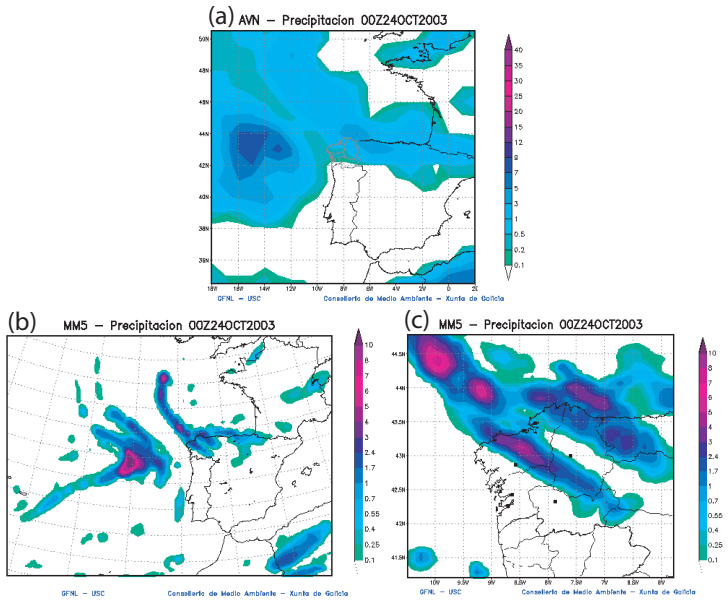

Figura 2.11: Campos de precipitación previstos para D+1 por (a) el modelo AVN del NCEP; y el modelo MM5 utilizado por Meteo Galicia con (b) 30 y (c) 10 km de resulución. (*FUENTE: Página web de Meteo Galicia*).

que utilizan en modelo MM5 para distintos estudios de modelización y predicción.

Predicción a Medio Plazo

Es la realizada entre los 4 y 15 días de alcance y, debido a la no linealidad de la atmósfera, la incertidumbre en la predicción determinista es tan alta que resulta más adecuado formular la misma en términos probabilísticos, utilizando las citadas técnicas de predicción por conjuntos (Ensemble Prediction Systems, EPS) que generan un conjunto de campos previstos para un mismo alcance.

En cuanto a los sistemas operativos globales de predicción por conjuntos, sólo los grandes centros de predicción numérica tienen la capacidad de cálculo necesaria para llevar a cabo las costosas integraciones. Por una parte, desde Noviembre de 2000, el ECMWF ha puesto en marcha un modelo de predicción por conjuntos a plazo medio (con alcance de 10 días) basado en 50 integraciones con condiciones iniciales perturbadas y una integración con condiciones sin perturbar de un modelo $T255L40$ (aprox. 80 km de resolución en latitudes medias). En este caso, las perturbaciones se obtienen con el método de vectores singulares (Buizza and Palmer, 1995). Por ejemplo, la Fig. 2.12 muestra el agrupamiento de los miembros previstos para el 01/07/2007. En este ejemplo se pueden apreciar sutiles diferencias entre unos

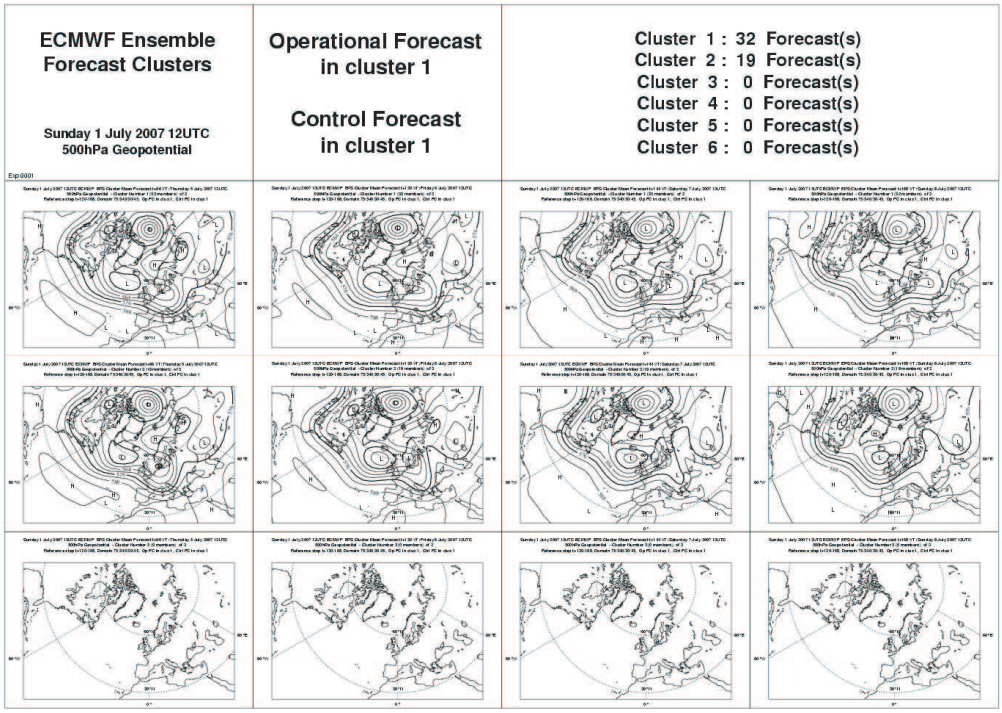

Figura 2.12: Agrupamiento de los miembros de la predicción por conjuntos del ECMWF. Se muestra el geopotencial en 500HPa. (*FUENTE: ECMWF*).

miembros y otros. Por otra parte, la Fig. 2.13 muestra la predicción probabilística obtenida a partir del EPS, que resume el efecto de la incertidumbre sobre los distintos patrones de circulación atmosférica previstos; en (Eckert et al., 1996) se estudian algunos aspectos relacionados con la clasificación de estos patrones.

Por otra parte, el NCEP también produce de forma operativa predicciones por conjuntos a medio plazo (hasta 15 días) utilizando el Global Forecast System model (GFS) (ver `www.nco.ncep.noaa.gov/pmb/products/`).

Predicción Estacional

Continuando con los distintos sistemas de predicción en orden creciente de alcance, la siguiente es la predicción estacional. Este tipo de predicción hace referencia a alcances superiores a un mes e inferiores a un año. A diferencia de las predicciones de plazos más cortos, no se trata de predecir valores instantáneos, sino valores estadísticos agregados por meses o trimestres del estado de la atmósfera. La principal fuente de incertidumbre para este tipo de predicciones procede de las condiciones iniciales y la evolución del océano, quedando la incertidumbre procedente de las condiciones iniciales en la componente atmosférica relegada a un segundo plano.

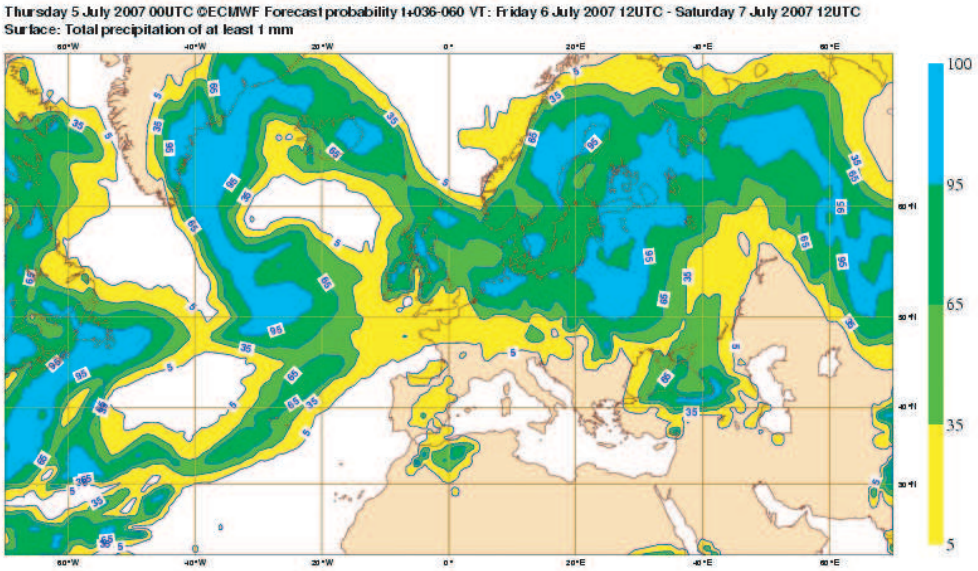

Figura 2.13: Predicción probabilística del evento *prec.* $> 1mm/24h$ obtenida con los 50 miembros del EPS (*FUENTE: ECMWF*).

Por tanto, los modelos numéricos apropiados para este tipo de predicción son modelos numéricos acoplados atmósfera-océano; en algunos casos, también se consideran modelos atmosféricos que tienen forzadas las temperaturas del agua del mar. Un ejemplo operativo de este tipo de predicciones es el *System2* del ECMWF, que integra mensualmente un modelo de predicción por conjuntos, con un horizonte de predicción hasta 180 días (desde hace unos meses ya está operativo el *System3*, nueva versión operativa de este sistema). Este tipo de modelos de predicción estacional son fruto de la experiencia llevada a cabo en distintos proyectos, como PROVOST y DEMETER; en este último se empleó, por primera vez, la predicción multi-modelo, utilizando siete modelos globales de predicción por conjuntos con distintos esquemas de perturbación (Palmer et al., 2004).

El paradigma de la predicción estacional es el fenómeno de *El Niño*; en concreto, la zona Niño-3 comprende un área del Océano Pacífico entre la latitudes 5N - 5S y longitudes 90W - 150W. El promedio de la temperatura del agua del mar en esta región se suele utilizar como indicador de la actividad de El Niño y las anomalías positivas y negativas de la temperatura en esta zona (ver Fig. 2.14) son indicadores de una actividad anómala del océano que implica una serie de efectos derivados en las condiciones meteorológicas de distintas zonas del globo (ver Philander, 1990, para más detalles).

La figura 2.15 ilustra las ventajas de una predicción por conjuntos frente a la

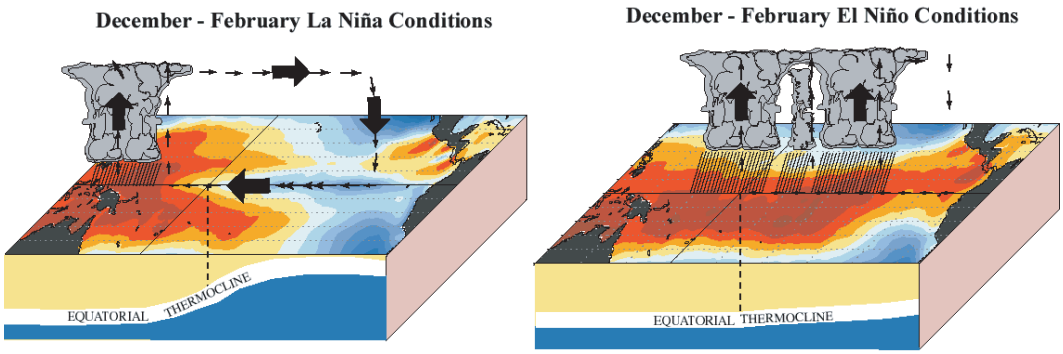

Figura 2.14: Esquema del fenómeno de El Niño en sus dos fases extremas (FUENTE: NOAA).

predicción determinista tradicional para este problema. La gran incertidumbre de las predicciones a escala estacional requiere disponer de alguna estimación de la confianza que se tiene en las predicciones realizadas, ya que en muchos casos las predicciones no serán fiables. Por otra parte, en la Fig. 2.16, se muestra el formato de la predicción estacional en la que se predicen las anomalías y su magnitud; también se suelen emitir mapas con las probabilidades de exceder los terciles superior e inferior.

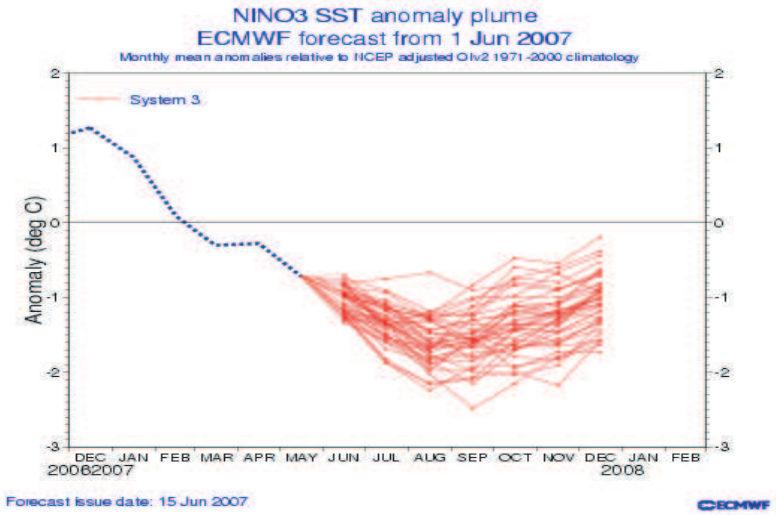

Figura 2.15: Predicción por conjuntos del fenómeno de El Niño con un conjunto de 51 miembros.(FUENTE: ECWMF).

Predicción Climática

Del mismo modo que los modelos descritos anteriormente, los modelos climáticos se basan en leyes físicas, representadas por ecuaciones matemáticas, que se resuelven

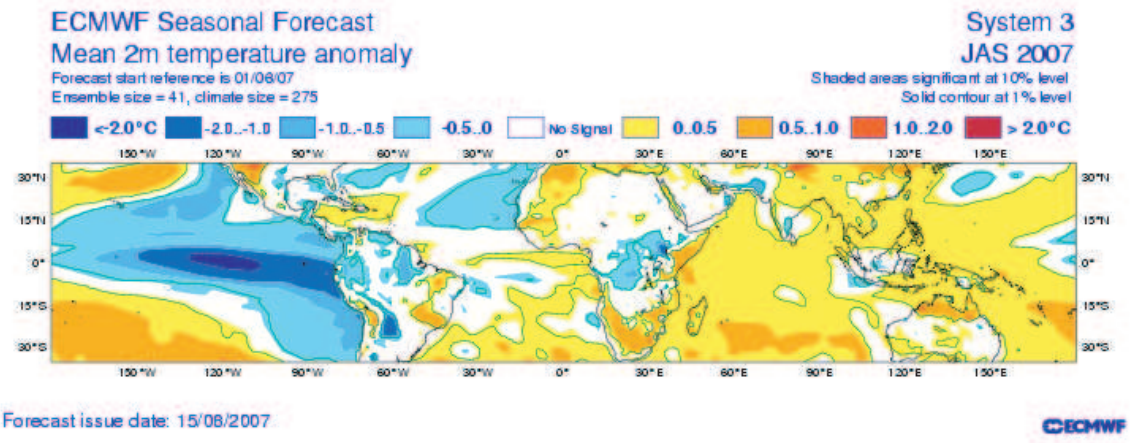

Figura 2.16: Ejemplo de predicción de la anomalía de temperatura media a 2m.(FUENTE: ECWMF).

utilizando una rejilla tridimensional sobre el globo terráqueo. El clima de un sistema es una síntesis estadística de los principales procesos que ocurren en su seno, sin embargo, a causa de su caracter acoplado, para simular el clima, es necesario tener en cuenta todos y cada uno de los componentes del sistema climático (la atmósfera, el océano, la superficie terrestre, la criosfera y la biosfera), junto con los procesos y retro-alimentaciones que ocurren entre ellos [1]. En los actuales modelos climáticos, cada componente es introducido de forma aislada con sus propias ecuaciones; en el módulo atmosférico, por ejemplo, se resuelven las ecuaciones que describen la evolución a gran escala del impulso, el calor y la humedad, descritas en la Sec. 2.1.1, siendo similares las ecuaciones que se aplican con respecto al océano. Para considerar el comportamiento conjunto del sistema, todos los componentes están acoplados entre sí mediante módulos de acoplamiento que introducen los intercambios de materia y energía entre los diferentes subsistemas. Esta interactividad obliga a que, a pesar de ser modelos de alcance decadal o secular, el paso de tiempo no pueda ser arbitrariamente largo, de manera que las ecuaciones han de ser resueltas aproximadamente cada media hora, durante décadas, siglos o milenios. Por su gran complejidad, estos modelos tienen fuertemente limitada su resolución, por ejemplo, para el módulo atmosférico, es del orden de 250 km en horizontal y de 1 km en vertical, y del orden de 150 km en horizontal y 300 m en vertical para el oceánico.

Esta complejidad influye además, en la manera en la que se formulan las predi-

[1] Al contrario que en otros tipos de modelos, los climáticos no pueden ignorar ningún proceso, por pequeño que este sea, ya que realizan cientos de miles de pasos de integración en los cuales se van acumulando errores que acaban dejando al sistema fuera de control.

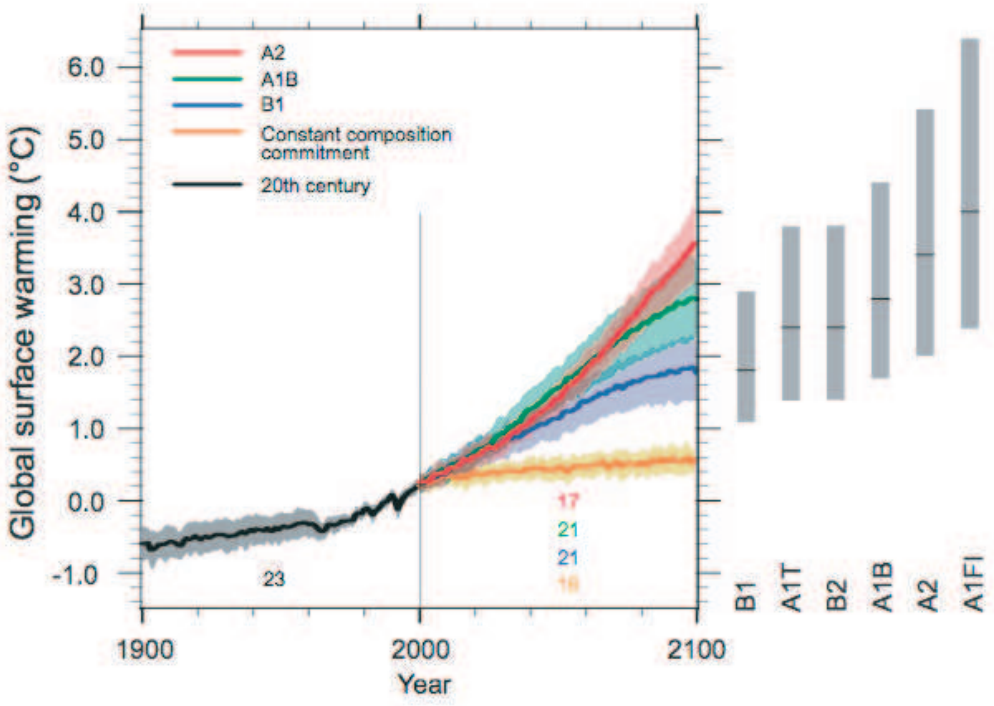

Figura 2.17: Proyecciones de temperatura en distintos escenarios de forzamiento futuro (*FUENTE: IPCC*).

ciones de estos modelos. Hemos visto que los modelos de corto plazo son los más sencillos, donde basta con explicar la componente atmosférica para obtener predicciones fiables que asocian estados de la atmósfera con instantes futuros, de forma determinista o probabilística a escala diaria e incluso horaria. En los modelos de medio plazo y estacionales esta asociación estado de la atmósfera con instante de tiempo se va diluyendo paulatinamente de forma que los modelos estacionales sólo formulan predicciones probabilísticas, en términos de anomalías sobre periodos de tiempo de escala decenal y mensual. Los modelos climáticos ocupan el escalón más alto en cuanto a complejidad y alcance de sus predicciones, y al mismo tiempo, el más bajo en cuanto a la formulación de las mismas, siendo únicamente capaces de ofrecer una caracterización estadística para periodos decadales y seculares, perdiéndose definitivamente la asociación estado del sistema con cierto periodo de tiempo futuro.

Este tipo de modelos, si bien inicialmente fueron concebidos para caracterizar y simular el clima, actualmente son utilizados para simular escenarios de clima futuro asociados con forzamientos o perturbaciones en el comportamiento del sistema

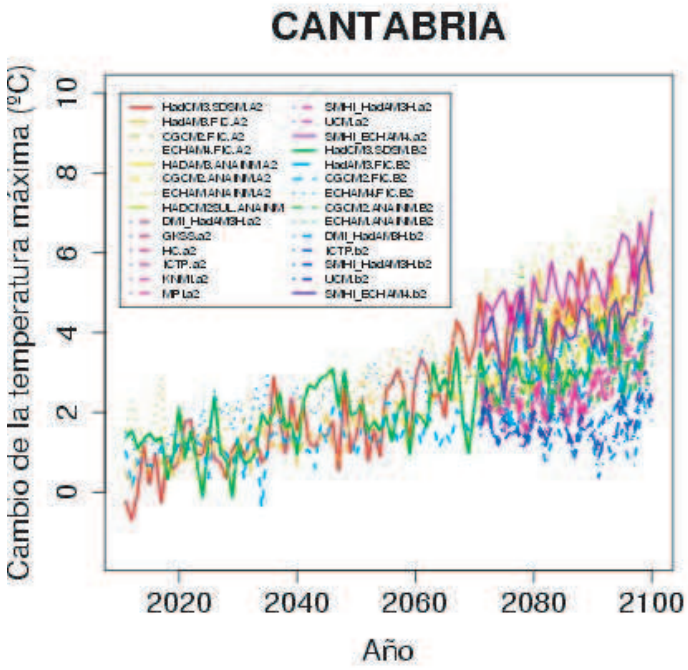

Figura 2.18: Proyecciones de temperatura para Cantabria obtenidas con distintos modelos (*FUENTE: Primer Informe de Escenarios Reginoalizados en España, AEMET*).

climático. Concretamente, se estudian los efectos derivados del forzamiento radiativo del sistema, provocado por el cambio en la capacidad calorífica de la atmósfera, inducido a su vez, por los cambios en su composición (incremento antropogénico de CO_2, etc.). Por ejemplo, en la Fig. 2.17, se muestra una típica salida de un modelo climático, donde se puede ver el periodo de calibración: 1900-2000, y el de proyección: 2001-2100. Estas predicciones llevan asociada una componente de incertidumbre que, a estas escalas, puede ser tanto o más importante que la propia proyección. Las más importantes proceden de inexactitudes tanto en los modelos climáticos, como en la magnitud del forzamiento, y son estimadas realizando simulaciones perturbadas. Por ejemplo, en la Fig. 2.17, se muestra el calentamiento global estimado bajo diferentes supuestos de forzamiento (A1, B1, A2, B2, etc.). Por otra parte, la incertidumbre procedente de los modelos climáticos se muestrea utilizando diferentes modelos[2]; por ejemplo, en la Fig. 2.18, se pueden ver los resultados obtenidos utilizando diferentes modelos climáticos, sobre la región de Cantabria. Como es lógico suponer, los efectos del calentamiento global y la vulnerabilidad a los mismos, tienen una variabilidad espacial muy superior a la que son capaces de resolver los modelos climáticos, por

[2]En los modelos climáticos, la incertidumbre asociada a las condiciones iniciales es irrelevante.

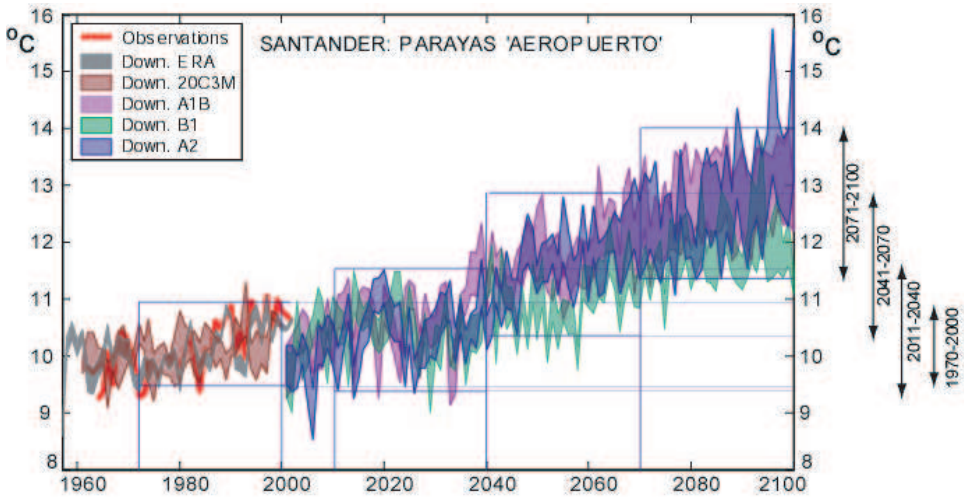

Figura 2.19: (*FUENTE: Informe preliminar del Proyecto: Escenarios Regionales de Cambio Climático en Cantabria, Universidad de Cantabria*).

lo que es necesario emplear métodos de regionalización para obtener estimaciones más precisas de su distribución espacial. Esto introduce una nueva fuente de incertidumbre, procedente de la técnica de regionalización utilizada. En la Fig. 2.19, se pueden ver los resultados obtenidos por diferentes métodos de regionalización para el Aeropuerto de Santander (Parayas).

2.3. Series de Observaciones y Simulaciones

Los registros históricos de observaciones son esenciales tanto en el pre- como en el post-proceso de la predicción numérica del tiempo. Las observaciones rutinarias de la atmósfera (en especial en superficie) han ido extendiéndose de forma continua desde la creación de la OMM en 1950, propiciando el marco intergubernamental para el establecimiento de redes internacionales de observación.

Para que las series de observaciones meteorológicas sean representativas de las características locales del clima, deben tener unos periodos mínimos que, dependiendo del elemento climático observado y de su entorno geográfico, oscilan entre 25 y 50 años para la precipitación y entre 10 y 25 años para la temperatura, aunque por conveniencia se suele fijar un período de referencia climatológico en 30 años (para más detalles ver Jagannathan and Arlqy, 1967). Además, este aspecto relativo a la variabilidad temporal debe ser combinado con una variabilidad espacial adecuada.

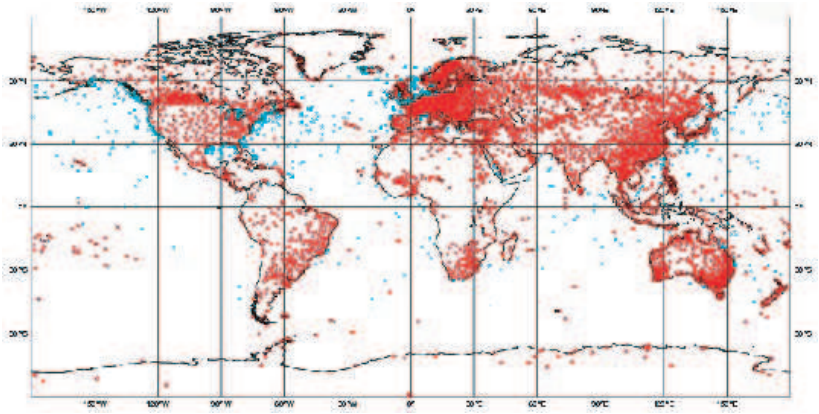

Figura 2.20: Observaciones utilizadas en la asimilación de datos del modelo numérico del ECMWF. (*FUENTE. Página Web del ECMWF*).

Por ejemplos, las nuevas redes de observación que se instalan en distintas regiones suelen presentar una excelente resolución espacial, pero son demasiado cortas para ser útiles para los modelos estadísticos pues no abarcan un período climatológico adecuado. Estas condiciones son todavía más estrictas cuando se trata de predecir eventos extremos.

2.3.1. Redes Globales de Observación

El GCOS (Global Climate Observing System) es el sistema coordinado que recopila y comprueba las observaciones de variables atmosféricas y oceánicas (en superficie) a nivel global. El núcleo de las observaciones superficiales consta de aproximadamente 10000 observatorios que realizan observaciones al menos cada seis horas, y a menudo horariamente. Las variables observadas son: presión atmosférica, viento, temperatura del aire, precipitación y humedad relativa. Además existen unas 1000 estaciones de radiosonda y más de 3000 aviones que realizan observaciones en varios niveles de la atmósfera. Unos 7300 barcos, 600 boyas a la deriva, 300 boyas fijas y 600 plataformas proporcionan observaciones en los océanos. Estos datos son utilizados diariamente por los modelos numéricos de circulación atmosférica, asimilando el estado de la atmósfera inicial en base a un conjunto de estas observaciones. Por ejemplo, la Figura 2.20 muestra las observaciones utilizadas un día particular en el ECMWF para el proceso de asimilación (inicialización) de su modelo operativo.

La longitud de las series de datos almacenadas es muy variable y oscila entre un par de cientos de años (para estaciones históricas), hasta unos pocos años, o meses. La información de un subconjunto de estas estaciones de superficie (GSN, GCOS

Surface Network) puede consultarse directamente en
(`www.gosic.org/ios/GCOS-main-page.htm`).

El GCOS también dispone de un subsistema espacial de observación, integrado por cuatro satélites en órbita polar y cinco geoestacionarios, con el objetivo principal de completar la cobertura de las observaciones de la red en superficie. El papel de las observaciones procedentes de satélites es cada vez más importante y la asimilación de estos datos en los modelos numéricos es un área de intensa investigación.

2.3.2. La Red de Observación de la AEMET

En este apartado se describe la red de observación de la Agencia Estatal de Meteorología (AEMET), cuyos datos se utilizarán en diferentes aplicaciones de esta Tesis, considerando subconjuntos específicos de la misma dependiendo de la resolución deseada.

Los distintos servicios meteorológicos nacionales disponen de su propia red de observación (algunas de cuyas estaciones se integran en el GCOS). Estas redes, mucho más densas, abarcan gran variedad de variables meteorológicas. Por ejemplo, AEMET dispone de una red de observatorios que abarca más de 6000 puntos geográficos en las distintas cuencas hidrográficas Españolas, como se muestra en la Fig. 2.21(d). En estos observatorios, se realizan mediciones diarias de precipitación y meteoros (tormenta, nieve, granizo, niebla, lluvia, calima, rocío y escarcha) que básicamente abarcan el período 1900-2000. Aunque las series tienen longitudes variables y contienen lagunas, en promedio se dispone de más de 30 años de información. Las temperaturas máxima y mínima se observan en un subconjunto de 2281 puntos. Además de la red secundaria, existen algunas estaciones en las que se realizan observaciones por parte de personal cualificado. Estas estaciones corresponden a la red principal y en la actualidad consta de 225 puntos de observación ver panel (c) de la Fig. 2.21(c), en los que se mide insolación, evaporación, recorrido, dirección y racha máxima de viento, temperaturas medias, y otras variables. Las observaciones se realizan diariamente, aunque en la red principal se realizan observaciones con mayor frecuencia (cada 10 min, cuatro veces al día, etc.). En total se tiene un gran volumen de información que es necesario tratar de forma eficiente.

A continuación se detallan las observaciones disponibles; salvo indicación expresa, las observaciones son diarias y hacen referencia a un periodo de 24 horas comprendidas de 07 UTC a 07 UTC del día siguiente:

- Observaciones en la red secundaria (Fig. 2.21(d)):

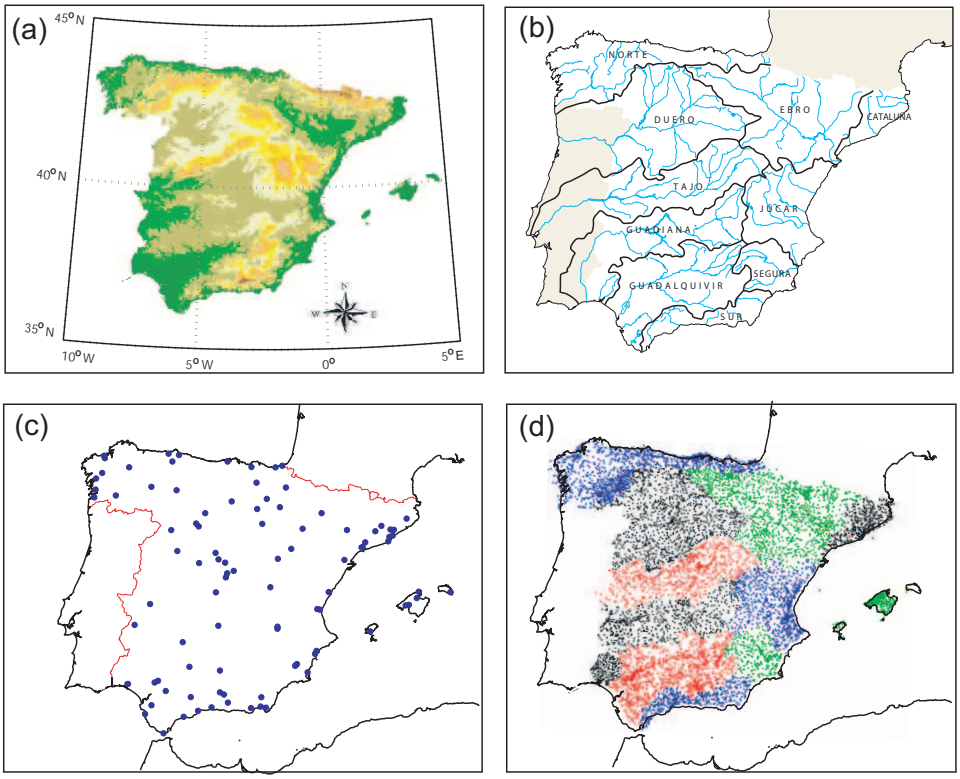

Figura 2.21: (a) Orografía de España peninsular y Baleares y sus (b) cuencas hidrográficas principales. (c) Red principal de estaciones de la AEMET, (d) red secundaria (pluviométrica y termopluviométrica).

1. *Temperaturas extremas* máxima y mínima a $2\,m$ del suelo (en ^oC), en un subconjunto de 2281 puntos.
2. *Precipitación* diaria acumulada (en $mm \equiv l\,m^{-2}$).
3. Ocurrencia de *meteoros*: niebla, tormenta, granizo, nieve, escarcha y rocío.

- Observaciones en la red principal (Fig. 2.21(c)). Además de incluir las observaciones anteriores, posee una serie adicional de mediciones:

1. *Racha máxima de viento*, medida de 0 a 24 horas a $10m$ del suelo (en $km\,h^{-1}$). La dirección de la racha máxima se mide en grados (tomando como origen el Norte).
2. *Insolación* medida de 0 a 24 horas (en horas diarias de sol). Si es relativa se mide en $\%$ respecto a la máxima teórica del día.

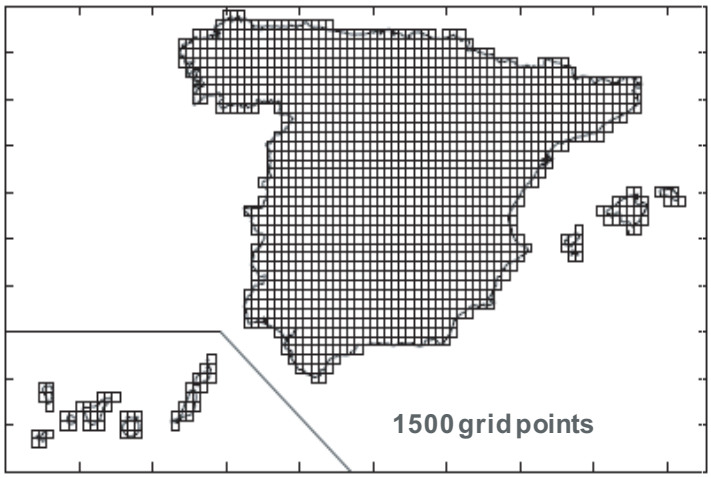

Figura 2.22: Ejemplo de rejilla realizada a 0.2º de resolución para precipitación y temperaturas extremas.

3. *Evaporación potencial*, considerada como la máxima evaporación posible, medida de 0 a 24 horas (en *mm*) (existen algunas medidas en la red secundaria).

4. En la red principal existen otras muchas variables que se podrían incluir entre las observaciones. Algunas de ellas son: Intensidad máxima y duración de precipitación, recorrido del viento, cantidad y tipo de nubosidad, humedad media, y visibilidad.

También se han elaborado bases de datos de observaciones sin lagunas y homogeneizadas sobre rejillas de observaciones interpoladas. Por ejemplo, (Chen et al., 2002) describe la elaboración de una base de datos con medias mensuales de precipitación en superficie para una rejilla de 2.5º de resolución sobre el globo. Por otra parte, Herrera et al. (2007) han elaborado recientemente una rejilla sobre España con resolución 0.2º para precipitación y temperaturas extremas (ver Fig. 2.22). Aunque las rejillas de observaciones interpoladas eliminan algunos de los problemas que surgen al tratar con datos de observaciones, introducen un elemento artificial en el proceso que, en la medida de los posible, es aconsejable evitar. Las técnicas de minería de datos que se utilizan en esta Tesis están preparadas para trabajar con lagunas, etc., por lo que en esta Tesis se abordarán los problemas considerando observaciones locales, evitando utilizar rejillas interpoladas de observaciones.

2.4. Simulaciones de Modelos Numéricos: Reanálisis

Del mismo modo que existen series de observaciones de la fenomenología meteorológica, interesa disponer de información paralela referente al estado de la atmósfera simulado por los modelos numéricos del Tiempo y el Clima (recordemos que estos modelos son los que proporcionan capacidad predictiva futura). Para este cometido, los proyectos de *reanálisis* (análisis retrospectivos) han permitido reconstruir el estado de la atmósfera en tiempos pasados asimilando toda la información observacional disponible a través de un modelo numérico. Los campos resultantes de la asimilación, y en algunos casos predicción, caracterizan el estado de la atmósfera sobre la rejilla del modelo considerado.

Hasta la fecha se han realizado varios proyectos de reanálisis que, salvo excepciones, están disponibles para la comunidad investigadora. Por ejemplo, el primer proyecto de reanálisis global llevado a cabo en el ECMWF se denominó ERA-15 y proporciona los campos de análisis y las predicciones a corto plazo cada seis horas obtenidas con un modelo $T106L31$ (1.75 grados de resolución) para el período comprendido entre Diciembre 1978 y Febrero 1994. Recientemente, este proyecto ha sido extendido y ya se dispone de información para el período Septiembre 1957 - Agosto 2002 (ERA-40) obtenida con un modelo de mayor resolución ($T159L60$, aprox. 1.125 grados). Esta base de datos dispone de variables en superficie y en altura, además de otras variables derivadas (que no son integradas directamente en el modelo). Estos datos, suponen un enorme volumen de información. El NCEP también dispone de un reanálisis llevado a cabo con un modelo $T62L28$ (2.5 grados de resolución) para el período 1958 hasta la actualidad (Kalnay et al., 1996); para más información sobre actualizaciones del reanálisis consultar `wesley.wwb.noaa.gov/reanalysis2`.

La lista completa de variables disponibles para el proyecto ERA puede consultarse en `www.ecmwf.org/research/era`, mientras que en el caso del NCEP puede consultarse `www.cpc.ncep.noaa.gov/products/wesley/`. Aunque no se utilizarán en esta Tesis, existen además reconstrucciones históricas de otras variables como SST (Sea Surface Temperature) (Smith et al., 1996), la cubierta de hielo marino (Rayner et al., 1996), etc.

CAPÍTULO 3

Minería de Datos. Redes Bayesianas

En las últimas décadas se han desarrollado numerosas técnicas de análisis y modelización de datos en distintas áreas de la estadística y la inteligencia artificial (Duda et al., 2000; Bishop, 2006). La Minería de Datos (MD) es un área moderna interdisciplinar que engloba a aquellas técnicas que operan de forma automática (requieren de la mínima intervención humana) y, además, son eficientes para trabajar con las grandes cantidades de información disponibles en las bases de datos de numerosos problemas prácticos. Estas técnicas permiten extraer conocimiento útil (asociaciones entre variables, reglas, patrones, etc.) a partir de la información cruda almacenada, permitiendo así un mejor análisis y comprensión del problema. En algunos casos, este conocimiento puede ser también post-procesado de forma automática permitiendo obtener conclusiones, e incluso tomar decisiones de forma casi automática, en situaciones prácticas concretas (*sistemas inteligentes*). La aplicación práctica de estas disciplinas se extiende a numerosos ámbitos comerciales y de investigación en problemas de predicción, clasificación o diagnosis (ver Witten and Frank, 2005; Cios et al., 2007, para una introducción teórica y práctica a esta disciplina).

La creciente disponibilidad de información meteorológica y climática (registros históricos, simulaciones, etc.; ver Sec. 2.3) hacen de las Ciencias Atmosféricas un área idónea para la minería de datos. En concreto, la minería de datos permite tratar de forma eficiente la gran redundancia espacial y temporal existente en los datos disponibles, así como obtener las distintas asociaciones y dependencias existentes entre las distintas variables observadas y simuladas en un problema práctico (ver Gutiérrez et al., 2004b, para más detalles). Sin embargo, la aplicación de la minería

de datos en este ámbito no ha sido tan profusa como en otras disciplinas, como la biología (Wong and Li, 2006), economía (Kovalerchuk and Vityaev, 2003), etc.

Entre las diferentes técnicas disponibles en minería de datos, en esta Tesis se utilizan las redes probabilísticas, que permiten modelizar de forma conjunta toda la información relevante para un problema dado, utilizando posteriormente mecanismos de inferencia probabilística para obtener conclusiones en base a la evidencia disponible (Pearl, 1988; Castillo et al., 1997). Como veremos en los capítulos posteriores, estos modelos se adaptan de forma natural a la concepción probabilística y local actual de la modelización en Meteorología y climatología. En particular, en esta tesis se describe, por primera vez, la aplicación de estas técnicas a la predicción probabilística local.

La organización del capítulo es la siguiente. En Sec. 3.1, se describe el esquema general, los elementos y el funcionamiento de un sistema experto. En la Sec. 3.3, se introducen las redes probabilísticas, destacando aspectos como la codificación de dependencias y los criterios de separación, para expresar la distribución de probabilidad conjunta de un sistema de manera tratable. En la Sec. 3.4, se motivan y definen las RBs, describiendo el proceso de inferencia probabilística, cuya utilidad consiste en hallar de forma eficiente, la probabilidad de cualquier nodo (o conjunto de nodos), dada una cierta información (evidencia). En la Sec. 3.5, se trata el importante problema del aprendizaje automático, que es donde se diseña la topología de dependencia de las diferentes variables que componen el modelo probabilístico (en este caso la RB), y se describen los dos tipos principales de algoritmos de aprendizaje a partir de datos como son: los basados en test de independencia y los basados en búsqueda. En la Sec. 3.6 se comentan algunas de las muchas estrategias de aprendizaje específicas para clasificadores por ser el tipo de RB más estudiado por la comunidad científica. Finalmente, la Sec. 3.7 describe un importante aspecto con incidencia en la práctica como es el sobreajuste.

3.1. Sistemas Inteligentes

En la Fig. 3.1, se muestra el esquema general de un Sistema Inteligente (SI) con sus dos elementos principales: la *base de conocimiento* y el *motor de inferencia o módulo de razonamiento*. La base de conocimiento almacena de forma estructurada (utilizando algún patrón de conocimiento adecuado, como reglas, grafos, etc.) la información relevante para un problema dado. En los antiguos SIs, este conocimiento era obtenido de un experto humano, mientras que en los modernos se extrae directamente de los datos utilizando algoritmos de aprendizaje automático y minería de

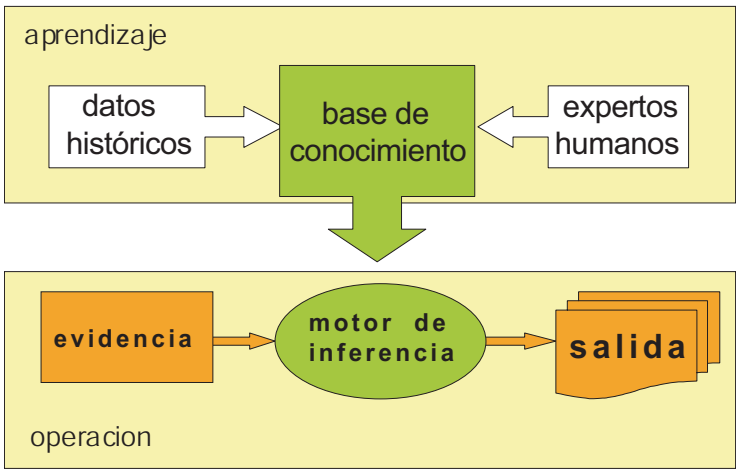

Figura 3.1: Esquema de un sistema inteligente, con sus elementos principales.

datos.

Una vez adquirido el conocimiento, el módulo de razonamiento ha de ser capaz de generar conclusiones acerca del estado del sistema a partir de una evidencia concreta. Por ejemplo, en el problema de la predicción meteorológica local, si se dispone de evidencia sobre la precipitación en un subconjunto de estaciones (está lloviendo en Santander y Bilbao), entonces el módulo de razonamiento tendrá que ser capaz de formular conclusiones sobre el estado de la precipitación en cualquier otra estación del sistema. En una aplicación práctica concreta, la *evidencia* es el subconjunto de variables de estado del sistema cuyo valor es conocido, y el mecanismo de *razonamiento*, *inferencia* o *propagación de evidencia* es el procedimiento que estima el valor del resto de variables utilizando la base de conocimiento y la evidencia disponibles.

El módulo de razonamiento está diseñado en función del tipo de patrón de conocimiento empleado en el sistema y, por tanto, la forma de obtener conclusiones será distinta en cada caso. Por ejemplo, un sistema basado en reglas tiene un módulo de razonamiento que consiste en ejecutar aquellas reglas cuya premisa se conozca (es decir, forme parte de la evidencia), proponiendo aquellas conclusiones obtenidas de aplicar las reglas ejecutadas; las conclusiones, a su vez pasarán a formar parte de la evidencia para tratar de seguir ejecutando reglas. Otros SIs más modernos utilizan patrones de conocimiento más abstractos, con lo cual el proceso de razonamiento es más complicado y menos intuitivo. Dos ejemplos notables de estas nuevas metodologías son los algoritmos inspirados en la Biología, como las em redes neuronales, y los algoritmos de aprendizaje estadístico, como las *redes probabilísticas*:

- Las *Redes Neuronales* se inspiran en la Biología para definir un patrón de conocimiento que simule la estructura de un cerebro, considerando un gran número de unidades de proceso simples (o neuronas) conectadas entre sí. En este caso, la base de conocimiento es abstracta, pues viene dada por los pesos de las conexiones que se establecen entre las neuronas. La forma de aprender estos pesos a partir de un conjunto de datos sigue diversas analogías con la Biología, como las técnicas de aprendizaje no supervisado basadas en el córtex visual. Por otra parte, en este caso el módulo de razonamiento es simple y se basa en la propagación por las distintas capas de neuronas de un conjunto de datos de entrada, hasta una capa final que proporciona la salida.

- Las *Redes Probabilísticas* automatizan el proceso de modelización probabilística utilizando la expresividad de los grafos. Los modelos resultantes combinan resultados de la teoría de grafos (para representar las relaciones de dependencia e independencia del conjunto de variables) y de la probabilidad (para cuantificar estas relaciones). Esta unión permite realizar de forma eficiente tanto el aprendizaje automático del modelo, como la inferencia a partir de la evidencia disponible. La base de conocimiento de estos sistemas es una estimación de la función de probabilidad conjunta de todas las variables del modelo, mientras que el módulo de razonamiento es donde se hace el cálculo de probabilidades condicionadas. Por ejemplo, una red probabilística permite definir un modelo conjunto para la precipitación en las estaciones de la red secundaria (miles de variables).

El estudio de estas dos técnicas proporciona una buena perspectiva global del problema del aprendizaje estadístico y la minería de datos, y permite entender mejor otras técnicas alternativas (Hastie et al., 2001; Duda et al., 2000).

En esta Tesis se consideran únicamente las redes probabilísticas, ya que se adaptan de forma natural al actual tratamiento probabilístico y discreto de muchos de los sistemas objeto de las Ciencias de la Atmósfera. Además, estas técnicas definen un único modelo global que permite propagar cualquier evidencia en el modelo (por ejemplo: está lloviendo en Santander y Bilbao), obteniendo las probabilidades condicionadas de precipitación en el resto de estaciones; este último aspecto supone una notable diferencia respecto de las redes neuronales, ya que estas últimas plantean un modelo específico para cada conjunto de variables evidenciales. La potencia de las redes probabilísticas viene de su capacidad para codificar las dependencias/independencias relevantes para un problema dado, considerando no sólo las dependencias marginales, sino también las dependencias condicionadas entre con-

juntos de variables.

3.2. Dependencia e Independencia de Variables

El concepto de probabilidad condicionada proporciona la forma de definir la dependencia estadística de dos conjuntos de variables aleatorias. Si X e Y son dos subconjuntos disjuntos de variables aleatorias, se dice que X es independiente de Y (que denotaremos $I(X,Y)$) si y solamente si

$$p(x|y) = p(x) \Leftrightarrow p(x,y) = p(x)p(y), \tag{3.1}$$

para todos los valores posibles x e y de X e Y, respectivamente; en otro caso, X se dice dependiente de Y (y se denota $D(X,Y)$). Esta ecuación indica que si X es independiente de Y, entonces la función de probabilidad conjunta de X e Y es igual al producto de sus marginales.

Los conceptos de dependencia e independencia se refieren a dos subconjuntos de variables, pero se puede generalizar el concepto cuando hay implicados más de dos conjuntos. Si X, Y y Z son tres conjuntos disjuntos de variables, entonces X se dice condicionalmente independiente de Y dado Z, $I(X,Y|Z)$, si y sólo si

$$p(x|z,y) = p(x|z) \Leftrightarrow p(x,y|z) = p(x|z)p(y|z), \tag{3.2}$$

para todos los valores posibles x, y, z. En otro caso X e Y se dicen condicionalmente dependientes dado Z, $D(X,Y|Z)$.

El concepto de independencia condicional $I(X,Y|Z)$ lleva en sí la idea de que una vez que es conocida Z, el conocimiento de Y no altera la probabilidad de X. En otras palabras, si Z ya es conocida, el conocimiento de Y no añade información alguna sobre X.

3.3. Redes Probabilísticas

Las redes probabilísticas se desarrollaron hace dos décadas (Pearl, 1988) para poder tratar probabilísticamente la incertidumbre en problemas que involucran un número elevado de variables $X_1, \ldots, X_n$. Una red probabilística es capaz de especificar una probabilidad conjunta para el problema a partir de las relaciones de dependencia relevantes existentes entre las variables del mismo, de forma que permite calcular probabilidades marginales de forma eficiente y actualizar éstas (probabilidades condicionadas) cuando se dispone de nueva información o evidencia. Considerando una red probabilística como un sistema inteligente, sus dos elementos principales serían:

- La base de conocimiento, dada en términos probabilísticos mediante la Función de Probabilidad Conjunta (FPC) $p(x_1, \ldots, x_n)$ que cuantifica las relaciones de dependencia existentes entre las variables del problema (de hecho la distribución conjunta se define factorizadamente a través de distribuciones locales, obtenidas de las dependencias existentes).

- El motor de inferencia, constituido por los algoritmos de cálculo de probabilidades condicionadas, que permite obtener conclusiones a partir de información particular (o evidencia) conocida en un instante concreto.

Salvo en casos muy concretos, hallar una factorización apropiada de la FPC que exprese exclusivamente todas y cada una de las relaciones de independencia requeridas para un cierto problema es una tarea compleja. Es decir, dado un conjunto de variables y un conjunto de relaciones de independencia definidas sobre ellas, no es trivial hallar un modelo probabilístico que incluya estas relaciones y sólo requiera especificar los parámetros libres del modelo.

Sin necesidad de que el número de variables de estado del sistema sea muy elevado, el número de parámetros necesario para definir un un modelo multinomial puede resultar astronómico. Por ejemplo, el número de parámetros necesario para definir una *FPC* de 50 variables binarias es $2^{50} > 10^{15}$, lo cual además de resultar computacionalmente prohibitivo, requiere un conjunto de entrenamiento enorme para estimar los parámetros. Por ello, en sistemas con muchas variables, un modelo global con todas las dependencias posibles que se podrían dar entre las variables, no resulta ni práctico ni fiable, siendo preceptivo simplificar su descripción, para reducir el número de parámetros, generando un modelo que además de ser eficiente, conserve las propiedades del sistema en la medida de lo posible.

Para resolver este problema, los primeros Sistemas Inteligentes, basados en probabilidad, utilizaban distintas soluciones. Cada uno de estos modelos asumía distintas hipótesis de independencia entre las variables para simplificar la FPC y, así, el número de parámetros. Por ejemplo, el modelo más simple se obtendría suponiendo que todas las variables son independientes entre sí. Por tanto, se tendría que la FPC se podría expresar como: $p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$; en este caso, el número de parámetros necesarios en un problema con 10 variables binarias se reduciría a 10, en contraposición a los $2^{10} = 1024$ necesarios en el caso general, que incluye todas las posibles dependencias entre las variables. Sin embargo, el modelo global de independencia es de muy poca utilidad, pues no permite expresar ninguna relación de dependencia entre las variables. también se pueden utilizar modelos intermedios que asumen ciertas hipótesis de independencia entre las variables, sin embargo es-

tos modelos son sólo válidos en problemas concretos que se adecúan a las hipótesis impuestas, pero no permiten construir modelos probabilísticos para un problema general.

Las redes probabilísticas fueron introducidas para resolver este problema, considerando grafos para codificar un cierto conjunto de relaciones de dependencia; los nodos del grafo representan a las variables y las aristas del grafo representan las relaciones de dependencia entre ellas. El grafo es una representación visual e intuitiva de las relaciones importantes entre las variables del problema. El grafo también proporciona un criterio para factorizar la FPC acorde a las relaciones indicadas en el mismo, incluyendo el mínimo número de parámetros necesarios. Por tanto, el grafo proporciona la descripción cualitativa del problema, mientras que la factorización de la probabilidad resultante proporciona la descripción cuantitativa, que permite propagar evidencia y obtener conclusiones utilizando la probabilidad como mecanismo de inferencia.

Existen distintos tipos de redes probabilísticas, dependiendo del tipo de grafo utilizado (dirigido, no dirigido, mixto), y de la forma de codificar las relaciones de dependencia en el mismo. Las más famosas son las Redes Bayesianas (RBs), dadas por grafos dirigidos acíclicos, y las redes de Markov, dados por grafos no dirigidos; esta tesis se centra básicamente en el primer tipo de modelos (ver Castillo et al., 1997; Gutiérrez et al., 2004b, para más detalles). Tanto la estructura gráfica como las probabilidades pueden ser aprendidas directamente de los datos con algoritmos de aprendizaje apropiados, y sin necesidad de supervisión humana. Por tanto, pese a su complejidad teórica, estos modelos son bastante intuitivos y fáciles de aplicar en situaciones prácticas por lo que actualmente son utilizadas como técnicas de minería de datos en numerosos problemas.

Llegado este punto, cabe preguntarse cómo se codifican las relaciones de dependencia expresadas en el grafo. En la siguiente sección se analiza este problema.

3.3.1. Codificación de Dependencias. Criterios de Separación

Los criterios que permiten conocer si una cierta relación de dependencia o independencia condicional se cumple en un grafo son conocidos como criterios de separación. El caso más simple es el de los grafos no dirigidos, donde las aristas entre nodos denotan dependencias y, por tanto, el criterio de separación más simple consiste en que dos nodos (o conjuntos de nodos) son independientes si no existe ningún camino que los une en el grafo, y son independientes dado un tercero si todos los caminos que los unen pasan por el tercer nodo. Por ejemplo, en el grafo de la

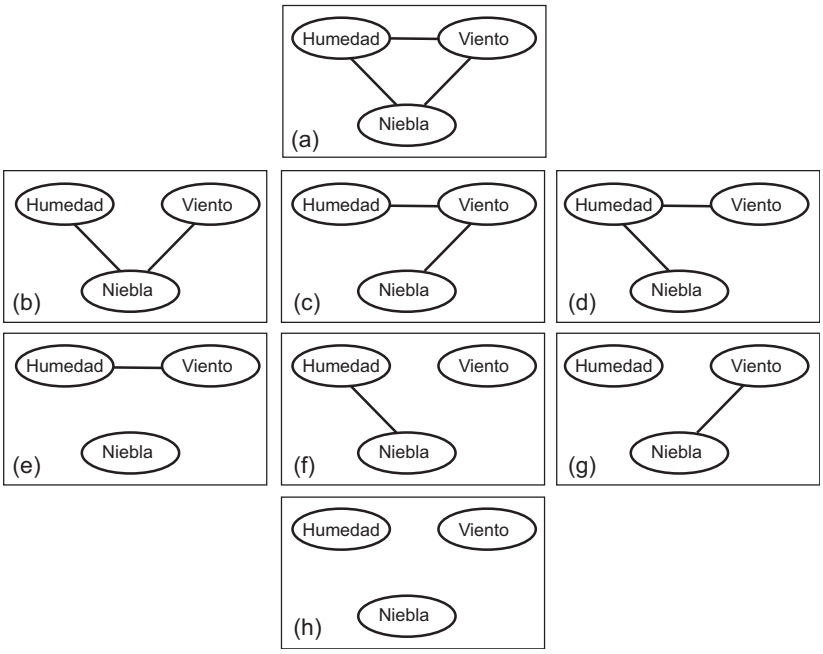

Figura 3.2: Posibles grafos no dirigidos para tres variables.

Fig. 3.2(a) se dan todas las posibles dependencias entre nodos. En el grafo de la Fig. 3.2(b) se expresa que $D(humedad, viento)$, y que además $I(humedad, viento|niebla)$; es decir, la *humedad* y la *velocidad del viento* son dependientes, pero son condicionalmente independientes dada la *niebla*. Sin embargo, en Meteorología se sabe que la humedad es prácticamente independiente de la velocidad del viento, mientras que si hay niebla, hay mucha humedad y poco viento, es decir son condicionalmente dependientes. El problema de los grafos no dirigidos es que no son capaces de expresar este tipo de relaciones: $I(humedad, viento)$ junto con $D(humedad, viento|niebla)$. Es decir, la capacidad expresiva de los grafos no dirigidos es limitada.

El criterio de separación para gafos dirigidos es más complejo, pero a la vez más potente, dando cabida a relaciones de dependencia más generales (como las que se cumplen realmente en el ejemplo anterior). El sentido de las aristas proporciona un grado de libertad más para definir el criterio de separación; por ejemplo, los vecinos de un nodo se dividen en *padres* (de los que sale una arista que confluye en el nodo) e *hijos* (a los que llega una arista desde el nodo). Los nodos que se alcanzan *corriente arriba* de uno dado son sus ancestros: padres, abuelos, etc; y todos los nodos que se encuentran *corriente abajo* son sus descendientes: hijos, nietos, etc.

En un grafo dirigido, dos nodos son dependientes si existe un camino que los une en el subgrafo ancestral moralizado (el subgrafo que sólo contiene a los nodos ancestrales de ellos en el cual, además, se han añadido enlaces adicionales entre

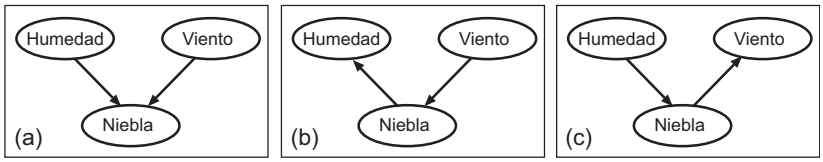

Figura 3.3: Tres grafos dirigidos definidos a partir del mismo grafo no dirigido.

padres con hijos comunes); de acuerdo con esto, en el grafo de la Fig. 3.3(c) todos los nodos son dependientes entre sí. Expresado de forma equivalente, dos nodos son condicionalmente dependientes dado un tercero si existe algún camino que los una y que no contenga al tercero, o que lo contenga en el vértice de una v-estructura (una *v-estructura* es un camino de la forma $A \rightarrow C \leftarrow B$, donde C es el vértice). Por tanto, las v-estructuras cumplen la misión de codificar relaciones de la forma: $I(A, B|\emptyset)$, $D(A, B|C)$, donde C es el vértice de una v-estructura entre A y B. Este tipo de relaciones se denominan transitivas y son importantes en la práctica. El criterio de separación en grafos dirigidos (denominado d-separación) fue definido de esa forma para incluir este tipo de relaciones (ver Castillo et al., 1997, para más detalles).

El número de grafos dirigidos se incrementa con respecto al de no dirigidos. Por ejemplo, la Fig. 3.3 muestra tres grafos dirigidos distintos asociados al mismo grafo no dirigido. En el grafo (a) se tiene que $I(humedad, viento)$ (pues no existen ancestros de ambos), pero $D(humedad, viento|niebla)$ (pues existe el camino $humedad \rightarrow niebla \leftarrow viento$). Éste es el modelo que mejor refleja la realidad. Los otros dos grafos (b)-(c) cumplen $D(humedad, viento)$, $I(humedad, viento|niebla)$; en realidad, ambos grafos son equivalentes, es decir, definen el mismo conjunto de relaciones de dependencia (en general, dos grafos son equivalentes si tienen asociado el mismo grafo no dirigido y poseen las mismas v-estructuras). En realidad, las v-estructuras son los elementos clave en un grafo dirigido que le confieren su capacidad para codificar modelos de dependencia con propiedades convenientes (ver Castillo et al., 1997, Cap. 4).

3.3.2. Factorización de la Probabilidad

Además del grafo, que representa la parte cualitativa del modelo, codificando las relaciones de dependencia relevantes para un problema dado, una red probabilística está compuesta por un modelo probabilístico (una FPC), que contenga las mismas dependencias/independencias que el grafo, con el criterio de separación correspondiente. El problema de construir esta FPC no es trivial ya que el conjunto de modelos

de dependencia asociados a una FPC no tiene porqué ser el mismo que el que se puede representar mediante un grafo, dirigido o no dirigido. Por tanto, en general, es posible que el modelo de dependencia asociado a un grafo no tenga una correspondencia exacta con el derivado de la FPC (en caso afirmativo el grafo se denomina un mapa perfecto); el problema consiste en hallar una factorización de la FPC que tenga todas las dependencias que contiene el grafo y quizás alguna adicional, lo que implica algún parámetro extra, pero no una pérdida de generalidad; estos modelos se denominan *I-mapas*. Las redes de Markov son los I-mapas de los modelos probabilísticos asociados a grafos no dirigidos, donde la FPC se factoriza en base a funciones locales definidas sobre conjuntos de nodos que se obtienen manipulando el grafo. La factorización dada por un grafo dirigido es mucho más intuitiva y los modelos resultantes se denominan redes Bayesianas.

En el siguiente ejemplo se ilustra la forma de construir un modelo probabilístico simplificado (factorizado) que contenga un conjunto de relaciones de independencia arbitrario.

Ejemplo 3.1 ***Modelo Probabilístico Factorizado.***

Considérese un problema definido con cuatro variables binarias $\{X_1, X_2, X_3, X_4\}$*, con un total de* 16 *estados distintos. Si no se conoce ninguna relación de independencia, tendrá* 15 *parámetros libres. Supóngase que se sabe que las variables del problema cumplen:*

$$I(X_3, X_1|X_2)\ y\ \ I(X_4, \{X_1, X_3\}|X_2).$$

Estas relaciones de dependencia implican:

$$p(x_3|x_1, x_2) = p(x_3|x_2),\ \ p(x_4|x_1, x_2, x_3) = p(x_4|x_2);$$

por tanto, los parámetros de la FPC del modelo resultante no son libres, sino que están sujetos a las restricciones anteriores. Sería deseable encontrar una forma de expresar la FPC con un número reducido de parámetros que fuesen libres para este problema (de esta manera, además de reducir el número de parámetros, se eliminaría redundancia y posibles fuentes de error). En este ejemplo, la regla de la cadena proporciona una solución simple. La regla de la cadena indica que una FPC siempre se puede factorizar de la siguiente forma

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3).$$

En este caso concreto, las condiciones anteriores permiten sustituir alguna de estas funciones, resultando:

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_2),$$

que es la expresión general de las funciones de probabilidad que cumplen las relaciones de independencia anteriores (los parámetros libres se reducen de 15 *a* 7*).*

Sin embargo, en general no es tan simple hallar una factorización apropiada de la FPC que contenga las condiciones de dependencia requeridas para un problema dado. Es decir, dado un conjunto de variables y un conjunto de relaciones de dependencia definidas sobre ellas, no es trivial hallar un modelo probabilístico (una factorización de la FPC) que incluya estas relaciones y sólo requiera especificar los parámetros libres del modelo. Las redes probabilísticas facilitan este proceso proporcionando mecanismos sencillos de factorización de la FPC que incluyen las dependencias representadas por el grafo.

3.4. Redes Bayesianas. Definición y Tipos

Las Redes Bayesianas (RBs) son los modelos de redes probabilísticas más populares y están definidas por grafos dirigidos acíclicos (DAGs) que codifican las relaciones de dependencia/independencia según el criterio para separación de grafos dirigidos definido en la Sec. 3.3.1. Estos grafos definen un modelo probabilístico con las mismas dependencias a través de un factorización mediante el producto de varias funciones de probabilidad condicionada:

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\pi_i), \tag{3.3}$$

donde Π_i es el conjunto de los padres del nodo X_i en el grafo.

A través de esta factorización, las independencias del grafo son inmediatamente traducidas al modelo probabilístico de una manera muy práctica.

3.4.1. Tipos de Redes Bayesianas

Dependiendo del carácter discreto o continuo de las variables, las funciones de probabilidad (3.3) son familias paramétricas específicas. Fundamentalmente hay tres tipos de RBs (en esta Tesis se considerará únicamente el primer caso):

- *RBs Multinomiales.* En una RB multinomial se considera que todas las variables son discretas, es decir, cada variable tiene un número finito de posibles estados. Asimismo se considera que la función de probabilidad de cada variable condicionada a sus padres es multinomial y por tanto, queda especificada por la tabla de probabilidades correspondientes a las diferentes combinaciones

de estados entre las variables implicadas. Nótese que el número de parámetros requeridos para especificar la FPC en el caso de dependencia total en el término de la izquierda de (3.3) es m^n, siendo n el número de variables y m el número de estados, mientras que la factorización definida por la RB sólo requiere $o(n\,m^k)$, donde k es el máximo número de padres en el grafo.

- *RBs Gaussianas.* En una RB Gaussiana se considera que las variables tienen una distribución normal multivariante, $N(\mu, \Sigma)$, cuya función de densidad conjunta viene dada por

$$f(x) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left\{-1/2(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \tag{3.4}$$

donde μ es el vector de medias n-dimensional, Σ es la matriz de covarianzas $n \times n$, $|\Sigma|$ es el determinante de Σ, y μ^T denota la traspuesta de μ. En una RB Gaussiana esta FPC se especifica como en (3.3) por:

$$f(x_i|\pi_i) \sim N\left(\mu_i + \sum_{j=1}^{i-1} \beta_{ij}(x_j - \mu_j),\ v_i\right), \tag{3.5}$$

donde β_{ij} es el coeficiente de regresión de X_j en la regresión de X_i sobre los padres de X_i, Π_i, y $v_i = \Sigma_i - \Sigma_{i\Pi_i}\Sigma_{\Pi_i}^{-1}\Sigma_{i\Pi_i}^T$ es la varianza condicional de X_i, dados $\Pi_i = \pi_i$, donde Σ_i es la varianza incondicional de X_i, $\Sigma_{i\Pi_i}$ es el vector de covarianzas entre X_i y las variables en Π_i, y Σ_{Π_i} es la matriz de covarianza de Π_i. Nótese que β_{ij} mide el grado de relación entre X_i y X_j. Si $\beta_{ij} = 0$, entonces X_j no es padre de X_i. Así, la RB *Gaussiana* está definida por un conjunto de parámetros $\{\mu_1, \ldots, \mu_n\}$, $\{v_1, \ldots, v_n\}$, y $\{\beta_{ij} \,|\, j < i\}$, como se mostró en (3.5). Más detalles, con aplicaciones acerca de este tema se encuentran en Geiger and Heckerman (1994).

- *RBs Mixtas.* En general, el problema de definir modelos de RBs que incluyan variables discretas y continuas al mismo tiempo es complicado. Un caso particular lo constituyen los modelos Gaussianos condicionados, donde se permite que una variable continua tenga padres discretos, pero no al contrario (Jordan, 1998b). También existen otras variantes que se han aplicado a ejemplos concretos. Por ejemplo, (Castillo et al., 1998) describen una aplicación de modelos mixtos con variables discretas y *Beta*.

En lo que resta de Tesis nos centraremos exclusivamente en las RBs multinomiales, de las cuales a continuación se describe un sencillo ejemplo ilustrativo, que será utilizado a lo largo de este capítulo.

Ejemplo 3.2 (Ejemplo de Red Bayesiana Multinomial) *Aunque este ejemplo tiene unas dimensiones mucho más reducidas que los modelos operativos, será utilizado para representar a pequeña escala las características fundamentales de la escala real. En este caso construimos un modelo que trata de explicar las relaciones entre diferentes fenómenos meteorológicos en el observatorio de Parayas (Aeropuerto de Santander):* lluvia*,* rocío*,* tormenta*,* granizo*,* nieve*,* escarcha*, y* niebla*, a partir de una base de datos binaria con las ocurrencias diarias de estos fenómenos durante el período* $1992 - 1995$*. El número de estados posibles de este sistema es* $2^7 = 128$*, resultando, al igual que ocurre en la escala real, demasiado elevado como para plantear un modelo general de dependencia, dado el número de observaciones disponible* [1]*.*

En la Fig. 3.4, se muestra una RB sencilla, realizada manualmente, donde sólo se representan las relaciones de dependencia relevantes, definiendo un modelo probabilístico a través de una factorización dada por 7 *funciones de probabilidad condicionada, una para cada variable, condicionadas al conjunto de sus padres. En la Tabla 3.2, se muestra la tabla de probabilidad correspondiente al nodo* tormenta *dado su único padre* lluvia*:* $p(tormenta|lluvia)$*. Los valores de la tabla han sido estimados a partir de los datos.*

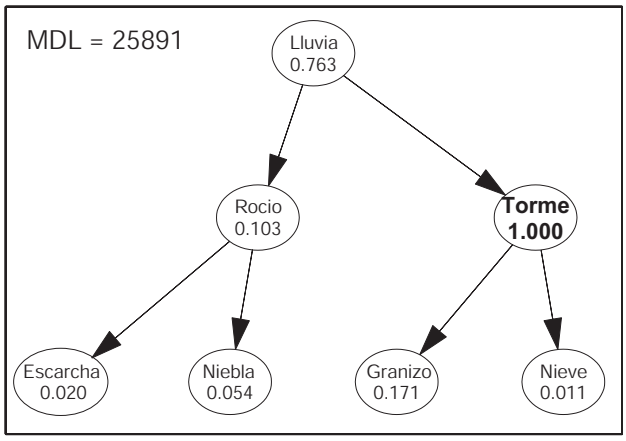

Figura 3.4: Grafo dirigido realizado manualmente que representa relaciones 'lógicas' existentes entre las variables, en el que se muestran las probabilidades de los diferentes meteoros cuando hay tormenta.

[1] En un caso real, el número de estados es del orden de $2^{100} > 10^{30}$, y el número de observaciones diarias es del orden de 15.000.

	tormenta	
π	*si*	*no*
lluvia=si	0.0264	0.4061
lluvia=no	0.0082	0.5593

Tabla 3.1: Tabla de probabilidad para el nodo *tormenta*, dados sus padres $\Pi = \{lluvia\}$.

3.4.2. Razonamiento Probabilístico. Inferencia

Una vez especificado el modelo completo (*grafo + FPC + parámetros*), la principal utilidad de las RBs consiste en hallar de forma eficiente la probabilidad de cualquier nodo (o conjunto de nodos), dada una cierta información (evidencia). Este proceso de razonamiento probabilístico permite cuantificar la incertidumbre de las distintas variables y eventos del problema a medida que se va teniendo nueva información, o evidencia. Por ejemplo, el grafo de la Fig. 3.4 define las relaciones entre los meteoros asociados a una estación (Parayas, Aeropuerto de Santander). Si la variable de interés es la *lluvia*, entonces la probabilidad $p(lluvia)$ proporciona información sobre su estado inicial, y la probabilidad condicionada $p(lluvia|e)$ proporciona información actualizada cuando se conoce que ha ocurrido cualquier evento e. En general, el cálculo de estas probabilidades es computacionalmente intratable para problemas con muchas variables; sin embargo, existen métodos eficientes de propagación de evidencia que utilizan la estructura de dependencia del grafo para realizar esta tarea (ver Castillo et al., 1997, Caps. 6 y 7). Así se pueden obtener las probabilidades marginales o condicionadas de cada uno de los nodos; por ejemplo, utilizando el grafo de la Fig. 3.4, y la tabla 3.2, tendríamos: $p(lluvia|tormenta) = p(lluvia\&tormenta)/p(tormenta) = 0.0264/0.0346 = 0.763$.

Conviene resaltar la dificultad de establecer, aún por personal experto, un conjunto apropiado de relaciones relevantes en un problema práctico dado. Cuando el modelo es elaborado por un experto humano, se pone de manifiesto la dificultad de establecer relaciones de dependencia que no se basen en asociaciones directas causa-efecto concebidas fundamentalmente a través de la experiencia y/o el conocimiento de las leyes físicas que interrelacionan las diferentes variables. Esto dificulta, o anula, la posibilidad de tener en cuenta dependencias condicionadas, derivadas de la acción combinada de varias variables sobre otra u otras diferentes. En la Sec. 3.5.2 se ofrecen algunos resultados que dan una idea del alcance de estas limitaciones en un caso real.

Un algoritmo de aprendizaje automático permite construir una solución de for-

ma relativamente rápida y simple, por lo que muchos elementos importantes del modelo (factorización de la FPC, tablas de probabilidad, etc.) permanecen ocultos durante el proceso, de manera que el usuario obtiene un modelo tipo "caja negra" como respuesta a su problema. Este hecho es muy común en otras técnicas de aprendizaje automático; sin embargo, en este caso, es posible analizar el modelo una vez obtenido, y elaborar una explicación física o estadística de la solución. Con otras técnicas (como las redes neuronales), esta interpretación a posteriori es más complicada (o incluso imposible) pues los patrones de conocimiento que emplean son más abstractos. En las siguientes secciones se describen con mas detalle las estrategias de aprendizaje automático estándar para obtener estos modelos directamente a partir de un conjunto de datos.

3.5. Algoritmos de Aprendizaje Automático

Para definir una RB es necesario especificar una estructura gráfica (un DAG) y una función de probabilidad conjunta definida por el producto de las probabilidades de cada nodo dados sus padres. Sin embargo, en muchos problemas prácticos no se conoce ni la estructura del grafo ni las probabilidades resultantes necesarias. Debido a este problema, se han desarrollado diferentes métodos de aprendizaje automático para obtener una RB a partir de un conjunto de datos observados (una base de datos). Estos métodos, como es lógico, se encuentran en continua revisión, con numerosas adaptaciones para cada problema específico. Por ello, en este capítulo nos limitaremos a dar una panorámica actual de los métodos estándar más utilizados, con la excepción de un apartado final donde se describen algunos métodos específicos para problemas de clasificación. Para ampliar información sobre métodos "clásicos" de aprendizaje, se pueden consultar algunas recopilaciones en Buntine (1996), Sanguesa and Cortes (1997) o Krause (1999); y para descripciones más detalladas se puede ver Neapolitan (2003), Jordan (1998a), o Heckerman (1996).

El aprendizaje automático de RBs se puede dividir en dos tareas distintas:

- *Aprendizaje estructural*, donde se establecen las relaciones de dependencia entre las variables del modelo a partir de un conjunto de datos, obteniendo el grafo que mejor los represente.

  El aprendizaje estructural es el más complejo de los dos y es el auténtico "cuello de botella" en el desarrollo y operatividad de las RBs aplicadas a las Ciencias Atmosféricas. La búsqueda de la estructura que mejor representa el conjunto de datos es un problema NP-completo (lo cual lo hace computa-

cionalmente intratable a partir de unas pocas variables) y requiere algoritmos eficientes que, aunque no garanticen la optimalidad, la aproximen con un coste computacional acotado (ver, por ejemplo, Neapolitan, 2003). La complejidad del aprendizaje estructural se deriva de la enorme cantidad de DAGs distintos para un determinado número de variables. Por ejemplo, para 3 variables existen 12, para 4 variables 543, y el número crece de forma combinatoria con el número de variables. Queda patente que este problema es especialmente crítico en Meteorología, ya que cualquier aplicación práctica, por simple y concreta que sea, involucra un gran número de variables, lo que dificulta la definición del modelo gráfico más conveniente. Por tanto, la aplicación de las RBs en sistemas atmosféricos depende de la disponibilidad de procedimientos de aprendizaje automático para inferir, a partir de los datos disponibles, una estructura gráfica conveniente.

- *Aprendizaje paramétrico*, donde se calculan los parámetros de acuerdo con la estructura del grafo, el tipo de RB elegida y la base de datos disponible. El aprendizaje paramétrico consiste en estimar los parámetros estadísticos necesarios (tablas de probabilidad, medias, varianzas, etc.) a partir de los datos, para lo que se utilizan diferentes estimadores estadísticos; en la actualidad existen numerosos algoritmos computacionalmente baratos y fiables por lo que en esta Tesis únicamente se dan las descripciones elementales refiriendo al lector a documentos más exhaustivos sobre el tema (ver, por ejemplo, Castillo et al., 1997).

Los métodos de aprendizaje estructural se pueden clasificar básicamente en dos tipos. Por un lado, están los métodos que utilizan *métricas* de complejidad-bondad de ajuste y *algoritmos de búsqueda*, en los cuales la *métrica* define la calidad de una red en función de los datos disponibles y el *algoritmo de búsqueda* explora el conjunto de todas las posibles redes tratando de maximizar la medida de calidad; dependiendo de la medida y la técnica de búsqueda empleadas, surge toda una gama de algoritmos de aprendizaje, desde métodos voraces simples (Cooper and Herskovitz, 1992) hasta métodos que emplean algoritmos genéticos (Larrañaga et al., 1996b).

Por otra parte, tenemos los métodos basados en la *detección de independencias*. En este caso, a partir de unos datos dados, se aplican tests estadísticos para determinar las relaciones de independencia condicionales más importantes que se derivan de dichos datos. Finalmente se trata de obtener la RB que mejor se ajusta a estas relaciones de independencia, tratando de representar de forma fiable el mayor número posible de las mismas. Aunque son algo más eficientes, son al mismo tiempo

más sensibles a fallos en el test, especialmente cuando se trata con muchas variables (Friedman et al., 1999), por lo que en esta tesis únicamente se utilizarán algoritmos de búsqueda, sin que ello suponga que se descarta el uso de otro tipo de algoritmos.

También han sido diseñados métodos híbridos, mezclando las dos estrategias básicas con el fin de optimizar la búsqueda (ver, por ejemplo de Campos, 2006); sin embargo, estos métodos no se describen en esta Tesis.

3.5.1. Algoritmos Basados en Test de Independencia

Estos algoritmos realizan de forma eficiente la comprobación de las distintas dependencias condicionales que se pueden formar combinando las variables del problema (se suelen considerar conjuntos reducidos de variables). Finalmente, las dependencias halladas se expresan de forma conjunta mediante un grafo. Aunque, como ya hemos mencionado, esta tésis se centra en los algoritmos de búsqueda, a continuación y como ejemplo ilustrativo, se ofrece una breve descripción del algoritmo PC, por ser uno de los métodos estándar en esta categoría.

Descripción del Algoritmo PC

La idea básica de este algoritmo es deducir el conjunto de independencias existentes en un problema dado a partir de una base de datos característica del mismo. Dichas independencias se representan en un grafo dirigido y constituyen la base estructural de la red Bayesiana.

El algoritmo PC se basa en la realización de tests de hipótesis $I(X, Y|Z)$, para distintas combinaciones de X, Y y Z. Para ello se utiliza la Ec. (3.2) para determinar si la dependencia se cumple con una confianza dada a partir de la muestra disponible en una base de datos. La medida utilizada en el algoritmo PC es la información mutua, concepto que deriva de la entropía y proporciona un criterio de medida de dependencia a partir de la cantidad de información que una variable contiene sobre otra.

La expresión matemática de la información mutua entre dos variables X e Y, condicionada a una tercera Z es:

$$MI(X, Y|Z) = \sum_z p(z) \sum_{x,y} p(x, y|z) log \frac{p(x, y|z)}{p(x|z)p(y|z)}, \tag{3.6}$$

donde las probabilidades se estiman tomando las frecuencias relativas observadas en la base de datos, que son los estimadores de máxima verosimilitud de la probabilidad. Una vez establecida la medida de referencia, la hipótesis $I(X, Y|Z)$ se comprueba

utilizando el estadístico:

$$G = 2\,N\,MI(X, Y|Z), \tag{3.7}$$

donde N es el tamaño de la muestra. Este estadístico posee una distribución χ^2 bajo la hipótesis de independencia, con $gl = (r_x - 1)(r_y - 1)r_z$ grados de libertad donde r_x, r_y y r_z representan la cardinalidad de cada variable (la Fig. 3.5 muestra la distribución χ^2 para diferentes grados de libertad). Fijando el nivel de significación α el test queda finalmente determinado por la región de aceptación $\{x|F_{\chi^2(gl)}(x) <= 1-\alpha\}$; es decir, si el estadístico G queda queda en la cola de la distribución, fuera de la región de aceptación, se rechaza la hipótesis de independencia con la significación dada. Por ejemplo, en la Fig. 3.5 se muestra en azul la región de rechazo para el caso de 15 grados de libertad con una significación del 5 % (confianza del 95 %).

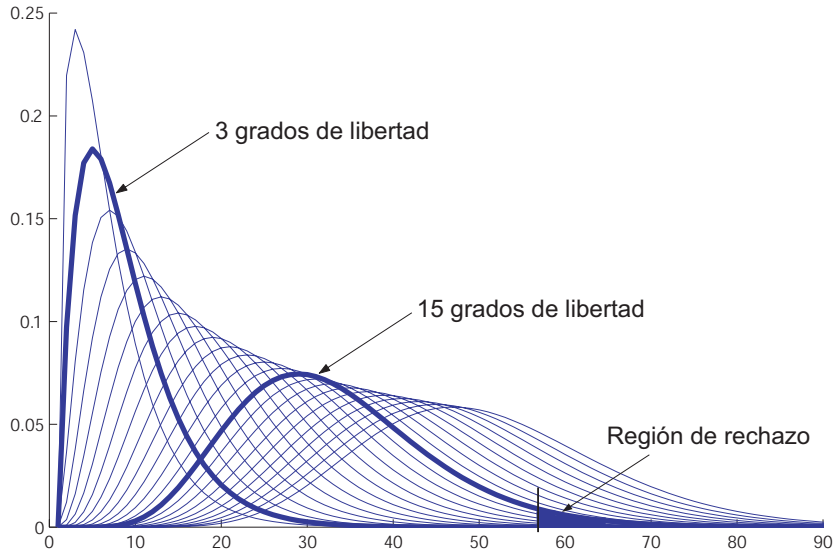

Figura 3.5: Aspecto de la distribución χ^2 para diferentes grados de libertad. Se han resaltado las distribuciones para 3 y 15 grados de libertad simplemente por marcar referencias visuales; también se muestra la región de rechazo para el caso de 15 grados de libertad.

Existen dos versiones de este algoritmo, una de ellas parte de un grafo completo no dirigido y la otra de un grafo completamente disconexo. Si se parte de un grafo completo, el algoritmo PC examina primero cada pareja de variables para comprobar si son incondicionalmente independientes, en cuyo caso se elimina la arista correspondiente. A continuación, vuelve a examinar cada par de variables para verificar si son condicionalmente independientes dada una tercera variable (que forma parte del conjunto de nodos adyacentes a alguno de los que se esta comprobando); si es

así, se elimina la arista correspondiente, y así sucesivamente. Una vez completado el grafo, se procede a dar la dirección a cada una de sus aristas. Para ello, primero se consideran todas las ternas de nodos, A, B y C tales que $D(A,B)$, $D(B,C)$ e $I(A,C|B)$, y si B no forma parte del conjunto que las separa, entonces se asigna la estructura $A \rightarrow B \leftarrow C$; las aristas restantes se dirigen de forma que no creen nuevas v-estructuras. La otra alternativa es partir de un grafo disconexo. En este caso, el funcionamiento es similar, pero en lugar de quitar enlaces, se van añadiendo en base a las pruebas realizadas.

Al igual que el resto de algoritmos basados en la detección de independencias, el algoritmo PC tiene un elevado coste computacional, dependiendo del número máximo de nodos que se considere en cada uno de los conjuntos del test.

Búsqueda de Independencias Específicas. CSI.

Entre las numerosas variaciones de los algoritmos de búsqueda de independencias se describe en esta sección el concepto de independencia específica de contexto (CSI, de las siglas en inglés *Context-Specific Independence*) por considerarlo de gran interés en el contexto del a Meteorología; para más detalles consultar Boutilier et al. (1996) y otros trabajos más recientes con ideas como la modularidad o la partición de los datos (Neil et al., 2000).

Formalmente decimos que D y B son contextualmente independientes dado C en el contexto A=a, si se cumple que:

$$p(D|a,B,C) = p(D|a,C) \tag{3.8}$$

Por ejemplo, en una red donde la variable D tiene tres padres A, B y C y presenta la siguiente tabla de probabilidades:

A	B	C	$p(D\|A,B,C)$
1	1	1	.95
1	1	0	.95
1	0	1	.20
1	0	0	.05
0	1	1	.00
0	1	0	.00
0	0	1	.00
0	0	0	.00

se ve claramente que cuando se produce el evento $A = 0$, la probabilidad de $D = 1$ es nula, independientemente de los valores que tomen las restantes variables.

De esta manera, el concepto CSI permite poner de manifiesto dependencias o independencias que existen sólo en ciertos contextos y que quedan ocultas al tratar el problema de forma global. Por ejemplo, en Meteorología existen distintas situaciones atmosféricas (o contextos) que pueden desencadenar distintas relaciones de dependencia entre las variables. El principal inconveniente de esta metodología es su elevado coste computacional, que no permite abordar de forma automática en un tiempo razonable problemas prácticos reales. Nótese que en este caso es necesario combinar los valores de cuatro conjuntos de variables distintos (los tres de la dependencia mas el contexto).

3.5.2. Algoritmos de Evaluación y Búsqueda

Otra categoría de métodos de aprendizaje estructural son los algoritmos de evaluación y búsqueda, que se componen de dos partes:

1. Una *medida de calidad*, usada para evaluar la calidad de cada red candidata. Esta medida es global y mide simultáneamente tanto la calidad de la estructura gráfica y del modelo probabilístico resultante.

2. Un *algoritmo de búsqueda*, utilizado para explorar el espacio de grafos posibles de manera eficiente, y así encontrar de forma iterativa una RB de buena calidad.

Medidas de Calidad

Existen diferentes medidas de calidad propuestas en la literatura para favorecer aquellas redes que mejor representan a los datos (término de bondad de ajuste) y penalizar al mismo tiempo las más complejas (término de penalización). Para ello, a cada red $B = (M, \theta)$, con estructura (grafo dirigido) M y parámetros θ, se le asigna una medida de calidad que es función de su probabilidad a posteriori dados los datos D. Esta distribución de probabilidad a posteriori $p(B|D)$ representa el término de bondad de ajuste y se calcula como sigue:

$$p(B|D) = p(M, \theta|D) = \frac{p(M, \theta, D)}{p(D)} \propto p(M)p(\theta|M)p(D|M, \theta), \tag{3.9}$$

Por ejemplo, en el caso de RBs multinomiales, asumiendo ciertas hipótesis acerca de las distribuciones a priori de los parámetros y una probabilidad inicial uniforme para todos los modelos, se puede obtener la siguiente medida de calidad Bayesiana

basada en el logaritmo de la verosimilitud, incluyendo un término de penalización (ver, por ejemplo, Heckerman, 1996):

$$Q(B|D) = \log p_B(D) - \frac{1}{2} Dim(B) \log N \tag{3.10}$$
$$\propto \sum_{i=1}^{n} \sum_{j=1}^{r_i} \sum_{k=1}^{s_i} N_{ijk} \log \frac{N_{ijk}}{N_{ik}} - \frac{1}{2} Dim(B) \log N,$$

donde N es el tamaño de la muestra, n es el número de variables, r_i es el cardinal de la i-th variable, s_i el número de estados del conjunto de padres Π_i, N_{ijk} es el número de realizaciones en la base de datos que son consistentes con $y_i = j$ y $\pi_i = k$, y N_{ik} el número de aquellas consistente con $\pi_i = k$. $Dim(B)$ denota la dimensión de la red, que es el número de parámetros libres necesarios para definir la FPC. Todos los parámetros de (3.10) se pueden obtener fácilmente a partir de los datos. Una de las principales ventajas de esta medida de calidad es que se puede factorizar de acuerdo con la estructura de la red; de esta manera, cada vez que se modifica un enlace en el proceso de aprendizaje, sólo es necesario recalcular un término de esta ecuación, lo cual supone una ventaja en términos de eficiencia.

Continuando con las redes multinomiales, otra medida de calidad muy utilizada, es MDL(*Minimum Description Length*), descrita en Lam and Bacchus (1994), donde para cada modelo de red B, se evalúan simultáneamente la simplicidad y la verosimilitud de los N datos de entrenamiento D, maximizando:

$$MDL(B|D) = \sum_{i=1}^{N} \log p_B(D) - r_B * log(N/2), \tag{3.11}$$

donde r_B, es el número de parámetros libres asociados a la FPC. No es dificil comprobar que 3.10 y 3.11 son equivalentes (Castillo et al., 1997), y por lo tanto, en lo sucesivo consideraremos únicamente MDL.

Existen otras medidas de calidad, como las basadas en la teoría de la información (Cheng et al., 1997), o medidas de calidad adaptadas al objetivo de la red (aprendizaje discriminativo); entre estas últimas, las más conocidas son las que maximizan la verosimilitud condicional, aunque presentan graves problemas de eficiencia computacional (Grossman and Domingos, 2004; Jing et al., 2005), y por el momento sólo son aplicables en sistemas de baja dimensionalidad. Posteriormente, en la Sec. 8.2, se volverá sobre este importante tema.

En el caso de RBs Gaussianas, se obtiene una medida de calidad muy similar considerando que la distribución de los parámetros es de tipo normal-Wishart (ver Geiger and Heckerman, 1994, para una descripción detallada).

Algoritmos de Búsqueda

En cuanto a los algoritmos de búsqueda en el espacio de las posibles redes Bayesianas, existen numerosos métodos y la bibliografía al respecto es muy extensa (ver Blanco et al., 2003; Neapolitan, 2003, y las referencias contenidas). Los más simples de estos métodos aplican técnicas voraces para explorar, de forma iterativa (por ejemplo, añadiendo enlaces a un grafo inicialmente disconexo), el espacio de RBs, sin reconsiderar ninguno de los pasos dados (a continuación se describen los algoritmos $K2$ y B). Otros métodos más sofisticados como *hill-climbing*, *simulated annealing*, etc., añaden una componente probabilística a la búsqueda permitiendo evitar óptimos locales del problema (Mohamadi et al., 2008). Por otra parte, los algoritmos bio-inspirados, como los algoritmos genéticos (P. Larra et al., 1999), evolutivos (Galea et al., 2004), colonias de hormigas (Middendorf et al., 2002), etc., evolucionan conjuntos de soluciones candidatas que van mejorando generación tras generación.

En la práctica los métodos voraces son muy utilizados en problemas con un gran número de variables y datos. Aunque estos métodos pueden converger a soluciones locales no óptimas, la complejidad del problema de búsqueda hace que otras opciones más generales para lograr un óptimo resulten excesivamente costosas (desde el punto de vista computacional).

Descripción del algoritmo k2

El algoritmo *k2* realiza una búsqueda voraz muy eficaz para encontrar una red de calidad en un tiempo razonable (Cooper and Herskovitz, 1992). Este algoritmo ordena previamente los nodos para evitar la generación de ciclos y comienza con un grafo vacío. A continuación, para cada variable Y_i, siguiendo el orden establecido, el algoritmo añade a su conjunto actual de padres Π_i (vacío inicialmente) aquel nodo cuyo orden es anterior al suyo propio y que además produce un mayor incremento en la medida de calidad de entre aquellos que mejoran la calidad del grafo previo. Este proceso se repite para cada nodo, mientras el incremento de calidad supere un cierto umbral preestablecido. También se puede limitar el número máximo de padres de cada variable en el proceso de búsqueda.

Considerando la descomponibilidad de la red, la contribución de la variable Y_i con padres Π_i a la calidad de la red viene dada por

$$\sum_{j=1}^{r_i} \sum_{k=1}^{s_i} N_{ijk} \log \frac{N_{ijk}}{N_{ik}} \tag{3.12}$$

por lo que cada iteración del algoritmo *k2* afecta a uno sólo de los términos del sumando en (3.10), reduciendo un orden la complejidad del algoritmo. Esto lleva a un algoritmo de aprendizaje iterativo muy simple de complejidad $o(n^2)$.

Esto lleva a un algoritmo de aprendizaje iterativo muy simple, cuyo pseudo-código es el siguiente:

Algorithm 1 Algoritmo *K2* para aprendizaje estructural de RBs.

Require: Una muestra de datos de un conjunto previamente ordenado de nodos $\mathbf{D} = \{\mathbf{X_1}, \ldots, \mathbf{X_n}\}$ y un grafo disconexo B.

```
1: for i = 1 : n − 1 do
2:    S = score(D, B);
3:    for k = i + 1 : n do
4:       M = score(D, B ∪ {i → k});
5:       if M > S then
            S = M; B(i, k) = 1;
6:       end if
7:    end for
8: end for
```

Ensure: Un grafo dirigido acíclico B y su medida de calidad S.

Por ejemplo, el grafo de la Fig. 3.6 ha sido obtenido aplicando este algoritmo de aprendizaje automático al mismo problema descrito en el Ejemplo 3.4, donde la estructura fué obtenida subjetivamente; se puede ver que la calidad obtenida con k2 (MDL=-25773), es ligeramente superior a la del grafo subjetivo (MDL=-25891).

Una de las críticas más importantes a este algoritmo es la necesidad de ordenar previamente los nodos; en muchos problemas prácticos esta tarea es subjetiva y, por tanto, complica la aplicación del algoritmo, aunque existen algunos trabajos que facilitan la obtención automática de "buenas" ordenaciones de los nodos (ver, por ejemplo, Larrañaga et al., 1996a). En el Cap. 8 se analiza este problema en más detalle.

Descripción del algoritmo de aprendizaje B

El algoritmo B elimina el problema de la dependencia de la ordenación previa de los nodos, a costa de incrementar la complejidad del algoritmo (Buntine, 1991). Al igual que el algoritmo *k2*, se inicia con conjuntos de padres vacíos y en cada etapa se añade aquel enlace que maximiza el incremento de calidad, eliminando aquellos enlaces que originen ciclos. Este proceso se itera hasta que no se consigue un incremento de calidad, o bien se obtiene una red completa.

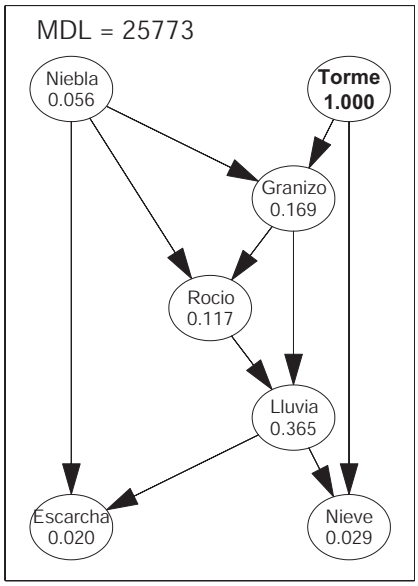

Figura 3.6: Grafo obtenido aplicando *k2* al Ejemplo 3.4.

La complejidad de este algoritmo es $o(n^3)$, por lo que su aplicación directa en problemas meteorológicos que incluyen gran cantidad de variables es muy limitada, quedando restringido a sistemas con muy pocas variables y estados. En el Cap. 8 se presenta una adaptación relativamente sencilla de este algoritmo que reduce su complejidad a cuadrática, haciéndolo útil para problemas con muchas variables. El pseudo-código es el siguiente:

Algorithm 2 Algoritmo B para aprendizaje estructural de RBs.

Require: Una muestra de datos de un conjunto de nodos $\mathbf{D} = \{\mathbf{X_1}, \ldots, \mathbf{X_n}\}$ y un grafo disconexo B.

1: **for** $i = 1 : n - 1$ **do**
2: $S = score(\mathbf{D}, B)$;
3: **for** $k = 1 : n$ **do**
4: $M = score(\mathbf{D}, B \cup \{i \rightarrow k\})$;
5: **if** $M > S \wedge acyclic(B \cup \{i \rightarrow k\})$ **then**
$S = M; B(i, k) = 1$;
6: **end if**
7: **end for**
8: **end for**

Ensure: Un grafo dirigido acíclico B y su medida de calidad S.

En la Fig. 3.7, se puede ver el grafo resultante de aplicar el algoritmo B que lógicamente obtiene el mayor valor de MDL.

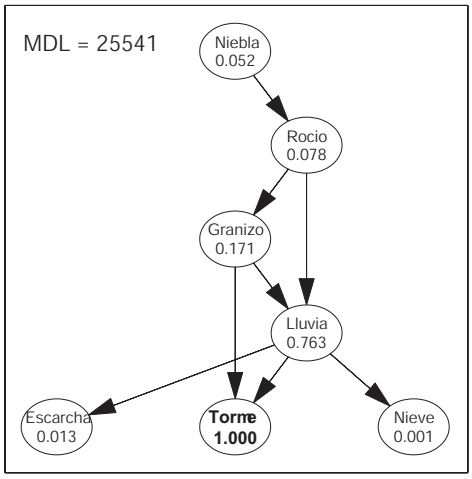

Figura 3.7: Grafo obtenido aplicando B al Ejemplo 3.4.

	Ll	Nv	Gr	Ro	Es
$p_{manual}(x\|Tr)$	.763	**.011**	.171	**.103**	.020
$p_B(x\|Tr)$	.763	.001	.171	.078	.013
$p_{manual}(x\|Nb)$	.401	.003	.015	.299	.019
$p_B(x\|Nb)$	.401	.003	.013	.299	.020
$p_{manual}(x\|Tr, Nb)$	**.739**	**.011**	.171	**.242**	.019
$p_B(x\|Tr, Nb)$	.767	.001	.156	.191	.013

Tabla 3.2: Probabilidades obtenidas con dos RBs distintas para el mismo problema.

Hombre vs Máquina

A continuación se muestra un ejemplo donde se pone de manifiesto la limitación del razonamiento humano, incluso en un sistema sencillo, para definir un grafo que represente tanto las dependencias incondicionales, como las condicionales de un modelo. La Tabla 3.5.2 muestra las probabilidades de ocurrencia para algunas variables, dadas distintas evidencias. p_{manual} denota las probabilidades obtenidas a partir del modelo generado por un experto *humano*, concretamente el definido por el grafo de la Fig. 3.4 y p_B las del *automático* definido en la Fig. 3.7. Se han resaltado en negrita aquellos valores sensiblemente discrepantes del modelo humano respecto del automático [2].

[2]Este es un caso muy sencillo en el que un modelo humano compite relativamente bien con modelos automáticos, pero cuando se consideran modelos más complejos, el diseño manual es, por lo general, muy inferior al automático.

3.6. Aprendizaje de Clasificadores

Los problemas de clasificación constituyen una tipología muy particular y extendida en la minería de datos, para la que se han desarrollado numerosos algoritmos específicos (Duda et al., 2000). En estos problemas, el conjunto de variables se descompone en una variable clase discreta C y un conjunto de atributos $X_1, \ldots, X_n$, y el objetivo es hallar el estado de la clase para una combinación de atributos dada. Dentro de los métodos probabilísticos, el esquema de clasificador más simple es el denominado clasificador Ingenuo de Bayes (IB) (en Inglés *Naïve Bayes*), en el cual se asume la hipótesis de que los atributos son condicionalmente independientes entre sí dada la clase:

$$I(X_i, X_j \,|\, C), \;\; \forall i, j. \tag{3.13}$$

Este modelo probabilístico puede ser expresado por una RB donde la variable clase es el único padre de todos los atributos, como se muestra en la Fig. 3.8. Esta red determina la siguiente factorización de la FPC, que caracteriza el modelo probabilístico conjunto asociado a este clasificador:

$$p(x_1, ..., x_n, c) = P(c) \prod_{i=1}^{n} P(x_i|c). \tag{3.14}$$

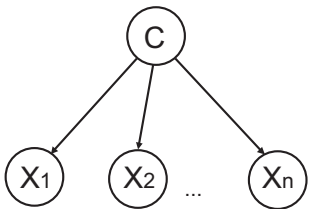

Figura 3.8: Grafo dirigido correspondiente a un modelo ingenuo de Bayes con una variable clase C y n atributos.

Este esquema de red tan sencillo ha resultado ser muy eficiente en numerosos problemas prácticos (incluso en problemas donde claramente se violan las hipótesis anteriores), obteniendo resultados similares a otros esquemas más complejos (ver, por ejemplo, Tsagalidis et al., 2008, para una aplicación en Meteorología). Por tanto, han sido numerosas las extensiones publicadas en la literatura para mejorar los resultados de este método, entre ellas cabe destacar dos grupos, el dedicado a los clasificadores ingenuos extendidos y, más reciente, el dedicado a los clasificadores ingenuos jerárquicos.

- Clasificadores Ingenuos Extendidos. Existe toda una familia de algoritmos de aprendizaje específicos para obtener clasificadores ingenuos extendidos (*augmented naive classifiers*). El más sencillo de todos es el TAN (*tree-augmented naive*), descrito en Friedman and Goldszmidt (1996), donde se eliminan la hipótesis (3.13) considerando dependencias entre atributos en forma de árbol, que pueden ser aprendidas de forma muy eficiente; ver, por ejemplo, para una introducción a clasificadores extendidos Friedman et al. (1997).

  Un algoritmo extendido más complejo es el denominado HCS (*hill-climbing search*), que consiste en:

  1 Se inicia con un *clasificador ingenuo* y se evalúa.

  2 Se considera cada enlace legal, segun las condiciones particulares, y se evalúa el nuevo clasificador.

  3 Si no hay mejora se elimina el enlace y se vuelve a 2.

  La complejidad de estos algoritmos sin restricciones es $O(n^3)$, resultando bastante ineficientes en sistemas de elevada dimensionalidad. Entre las numerosas estrategias para mejorar su eficiencia, el algoritmo SP (*Super-Parent*) de Keogh and Pazzani (1999), es particularmente interesante, y consiste en dividir los nodos en dos grupos: denominados *superpadres* y *huérfanos*. De esta manera, la búsqueda se divide en dos etapas, con lo que se consigue que el algoritmo SP tenga una complejidad $O(N^2)$ y una calidad similar a la del algoritmo HCS.

- Clasificadores Ingenuos Jerárquicos. Los *Hierarchical Naive Bayes* (HBN) descritos en Zhang (2002) y Langseth and Nielsen (2002), son una alternativa a los clasificadores extendidos descritos anteriormente, en los cuales se introducen variables latentes para relajar las hipótesis de independencia entre atributos, y así evitar el doble recuento por redundancia (*double-counting*) en la clasificación, ver Fig. 3.9.

  El algoritmo principal para el aprendizaje de HBN, se describe en Zhang (2002) y tiene las siguientes características:

  - Realiza una búsqueda voraz, sobre una estructura de árbol.
  - Las variables latentes son introducidas considerando la dependencia condicional entre los atributos dada la clase, seleccionando la pareja de atributos con mayor información mútua; para n atributos hay $(n^2 - n)/2$ posibilidades.

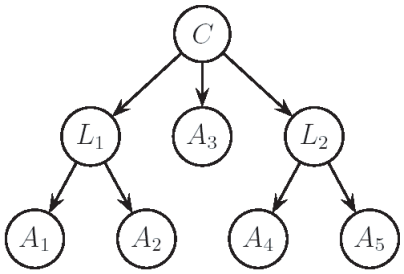

Figura 3.9: Ejemplo de HBN, tomado de Langseth and Nielsen (2002), donde C es el clasificador, A_j son los atributos y L_i son las variables latentes.

- Se determina la cardinalidad de la variable latente seleccionada, colapsando estados y evaluando la ganancia; se evalúa con un MDL condicionado.

- El proceso se repite hasta que no se aceptan más candidatos.

Existen también otros algoritmos que imponen ciertas restricciones estructurales derivadas, por ejemplo, del conocimiento previo, como pueden ser la presencia/ausencia de determinados enlaces o el orden de las variables (de Campos and Castellano, 2007), o considerando clases equivalentes (no solo en términos de independencia condicional sino en términos de clasificación) para reducir el espacio de búsqueda (Acid et al., 2005). También existe una profusa literatura sobre mezclas o comités de expertos, que combinan los resultados de distintos clasificadores (ver, por ejemplo, Langseth and Nielsen, 2006; Webb et al., 2005).

Por otra parte, si se emplea una medida de calidad adaptada específicamente al problema de interés, aunque se pierde capacidad descriptiva del conjunto de datos, se puede mejorar la calidad clasificadora del modelo. En Bradley (1997), se demuestra que AUC (*Area Under ROC Curve*) es una medida de calidad apropiada y sencilla para medir la calidad de un clasificador; se pueden encontrar ejemplos de su utilidad en Grossman and Domingos (2004), o en el *Full Bayesian network Classifier* (FBC) de Su and Zhang (2006), entre otros.

Para terminar, en Larrañaga et al. (2005), se puede encontrar un compendio de artículos ofreciendo una visión actual y completa del problema del aprendizaje de clasificadores, y en Larrañaga et al. (2003) se pueden descargar algunos artículos recientes *on-line*.

3.7. Sobreajuste

Después de describir las redes Bayesianas y la forma en la que pueden construirse a partir de los datos, conviene hacer mención al problema del sobreajuste que es crítico para el uso correcto de esta metodología en problemas prácticos, sobre todo cuando se tienen muchas variables y se dispone de pocos datos (que es una situación común en Meteorología). Incluso aunque la mayoría de los métodos de aprendizaje automático incluyen términos de penalización para evitar este problema, es necesario adoptar una serie de medidas de sentido común para tratar de adaptar la dimensionalidad del problema de aprendizaje que se plantee al tamaño del a muestra disponible. En general, hay tres formas de reducir el número de parámetros para evitar el sobreajuste:

1. **Eliminando relaciones de dependencia poco relevantes**. Esto se consigue limitando el número total de enlaces y el número máximo de padres por nodo en los métodos de aprendizaje estructural, estableciendo únicamente las dependencias más relevantes; así se obtiene una FPC en la que no figuran todos y cada uno de los estados posibles del sistema.

2. **Reduciendo dimensiones**. Las variables que definen el estado de la atmósfera, presentan un elevado grado de dependencia que puede ser utilizada para representar dicho estado por un número menor de variables independientes. El análisis de Componentes Principales (CPs), también conocido en las Ciencias Atmosféricas como análisis de Funciones Ortogonales Empíricas (EOF), es una técnica estándar para eliminar la información redundante con la mínima pérdida de variabilidad.

3. **Reduciendo el número de estados**. También se puede reducir el número de parámetros en las RB Multinomiales, restringiendo el número de estados de cada variable. En nuestro caso, es necesario reducir al máximo el número de parámetros, por lo que nos centraremos en el caso de predicciones probabilísticas de variables binarias; de hecho esto resta poca utilidad a los modelos ya que actualmente en Meteorología es muy común formular las predicciones en estos términos. Para ello, basta con transformar las variables en binarias, en base a un cierto umbral crítico de su rango (por ejemplo $precip > 10\,mm$, $viento < 50\,km/h$, etc.). Estos recursos se utilizan con mucha frecuencia para simplificar la predicción y la verificación y desde luego también serán empleados en este trabajo. Lo mismo se puede decir para la variable estado de la atmósfera, ver Sec. 5.3.

CAPÍTULO 4

Medidas de Verificación de Predicciones Probabilísticas

Existen numerosas medidas y métodos de verificación diferentes, que se utilizan en función del tipo de variable a predecir y del método de predicción empleado. En este capítulo nos centraremos en las predicciones de tipo *probabilístico* en las cuales el predictando se estima mediante una distribución de probabilidad sobre un rango de valores o categorías del mismo. La forma específica de esta distribución vendrá dada por las características de la variable analizada, que puede ser continua, discreta, o mixta; por ejemplo, la temperatura sigue aproximadamente una distribución normal y, por tanto, para especificar su distribución de probabilidad prevista bastará con estimar la media y la varianza; en otros casos, como el de la precipitación, la situación es más complicada pues la naturaleza de la variable es mixta con carácter discreto (lluvia/no lluvia) y continuo (cantidad de lluvia). En estos casos, se suele discretizar la variable en todo su rango para considerar una predicción categórica (ver, por ejemplo, Fig. 4.1). También es común considerar eventos binarios asociados a la superación o no de un cierto umbral (ver, por ejemplo, Fig. 4.2). Estos eventos binarios también surgen de forma natural cuando se estudia la ocurrencia de los diferentes meteoros, puesto que al no ser fácilmente cuantificables, directamente suelen ser observados como eventos binarios (niebla, tormenta, granizo, nieve, etc.).

En este capítulo se analizan las diferentes medidas de validación relacionadas con la predicción probabilística de variables categóricas, especialmente las referidas a eventos binarios, algunas de las cuales se utilizarán asiduamente a lo largo de la Tesis.

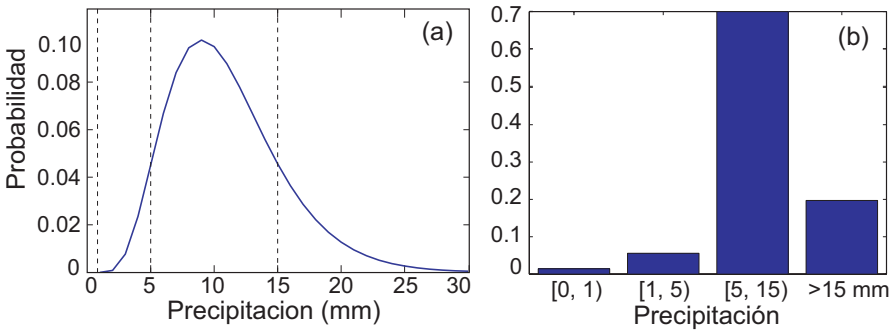

Figura 4.1: Predicción probabilística de la cantidad de lluvia, considerada como variable (a) continua y (b) discretizada.

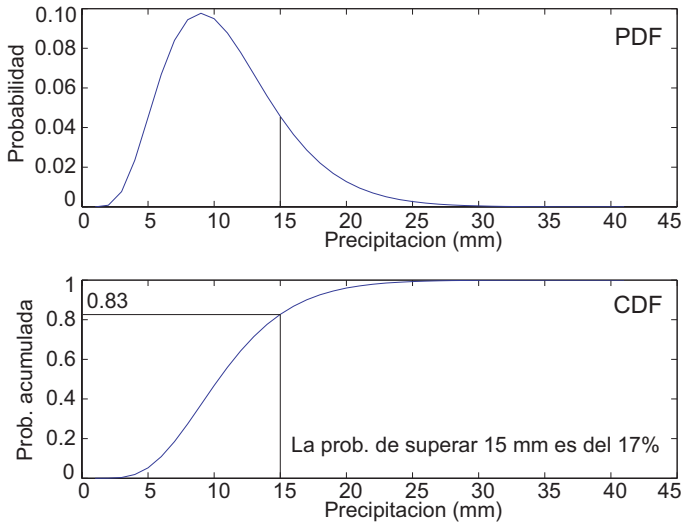

Figura 4.2: Funciones de densidad y distribución de la variable cantidad de lluvia. La probabilidad de que la precipitación supere un umbral (e.g. 15 mm) se obtiene fácilmente a partir de la función de distribución.

A modo de referencia sobre la que relativizar las medidas de calidad, se suelen utilizar sistemas de predicción probabilística elementales, como la *persistencia*, donde la predicción de la variable para el instante t viene dada por el valor observado en $(t-1)$, o la *climatología* (que puede ser estacional, mensual, diaria, etc.), donde la distribución para el instante t viene dada por un estadístico de los valores ocurridos en un cierto subconjunto del registro histórico. De esta manera, cualquier sistema de predicción puede ser verificado respecto a estos sistemas de referencia.

La organización del capítulo es la siguiente. En primer lugar, en la Sec. 4.1, se definen los diferentes aspectos a tener en cuenta a la hora de evaluar la calidad de una predicción, y en la Sec. 4.2, se describen algunas de las medidas más uti-

lizadas actualmente. En Jolliffe and Stephenson (2003) y en Gutiérrez et al. (2004b) se pueden encontrar descripciones completas de diferentes métodos de verificación utilizados actualmente en Meteorología y minería de datos.

4.1. Aspectos de la Calidad de una Predicción

En general, para evaluar la calidad de una predicción no basta con medir el error cometido, ya que existen aspectos, como por ejemplo la fiabilidad o la resolución, que también deben ser tenidos en cuenta, especialmente en los sistemas de predicción probabilística. Por este motivo, no hay una única medida capaz de proporcionar una información completa de la calidad del sistema, de forma que existen diversos índices de validación que se utilizan para describir diferentes atributos de la relación entre observaciones y predicciones. Algunos de estos atributos caracterizan a un único sistema, mientras que otros son comparativos e indican la diferencia entre sistemas de predicción distintos (o entre un sistema y otro sistema de referencia como la climatología o la persistencia).

Murphy and Winkler (1987) establece que toda la información de una validación está contenida en la distribución conjunta $P(o, \hat{o})$ de predicciones o y observaciones $\hat{o}$. En el caso particular de variables discretas (con estados o categorías C_0, ..., C_d), esta información se puede representar mediante una tabla de contingencia:

		Observación			
		C_0	C_1	...	C_d
Predicción	C_0	P_{00}	P_{01}	...	P_{0d}
	C_1	P_{10}	P_{11}	...	P_{1d}
	...		...		
	C_d	P_{d0}	P_{d1}	...	P_{dd}

que permite obtener distintos índices de validación en base a la información condicional o marginal asociada a la misma: $P(o|\hat{o})$, $P(\hat{o}|o)$, $P(\hat{o})$, $P(o)$; a pesar de las relaciones obvias que existen entre estas probabilidades, su interpretación y utilidad meteorológica es muy distinta y, por ello, cada índice aporta información interesante desde un punto de vista diferente.

Una vez que se tiene la distribución conjunta formada por las predicciones y las observaciones $P(o, \hat{o})$, se pueden examinar sus características de varias formas: (más detalles en Wilson, 2001):

- Se pueden analizar globalmente las correspondencias entre pares de observación-predicción; es decir, se analiza la distribución conjunta total. El *Brier Score* pertenece a este grupo.

- Se puede condicionar el análisis a valores concretos de la predicción. La *fiabilidad*, la *resolución* y la *definición* pertenecen a este grupo.

$$P(o, \hat{o}) = P(o|\hat{o})P(\hat{o}). \tag{4.1}$$

- Se puede condicionar el análisis a valores concretos de la observación. La *discriminación* (área ROC, etc.) es de este tipo.

$$P(\hat{o}, o) = P(\hat{o}|o)P(o). \tag{4.2}$$

Estos tres tipos de análisis de la distribución conjunta permiten examinar diferentes aspectos de la calidad de las predicciones, llamados *atributos de la predicción*. Murphy (1993), identifica una serie de atributos importantes para una predicción sea determinista o probabilística:

- **Sesgo o desviación sistemática** (*bias*): Referido a la concordancia entre la predicción media y la observación media; un sesgo positivo indica una sobreestimación del valor a predecir, mientras un sesgo negativo indica una subestimación (por ejemplo, se predice menos cantidad de lluvia de la que realmente ocurre).

- **Asociación**: Indica el grado de relación lineal entre observación y predicción. La covarianza y la correlación, son medidas de asociación.

- **Precisión**(*accuracy*): Relativa a la concordancia entre el valor previsto y el observado realmente, promediada sobre una muestra de parejas individuales de predicciones y observaciones. Medidas de precisión son, por ejemplo, el error absoluto medio, el error cuadrático medio y el Brier Score (BS).

- **Habilidad**(*skill*) o precisión relativa: Es la precisión de un sistema respecto a la obtenida con otro de referencia (por ejemplo, climatología o persistencia). Se definen de modo que un valor positivo (negativo) indica que el sistema es más (menos) preciso que el de referencia. El Brier Skill Score (BSS) es el más conocido.

- **Fiabilidad**(*reliability*) o *bias* condicional: Para que un sistema sea fiable, la probabilidad prevista y la observada deben coincidir lo máximo posible en todo el rango de probabilidad; cuando no es así se habla de bias condicionado. La fiabilidad se refiere a un determinado evento y se representa gráficamente dibujando la curva de probabilidad prevista frente a probabilidad observada

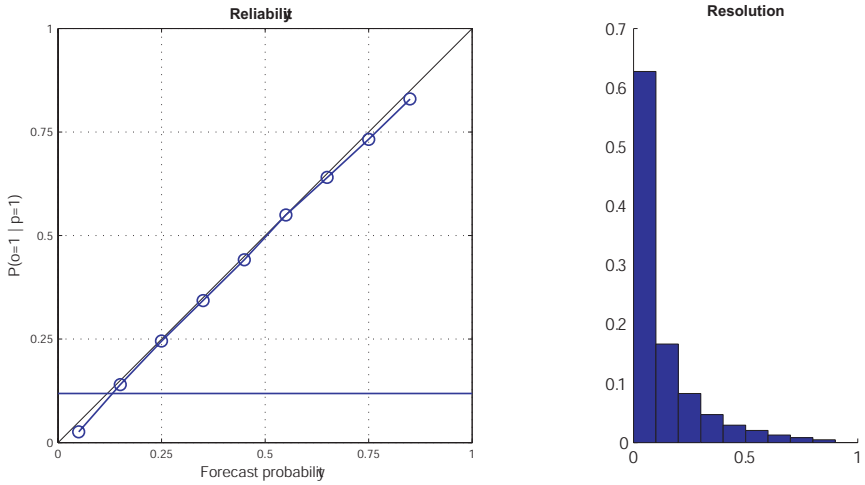

Figura 4.3: El gráfico de la izquierda muestra la fiabilidad para los diferentes umbrales de probabilidad prevista, marcados con circulos. En la derecha aparece el correspondiente a la resolución. Ambos han sido realizados para las 42 estaciones de la mitad oriental de la Cuenca Norte que aparecen en la tabla 5.1, para el evento $prec > 10mm/24h$, previsto con un método de análogos.

(climatológica) para intervalos de 5 ó 10 % (ver parte izquierda de la figura 4.3). Cuanto más cerca esté de la diagonal más fiable será el sistema de predicción probabilística, mientras que las desviaciones respecto a la diagonal indican que, para la probabilidad prevista, el sistema sobreestima o subestima según se encuentre por debajo o por encima de la diagonal. Es muy importante estudiar la fiabilidad de los sistemas para diferentes eventos, ya que los eventos raros casi nunca se predicen con probabilidades altas, manifestando una fuerte pérdida de fiabilidad que debe ser corregida. Existen diversas técnicas para realizar estas correcciones, entre las que se encuentran las de *inflado* que básicamente consisten en hacer un cambio de escala para aumentar la varianza estimada hasta su valor real. Existen versiones más sofisticadas, como la del *expanded downscaling*, donde la corrección se realiza mediante correlación canónica (Burguer, 1996). Estas técnicas paramétricas de inflado tienen como contrapartida que el error del estimador corregido siempre es mayor que el del estimador original (ver von Storch, 1999, para más detalles).

Cuando se analiza la fiabilidad de distintas predicciones (por ejemplo, distintos observatorios), utilizar el promedio es un concepto relativamente engañoso, ya que las fiabilidades individuales generalmente son peores, como ocurre cuando se promedian errores. La fiabilidad media únicamente indica si el sistema

subestima o sobreestima en promedio. En la Fig. 4.3 se muestran las fiabilidades obtenidas al aplicar un método de predicción local a 42 estaciones de la mitad oriental de la Cuenca Norte que aparecen en la tabla 5.1. Las dos figuras están asociadas al evento $precip > 10mm/24h$, y las predicciones han sido realizadas con el método de análogos.

- **Resolución**: Es la distribución de las probabilidades predichas. La resolución se representa mediante un histograma que representa las frecuencias de las probabilidades previstas. Es deseable que el histograma tenga forma de 'U' de manera que prediga pocas veces valores ambiguos en torno a 0.5. (ver parte derecha de la figura 4.3).

4.2. Medidas Estándar de Validación

Si se tiene un conjunto de observaciones $o_1, o_2, ..., o_n$ sobre las que se realizan las predicciones $\hat{o}_1, \hat{o}_2, ..., \hat{o}_n$, los índices de validación más habituales son los de precisión y habilidad, ya que proporcionan una medida numérica de la calidad de la predicción realizada. De todos ellos únicamente describiremos el *Brier Score*(BS), remitiendo al lector a Jolliffe and Stephenson (2003) y Gutiérrez et al. (2004b), para más detalles acerca de estas y otras medidas.

- **Brier Score (BS)**: El BS fue introducido por Brier (1950), y es una medida clásica de calidad en términos de precisión de una predicción probabilística para una variable discreta. En realidad esta medida corresponde al error cuadrático medido en unidades de probabilidad para la ocurrencia de las distintas categorías de la variable.

$$BS =< (p_i - o_i)^2 >_i, \tag{4.3}$$

donde $p_i = P(\hat{o}_i = 1)$, y o_i es 1 si ocurre el evento y 0 en caso contrario. El valor de BS es nulo para una predicción perfecta.

Es muy usual utilizar la descomposición del BS como suma de tres componentes; para realizar la descomposición se utiliza el siguiente argumento (véase Murphy, 1973):

$$(p_i - o_i)^2 = f_i(p_i - 1)^2 + (1 - f_i)p_i^2, \tag{4.4}$$

donde $f_i = p(o_i = 1|p_i)$ es el número de casos observados de entre los previstos con probabilidad p_i. Por tanto,

$$BS = BS_f - BS_r + I, \tag{4.5}$$

donde BS_f es la componente de fiabilidad cuya expresión es:

$$BS_f =< (p_i - f_i)^2 > \tag{4.6}$$

BS_r es la componente de resolución cuya expresión es:

$$BS_r =< (f_i - p_c)^2 > \tag{4.7}$$

p_c es la probabilidad climática del evento. Por último I es la componente de incertidumbre cuya expresión es:

$$I = \frac{1}{M}\sum_{i=1}^{M}(p_c - o_i)^2 \ = P_c(1 - P_c). \tag{4.8}$$

I es el *Brier Score* de un sistema de predicción climatológico.

El BS da una idea del error cometido en una predicción, pero no dice nada acerca de la dificultad de dicha predicción. La dificultad de una predicción está muy asociada con su probabilidad climatológica, cuanto más infrecuente sea, es más difícil de predecir; por ejemplo, es mucho mas difícil predecir lluvia en Almería que en Santander. Por otra parte, si el evento es muy raro, como por ejemplo la lluvia en Almería, un sistema sin utilidad que siempre diga que no va a llover (como la climatología o la persistencia) tendría un *Brier Score* casi nulo por lo que sería casi imposible de superar por un sistema de mayor utilidad. Es decir, cuando están muy descompensadas las frecuencias de ocurrencia y de no ocurrencia, el *Brier Score* sólo es representativo del evento más frecuente.

- **Brier Skill Score**: El BS da una idea muy particular del comportamiento de un sistema de predicción ya que depende fuertemente de la frecuencia del evento, de la localidad donde se aplica y de la serie utilizada. Por tanto, si se quiere conocer correctamente la calidad del sistema, se tendrían que especificar todos y cada uno de los valores del BS para cada localidad y para cada evento, resultando una matriz de datos diferente para cada período considerado. Para evitar este problema se puede obtener una medida de validación relativa, en lugar de absoluta, considerando un sistema de predicción de referencia (como la climatología o la persistencia) (Talagrand, 1997). A estas medidas relativas

se las denomina *índices de pericia* (Skill Scores, SS) y se calculan de la siguiente forma:

$$SS = \frac{P(o = 1|p) - P(o = 1|p_c)}{1 - P(o = 1|p_c)} \tag{4.9}$$

donde p es el sistema de predicción que se desea validar, p_c es el sistema de referencia, y $P(o = 1|p)$ es la probabilidad de que se produzca el evento dado que el sistema lo predice con probabilidad p.

Teniendo en cuenta que el *Brier Score* es equivalente a $1 - P(o = 1|p)$, se puede obtener un índice de pericia a partir de (4.9). Este índice se denomina *Brier Skill Score* (BSS):

$$BSS = 1 - \frac{BS_p}{BS_c}, \tag{4.10}$$

donde BS_p es el BS del sistema de predicción y BS_c el BS de la climatología (u otro sistema de referencia). La interpretación es sencilla: Si $BSS > 0$, entonces el sistema de predicción es mejor que el climatológico; en cambio, si $BSS < 0$ el sistema no mejora la climatología. En el caso de que $BSS = 1$ se tendría una predicción perfecta.

Como se ve en la Fig.4.4, el *Brier Skill Score* disminuye casi linealmente con la probabilidad climatológica del evento, de manera que a medida que el error es inferior al error climatológico el BSS aumenta, obteniendo sus valores máximos cuando $p_c = 0.5$.

4.2.1. Medidas de Validación de Predicciones Categóricas

Cuando la predicción probabilística se refiere a la ocurrencia o no-ocurrencia de las categorías de una variable discreta, se habla de predicciones categóricas. Estas se verifican mediante tablas de contingencia, que se construyen combinando todas las posibilidades entre categoría prevista y categoría observada. Las predicciones categóricas son también una de las principales tareas de la minería de datos, concretamente en el caso particular de la clasificación, razón por la que ámbas áreas comparten algunas medidas de validación.

El caso más sencillo y utilizado corresponde al caso de variables binarias, ya que basta un único umbral de probabilidad u_p para determinar la ocurrencia del evento (cuando $P(prec > 0mm/24h) > u_p$) y la no ocurrencia en caso contrario. En el caso binario, las tablas de contingencia están determinadas por cuatro parámetros:

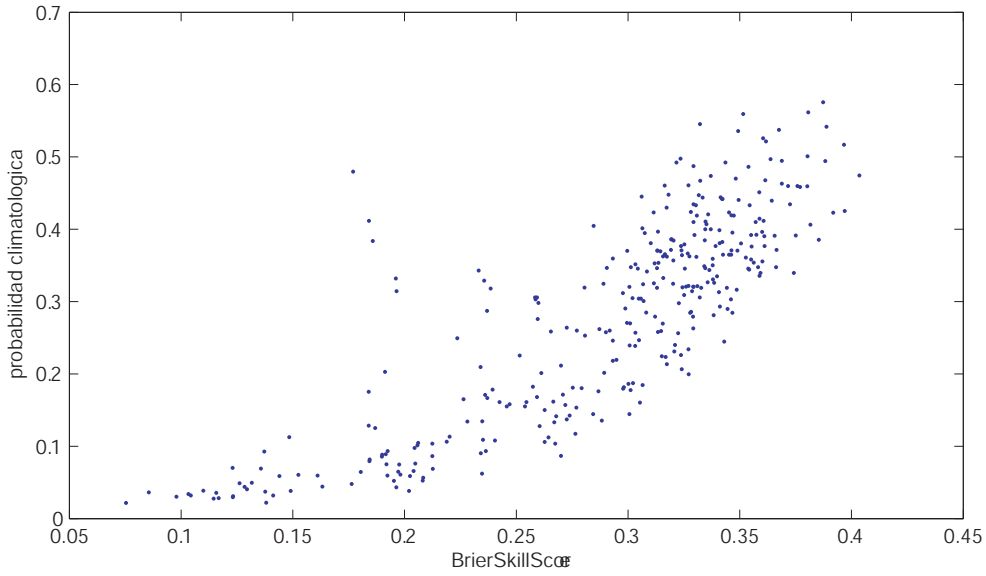

Figura 4.4: *Brier Skill Scores* para las 42 estaciones de la mitad oriental de la Cuenca Norte que aparecen en la tabla 5.1, para el evento $prec > 10mm/24h$, previsto con un método de análogos.

		Observado	
		Si	No
Previsto	Si	α	β
	No	γ	δ

donde evidentemente, α, β, γ y δ dependen de u_p.

Cuando se toman decisiones, se pueden cometer dos tipos de errores: *falso positivo* y *falso negativo*, también conocidos como errores de Tipo I y Tipo II respectivamente. Generalmente se considera que el *falso negativo* es un error más grave que el *falso positivo*, aunque la importancia relativa de estos dos errores es diferente para diferentes usuarios. Por ello, hay que tener en cuenta los requerimientos específicos del usuario a quien va dirigida la predicción; en este caso, las medidas globales como el BSS pueden inducir a equivocaciones a la hora de elegir el mejor sistema.

A continuación se describen algunas de las medidas más empleadas. Para indicar su rango de valores y el valor óptimo, se utilizará la notación: $[a, \mathbf{b}]$ donde se indica el rango y el valor óptimo (en negrita).

- **Precisión** $[0, \mathbf{1}]$:

$$ACC = \frac{\alpha + \delta}{\alpha + \beta + \gamma + \delta} \tag{4.11}$$

Realmente mide la proporción de aciertos, y puede ser maximizada prediciendo siempre la categoría más común. En regiones donde el evento es muy raro, se hace prácticamente la unidad debido al gran número de aciertos negativos (no aporta información para eventos raros). La precisión de un sistema que nunca predice el evento es:

$$ACC_{no} = \frac{\delta}{\gamma + \delta} = 1 - p_c \tag{4.12}$$

Como ocurría con el BSS, este índice queda fuertemente sesgado en favor de las regiones donde el evento es más raro, por lo que el empleo de éste índice no es aconsejable para ello.

- **Critical Success Index ó Threat Score** $[0, \mathbf{1}]$:

$$CSI = \frac{\alpha}{\alpha + \beta + \gamma} \tag{4.13}$$

Es similar a la precisión, pero se han quitando los aciertos negativos. Aunque es más equilibrado que el ACC, sus mayores valores se dan en aquellas localidades donde el evento es más común. El CSI de un sistema que nunca predice el evento es nulo, mientras que el CSI de un sistema que siempre lo predice es:

$$CSI_{si} = \frac{\alpha}{\alpha + \beta} = p_c \tag{4.14}$$

Al igual que la precisión, es muy inestable para eventos raros.

- **Hit Rate(HIR)** $[0, \mathbf{1}]$:

$$HIR(u_p) = P(p \geq u_p | o = 1) = \frac{\alpha}{\alpha + \gamma} \tag{4.15}$$

Es una medida de precisión, que mide la probabilidad de detección; es decir, de predecir eventos que realmente se han observado.

- **False Alarm Rate(FAR)** $[\mathbf{0}, 1]$:

$$FAR(u_p) = P(p \geq u_p | o = 0) = \frac{\beta}{\beta + \delta} \tag{4.16}$$

También es una medida de precisión, que mide la proporción de fallos cuando se predice positivamente un evento. Obsérvese que el concepto habitual de falsa alarma es un suceso que se ha predicho, pero luego no ha ocurrido; sin embargo, en este caso, el evento que condiciona es la no ocurrencia del evento. Con este cambio se consigue que cualquier usuario tendrá un criterio de

comparación referido siempre a la incertidumbre climatológica del evento, que es un invariante estadístico independiente del tipo de sistema de predicción empleado.

- **Curvas ROC (Relative Operating Characteristics)**

  También denominadas *Receiver Operating Characteristics*, se crearon para evaluar radares en la II guerra mundial, posteriormente en diagnóstico médico y en la actualidad son muy populares en minería de datos. Las Curvas ROC permiten validar de forma uniforme los sistemas de predicción categóricos basados en umbrales de probabilidad. Se obtienen representando los valores FAR frente a los HIR para diferentes umbrales de probabilidad u_p. Sobre ellas se pueden comparar dos sistemas diferentes, bien globalmente o bien para determinados niveles relativos de falsas alarmas y omisiones que más convengan al usuario, concretando un umbral de predicción u_p.

  Cuando se opera con un umbral inferior a la probabilidad climatológica, el sistema comete más falsas alarmas que aciertos, ya que la pendiente de la curva ROC se hace menor que la diagonal, por ello en todos los sistemas de predicción existe un valor de $u_p = p_c$ (donde la pendiente de la curva ROC es la unidad) para el cual las prestaciones del sistema son máximas, ver Gutiérrez et al. (2004b) para más detalles. En la Fig. 4.5 se muestran dos ejemplos de curvas ROC para dos eventos de probabilidad muy diferente; en la misma se muestran los valores de u_p utilizados para construir la curva, así como el valor mínimo de dicho umbral por debajo del cual no existe suficiente información estadística en el conjunto de entrenamiento (en la figura lo denominamos límite operativo), a partir del cual la curva es interpolada en línea recta hasta el valor teórico [1,1].

  Cuando se comparan dos sistemas de predicción diferentes para un mismo evento mediante las curvas ROC, se observa que para cada umbral de probabilidad, el mejor sistema será aquel cuya tasa HIR sea más alta para un mismo nivel de FAR. El criterio más empleado es el del área bajo la curva ROC (*Area Under Curve*, AUC), de forma que es mejor aquel sistema que mayor área cubre bajo su curva ROC. En realidad la medida de calidad AUC, como medida global, evalúa la capacidad del sistema para ordenar correctamente las probabilidades previstas.

  Por ejemplo, la Fig. 4.6 muestra los valores de AUC para eventos de diferente probabilidad climatológica, constatando que no es tan sensible a la probabili-

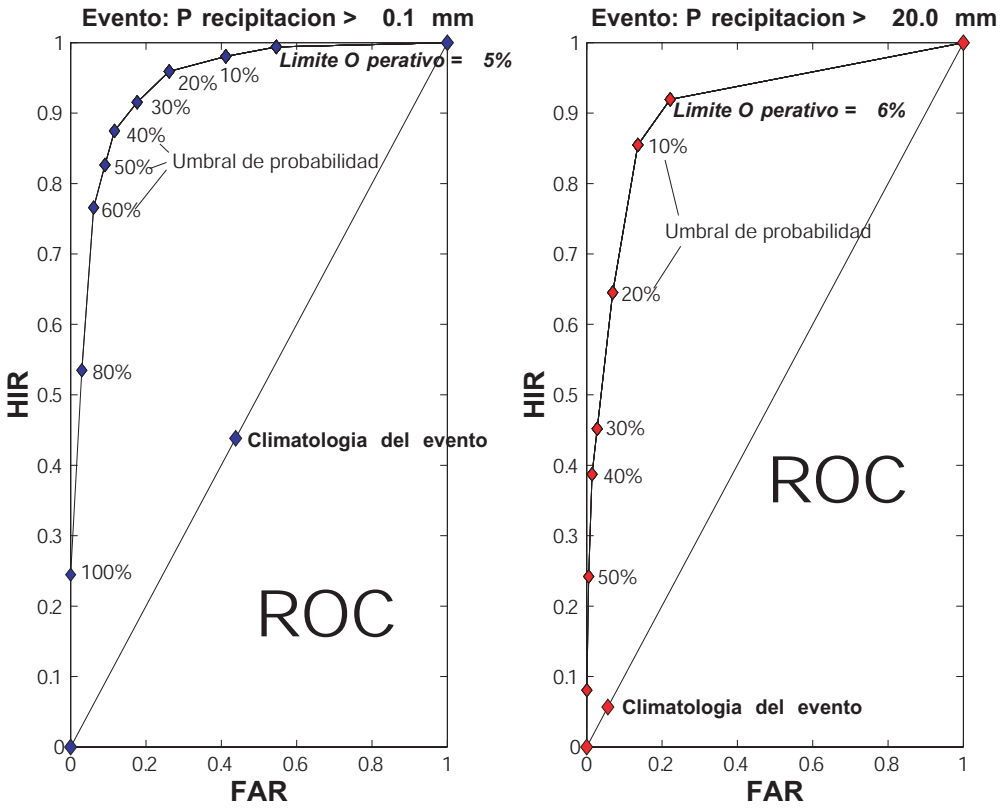

Figura 4.5: Dos ejemplos de curvas ROC para diferentes eventos.

dad climatológica del evento como lo es el índice BSS, excepto para los casos más extremos, que corresponden a la parte inferior de la figura, por este motivo será la medida que más se utilizará a lo largo de esta Tesis.

AUC también es una medida muy utilizada en la evaluación de clasificadores en minería de datos, donde también es preferida a las medidas de precisión por motivos similares a los aquí expuestos, ver Bradley (1997) para más detalles.

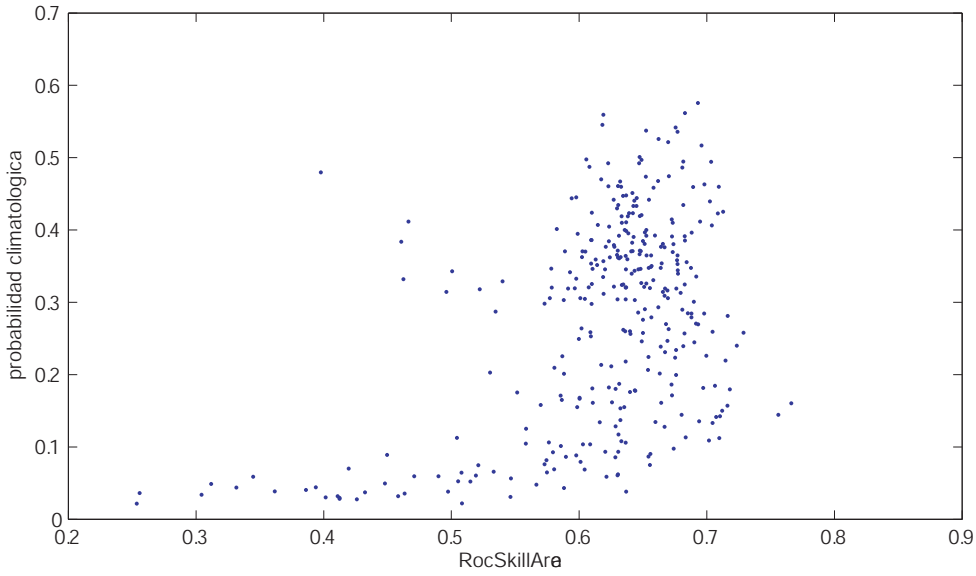

Figura 4.6: Valores de AUC frente a la frecuencia climatológica para diferentes eventos de precipitación, previstos con un método de análogos en varias localidades de la Cuenca Norte. $RocSkillArea = 2 * AUC - 1$

Parte II

Aportaciones

CAPÍTULO 5

Dominios de Estudio y Datos Utilizados

Antes de entrar de lleno en las aportaciones de esta Tesis, y dado que muchas de ellas se apoyan en aplicaciones sobre ejemplos prácticos, se comienza esta parte dedicando un capítulo a la descripción de las áreas de estudio y de los datos que se utilizarán en dichos ejemplos. En primer lugar se describen los registros históricos utilizados de observaciones en superficie, principalmente de precipitación, que es la variable de interés en esta Tesis (Sec. 5.1); a continuación, se describen los datos procedentes de modelos numéricos de simulación de la atmósfera, en particular datos de reanálisis (Sec. 5.3). En la mayor parte de la Tesis, los primeros datos se considerarán predictandos (desconocidos), mientras que los últimos serán predictores (conocidos, habitualmente proporcionados por algún modelo numérico operativo; ver Sec. 2.2.3).

5.1. Áreas de Estudio y Observaciones

El área de estudio considerada en esta Tesis es la Península Ibérica y Baleares, mostrada en la Fig. 5.1(a). Sobre esta región se consideran los datos diarios de precipitación en una red de 100 observatorios (indicados por las cruces de la Fig. 5.1b) suministrada por la Agencia Estatal de Meteorología, AEMET. Los datos cubren el periodo $1957-2002$ y son representativos de la climatología global de dicha región. Estas 100 estaciones han sido elegidas aplicando criterios de homogeneidad (todas las estaciones son homogéneas, según el test de Alexandersson, 1986, con un nivel de confianza del 95 %), y poseen un mínimo del 90 % de observaciones válidas

(es decir, tienen menos de un 10 % de lagunas en el período total de la serie). Debido al coste computacional de los algoritmos de aprendizaje automático de redes Bayesianas para series incompletas, se ha optado por utilizar series completas. Para ello, se han filtrado los días (realizaciones) para los cuales falta la observación de alguna de las estaciones, resultando un período común de 7654 días no consecutivos comprendidos entre el 1/1/1961 y el 30/6/2000 (ver Fig. 5.2).

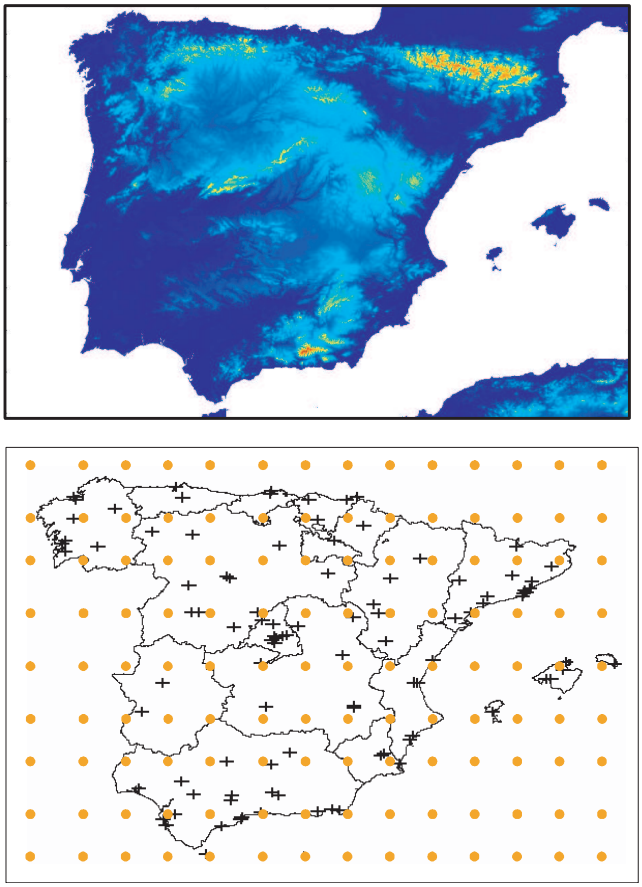

Figura 5.1: Arriba. Orografía del dominio de baja resolución cubriendo la Península Ibérica. Abajo. Estaciones (cruces) y puntos de rejilla del reanálisis (puntos) en la región.

En algunos experimentos se considera la sub-región de alta resolución correspondiente a la Cornisa Cantábrica, mostrada en la Fig. 5.3, sobre la que se dispone de observaciones de precipitación para 42 observatorios (ver Tabla 5.1). Estas estaciones fueron elegidas aplicando filtros de homogeneidad y cantidad de datos a las 16435 realizaciones (días) del periodo 1957-2002, sobre las 625 estaciones disponibles de la red pluviométrica de la AEMET en la zona (ver Fig. 5.3b), resultando finalmente 42 estaciones homogéneas, con 4816 días sin lagunas que corresponden al periodo

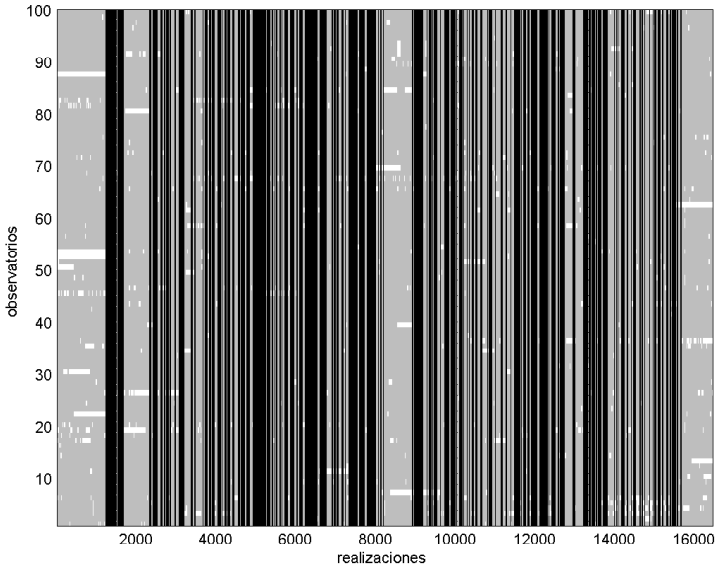

Figura 5.2: Inventario de observaciones para las 100 estaciones en el período 1/9/1957 al 30/8/2002. En gris se muestran las fechas con observaciones disponibles (para los distintos observatorios), mientras que en blanco se muestran los datos faltantes. Las líneas negras indican las fechas para las cuales se dipone de datos en todos los observatorios.

comprendido entre el 1/1/1974 y el 31/12/1991 (ver Fig. 5.4). En la Tabla 5.1, se ofrecen algunos datos geográficos y climatológicos básicos sobre las 42 estaciones seleccionadas.

En esta región también se analiza el caso de conjuntos de datos inhomogéneos (Capítulo 10). Para ello, se consideran 12 estaciones "históricas" con series de datos más largas (en particular las 12 primeras estaciones de la lista, mostradas en rojo en la Fig. 5.3b), que poseen 9860 datos completos, y 30 estaciones "nuevas" (mostradas en azul en la Fig. 5.3b) que disponen de 4816 datos completos (aunque se considerarán números inferiores, hasta un año, para recrear escenarios con distintos niveles de inhomogeneidad).

5.2. Predictandos: Observaciones en Superficie

Como se ha visto en la Sec. 2.3.2, existen algunos fenómenos meteorológicos que por sus especiales características no pueden ser cuantificados de forma estándar y, por ello, su observación se limita a cifrar su ocurrencia, o no. Tal es el caso de ciertos meteoros como las *tormentas*, *nieblas*, *heladas*, *escarcha*, etc. Asimismo, cualquier fenómeno que se pueda cuantificar de forma continua, como por ejemplo

Algunos datos geográficos y climatológicos básicos sobre las 42 estaciones seleccionadas en el dominio de alta resolución. Las primeras 12 corresponden a las estaciones con series largas, que aparecen en color rojo en la figura 5.3, y el resto corresponden a las estaciones con series cortas, que aparecen en azul en la misma figura. La tabla muestra la probabilidad de que la variable precipitación supere ciertos umbrales u en cada estación.

Localidad	alt	lon	lat	$u = 0$	$u = 10$
FUENTERRABIA 'AEROPUERTO'	8	-1.79	43.35	0.47	0.165
SAN SEBASTIAN 'IGUELDO'	252	-2.03	43.30	0.5	0.15
ELDUAYEN	256	-2.00	43.13	0.38	0.18
AMURRIO 'INSTITUTO'	219	-3.00	43.05	0.32	0.1
BILBAO 'AEROPUERTO'	39	-2.90	43.29	0.46	0.11
EL MERCADILLO DE LIERGANES	90	-3.74	43.34	0.45	0.14
VILLACARRIEDO	212	-3.80	43.23	0.39	0.14
ROZADIO	210	-4.38	43.22	0.43	0.09
RANON 'AEROP. DE ASTURIAS'	127	-6.03	43.56	0.46	0.1
RIOSECO DE SOBRESCOBIO	390	-5.45	43.22	0.37	0.13
OVIEDO 'EL CRISTO'	336	-5.87	43.35	0.44	0.08
ZARDAIN	410	-6.55	43.39	0.46	0.11
SAN SEBASTIAN 'ATEGORRIETA'	8	-1.95	43.32	0.49	0.16
LASARTE MICHELIN	85	-2.02	43.27	0.44	0.16
EIBAR BANCO DE PRUEBAS	121	-2.47	43.18	0.46	0.14
ECHEVARRIA	100	-2.47	43.25	0.47	0.14
ABADIANO 'MENDIOLA'	160	-2.61	43.14	0.44	0.14
DURANGO VIVERO	280	-2.64	43.15	0.37	0.13
GURIEZO 'G.C.'	77	-3.32	43.35	0.36	0.14
CARRANZA	220	-3.35	43.22	0.36	0.12
COTERILLO DE AMPUERO	25	-3.43	43.33	0.43	0.13
MIRONES	200	-3.70	43.29	0.49	0.18
SANTANDER 'CENTRO'	64	-3.81	43.46	0.46	0.11
ESCOBEDO DE VILLAFUFRE	180	-3.91	43.26	0.37	0.13
TORRELAVEGA 'SNIACE'	70	-4.03	43.36	0.41	0.11
CELIS	100	-4.43	43.28	0.43	0.1
CAMIJANES	115	-4.48	43.32	0.42	0.11
ESPINAMA	877	-4.77	43.13	0.33	0.09
TAMA	270	-4.59	43.18	0.3	0.05
AMIEVA 'RESTAÑO'	700	-5.02	43.22	0.42	0.16
CANGAS DE ONIS	80	-5.12	43.35	0.36	0.09
RIBADESELLA 'FARO'	117	-5.08	43.47	0.4	0.1
GIJON	3	-5.64	43.53	0.44	0.08
BEZANES	654	-5.29	43.15	0.31	0.13
SOTO DE RIBERA	127	-5.87	43.31	0.38	0.09
MERES DE SIERO	180	-5.73	43.38	0.43	0.09
GRADO	60	-6.06	43.38	0.4	0.09
GENESTAZA	680	-6.35	43.20	0.38	0.13
SOTO DE LA BARCA	213	-6.39	43.29	0.34	0.08
PRESA DE LA BARCA	214	-6.30	43.31	0.41	0.09
SOTO DE LOS INFANTES	127	-6.27	43.35	0.41	0.11
CUEVAS DE ALTAMIRA	161	-4.11	43.37	0.42	0.11

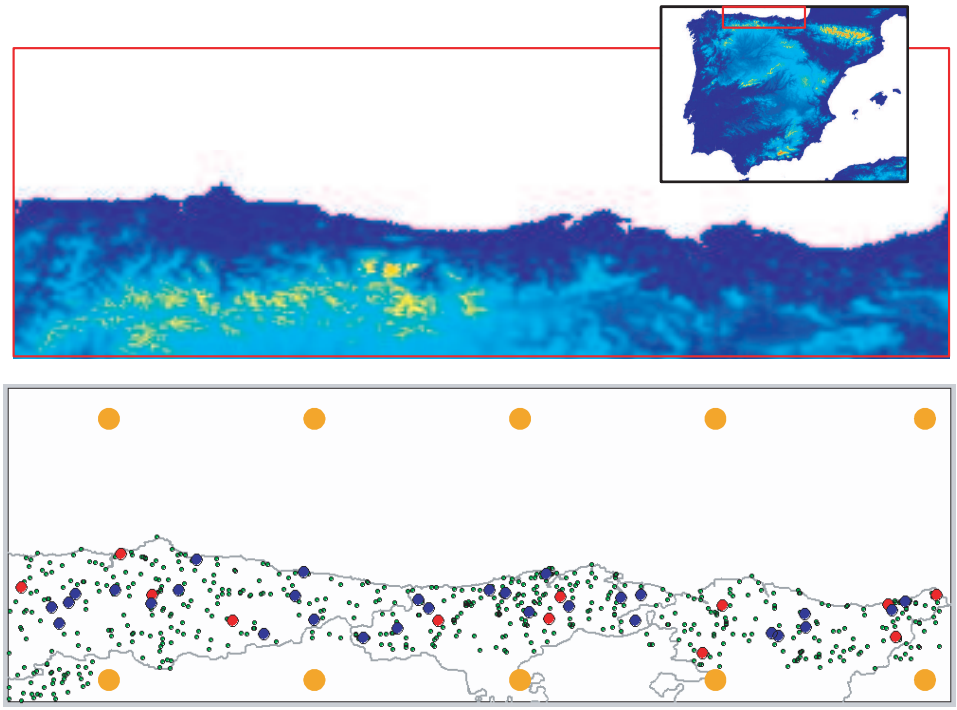

Figura 5.3: (a) Orografía del dominio de alta resolución cubriendo la Cornisa Cantábrica de la Península Ibérica. (b) Estaciones de la red pluviométrica en la región (puntos verdes) y conjunto de 42 estaciones elegidas para el estudio (puntos rojos y azules); también se muestran los puntos de rejilla del reanálisis (puntos amarillos) en la región.

la temperatura y la precipitación, también es susceptible de ser tratado como una variable binaria sin más que considerar un umbral apropiado u, dando lugar a dos posibles valores: $precip \leq u$ y $precip > u$.

En esta Tesis se analiza el problema de la predicción local de precipitación. Por tanto, se han considerado dos umbrales representativos de la variabilidad de la precipitación en condiciones no extremas (ver Tabla. 5.1): $prec > 0mm/24h$ y $prec > 10mm/24h$; por tanto, cada evento es una variable binaria con los estados $\{0{=}no\ ocurrencia,\ 1{=}ocurrencia\}$. Esta cuantización binaria de las observaciones, además de reducir notablemente el número de parámetros en los modelos, resulta muy apropiada para la predicción probabilística de numerosos eventos. Como se vio en el Cap. 4, estas variables son de uso común en predicción probabilística, existiendo numerosas medidas de validación para este tipo de predicciones.

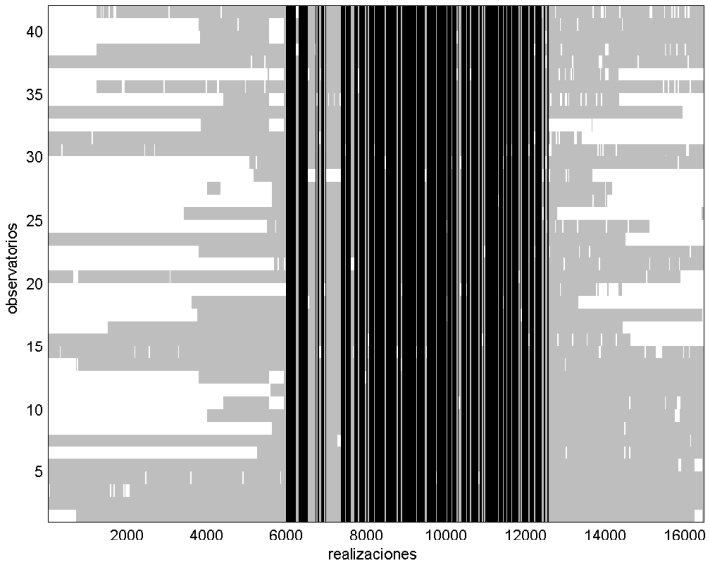

Figura 5.4: Inventario de observaciones para las 100 estaciones en el período 1/9/1957 al 30/8/2002. En gris se muestran las fechas con observaciones disponibles (para los distintos observatorios), mientras que en blanco se muestran los datos faltantes. Las líneas negras indican las fechas para las cuales se dipone de datos en todos los observatorios. El orden es el mismo en el que aparecen en la Tabla 5.1.

5.3. Predictores: Patrones Atmosféricos

Además de las observaciones de precipitación en superficie, que son los predictandos a predecir en esta Tesis, también se consideran como predictores los patrones del estado de la atmósfera, dados por modelos numéricos que simulan la dinámica atmosférica sobre una rejilla 3D que cubre el globo (con resoluciones típicas como las mostradas en la rejilla de la Fig. 5.1). En particular los proyectos de reanálisis, como ERA-40 (ver 2.4), proporcionan una serie histórica de estados de la atmósfera (asimilados a la rejilla de un modelo numérico), caracterizados por el valor del geopotencial (Z), temperatura (T), dirección U y V del viento y humedad (H), principalmente, sobre la rejilla del modelo; estas variables son responsable de la dinámica de la atmósfera y, por tanto, caracterizan su estado, de forma genérica, para una fecha dada.

Dado que en esta Tesis se consideran exclusivamente fenómenos cuya escala de agregación temporal es diaria (de 07 a 07 horas), los patrones atmosféricos deben representar este rango horario. Por tanto, se consideran los valores del reanálisis entre las 06 y 30 horas (que son las más cercanas a las dos horas de medida, ya que

se dispone de datos de reanálisis de cada 6 horas). De acuerdo con esto, dependiendo de la escala espacial del problema, se pueden plantear diferentes patrones para caracterizar el estado de la atmósfera sobre la región de interés:

- *Patrón 1:* Rejilla de 2.5° × 2.5° de longitud y latitud mostrada en la Figura 5.5(a) y que corresponde a la zona del Atlántico Norte, que influencia en gran medida el clima en la Península. En este caso, los patrones se obtienen combinado los campos T, H, Z, U y V a las 12h en los niveles $1000mb$, $925mb$, $850mb$, $700mb$, $500mb$ y $300mb$:

$$\mathbf{c_{12}} = (T_{12}^{1000}, \ldots, T_{12}^{300}, H_{12}^{1000}, \ldots, H_{12}^{300}, \ldots, V_{12}^{1000}, \ldots, V_{12}^{300}), \tag{5.1}$$

  donde C_i^j denota el campo de la variable C a la hora i en el nivel j. Evidentemente este patrón no está orientado a predictandos de alta resolución espacial.

- *Patrón 2:* Rejilla de 1.0° × 1.0° mostrada en la Figura 5.5(b), cubriendo la zona de estudio. En este caso, se consideran las mismas variables y niveles de presión (altura), pero incluyendo una componente temporal (se toman los campos a las 06h y las 30h). Esta componente temporal compensa la reducción de escala de la rejilla teniendo en cuenta efectos de borde y de contorno que podrían alcanzar la zona de la rejilla durante el período de interés.

$$\mathbf{c} = (\mathbf{c_{00}}, \mathbf{c_{30}}). \tag{5.2}$$

  Este patrón está orientado a predictandos de baja resolución como, por ejemplo, las estaciones de la red principal mostradas en la Fig. 5.1b.

- *Patrón 3:* Rejilla de 1.0° × 1.0° mostrada en la Figura 5.5(c). En este caso se considera un patrón atmosférico concreto para cada una de las doce cuencas hidrográficas españolas. Para ello se combinan los campos anteriores en un dominio temporal de mayor resolución: 06, 12, 18, 24, y 30 UTC. En este caso se cubre el período de predicción con toda la información disponible:

$$\mathbf{c} = (\mathbf{c_{06}}, \mathbf{c_{12}}, \mathbf{c_{18}}, \mathbf{c_{24}}, \mathbf{c_{30}}). \tag{5.3}$$

  Este es el patrón específicamente orientado a predictandos de alta resolución como, por ejemplo, las estaciones de la red secundaria mostradas en la Fig. 5.3b.

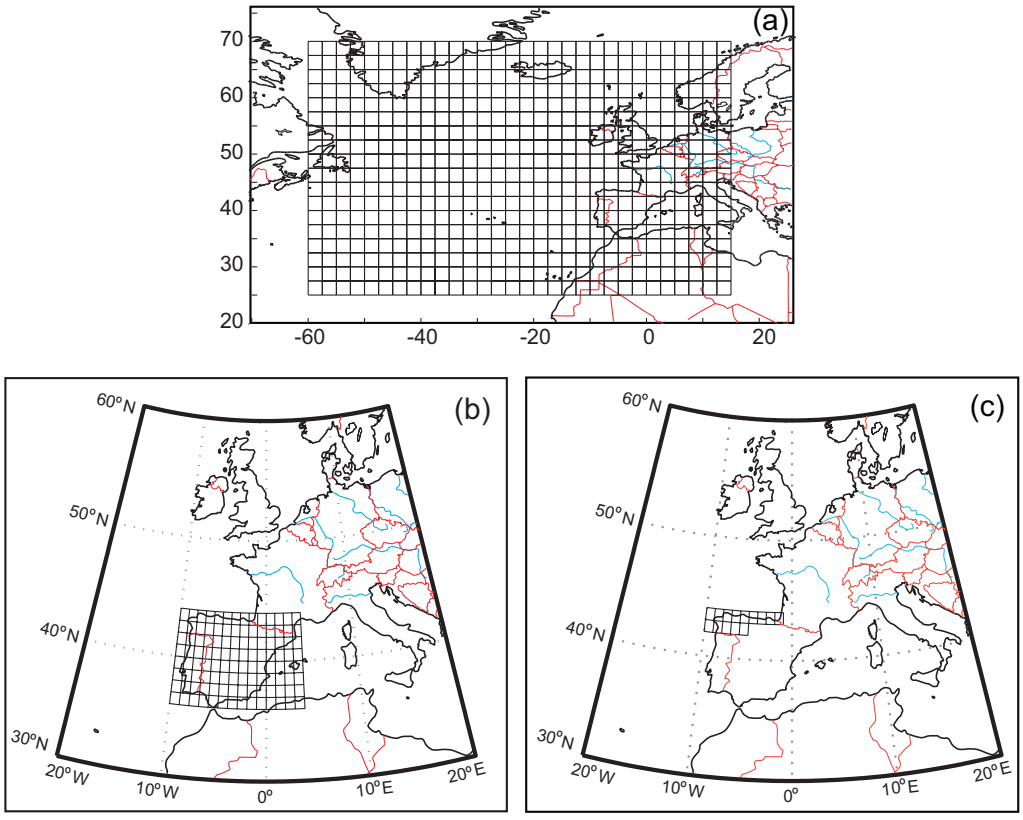

Figura 5.5: Distintas áreas que cubren la península Ibérica: (a) Rejilla de escala (macro-β) de $2.5° \times 2.5°$ de latitud y longitud correspondiente al Atlántico Norte; (b) rejilla peninsular (meso-α) $1° \times 1°$ de latitud y longitud; (c) rejilla meso-β $1° \times 1°$ para la cuenca Norte de la Pensínsula Ibérica (cada una de las doce cuencas tendría su propia rejilla en este caso).

Teniendo en cuenta que esta Tesis está orientada a la predicción local, sólo se utilizarán los patrones correspondientes a los Patrones 2 y 3, ya que el primer patrón resultó ser muy ineficiente para este problema (ver Gutiérrez et al., 2004a, para una evaluación detallada de los patrones para el problema de la predicción local). También se han realizado experimentos con patrones específicos en los que se han seleccionado únicamente variables relevantes para precipitación, pero en esta Tesis se ha optado por mantener un patrón atmosférico genérico.

Concretamente, el Patrón 2 se utilizará para la región y observaciones mostradas en la Fig. 5.1, mientras que el Patrón 3 se utilizará para las observaciones de alta resolución mostradas en la Fig. 5.3.

Discretización en Tipos de Tiempo

Los patrones presentados en la sección anterior tienen un elevado número de dimensiones, lo que, en ocasiones, dificulta el tratamiento de los mismos. Por ejemplo, el estado de la atmósfera generado por el Patrón 2 sería un vector con 9×15 (grid) $\times 6$ (niveles) $\times 5$ (variables) $= 4050$ dimensiones. Sin embargo, estos patrones presentan un elevado grado de dependencia espacial y temporal, que puede ser utilizada para representar dicho estado por un número menor de variables independientes, puesto que su dimensión efectiva es mucho menor. Para ello pueden aplicarse técnicas de reducción de dimensionalidad, como la popular técnica de Componentes Principales (ver, por ejemplo, Gutiérrez et al., 2004b), o técnicas de cuantización o agrupación, que proporcionan un conjunto representativo reducido de patrones (tipos de tiempo, o *weather types*, en inglés) que proporcionan una única variable discreta C que caracteriza la configuración atmosférica de forma cuantizada, pudiendo tomar sólo un número predefinido de valores $1, \ldots, k$, cada uno de las cuales corresponde a una clase o tipo de tiempo (ver, por ejemplo, Gutiérrez et al., 2004b, Cap. 2 para más detalles).

Esta discretización se suele realizar mediante técnicas de agrupamiento, que son utilizadas en numerosas disciplinas para dividir o segmentar un conjunto de datos en subconjuntos homogéneos siguiendo algún criterio de similitud (ver Anderberg, 1973, para una descripción detallada de estos métodos). El método más común es el algoritmo de las k-medias (ver, por ejemplo, Gutiérrez et al., 2004b) en el cual, dado un grupo de vectores o patrones reales d-dimensionales $X = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$, y un número predefinido de grupos k, se calculan un conjunto de prototipos d-dimensionales, $\{\mathbf{v_1}, \ldots, \mathbf{v_k}\}$, o centroides, cada uno de ellos caracterizando a un grupo de datos $C_i \subset X$ formado por los vectores para los cuales $\mathbf{v_i}$ es el prototipo

más cercano, según alguna métrica elegida (normalmente la distancia euclídea). Esta tarea es realizada siguiendo un procedimiento iterativo, el cual comienza con un conjunto inicial de centroides $\mathbf{v_1^0}, \ldots, \mathbf{v_m^0}$, elegidos aleatoriamente (ver Peña et al., 1999, para una descripción y comparación de diferentes procedimientos de inicialización). El objetivo del algoritmo es minimizar globalmente la distancia intra-grupos

$$\sum_{i=1,\ldots,m} \sum_{\mathbf{x_j} \in C_i} \|\mathbf{x_j} - \mathbf{v_i}\|^2. \tag{5.4}$$

Como una búsqueda exhaustiva del mínimo es prohibitiva, se calcula un mínimo local mediante un ajuste iterativo de los centroides de los grupos, re-asignando cada patrón al centroide más cercano. En la iteración $(r+1)$-ésima, cada uno de los vectores $\mathbf{x}_j$ es asignado al grupo i-ésimo, donde $i = argmin_c \parallel \mathbf{x}_j - \mathbf{v_i^c} \parallel$, y los prototipos son actualizados por medio de los correspondientes patrones:

$$\mathbf{v}_i^{r+1} = \sum_{\mathbf{x_j} \in C_i} x_j / \#C_i,$$

donde $\#C_i$ denota el número de elementos en C_i. Bajo ciertas condiciones, el proceso iterativo anterior converge después de R iteraciones, y los centros finales $\mathbf{v}_i^R$ son los prototipos (centroides). Cada uno de los centroides $\mathbf{v}_i$ representa un grupo C_i formado por los patrones más cercanos a $\mathbf{v}_i^R$ que a cualquier otro centroide; este grupo caracteriza un "tipo de tiempo" dado. El algoritmo de k-medias consiste en los siguientes pasos:

1. Seleccionar el número de grupos deseados k.
2. Inicializar los centroides de los grupos (p.e. aleatoriamente).
3. Repetir:
   a) Asignar cada vector (patrón atmosférico) al grupo más cercano.
   b) Re-calcular los centroides de cada grupo, para que sean la media de los patrones asignados a ese grupo.

Por lo tanto se puede describir el estado de la atmósfera mediante un escalar discreto C^k obtenido clasificando los patrones en k clases, utilizando el algoritmo *k-medias*, descrito anteriormente. En las figuras 5.6(a) y (b) se muestran dos ejemplos de la agrupación del estado de la atmósfera en 100 clases, correspondientes a los Patrones 1 y 3, mostrados en la Figura 5.5(a) y (d), respectivamente. En la primera de estas figuras se puede observar claramente la circulación de verano e invierno

asociada a ondas largas diferenciadas en la parte izquierda y derecha de la figura. Este ejemplo ilustra la resolución espacial de estas variables de estado C^k.

Hay que tener en cuenta que el número de estados k debería de estar asociado con la variabilidad espacial que es capaz de resolver el modelo numérico para el patrón considerado; este aspecto se analizará en el Capítulo 7 comparando, para un mismo dominio, los resultados obtenidos con dos patrones distintos, C^{10} y C^{100}, con 10 y 100 estados, respectivamente.

En algunos experimentos aislados se utilizarán los datos del patrón atmosférico en los distintos puntos de rejilla, pero en la mayor parte de la Tesis se considerará el estado de la atmósfera mediante las variables discretas C^{10} o C^{100}, caracterizando un número discreto de tipos de tiempo.

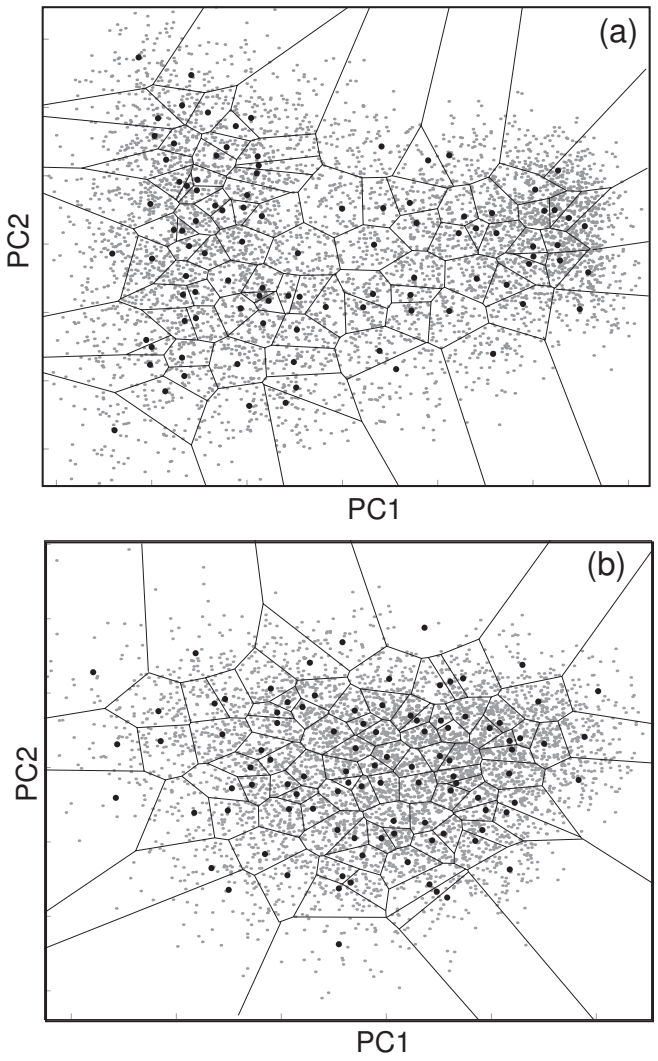

Figura 5.6: Agrupamiento del reanálisis ERA-40 con el algoritmo de *k-medias* considerando 100 grupos para (a): Patrón 1 y (b): Patrón 3. El grafo muestra los patrones diarios y los centroides proyectados en el espacio de las dos primeras componentes principales. Las líneas de separación entre diferentes grupos corresponden al *diagrama de Voronoi* asociado a los centroides.

CAPÍTULO 6

Aplicaciones de las Redes Bayesianas en Meteorología

En los últimos años las redes Bayesianas (RBs) se han aplicado con éxito en distintos campos, como la genómica (Friedman et al., 2000), economía (Jangmin et al., 2004), ingeniería (Nadi et al., 1991), ecología (McCann et al., 2006), diagnóstico médico (Nikovski, 2000), etc. (ver Mittal and Kassim, 2007; Pourret et al., 2008, para una descripción más detallada de aplicaciones en distintos ámbitos). Es de destacar, en el campo de la hidrología, la Tesis presentada por Mediero Orduña (2007), en la que se describe un sistema de predicción donde se combinan un modelo para simular el comportamiento hidrológico de una cuenca fluvial, y una red Bayesiana (aprendida con datos sintéticos) para capturar la variabilidad e incertidumbre de los procesos hidrológicos desde un punto de vista probabilístico. Sin embargo, las aplicaciones de estos modelos en Meteorología son relativamente escasas, con numerosos casos de estudio similares al mostrado en el Ejemplo 3.2. Por ejemplo, Abramson et al. (1996) describen el sistema *Hailfinder* para la predicción de granizo, Kennett et al. (2001) describen un modelo para la predicción de brisa marina, y Hruschka and Ebecken (2005) presentan un sistema para la predicción de niebla.

Por tanto, un primer objetivo en el desarrollo esta Tesis fue realizar un análisis de las aplicaciones potenciales de las RBs en este campo, explorando distintos problemas meteorológicos de interés en los que esta técnica de minería de datos podría ofrecer ventajas frente a otras metodologías. En este capítulo no se realizará ningún

desarrollo teórico en el área de las RBs, sino que se utilizarán los algoritmos estándar de aprendizaje y propagación de evidencia descritos en el capítulo 3 como una aproximación inicial a la Meteorología. Parte de los resultados que aparecen en este capítulo se encuentran en una publicación que describe, desde un punto de vista genérico, las posibles aplicaciones de las RBs en Meterorología (Cano et al., 2004). Como herramienta de trabajo, se ha utilizado la *toolbox* de RBs para *Matlab* (Murphy, 2001), cuyos detalles se encuentran en `http://bnt.sourceforge.net/`.

Este capítulo está dividido en tres secciones con sendas aplicaciones básicas de las RBs, planteadas sobre problemas de Meteorología que se adecúan al tratamiento probabilístico. En cada una de ellas, primero se introduce el problema meteorológico, después se establecen las limitaciones de la metodología al uso y finalmente se proponen las soluciones (siempre ilustradas con un sencillo ejemplo de aplicación), con el único fin de poner de manifiesto las ventajas de las RBs cuando se aplican en dichos problemas. En la Sec. 6.1, se desarrolla un modelo de RB para caracterizar la *distribución espacial de precipitación*, constatando la dificultad existente para modelizar las relaciones espaciales de dependencia entre distintos observatorios (o localidades), y demostrando que las RBs proporcionan un modelo capaz de caracterizar conjuntamente las relaciones a priori y condicionales más relevantes en un único modelo conjunto. En la Sec. 6.2, se analizan los *generadores de tiempo* para la simulación estocástica de fenómenos meteorológicos, mostrando que las RBs son capaces de considerar el comportamiento colectivo, obteniendo realizaciones espacialmente consistentes. En la Sec. 6.3, se trata la *predicción local de precipitación* utilizando la aproximación más inmediata, que consiste en relacionar las predicciones de los modelos numéricos de circulación atmosférica (ver Sec. 2.2) con las observaciones. Los resultados obtenidos en este caso son prometedores, pero manifiestamente mejorables, dando pie a a las aportaciones que se muestran posteriormente en la Tesis.

6.1. Estructura Espacial de la Precipitación

La estructura espacial de la precipitación es un caso paradigmático de las complejas relaciones espaciales que presentan, en general, las variables meteorológicas (sobre todo aquellas más influenciadas por la orografía). Por ello, es un buen ejemplo que nos va a servir para ilustrar algunas de las principales ventajas de la aplicación de las RBs en meteorología. En este caso el área y los datos utilizados corresponden al dominio de baja resolución sobre la Península Ibérica descrito en la Sec. 5.1.

La compleja estructura espacial de la precipitación se ilustra en la Fig. 6.1, donde

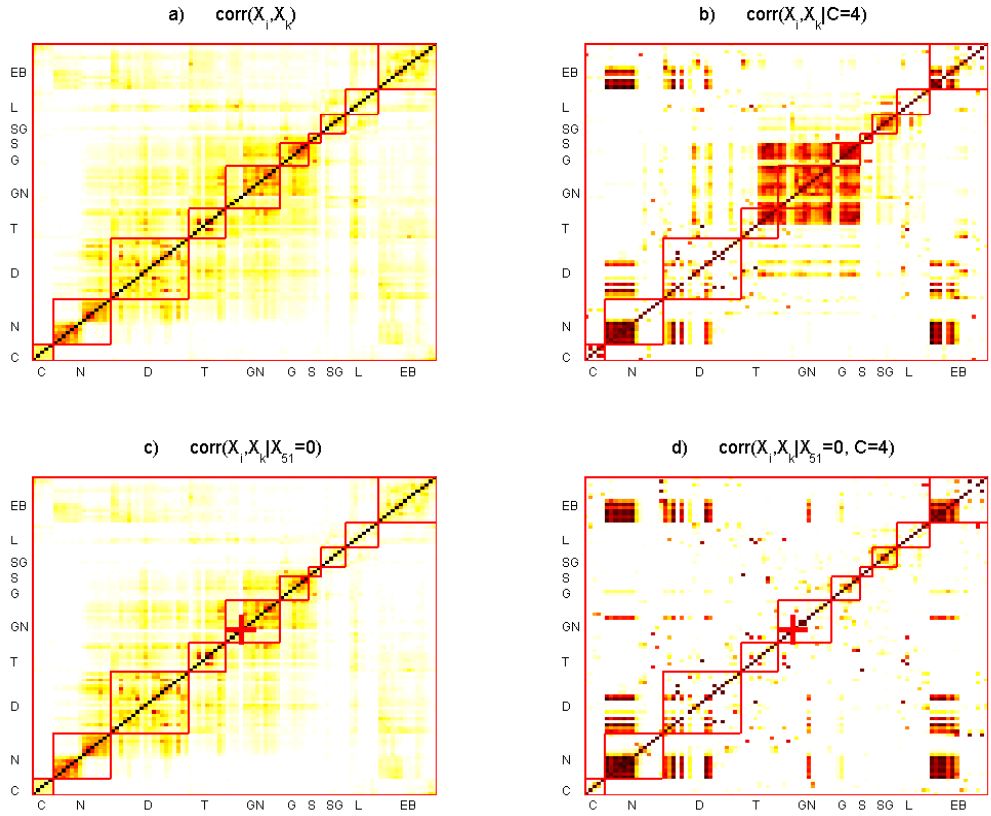

Figura 6.1: Correlación entre estaciones en el dominio de baja resolución. (a) Marginal. (b) Condicionada al caso *Estado de la atmósfera C=4*. (c) Condicionada a *Precipitación en Munera=0*. (d) Condicionada a *Precipitación en Munera=0* y *C=4*.

se muestra la correlación entre cada par de estaciones para cuatro planteamientos típicos de meteorología, como son: (a) global entre cada par de estaciones, (b) condicionada al estado de la atmósfera, para un tipo de tiempo concreto, (c) condicionada al estado de una estación determinada y (d) condicionada simultáneamente al estado de la atmósfera y una estación determinada. Además, para una mejor comprensión, se han separado con rectángulos rojos las diferentes cuencas hidrográficas de este dominio (sólo se muestran las 10 cuencas hidrográficas de la península), que han sido etiquetadas como C (Catalana), N (Norte), D (Duero), T (Tajo), GN (Guadiana), G (Guadalquivir), S (Sur), SG(Segura), L (Levante) y E (Ebro). En la Fig. 6.1 (a) se observa que, en general, las estaciones que corresponden a una misma zona hidrográfica presentan una dependencia más fuerte; incluso se pueden llegar a distinguir diferentes zonas dentro de una misma cuenca como, por ejemplo, la que se observa en la Cuenca Norte entre la parte derecha (Galicia) y la parte izquierda (Cantábrico). Además de esta relación a priori (o marginal) de las variables, existen otras depen-

dencias condicionales que hacen aún más complicada la visión global del problema y la comprensión de la dependencia espacial, como se puede ver en la Fig. 6.1(b), donde se muestra el aspecto que tiene la correlación entre cada pareja de estaciones cuando el cálculo se condiciona a un determinado estado de la atmósfera, mostrando cambios que pueden ser muy significativos; como ejemplo, se ha buscado un caso muy diferente respecto a la correlación a priori, que corresponde al estado $C = 4$ (un tipo de tiempo particular; ver Sec. 5.3). Análogamente, en la Fig. 6.1(c), se muestran las correlaciones condicionadas al evento *Precipitación en Munera=0mm/24h* (destacada con un + en rojo); a esta resolución el impacto de una estación sobre el resto es muy pequeño debido a la gran variabilidad espacial de la precipitación, posteriormente veremos que esto es muy diferente en el dominio de alta resolución. Finalmente, la Fig. 6.1(d) muestra las correlaciones condicionadas a los dos eventos anteriores *Precip. Munera=0mm/24h y C=4*; en este caso, las correlaciones vuelven a ser marcadamente diferentes respecto a los casos anteriores. Esta complejidad, referida a parejas de estaciones, es aún mayor en el caso de considerar grupos de estaciones (normalmente una cuenca entera), dejando patente la dificultad existente para representar y visualizar globalmente las relaciones de dependencia espacial de ciertas variables, como la precipitación acumulada diariamente. Como se verá a continuación, las RBs proporcionan un modelo capaz de caracterizar conjuntamente las relaciones a priori y condicionales más relevantes considerando un único modelo conjunto.

6.1.1. Ejemplo de Aplicación de Redes Bayesianas

Las redes Bayesianas permiten modelizar las relaciones espaciales de dependencia entre los distintos observatorios a partir del registro histórico de datos observados. Por ejemplo, la Fig. 6.2 muestra un grafo dirigido acíclico correspondiente a la precipitación discretizada de acuerdo con los umbrales establecidos en la sección 5.1, obtenido con un método de aprendizaje automático estándar (el algoritmo B), utilizando el registro histórico disponible, ver sección 3.5. Las estaciones (nodos) están conectadas mediante aristas que representan relaciones de dependencia; de modo que si exite una arista del nodo y^j al nodo y^k, se dice que y^j es un padre de y^k, o que y^k es un hijo de y^j. Por ejemplo, en la Figura 6.2, los nodos *Vitoria* y *Logroño* son ambos hijos de *Pamplona* y *Huesca* es padre de *Lérida*. Como se vió en la secc. 3.3.1, el criterio de d-separación permite saber si dos estaciones son o no condicionalmente dependientes dada otra sin más que comprobar los distintos caminos que unen las dos variables en cuestión en el grafo. Por otra parte, como se

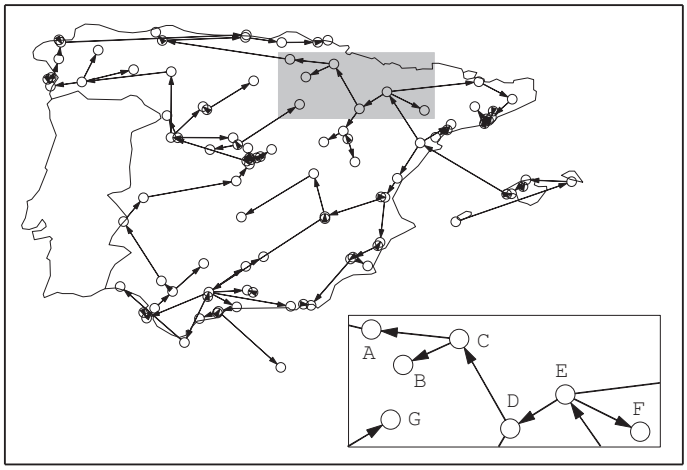

Figura 6.2: Grafo dirigido acíclico que codifica las relaciones de dependencia entre la precipitación entre las estaciones del dominio de baja resolución utilizado en este ejemplo. La zona ampliada muestra en detalle los enlaces en las estaciones: $A = Vitoria$, $B = Logroño$, $C = Pamplona$, $D = Zaragoza$, $E = Huesca$, $F = Lérida$ y $G = Soria$.

vió en la secc. 3.3.2, el grafo implica una factorización de la probabilidad conjunta (FPC) $P(X_1, X_2, ..., X_{100})$ mediante el producto de 100 funciones de probabilidad condicionada (CPDs), una para cada nodo, dados sus padres (por ejemplo, la Tabla 6.1 muestra una de estas probabilidades condicionadas, que han sido estimadas a partir del registro histórico de datos). El modelo factorizado resultante permite calcular la probabilidad de los distintos estados de precipitación en cada estación, y actualizar estas probabilidades cuando se dispone de evidencia. Por ejemplo, las probabilidades a priori de las variables se obtienen calculando $P(y_i = s), s = 0, 1$; $i = 1, \ldots, 100$, tal como muestra la Tabla 6.2(a) para algunas de ellas. Esta tabla

	Estado de Pontevedra			
$\Pi_{Pontevedra}$	0	1	2	3
$\{0, 0\}$	0.9377	0.0552	0.0047	0.0024
$\{0, 1\}$	0.5909	0.3727	0.0273	0.0091
$\{1, 0\}$	0.4026	0.4935	0.0909	0.0130
$\{1, 1\}$	0.1736	0.5069	0.2500	0.0694
$\{2, 0\}$	0.0909	0.4545	0.4545	0.0000
$\{2, 1\}$	0.1429	0.3333	0.3095	0.2143
$\{3, 0\}$	0.5000	0.1000	0.4000	0.0000
$\{3, 1\}$	0.0000	0.0357	0.3929	0.5714

Tabla 6.1: Probabilidad del nodo *Pontevedra*, condicionada a su conjunto de padres $\Pi = \{Vigo, Marín\}$.

(a) Probabilidad marginal (inicial), $P(y^k)$

Estado	Santander	Oviedo	Avilés	Lugo	Coruña	Santiago	Vigo
0	0.680	0.667	0.648	0.707	0.596	0.545	0.570
1	0.320	0.333	0.352	0.293	0.404	0.455	0.430

(b) $P(y_i|Santander = 1)$

Estado	Santander	Oviedo	Avilés	Lugo	Coruña	Santiago	Vigo
0	**0.000**	0.250	0.125	0.211	0.334	0.368	0.429
1	**1.000**	0.750	0.875	0.789	0.666	0.632	0.571

(c) $P(y^k|Santiago = 1)$

Estado	Santander	Oviedo	Avilés	Lugo	Coruña	Santiago	Vigo
0	0.556	0.502	0.482	0.466	0.142	**0.000**	0.052
1	0.444	0.498	0.518	0.534	0.858	**1.000**	0.948

(d) $P(y^k|Santander = 1, Santiago = 0)$

Estado	Santander	Oviedo	Avilés	Lugo	Coruña	Santiago	Vigo
0	**0.000**	0.365	0.210	0.361	0.657	**1.000**	0.894
1	**1.000**	0.635	0.790	0.639	0.343	**0.000**	0.106

Tabla 6.2: (a) Probabilidades marginales (iniciales) de los nodos; (b)-(d) probabilidades condicionadas a la evidencia: $e_1 = \{Santander = 1\}$, $e_2 = \{Santiago = 1\}$, y $e_3 = \{Santander = 1, Santiago = 0\}$. La evidencia se indica en negrita y la probabilidades que más varían se han subrayado.

muestra que los regímenes de precipitación en el área de estudio son similares, dado que todas ellas pertenecen a una misma cuenca hidrográfica. A continuación, a medida que se recibe cierta información o evidencia e relevante para el problema se pueden calcular las nuevas probabilidades $P(y_i|e)$ para actualizar el conocimiento. Por ejemplo, las Tablas 6.2(b)-(d) muestran el efecto producido por distintas evidencias. Comparando estas probabilidades con las probabilidades iniciales se pueden analizar los efectos de la información sobre las relaciones entre las distintas variables, mostrando que las RBs permiten modelar de forma eficiente, en un único modelo, las relaciones probabilísticas de dependencia espacial existente entre un gran número de estaciones pluviométricas. Estos modelos pueden contribuir en gran medida a la realización de estudios estadísticos que requieran una coherencia espacial. Posteriormente mostramos dos ejemplos que utilizan esta idea para desarrollar tanto generadores estocásticos de tiempo como modelos de predicción local, conservando la coherencia espacial.

6.2. Generadores de Tiempo (Weather Generators)

Los generadores de tiempo (en inglés, *weather generators*) se han utilizado en distintos ámbitos, principalmente en hidrología y agricultura, para simular series de datos meteorológicos a escala diaria (Wilby and Wilks, 1999; Wilks, 1999a). La simulación de una determinada variable se realiza a partir de su función de distribución o, en su defecto, a partir de algún conjunto de estadísticos representativos que permitan caracterizar ésta; por ejemplo, estos modelos permiten obtener simulaciones de la precipitación diaria en una localidad concreta a partir de la media y varianza anual, o mensual, y la probabilidad de ocurrencia de precipitación. Estos estadísticos pueden ser estimados a partir de registros históricos (los denominados valores normales climatológicos), y/o a partir de valores anómalos previstos (por ejemplo, la precipitación mensual prevista por una predicción estacional, o la anomalía de temperatura procedente de una proyección regional de cambio climático). Por tanto, los generadores de tiempo pueden considerarse técnicas de desagregación temporal que simulan series de tiempo diario respetando las propiedades climatológicas conocidas de la variable (distribución, probabilidades de ocurrencia, etc.).

La precipitación es una de las variables más importantes, con enorme influencia en agricultura, hidrología, etc., ya que otras variables como la evapotranspiración, la radiación solar y la temperatura dependen de ella. Por tanto, esta variable desempeña un papel destacado en los generadores de tiempo que, en su gran mayoría, están específicamente adaptados para trabajar con ella. Por ejemplo, Richardson

(1981) propuso un método que primero simula la precipitación, y en función de ella, la temperatura y la radiación solar. Una dificultad añadida para trabajar con la precipitación es su carácter mixto discreto/continuo, pues sólo el evento $prec = 0$ tiene probabilidad no nula. Por tanto, se puede distinguir entre una variable binaria *ocurrencia de precipitación* y una variable continua *cantidad de precipitación* (condicionada a $prec > 0$). La mayoría de los generadores de tiempo tratan estas dos variables de forma separada, desarrollando métodos apropiados para simular la ocurrencia del evento, por una parte, y su cantidad o intensidad por otra (ver Gutiérrez et al., 2004b, , Cap. 2, para más detalles).

En esta sección únicamente nos referiremos a eventos binarios (ocurrencia de precipitación).

6.2.1. Cadenas de Markov

Las cadenas de Markov (ver Fig. 6.3) fueron los primeros modelos aplicados en la simulación de series binarias de ocurrencia de precipitación. Según este modelo, la ocurrencia o no de precipitación para una localidad dada Y_i se simula en base a la probabilidad:

$$p(y_i^t | y_i^{t-1}), \tag{6.1}$$

donde el número de fechas "pasadas" a las que se condiciona constituye el *orden* del modelo. Un resultado interesante es que un sencillo modelo de primer orden (como el mostrado en la ecuación 6.1) permite simular correctamente la naturaleza persistente de la precipitación diaria, es decir, la autocorrelación de la serie (Gabriel and Neumann, 1962). Esto permite considerar generadores de tiempo muy eficientes para simular series aisladas de ocurrencia de precipitación, preservando la correlación temporal y las propiedades estadísticas del observatorio en estudio. Por ejemplo, la Fig. 6.4 ilustra la coincidencia de las frecuencias relativas observadas y simuladas de precipitación diaria, de longitud 10 años, en las distintas estaciones del dominio de alta resolución considerado en esta Tesis (la Cornisa Cantábrica, ver Fig. 5.3). Además, la Fig. 6.5 muestra que, a pesar de que se tiene información únicamente del día anterior, el modelo de primer orden también es capaz de simular adecuadamente las rachas de días secos y lluviosos.

En este tipo de modelos, la serie de cada observatorio se simula de forma independiente y, por tanto, cuando se aplican a un grupo de observatorios de una cierta región, se generan realizaciones sin coherencia espacial. Sin embargo, como ya se explicó en la Sec. 6.1, aunque la precipitación presenta una gran variabilidad espacial y temporal debido a la interacción atmósfera-suelo, al mismo tiempo existe una mar-

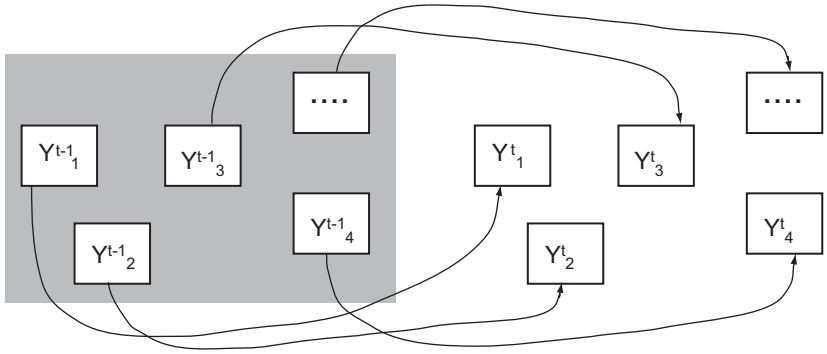

Figura 6.3: Esquema de un modelo de Markov de primer orden para un conjunto de localidades $Y_1, Y_2, \ldots$. La zona sombreada corresponde a valores "pasados".

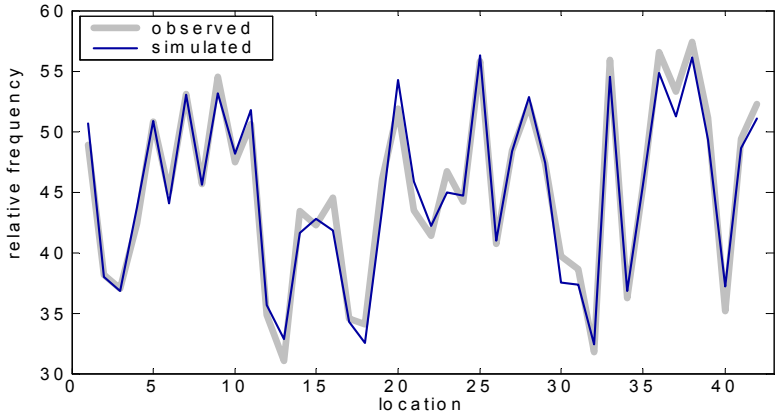

Figura 6.4: Ejemplo de aplicación del modelo de Markov de primer orden para el evento ocurrencia de precipitación sobre los 42 observatorios en la Cornisa Cantábrica.

cada estructura de dependencia espacial a escala local, como mestra el aspecto de la matriz de dependencias del conjunto de estaciones mostrada en la Fig. 6.6. Esta distribución espacial es relevante en numerosos y variados problemas relacionados con la meteorología para los que este tipo de métodos no son apropiados; por ejemplo, en los modelos hidrológicos el comportamiento colectivo de toda una cuenca fluvial es un aspecto crítico a la hora de evaluar los impactos potenciales de determinadas anomalías previstas (ver, por ejemplo, Wood, 2004; Bertacchi Uvo et al., 2001).

Para responder a este problema, recientemente han sido desarrollados algunos métodos que introducen la variabilidad espacial en el modelo. Por ejemplo, Wilks (1999b) realiza una simulación conjunta de series para un grupo de estaciones considerando una distribución normal conjunta. Por otra parte, se han propuesto modelos con variables ocultas que tienen en cuenta la estructura espacial del problema (como los modelos ocutos de Markov); estas técnicas parecen ser más apropiadas

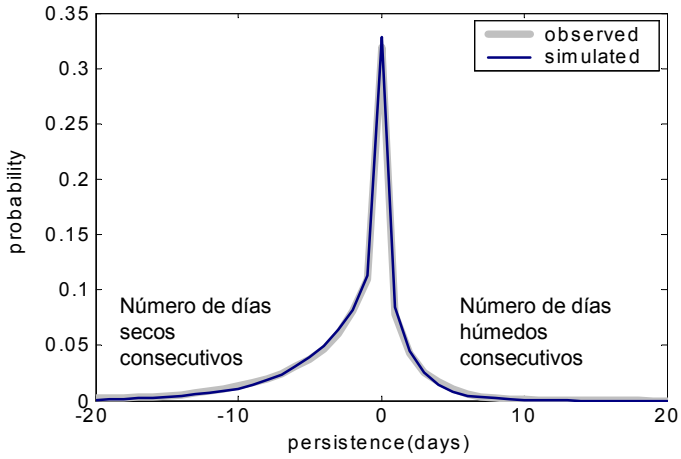

Figura 6.5: Ejemplo de aplicación del modelo de Markov de primer orden para la reproducción de rachas de lluvia y sequía.

para este problema.

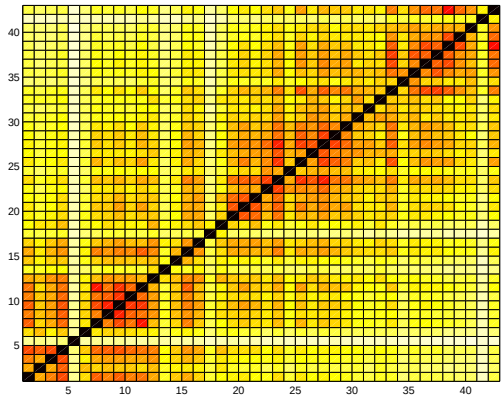

Figura 6.6: Matriz de dependencia espacial marginal entre los 42 observatorios de la Cornisa Cantábrica, donde se aprecian zonas claramente diferenciadas.

6.2.2. Modelos de Markov Ocultos (HMM)

Los Modelos de Markov Ocultos (en inglés *Hidden Markov Models*, HMM), son una variante de los modelos de Markov en la cual se introducen variables no observables que condicionan al conjunto de variables observadas (ver Fig. 6.7), y cuyos parámetros se infieren a partir de éstas mediante algoritmos iterativos (como el algoritmo de Maximización de la Esperanza, EM). En estos modelos la probabilidad de ocurrencia de precipitación en cada localidad Y_i se simula en base a:

$$p(y_i^t | y_i^{t-1}, H_t) = p(y_i^t, H_t). \tag{6.2}$$

Estos modelos son muy utilizados en reconocimiento de patrones, ya que obtienen conocimiento global del sistema codificado en las variables oculatas resultantes del proceso de aprendizaje, razón por la que su uso está cada vez más extendido, con numerosas aplicaciones en meteorología, especialmente en la simulación de la variabilidad espacio-temporal de la precipitación (Hughes et al., 1999; Sansom, 1999), y en el estudio de patrones sinopticos (Bellone et al., 2000).

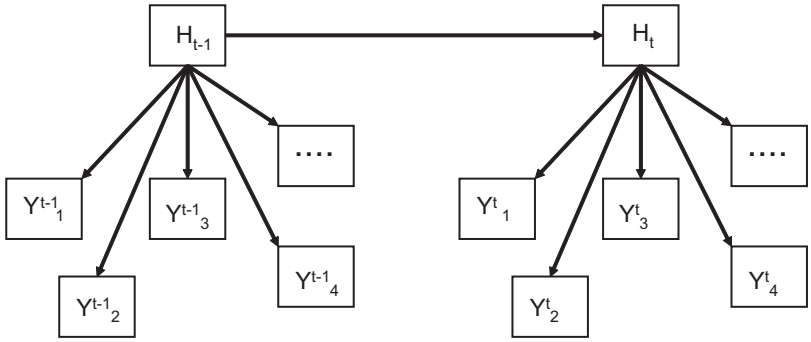

Figura 6.7: Esquema de un modelo de Markov oculto, donde todas las estaciones están condicionadas a la variable oculta H, que conecta el sistema en un instante de tiempo con el estado anterior.

La capacidad de expresión de la variabilidad espacial de los HMM es relativamente alta respecto del número de estados de la variable oculta. Por ejemplo, la Fig. 6.8 muestra la estructura de dependencia espacial de las series simuladas con un modelos de Markov con una variable oculta de 8 estados; esta figura reproduce las estructuas (modos) espaciales de precipitación de la serie original. Como es lógico suponer, la capacidad de expresión de la variabilidad espacial de los HMM mejora con el número de estados de la variable oculta aunque, por otra parte, no se pueden utilizar demasiados estados ya que el número de parámetros crece multiplicativamente, y el modelo se sobreajusta a los datos.

A continuación se propone una aplicación de las Redes Bayesianas que permite mantener la expresividad del modelo utilizando una estructura probabilística más simple, con menos parámetros.

6.2.3. Ejemplo de Aplicación de Redes Bayesianas

Como se vio en la Sec. 6.1, las redes Bayesians (RBs) permiten modelizar de forma conjunta la función de probabilidad de un conjunto de estaciones $Y_1, \ldots, Y_n$:

$$p(y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i|\pi_i). \tag{6.3}$$

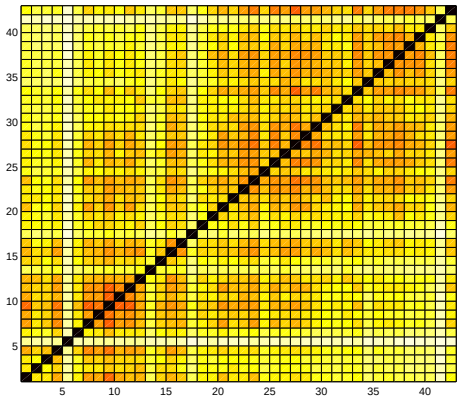

Figura 6.8: Matriz de dependencia espacial entre las 42 estaciones obtenida a partir de un HMM.

Por tanto, a partir de esta FPC se pueden simular series de precipitación para todas las estaciones que serían coherentes, tanto temporal, como espacialmente. Para ello, se puede utilizar un método de simulación simple e intuitivo, que consiste en simular las variables siguiendo un orden ancestral (simulando ordenadamente, de padres a hijos), de forma que sólo se simula el valor de una variable cuando han sido simulados todos y cada uno de sus padres (Henrion, 1988; Castillo et al., 1997); de esa manera, se puede utilizar la probabilidad de la variable en una localidad concreta, condicionada al valor simulado de los padres, para simular la ocurrencia o no de precipitación. Es decir, la ocurrencia o no de precipitación para una variable concreta Y_i^t, en el instante de tiempo t, se simularía a partir de $p(y_i|\pi_i^t)$, donde π_i^t es el valor de los padres, para el instante de tiempo t, que ya ha sido simulado en un paso anterior del algoritmo.

Este algoritmo actúa como un generador de tiempo con consistencia espacial ya que, en lugar de realizar simulaciones independientes para cada variable, se tienen en cuenta todas y cada una de las dependencias espaciales codificadas en el grafo. Por ejemplo, la Fig. 6.9 muestra el grafo obtenido para el conjunto de datos de la Cornisa Cantábrica, utilizando los algoritmos de aprendizaje automático descrito en la Sec. 3.5 (en particular, este grafo se ha aprendido con el algoritmo *localB* que se describe en la Sec. 8.1).

En la Fig. 6.10 se ilustra la diferencia entre una simulación realizada con un modelo de Markov (izquierda) frente a una simulación obtenida a partir de la red Bayesiana (derecha). Esta figura muestra tanto la probabilidad utilizada para simular la presencia de lluvia para cada estación (valores azules o verdes), como los valores simulados resultantes (rojos). Los valores en rojo simulados a partir de los azules

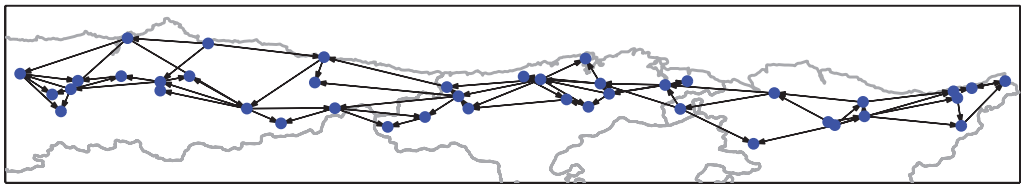

Figura 6.9: Grafo dirigido acíclico obtenido a partir de los datos con un algoritmo de aprendizaje *localB*.

plantean un modo de precipitación incoherente; en cambio en la parte derecha, los valores binarios obtenidos a partir de las probabilidades dadas por la red Bayesiana de forma secuencial, generan un modo de precipitación más realista respecto a la climatología. Por tanto, mientras el modelo de Markov produciría series simuladas de precipitación incorreladas entre unas localidades y otras, la red Bayesiana produciría series con la estructura espacial propia del problema. En la Fig. 6.11 se muestra la matriz de dependencia espacial obtenida al simular 10 años de datos con la red Bayesiana (observar la similitud con la matriz de correlación para observaciones, mostrada en la Fig. 6.6).

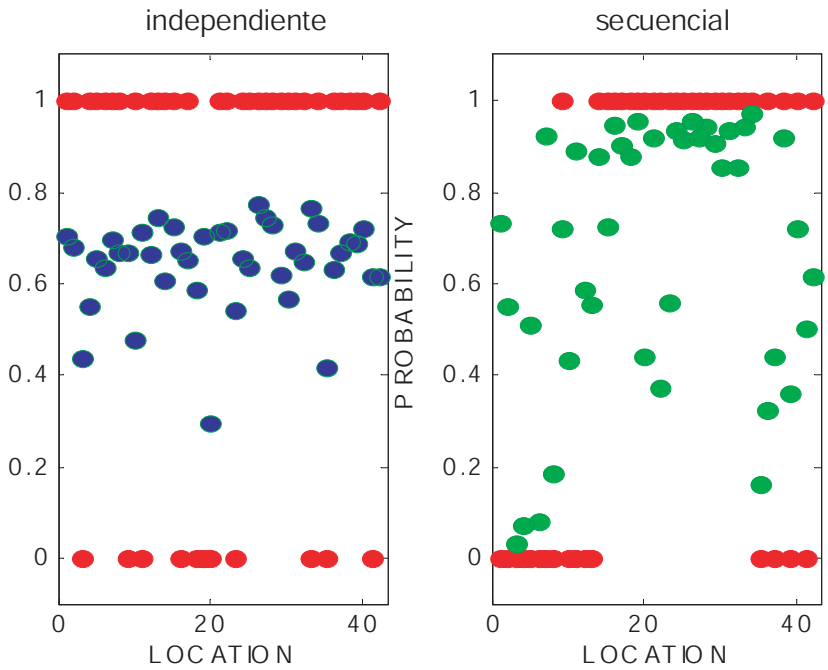

Figura 6.10: Simulación de una realización (un día) de precipitación para las 42 estaciones de la Cuenca Norte, considerando un modelo de Markov (izquierda) y una red Bayesiana (derecha). Los valores binarios (en rojo) son obtenidos a partir de los probabilísticos (en azul o verde).

La principal limitación del modelo basado en redes Bayesianas es que no posee la

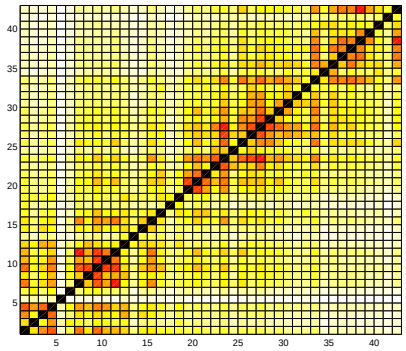

Figura 6.11: Matriz de dependencia espacial entre las 42 estaciones obtenida a utilizando la red Bayesiana.

componente de persistencia necesaria para simular las rachas de ausencia o presencia de lluvias, es decir, le falta la componente de Markov. Para ello, se pueden generalizar las redes Bayesianas considerando una componente dinámica formada dos grafos idénticos que representan el valor de las variables en instantes de tiempo $t-1$ y t, respectivamente (ver Fig. 6.12); en esta red se pueden introducir enlaces entre los dos instantes de tiempo de cada variables (a modo de transiciones de Markov). El modelo final posee enlaces intra-temporales (mostrados en gris en la figura) y intertemporales (mostrados en negro). Existen modelos más complicados, donde se permite que la estrucutra intra-temporal de la red cambie en el tiempo, pero en esta sección se ilustra la aplicación de los modelos más sencillos.

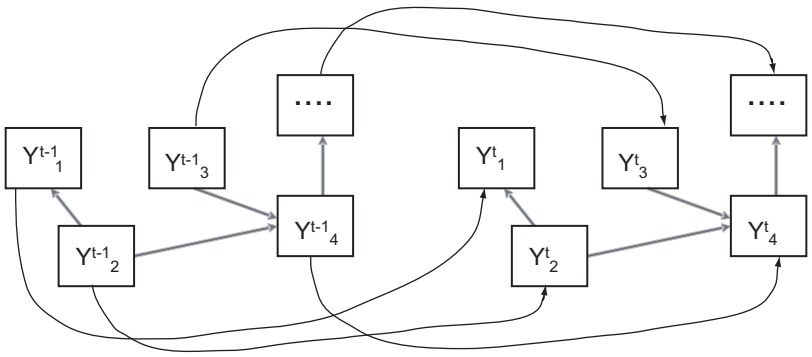

Figura 6.12: Estructura esquemática de la red Bayesiana dinámica utilizada; en gris aparecen los enlaces que representan la estructura espacial.

Esta misma idea puede también aplicarse para generalizar los modelos ocultos de Markov (HMM), considerando una estructura de dependencia espacial entre las localidades, para representar aquellas dependencias espaciales que no son representadas de forma apropiada por los estados (modos) de la variable oculta. Esto se lleva

a cabo planteando modelos *HMM extendidos* (HMME), de manera que la dependencia espacial dada por una red Bayesiana es añadida a la estructura del HMM, como se muestra en la Fig. 6.13.

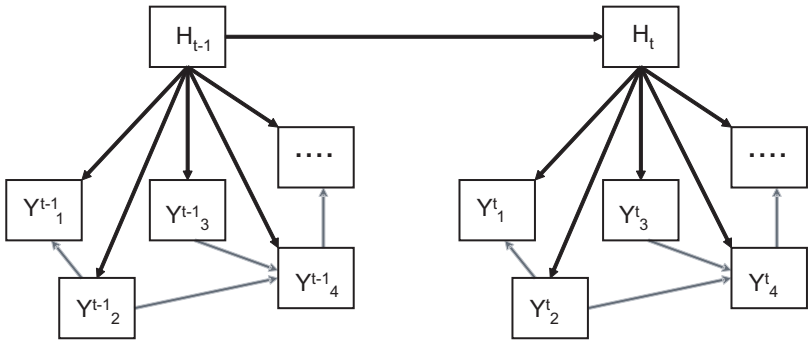

Figura 6.13: Esquema de un HMME, donde se muestran, en gris, los enlaces espaciales aprendidos con una red Bayesiana.

Para comparar los resultados del modelo extendido, HMME, y el modelo estándar, HMM, se ha procedido a realizar un agrupamiento de los distintos patrones de precipitación sobre la Cornisa Cantábrica, utilizando los registros históricos disponibles. Para ello se ha utilizado una red autoorganizativa (en inglés, Self-Organizing Model, SOM) para obtener 25 tipos o "modos" de precipitación, representada por los centroides resultantes del proceso de agrupamiento, que es similar al algoritmo de k-medias descrito en la Sec. 5.3. La Fig. 6.14 muestra la frecuencia de cada uno de los 25 tipos de tiempo para las observaciones (izquierda), simulación con HMM (central) y simulación con HMME (derecha), considerando series simuladas de la misma longitud que las series originales de cada localidad. Esta figura muestra que los modos simulados por el modelo HMME representan mejor la climatología de modos observada que el modelo HMM, ya que este último simula peor los modos de menor frecuencia debido a la limitación a 8 estados de la variable oculta.

Obsérvese que a la hora de combinar la variable oculta con las dependencias espaciales existen distintas alternativas, según se aprendan primero unos enlaces u otros; en esta Tesis no se exploran estas posibiliades y se dejan para un trabajo futuro.

6.3. Predicción Local

En la actualidad, la principal herramienta para la predicción meteorológica son las simulaciones del estado futuro de la atmósfera generadas por los modelos numéricos de circulación atmosférica, ya sean globales o locales (ver Sec. 2.2.3). Estos mo-

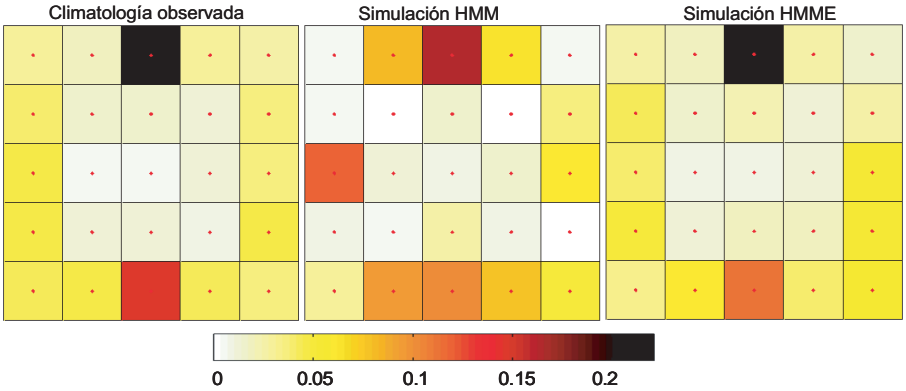

Figura 6.14: Comparación entre la simulación realizada por un HMM y un HMME.

delos deben parametrizar aquellos procesos cuya escala es inferior a la resolución utilizada en la simulación y que, generalmente, tienen gran influencia en la variabilidad de la fenomenología en superficie, a escala local (es decir, además de la componente dinámica, todos los modelos incluyen términos empíricos). Esta limitación constituye una de las principales razones por la que se han desarrollado técnicas estadísticas de postproceso de las predicciones, combinando esta información con las observaciones de las variables de interés en superficie (temperatura, precipitación, etc.), de forma que éstas se puedan infererir a partir de las variables de gran escala de los modelos, evitando el uso de otras variables que son más dependientes de las parametrizaciones (como la propia precipitación simulada por los modelos).

Desde un punto de vista histórico, las primeras técnicas estadísticas de predicción local no utilizaban las salidas de los modelos de circulación atmosférica, sino que se basaban en el análisis de series temporales, que ajustan los parámetros del modelo utilizando únicamente los registros históricos de observaciones, para simular la dinámica oculta en la serie temporal. Entre ellas se encuentran, los modelos autorregresivos (Zwiers and von Storch, 1990), incluyendo métodos no lineales basados, por ejemplo, en redes neuronales (Gardner and Dorling, 1998), y los métodos de reconstrucción de series caóticas basadas en la técnica de *embedding* (Pérez-Muñuzuri and Gelpi, 2000), por citar las más importantes. Sin embargo, estas técnicas estadísticas que utilizan únicamente las series de observaciones pierden su capacidad predictiva muy rápidamente con el alcance de la predicción, resultando competitivas únicamente en el muy corto plazo en comparación con otros métodos (ver, por ejemplo Cofiño, 2004).

Afortunadamente, las técnicas estadísticas ya no tienen porque restringirse exclusivamente a las series de observaciones, ya que existen cada vez más bases de datos históricas con reanálisis de modelos de circulación atmosférica (ver Sec. 2.4) que

pueden ser tratadas junto con las series de observaciones en los modelos estadísticos; en la Fig. 6.15 se muestra esquemáticamente este proceso. De esta manera, se puede aumentar la capacidad predictiva a escala local de los modelos de circulación para diferentes alcances de predicción. Por ejemplo, en esta Tesis se utilizan los datos del proyecto de reanálisis ERA-40 realizados con el modelo de circulación IFS del ECMWF, que proporciona una gran cantidad de campos meteorológicos (temperatura, geopotencial, viento, humedad, precipitación convectiva y a gran escala, etc.) en diferentes niveles de presión (ver Sec. 5.3, donde se describen en detalle los diferentes patrones y dominios que se utilizarán en esta Tesis).

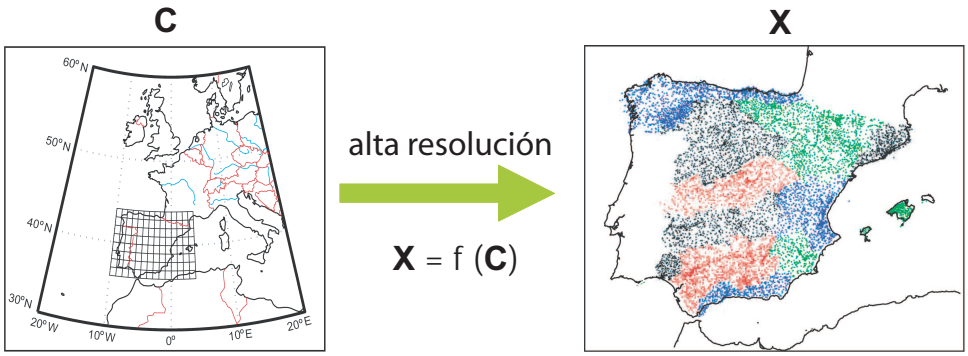

Figura 6.15: Esquema básico de la predicción local, donde se proyectan los campos C, generados por el modelo numérico en un cierto dominio, sobre los fenómenos observados a escala local o predictandos X.

En esta sección se describe un primer ejemplo de aplicación directa de las redes Bayesianas en este problema, considerando un sistema de predicción local de precipitación para la Cornisa Cantábrica.

6.3.1. *Ejemplo de Aplicación de Redes Bayesianas*

La aproximación más inmediata para el problema de predicción local consiste en combinar en un mismo modelo de probabilidad conjunta las salidas directas de un modelo numérico de circulación (predictores), en los puntos de rejilla del modelo sobre la zona de estudio, con las ocurrencias de precipitación en un conjunto de observatorios (predictandos). Es decir, se trata de obtener un modelo para estimar la distribución de probabilidad del vector de predictandos $\mathbf{Y}$ dado el vector de variables predictoras $\mathbf{C}$:

$$P(\mathbf{y}|\mathbf{c}) = P(y_1, ..., y_n \mid c_1, ..., c_m), \tag{6.4}$$

donde cada uno de los y_i hace referencia a las probabilidades de los diferentes estados de precipitación en los nodos predictandos y cada uno de los c_j hace referencia a las salidas del modelo numérico en los nodos del dominio seleccionado.

En esta sección se muestran los resultados de un caso simple, en el que las únicas variables del modelo numérico que se consideran son las predicciones de precipitación (variable que se desea predecir localmente), sin utilizar otras variables más robustas (como el viento, geopotencial e incluso la humedad, que están más libres de parametrizaciones y podrían predecir indirectamente la precipitación ocurrida). Este ejemplo describe, por tanto, una primera aproximación al problema de la predicción meteorológica local, que será analizado en mayor detalle en los siguientes capítulos de la Tesis, al considerarse el problema central de la misma.

En la Fig. 6.16 se muestra la red obtenida aplicando el algoritmo B (ver Sec. 3.5) a los datos de reanálisis (precipitación prevista por el modelo en los puntos de la rejilla del mismo, representados en color rojo) y observaciones (precipitación ocurrida en cada localidad, indicada en azul), para los días en que se dispone de observaciones en todas las localidades (ver Sec. 5.2). En este ejemplo se ha considerado el evento $Precip > 2mm$ y los datos han sido discretizados a 0 o 1 según la ocurrencia, o el valor previsto por el modelo, hayan sido inferiores o superiores a ese umbral. Para simplificar el proceso de aprendizaje se han impuesto dos restricciones adicionales: por una parte, los predictores no se pueden enlazar entre sí; por otra, no se permiten más de tres padres por predictando. De esta manera se obtiene un modelo que a partir de los valores discretizados de precipitación prevista por el modelo numérico en los puntos de grid, permite obtener las probabilidades de precipitación adaptadas a las distintas localidades.

Una vez entrenado el modelo considerando datos de reanálisis (asimilaciones y predicciones en tiempo pasado), se puede utilizar operativamente considerando como evidencia las predicciones dadas, día a día, por un modelo numérico de predicción similar al utilizado en el reanálisis. En este ejemplo se han considerado los valores de precipitación previstos con un día de antelación (D+1) por el modelo de predicción a plazo medio del ECMWF, durante el período del 1/12/1999 al 28/2/2000. A partir de estos valores se han inferido las probabilidades de precipitación en los 42 observatorios considerados y se han comparado con las ocurrencias reales. La Fig. 6.17 muestra los valores de Brier Skill Score (BSS) y las curvas ROC (ver Sec. 4.2) obtenidas en este caso (ver Cofiño et al., 2002, para más detalles). Hay que destacar que este tipo de sistemas híbridos estadístico-numéricos, son muy eficientes y se ejecutan prácticamente en tiempo real, por lo que pueden ser directamente incluidos en la cadena operativa de predicción para las diferentes pasadas y alcances

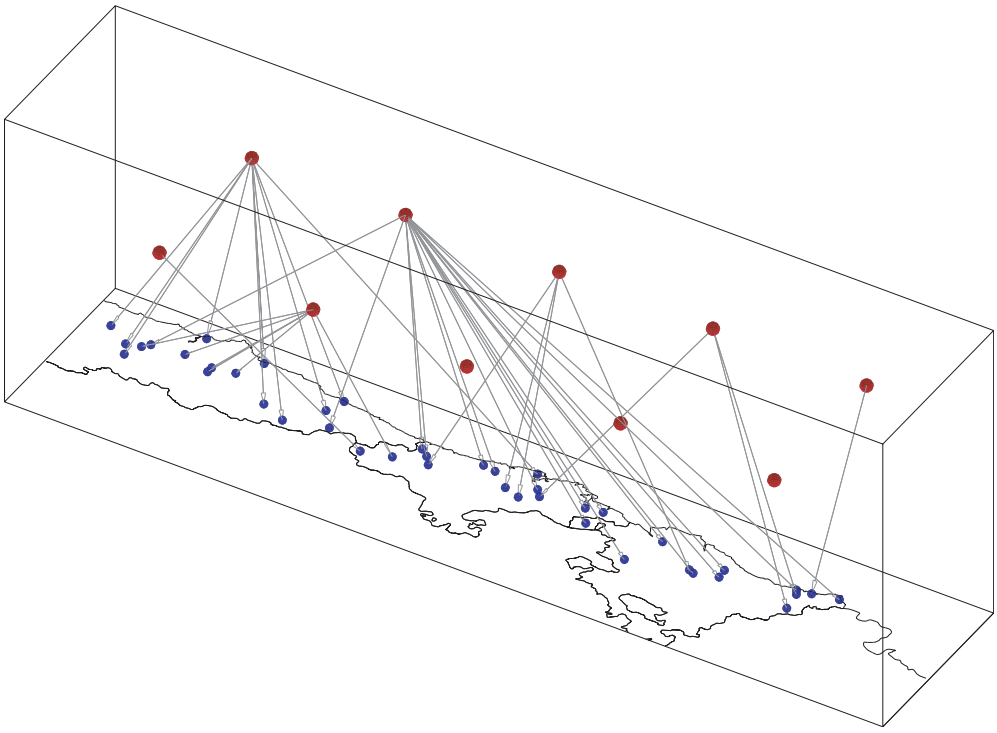

Figura 6.16: Red Bayesiana que conecta los predictores (puntos de rejilla del modelo numérico de circulación) con los predictandos (ocurrencias locales).

desarrollados en los servicios meteorológicos.

A pesar de los resultados y de su sencillez, este planteamiento es claramente mejorable, sobre todo desde un punto de vista práctico, ya que el número de variables y estados en un sistema razonablemente realista resultaría excesivo para estimar correctamente los parámetros, dadas las bases de datos disponibles (ver Sec. 3.7); además de implicar un enorme coste computacional en su fase de aprendizaje.

Dado que los resultados obtenidos resultan prometedores, en lo que resta de Tesis se abordará este problema con mayor detalle, planteando al mismo tiempo un sistema más eficiente. Para ello, por un lado, se reconsiderará la representación del sistema mediante modelos más sencillos (con menos parámetros, utilizando *weather types*), y por otro, se adaptarán las técnicas de aprendizaje descritas en 3.5, al *downscaling* estadístico, tema al que se dedicará por completo el Capítulo 8.

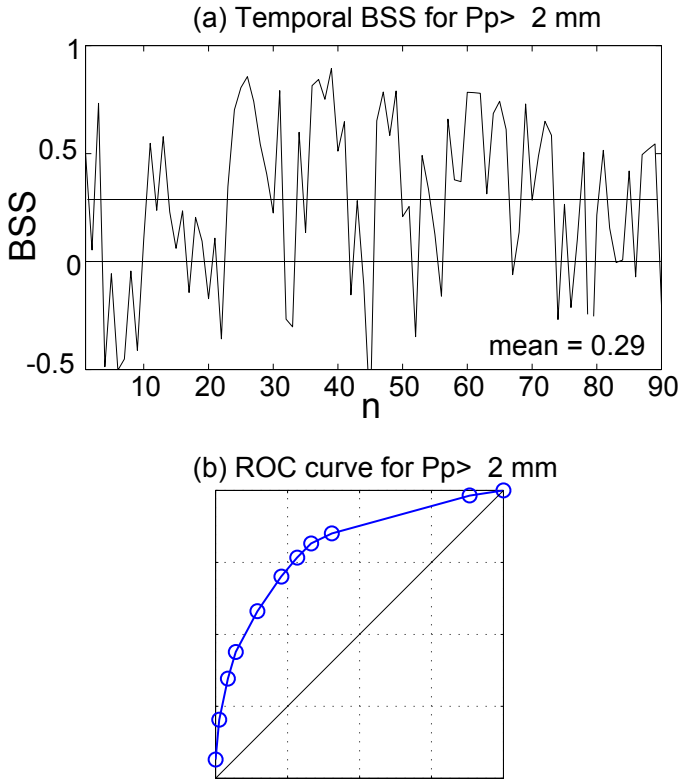

Figura 6.17: BSS y curvas ROC para la validación de las predicciones en el periodo DJF 1999 (90 días). Se han utilizado como evidencia los patrones de precipitación prevista por el modelo en puntos de rejilla.

CAPÍTULO 7

Paradigmas de Predicción Probabilística Local

En el capítulo anterior se han mostrado algunas aplicaciones genéricas de las redes Bayesianas (RBs) en distintos problemas de meteorología. El resto de la Tesis se centra en la aplicación de las RBs en un problema concreto: la *predicción probabilística local* de precipitación, pues es uno de los productos más demandados por los usuarios (Harris et al., 2001) y por lo tanto, una de las líneas de investigación más activas en la actualidad. Como primera aproximación al problema, en el presente capítulo se presentan los distintos paradigmas de predicción local existentes, planteados en un mismo marco probabilístico en términos de cálculo de probabilidades condicionadas, según la evidencia disponible. Por tanto, este capítulo servirá para centrar teóricamente el problema a tratar en capítulos posteriores: el *downscaling estadístico*, en el cual la única evidencia disponible son las predicciones de un modelo numérico de circulación.

La organización es la siguiente. En la Sec. 7.1 se presentan los problemas paradigmáticos de la predicción local, y se analizan las correspondientes topologías de grafos y la evidencia que son de aplicación para los mismos. En la Sec. 7.2 se muestra una perspectiva general de uno de estos paradigmas, el downscaling estadístico, desde el punto de vista de la predicción local. Finalmente, en la Sec. 7.3, se estudia la topología más sencilla para abordar este problema, heredada del *clasificador ingenuo*, que ya fué descrito en la Sec. 3.6; en este caso se analiza su utilidad para la predicción local.

7.1. Paradigmas de Predicción Local

En la Sec. 6.3 se introdujo el problema de la predicción local con RBs, utilizando como predictores los valores simulados/predichos por un modelo de circulación atmosférica para la variable de interés (precipitación) en los puntos de rejilla, $C_1, \ldots, C_m$, discretizados de forma conveniente; sin embargo, se comentó que era conveniente utilizar otras variables del modelo de circulación para mejorar la predicción. Por otra parte, en la Sec. 5.3 se analizó la discretización conjunta de todo el patrón de circulación utilizando *tipos de tiempo*, obtenidos mediante una técnica de agrupamiento (k-medias o redes auto-organizativas). Esta discretización permite manejar la información dada por distintas variables del modelo de ciculación a través de una única variable de estado discreta, con un número predefinido de estados (los tipos de tiempo).

Por tanto, a partir de ahora, se considerará el problema de predicción local desde una perspectiva más amplia, considerando un conjunto de localidades $X_1, \ldots, X_n$, con valores binarios (ocurrencia o no de un evento de precipitación), junto con el estado previsto por un modelo de circulación para la clase o tipo de tiempo C; esta variable es categórica con un número prescrito de estados (entre 10 y 100, dependiendo de la resolución que se desee para la predicción). Por tanto, la Función de Probabilidad Conjunta (FPC) que caracteriza genéricamente el problema de predicción local puede ser expresada como $p(\mathbf{x}, c) = p(x_1, \ldots, x_n, c)$, donde tanto C como $\mathbf{X}$ (o un subconjunto suyo) pueden ser evidenciales o no (es decir, se puede disponer de su valor estimado para efectuar la predicción). Obsérvese que en meteorología es común tener conocimiento parcial de $\mathbf{X}$ (por ejemplo, algunas estaciones automáticas proporcionan información en tiempo real). Por lo tanto, el conjunto de localidades puede ser separado en dos grupos, $\mathbf{X} = \{\mathbf{Z}, \mathbf{Y}\}$, según sean evidenciales o no, respectivamente. Las RBs permiten propagar cualquier evidencia, sin distinción entre sus variables. Por tanto, los distintos paradigmas de predicción local pueden ser formulados en base a las distintas combinaciones de evidencia disponibles según el un conjunto de *predictores* y *predictandos* en cada caso.

Si se denomina $\mathbf{Q}$ al conjunto de predictandos y $\mathbf{E}$ al conjunto de predictores, los distintos paradigmas de predicción local están determinados por el tipo de predictores ($\mathbf{E} = \{C\}$, $\mathbf{E} = \{\mathbf{Z}\}$ o $\mathbf{E} = \{\mathbf{Z}, C\}$), y el tipo de predictandos ($\mathbf{Q} = \{\mathbf{Y}\}$, $\mathbf{Q} = \{C\}$ o $\mathbf{Q} = \{\mathbf{Y}, C\}$) disponibles en la práctica para realizar la predicción; es decir, para obtener $p(\mathbf{q}|\mathbf{e})$. Ordenados por casos, se tienen los siguientes paradigmas:

1. Cuando la evidencia procede de hechos observados, $\mathbf{E} = \{\mathbf{Z}\}$, se habla de paradigma de *diagnóstico* y, en el caso general se tiene: $p(\mathbf{q}|\mathbf{e}) = p(c, \mathbf{y}|\mathbf{z})$;

existen, además, los siguientes casos particulares:

   a) *Clasificación*: $p(c|\mathbf{z})$.

   b) *Interpolación*: $p(\mathbf{y}|\mathbf{z})$.

2. La única variable susceptible de ser anticipada en el tiempo es el estado de la atmósfera (usando las predicciones de los modelos de circulación). Por tanto, cuando $\mathbf{E} = \{C\}$, o $\mathbf{E} = \{\mathbf{Z}, C\}$), se dice que la estimación tiene capacidad predictiva y se denomina paradigma de *predicción*, con $p(\mathbf{q}|\mathbf{e}) = p(\mathbf{y}|c, \mathbf{z})$; en esta situación, existe un caso particular:

   a) *Downscaling*: $p(\mathbf{y}|c)$.

Como nuestro interés se centra en predecir los valores locales, sólo se consideran los casos en los que $\mathbf{Q} = \{\mathbf{Y}\}$, es decir, *interpolación*, *predicción* y *downscaling*. Por otra parte, cuando se aplican las RBs a estos problemas, siguiendo un paralelismo con el estudio de clasificadores basados en RBs (ver Sec. 3.6) se pueden construir tres tipos de red, dependiendo de su topología:

- *Red Ingenua*, que es el tipo básico de RB para clasificación, con estructura *ad-hoc*, en la cual el predictor se enlaza a todos los predictandos, y estos no se enlazan entre sí.

- *Redes Extendidas*, que se obtienen a partir de la red *ingenua*, agregando enlaces entre predictandos.

- *Redes Genéricas*, que es una RB aprendida sin restricciones, es decir, sin forzar la estructura *ingenua* inicialmente.

En este contexto, si combinamos todas las opciones, restringidas a $\mathbf{Q} = \{\mathbf{Y}\}$, resultan 9 formas diferentes de plantear el problema de la predicción local, cuyos esquemas se muestran en la Fig. 7.1.

7.2. Downscaling Estadístico

El problema de predicción local descrito en la Sec. 6.3 se corresponde con el paradigma del *downscaling* descrito en la sección anterior, y es el que tiene más interés desde un punto de vista operativo, ya que sólo se dispone de predicciones futuras del estado de la atmósfera, a partir de un modelo de circulación. En la actualidad existen numerosos métodos para abordar el problema del downscaling, aunque

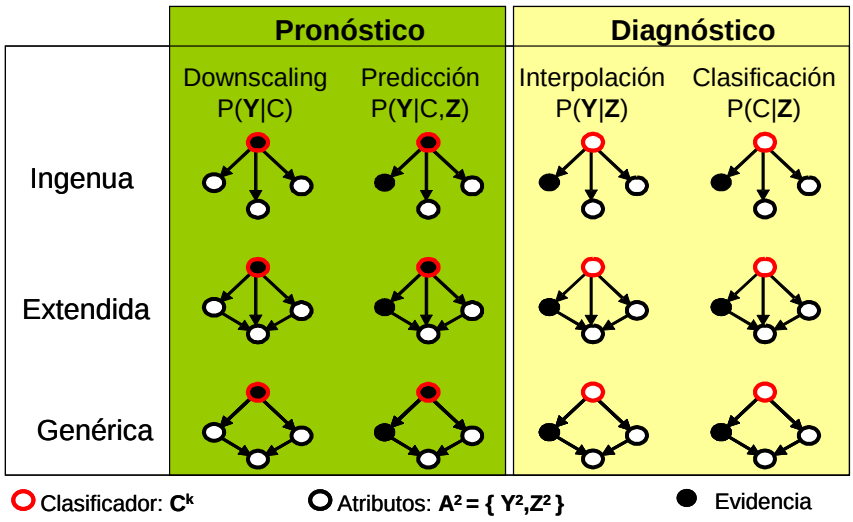

Figura 7.1: Diferentes topologías de red combinadas con diferentes evidencias **E**. Para el caso general del pronóstico local, aparecen restringidas al caso **Q={Y}**.

pueden clasificarse en dos grandes grupos dependiendo de si, además de los modelos globales de circulación atmosférica (ACM), se utilizan modelos dinámicos regionales (o de área limitada, LAM), descritos en la Sec. 2.2.3, o si además se utilizan modelos estadísticos empíricos obtenidos a partir de las observaciones disponibles en la región o localidad de interés (Murphy, 1999). En el primero de los casos los métodos de downscaling se denominan *dinámicos*, mientras que en el segundo se denominan *estadísticos*. La Fig. 7.2 muestra esquemáticamente ambos tipos de downscaling.

A su vez, los métodos estándar de downscaling estadístico se pueden clasificar básicamente en tres tipos: *funciones de transferencia*, *métodos basados en análogos* y *generadores de tiempo*; los dos primeros aumentan la resolución espacial (ver Wilby and Wigley, 1997; Zorita and von Storch, 1999, para una introducción a estos métodos) y el último la temporal (este grupo ya se ha abordado en la Sec. 6.2). A su vez, algunos de estos métodos son deterministas (proporcionan una estimación numérica del valor de la variable), mientas que otros son probabilísticos (predicen la probabilidad de un evento, o la función de distribución de una variable). En esta Tesis se analiza el downscaling probabilístico formulando el problema en los términos propuestos en la sección anterior, e introduciendo por primera vez las redes Bayesianas en este ámbito. Todos los métodos estadísticos asumen la hipótesis estacionaria, que consiste en suponer que los parámetros estadísticos obtenidos en el periodo de entrenamiento (o calibración), son también válidos para otros periodos de test o predicción, que normalmente corresponden a un tiempo futuro. Esta hipótesis es crítica cuando se aplican estos métodos a proyecciones de cambio climático. En

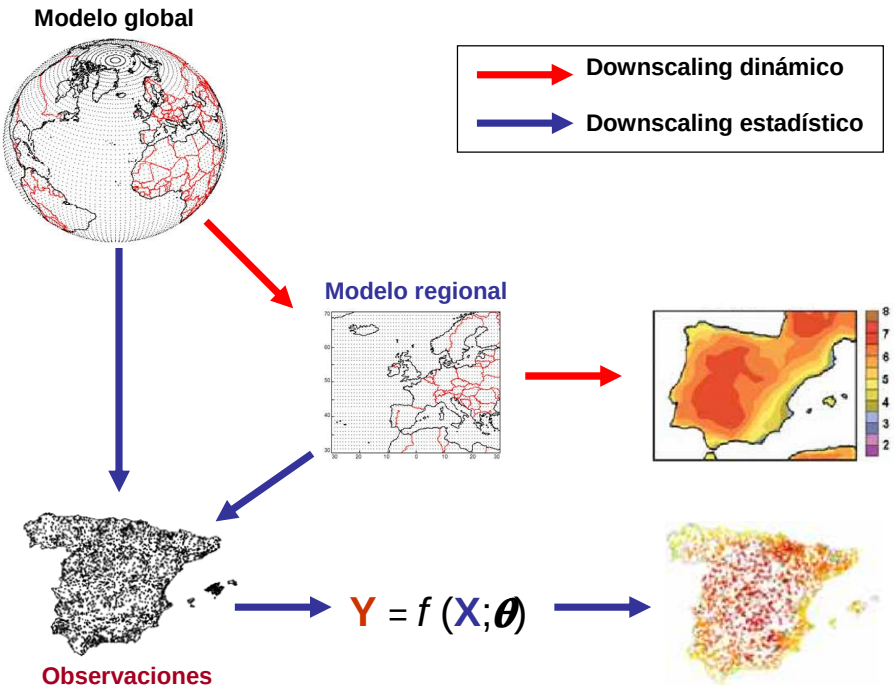

Figura 7.2: Esquema de los distintos procesos de *downscaling*.

esta Tesis, además, se analiza el problema bajo la hipótesis de "modelo perfecto" (es decir, las predicciones globales del modelo de circulación se consideran perfectas, y no se tienen en cuenta términos que corrijan los errores del propio modelo de circulación). Para ello, en lugar de considerar los campos predichos por un modelo para el estado de la atmósfera, se usan datos de reanálisis que corresponden a la condición inicial de las predicciones (es decir, al estado real asimilado al espacio del modelo). Por tanto, para la aplicación operativa de estos modelos, es necesario añadir el término del error de la predicción del modelo global.

7.2.1. Métodos Basados en Funciones de Transferencia

Estos métodos se basan en una función de transferencia $\boldsymbol{x} = f(\boldsymbol{c}; \theta)$ entre los predictores y predictandos, obtenida ajustando los parámetros del modelo θ considerando datos pasados; en este caso se suelen considerar como predictores los valores del modelo en los puntos de rejilla, o las componentes principales (ver Wilks, 1995; Gutiérrez et al., 2004b, para una introducción general a estos métodos). Los métodos más sencillos son los basados en regresión lineal múltiple Glahn (1974); Ayuso (1994) y fueron utilizados en los primeros sistemas operativos de downscaling; estos sistemas se denominaron Model Output Statistics (MOS) y actualizaban los

parámetros del modelo de regresión de forma casi continua para tener en cuenta los cambios del modelos atmósferico de circulación. Más recientemente, se han propuesto métodos no lineales más sofisticados, como las redes neuronales (Gutiérrez et al., 1999; Michaelides et al., 2001) o las *Support Vector Machine* (SVM), inicialmente propuestas para abordar probabilísticamente problemas de clasificación, han sido adaptadas para la regresión por (Yu et al., 2006). Estos métodos pueden aplicarse tanto en modo determinista, como en modo probabilístico. Por ejemplo, una regresión logística produce una salida acotada entre 0 y 1, que puede modelizar directamente probabilidades (Sanchez and Marcos, 2001); las redes neuronales también permiten modelizar problemas probabilísticos mediante la inclusión de funciones de activación sigmoidales en la última capa.

Estos métodos de *downscaling* estadístico no tienen en cuenta la dependencia espacial de los predictandos y, por tanto, la predicción de cada localidad se realiza de forma independiente. Sin embargo, existen distintas variantes de esos métodos, de mayor complejidad, que tienen en cuenta este problema, proporcionando predicciones espacialmente consistentes; estas técnicas se denominan *multi-localidad* (en ingles, *multi-site*). Para el caso lineal, el Análisis de Correlación Canónica (ACC), es una generalización de la regresión lineal que tiene en cuenta distintos "modos" de predictandos y predictores, dando lugar a predicciones consistentes espacialmente (Barnett and Preisendorfer, 1987). Sin embargo, para problemas que no admiten un tratamiento lineal, como por ejemplo la precipitación diaria, no existen técnicas tan eficientes que permitan resolver el problema. Por ejemplo, en Wilby et al. (2003), se utiliza una regresión múltiple para obtener promedios areales, que posteriormente son utilizados para generar valores individuales de precipitación en las estaciones del área, mediante técnicas de muestreo por aceptación-rechazo. Por otra parte, Kramer (1991) y Hsieh (2000) han desarrollado generalizaciones de las redes neuronales para obtener modelos de ACC no lineales, pero resultan demasiado complejos y costosos computacionalmente para ser aplicados en la práctica. Por tanto, el downscaling multi-localidad para problemas no lineales (por ejemplo para predicción de precipitación diaria) es todavía un problema abierto. En esta Tesis se aborda este problema con una nueva técnica: las redes Bayesianas.

En la Fig. 7.3 se muestran dos ejemplos de aplicación de las técnicas lineales descritas anteriormente (regresión y CCA) al conjunto de estaciones sobre la Península Ibérica descritas en la Fig. 5.1. El panel (a) muestra el resultado obtenido al realizar una regresión para dos predictandos diferentes a escala diaria: la temperatura máxima (izquierda) y la precipitación (derecha). Esta figura pone de manifiesto las limitaciones de estos métodos cuando se aplican en problemas con datos no normales y

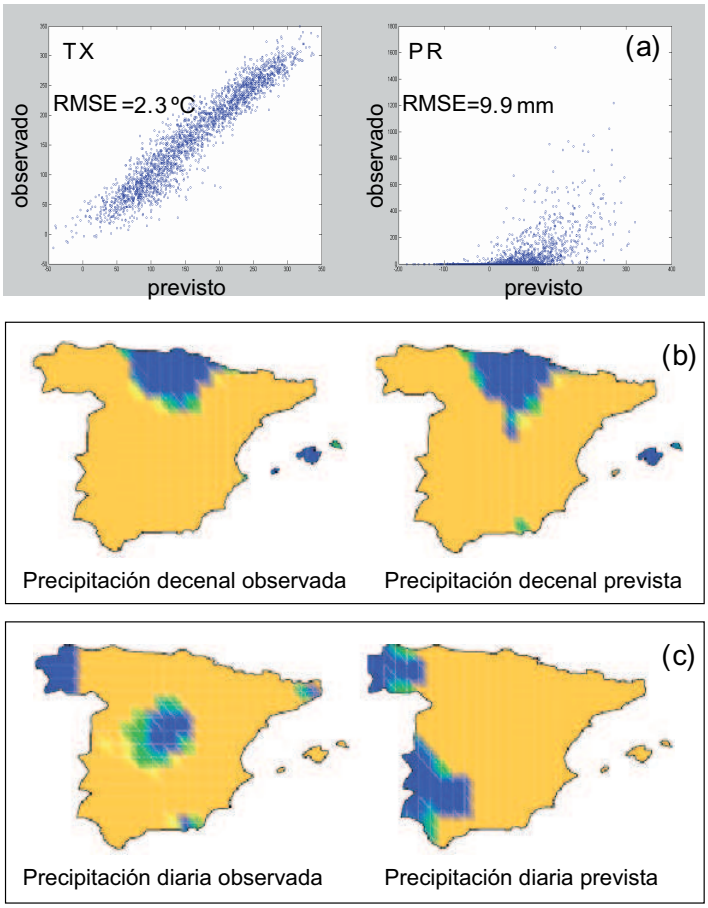

Figura 7.3: Algunos ejemplos de aplicación de métodos lineales basados en funciones de transferencia.

relacciones no lineales. Los paneles (b) y (c) muestran los resultados al aplicar ACC a la precipitación (agregada decenalmente y la variable original a escala diaria, respectivamente); esta figura muestra que la precipitación integrada decenalmente tiene un mejor comportamiento, al ser más apropiada la aproximación normal, dado su carácter agregado; sin embargo, para la precipitación diaria este método tiene un comportamiento muy deficiente.

7.2.2. Métodos basados en Análogos

Las técnicas de análogos (Lorenz, 1969) son métodos algorítmicos, donde no se calcula ningún modelo sino que se aplica un procedimiento predefinido para inferir una predicción en base a las ocurrencias históricas en la localidad de interés, utilizando aquellas en las que el patrón de la atmósfera era similar al patrón del día

a predecir. La hipótesis básica de estos métodos es que patrones de la atmósfera (predictores) similares producen predictandos similares. Por ejemplo, con ERA-40 se dispone de 16436 patrones atmosféricos diarios para buscar días similares (análogos). Por tanto, para cada situación prevista por el modelo numérico se obtiene un *conjunto de análogos* en base a los patrones atmosféricos, que está formado por los k días más próximos al patrón previsto en la base de datos del reanálisis. (ver Gutiérrez et al., 2004b, para una revisión de estos métodos).

La Fig. 7.4 muestra un esquema genérico de este método. En esta figura puede observarse que es necesaria una base de datos con patrones atmosféricos (por ejemplo ERA40) y otra simultánea con los registros históricos de observaciones (por ejemplo, las observaciones sobre la Península Ibérica descritas en la Sec. 5.1). El algoritmo parte de una predicción del patrón atmosférico previsto, y obtiene una predicción local para los observatorios disponibles. En la literatura han sido descritas distintas implementaciones de este tipo de métodos, tanto para la predicción de anomalías climáticas (Zorita and von Storch, 1999; Wilby and Wigley, 1997), como para la predicción estacional (Gutiérrez et al., 2005), y la predicción a medio y corto plazo (Gutiérrez et al., 2004a; van den Dool, 1989; Beersma and Buishand, 2003).

En general, se ha demostrado que el método de análogos funciona tan bien como otras técnicas más complicadas de downscaling (ver Zorita and von Storch, 1999), indicando que estos métodos de "hombre pobre" son alternativas eficientes para diversos problemas de downscaling.

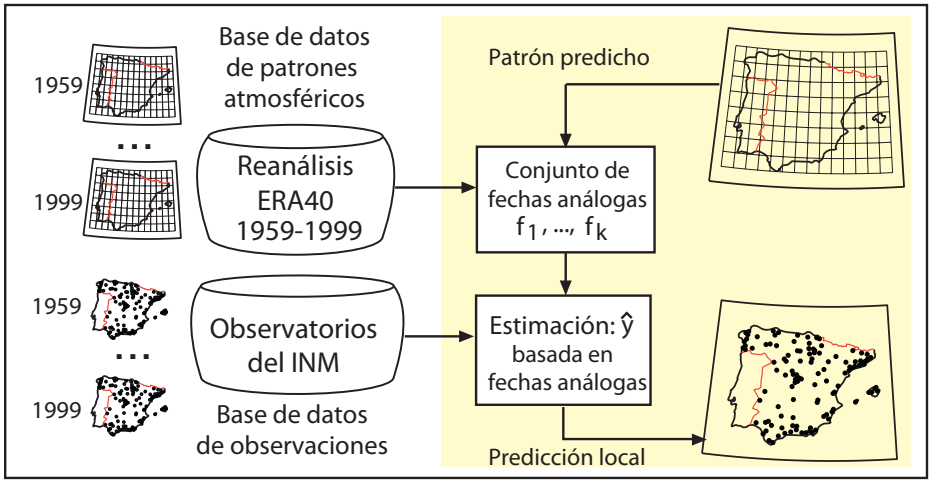

Figura 7.4: Esquema genérico del algoritmo estándar de downscaling por análogos.

El método de análogos, introducido originalmente por Lorenz (1969) en el marco

de la predicción de series temporales, es un caso particular de una metodología más general llamada técnica de vecinos próximos (Nearest Neighbors, NN) (Hastie et al., 2001). La ventaja más importante de esta metodología es que las distintas localidades no se predicen de forma independiente, sino que las predicciones se obtienen a partir de las ocurrencias de un mismo conjunto de días históricos y, por tanto, tienen cierta consistencia espacial. Por tanto, esta metodología puede considerarse directamente de tipo *multi-localidad*, aunque este problema no se considere de forma directa, y existan ciertas limitaciones intrínsecas. Una limitación importante de esta metodología es que considera siempre la misma cantidad de análogos sin tener en cuenta la estructura del espacio donde se buscan los vecinos del patrón previsto. Por ello, recientemente han sido propuestas una serie de alternativas que discretizan el espacio de búsqueda (el espacio de patrones atmosféricos), considerando un conjunto reducido de *tipos de tiempo* (ver Sec. 5.3), o clases de patrones atmosféricos. Estas técnicas permiten dividir automáticamente la base de datos de reanálisis en grupos C_i, caracterizados por un patrón prototipo $\mathbf{v}_i$. Cada grupo resultante puede ser utilizado como conjunto de análogos para aquellos patrones que sean posteriormente clasificados en ese grupo (patrones más cercanos a $\mathbf{v}_i$, que a cualquier otro prototipo). Por tanto, el cálculo de distancias sólo involucra al patrón previsto y a los prototipos, lo que reduce significativamente el tiempo de computación. En la literatura se han aplicado distintas técnicas de agrupamiento en este problema. Por ejemplo, Hughes et al. (1993) describe una algoritmo de downscaling que utiliza una técnica de agrupamiento llamada CART (ver también Zorita et al., 1995), que se ha mostrado eficaz en problemas con patrones atmosféricos simples; otra aproximación a este problema ha sido realizada aplicando redes auto-organizativas (SOM) como algoritmo de agrupamiento (Eckert et al., 1996; Cavazos, 1997, 2000), considerando también patrones simples y un número reducido de grupos. Recientemente, Gutiérrez et al. (2004a) han descrito un método genérico eficaz con patrones complejos (distintas variables en distintos niveles) y apropiado para la predicción por conjuntos. Este método utiliza la técnica de agrupamiento de k-medias para seleccionar los grupos de patrones análogos en la base de datos de reanálisis, según se describe en la Sec. 5.3.

Como una primera etapa de preproceso, se realiza el agrupamiento de la base de datos de reanálisis utilizando el método de k-medias (ver, por ejemplo, Gutiérrez et al., 2004a). Esta etapa permite prescindir de los datos de reanálisis en el resto del proceso, pues éstos son reemplazados por un número determinado de grupos $C_1, \ldots, C_m$ con sus correspondientes prototipos $\mathbf{v}_1, \ldots, \mathbf{v}_m$. En la fase operativa, las únicas distancias que es necesario calcular para obtener el conjunto de análogos de un patrón

de predicción dado $\mathbf{x}_t$, son entre $\mathbf{x}_t$ y $\mathbf{v}_i$, $i = 1, \ldots, m$, y en base a ellas se halla el grupo al que pertenece el patrón. Una vez calculado el grupo $C_k = \{\mathbf{x_{i_1}}, \ldots, \mathbf{x_{i_{q(k)}}}\}$, donde $q(k)$ es el número de patrones del grupo k-ésimo e $i_1, \ldots, i_{q(k)}$ se refieren a las fechas de los días en el período de reanálisis, se puede obtener una predicción probabilística local para una variable y y una estación s dadas. Esta predicción se puede obtener a partir de la función de probabilidad empírica de las observaciones análogas $\{y^s_{i_1}, \ldots, y^s_{i_{q(k)}}\}$. Por ejemplo, la probabilidad de que la variable supere un cierto umbral u puede ser estimada como:

$$P^s_i = P(y^s > u|C_i) = \frac{\#\{y^s_{i_j} > u;\ j = 1, \ldots, q(i)\}}{q(i)}. \tag{7.1}$$

Existen también modelos híbridos, como el propuesto por Cannon (2007), formado por una red neuronal y un método de análogos, donde los predictores genéricos, obtenidos de un modelo de circulación, son transformados por la red neuronal en predictores específicos que son utilizados por el método de análogos para obtener las predicciones. El acoplamiento entre ambos modelos estriba en que los parámetros de la red neuronal son entrenados minimizando el error de las predicciones realizadas por el método de análogos. Es decir, la red neuronal opera como un buscador de predictores óptimos para el método de análogos. Por tanto la capacidad de preservar las relaciones de dependencia espacial de este método depende del número de análogos considerado.

En las sección siguiente se muestra que el esquema de downscaling probabilístico basado en análogos es equivalente al método de downscaling utilizando redes Bayesianas ingenuas descrito en la Sec. 7.1. Por tanto, los métodos basados en redes Bayesianas que se proponen en esta Tesis pueden considerarse una generalización de esta metodología (Gutiérrez et al., 2002). De hecho, las técnicas de análogos son métodos *no generativos* o *discriminativos*, que obtienen directamente la probabilidad de cada predictando dado el predictor sin utilizar un modelo probabilístico conjunto. Sin embargo, las redes Bayesianas son métodos *generativos* y permiten calcular cada probabilidad condicionada $p(\mathbf{q}|\mathbf{e})$ a partir de un único modelo conjunto de todas las variables $p(\mathbf{q}, \mathbf{e})$, dando así una mayor consistencia y solidez a la metodología.

7.3. Downscaling con Redes de Estructura Ingenua

Dedicamos este apartado a estudiar la topología más sencilla de las mostradas en la Fig. 7.1, marcada como *ingenua*, heredada de la estructura de clasificador más

simple, conocida como *clasificador ingenuo*, descrita en la Sec. 3.6. En esa sección se mostró que el objetivo de un clasificador es hallar el valor de la variable de clase C (el tipo de tiempo), conocidos ciertos atributos $\mathbf{Z}$ (ocurrencias en localidades). Sin embargo, en el problema de downscaling esta estructura de red se usa de forma inversa, es decir, estimando los atributos dada la clase; por eso se habla de *red de estructura ingenua*, en lugar de clasificador ingenuo. A pesar de su sencillez, como se verá más adelante, la redes ingenuas poseen excelentes cualidades para la predicción meteorológica de todo tipo de fenómenos, mostrando una capacidad predictiva comparable a la de métodos más complicados. Por este motivo, estos modelos se utilizan como una referencia y punto de partida para comparar los resultados de métodos más complejos. En esta sección se analizan las propiedades de las redes ingenuas para el problema de downscaling.

La hipótesis de independencia condicional de la red ingenua, que determina la estructura de la Función de Probabiliad Conjunta (FPC) dada en (3.14), establece la independencia de las ocurrencias en distintas localidades, una vez conocido el estado de la atmósfera; es decir, $I(X_i, X_j|C)$, para cualquier par de localidades i y j. Aunque esta hipótesis puede resultar excesivamente rigurosa en otros campos, no lo es tanto en la predicción meteorológica, pues equivale a decir que las relaciones de dependencia espacial presentes en la fenomenología meteorológica desaparecen cuando se concreta el estado de la atmósfera. Esto es estrictamente cierto, al menos en teoría, ya que dicha relación de dependencia local se deriva exclusivamente del hecho de compartir en tiempo y espacio escenarios atmosféricos similares y, por lo tanto, solo cabe plantearse modelos más complejos para aquellos casos en los que circunstancialmente, las relaciones de dependencia local no desaparezcan totalmente dado el estado de la atmósfera. Estas circunstancias solo pueden proceder de los defectos del sistema de observación, de donde no cabe esperar ninguna capacidad predictiva útil, o del hecho de no disponer de un predictor (estado de la atmósfera) perfecto. Este último hecho es especialmente importante en este caso, ya que la discretización del estado de la atmósfera en tipos de tiempo introduce una incertidumbre que afecta, en menor o mayor medida, a las relaciones de dependencia condicional entre las ocurrencias locales. Dada su importancia, a continuación describimos con detalle este último aspecto.

Cuando se analiza el comportamiento de la red ingenua sobre la predicción local de eventos de meteorológicos, hay que tener en cuenta que las variables que intervienen en el problema son $\{\mathbf{X}, \mathbf{H}\}$, donde $\mathbf{H}$ es el estado de la atmósfera verdadero (que obviamente es una variable desconocida u oculta con gran dimensión). En este caso, se verifica la propiedad $I(X_i, X_j|\mathbf{H})$, para cualquier pareja $\{X_i, X_j\}$ de $\mathbf{X}$; es

decir, **H** contiene todas las dependencias posibles entre atributos, pues determina las distintas ocurrencias locales posibles. En ese caso, la probabilidad conjunta $p(\mathbf{x}, \mathbf{h})$ es:

$$p(\mathbf{x}, \mathbf{h}) = p(\mathbf{h}) \prod_{i=1}^{n} p(x_i \mid \mathbf{h}). \tag{7.2}$$

Sin embargo, la realidad es que **H** no es conocido y en su lugar se utiliza una variable C, que representa el estado de la atmósfera simulado a partir de cierto modelo numérico y discretizado en un número arbitrario de estados, con la consiguiente pérdida de resolución. En estas condiciones, parece lógico suponer que, debido a las imperfecciones en C (que es una aproximación de **H**), ya no se cumplirá la relación de independencia condicional, y por lo tanto cabe sospechar que:

$$p(\mathbf{x}, c) \neq p(c) \prod_{i=1}^{n} p(x_i|c). \tag{7.3}$$

en cuyo caso, tendría sentido buscar topologías alternativas a la red ingenua. Por ejemplo, la Fig. 7.5 muestra el error que se comete para estaciones próximas al utilizar una red ingenua, como por ejemplo San Sebastian y Fuenterrabía, constatando que la dependencia condicional en este tipo de sistemas provoca errores importantes en los modelos ingenuos. Por lo tanto, en estos casos es necesario utilizar métodos capaces de considerar estas relaciones de dependencia, justificando la utilización de las redes Bayesianas en este campo, de forma que se puedan crear redes más complejas, capaces de considerar relaciones de dependencia entre los predictandos.

Modelo Ingenuo p(a,b|c)=p(a|c)p(b|c)

p(a,b|c=1)	a=0	a=1
b=0	.35	.31
b=1	.17	.15

Modelo General p(a,b|c)=p(a|c)p(b|a,c)

p(a,b|c=1)	a=0	a=1
b=0	.50	.17
b=1	.03	.30

Figura 7.5: Función de probabilidad conjunta para el evento de precipitación $prec > 0mm/24h$ para A=San Sebastian y B=Fuenterrabía, considerando una red de estructura ingenua (arriba) y una red general, sin asumir ninguna hipótesis de independencia (abajo).

Sin embargo, tampoco se puede descartar que existan circunstancias bajo las cuales la red ingenua se mantenga como la opción más correcta, como ocurre por ejemplo, en sistemas cuyos predictandos son de baja resolución respecto al evento a predecir. Por lo tanto conviene explorar acerca de los límites de aplicación de este tipo de topología.

7.3.1. Dependencia Espacial

En la Sec. 6.1 se ha visto la gran complejidad que puede alcanzar la dependencia espacial marginal y condicionada en aplicaciones meteorológicas. Por tanto, conviene comprobar hasta que punto se cumple la independencia condicional de predictandos dado el estado de la atmósfera, cuando éste se representa por un número finito de tipos de tiempo (en esta Tesis se consideran aplicaciones con 10 y 100 tipos, para analizar el impacto de este factor). Por tanto, se utiliza el concepto de información mútua para comprobar si X_i e X_j son independientes o no dado C, ya que $MI(X_i, X_j|C)$ será tanto más pequeña cuanto más cierta sea la hipótesis de independencia, donde:

$$MI(X,Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$
$$MI(X,Y|C) = \sum_{x,y} p(x,y|c) log \frac{p(x,y|c)}{p(x|c)p(y|c)} \quad (7.4)$$

De acuerdo con esta idea, se han realizado dos experimentos para obtener una estimación cuantitativa de este problema considerando el ejemplo de las 100 localidades en la Península Ibérica descrito en la Sec. 5.2.

En primer lugar, con objeto de evaluar el impacto de la calidad del predictor en los resultados, en la Fig. 7.6 se comparan las informaciones mutuas condicionadas para dos predictores diferentes: uno con 10 clases, C^{10}, que sólo captura variabilidad correspondiente a la escala sinóptica (las diferentes escalas se describen en la Sec. 2.1.2), y otro con 100 clases, C^{100}, cuya variabilidad es mesoescalar. Esta figura muestra que cuanto mejor es el predictor (es decir, cuanto más se aproxima al estado real de la atmósfera), más apropiado es el modelo de red ingenua; por tanto, el aumento de resolución en los modelos de predicción numérica favorece la competitividad del modelo ingenuo. Nótese que la hipótesis del modelo ingenuo se cumpliría de forma perfecta si la información mútua condicionada fuese nula; sin embargo, se ve que para estaciones cercanas, persiste cierta dependencia entre las localidades. Por tanto, la hipótesis ingenua es tanto más correcta cuanto más lejanas estén las estaciones, confirmando que los modelos ingenuos son una opción acertada

para estimar la probabilidad conjunta en sistemas cuyos predictandos tienen baja resolución. Hay que tener en cuenta que la resolución de una red de predictandos depende de la escala del evento; por ejemplo una red de 30km, se considera de alta resolución para el evento $prec > 0mm$, y al mismo tiempo de baja resolución para el evento *tormenta*. Esto explica en buena parte el éxito de los métodos de análogos en redes de baja resolución.

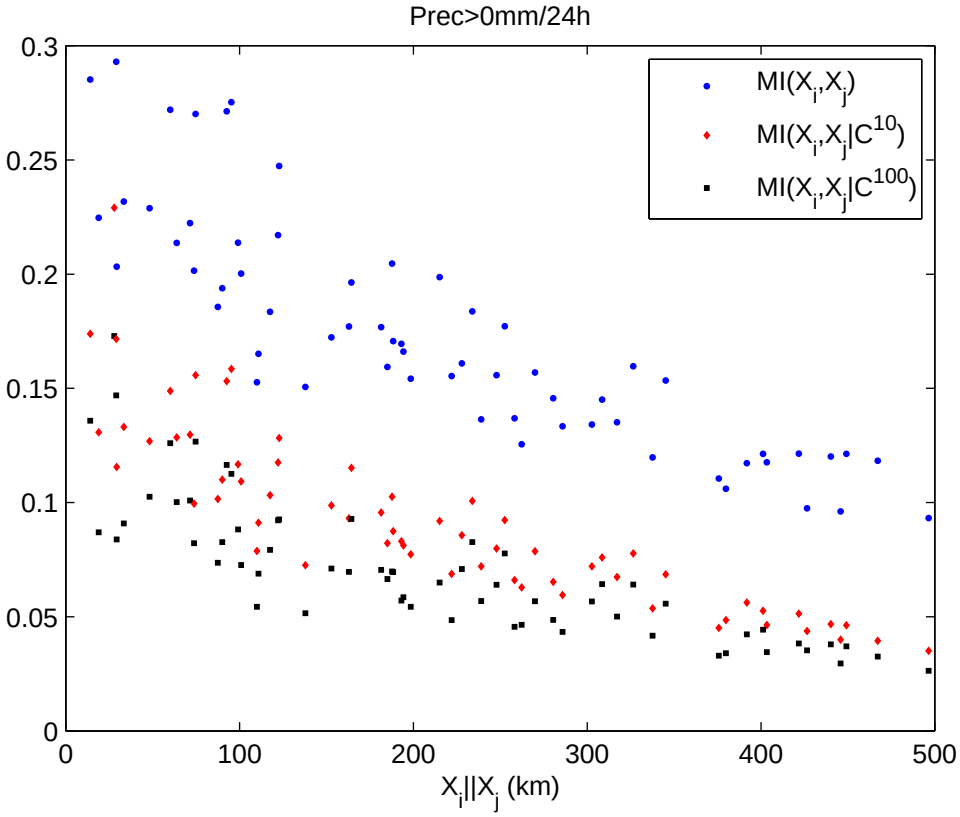

Figura 7.6: Información mútua en función de la distancia entre localidades, bajo tres supuestos, en azul sin condionar, en rojo condicionada a un patrón atmosférico de 10 clases y en negro condicionada a un patrón atmosférico de 100 clases.

En segundo lugar, para evaluar el impacto de la escala del evento sobre la validez de la hipótesis ingenua, la Fig. 7.7 muestra la información mutua condicionada en función de la distancia entre localidades para los eventos $prec > 0mm/24h$ y $prec > 10mm/24h$ (en este caso, solo para el predictor C^{100}). Esta figura muestra que el modelo ingenuo es más adecuado para eventos de escala menor, como el de $prec > 10mm/24h$, cuya dependencia espacial queda pobremente capturada en las observaciones.

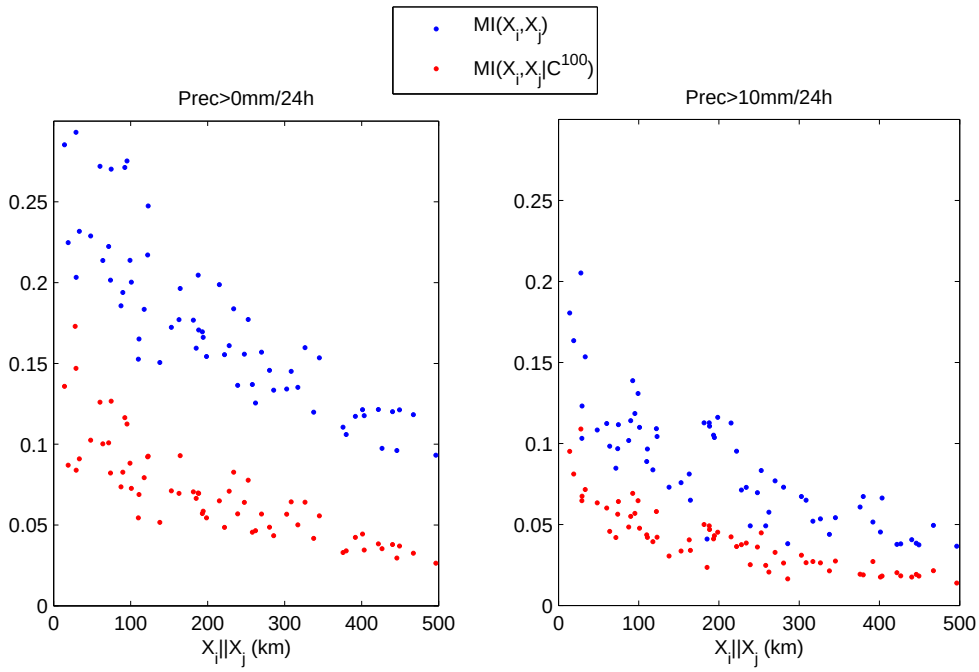

Figura 7.7: Representación de la información mútua condicionada al patrón atmosférico en función de la distancia. A la izquierda el evento $prec > 0mm/24h$ y a la derecha el evento $prec > 10mm/24h$. Se utiliza un patrón atmosférico agrupado en 100 tipos de tiempo.

Se puede decir, por tanto, que el modelo ingenuo es la mejor opción cuando la escala del fenómeno escapa a la resolución de la red de observación; por ejemplo, si la resolución de la red de observación es baja para estudiar tormentas, ningún modelo predice, ni individual ni colectivamente mejor que el modelo ingenuo, y la única forma de que otros modelos puedan mejorar, es aumentando la resolución de la red de observación. Estas ideas se resumen en la Fig. 7.8, donde se puede ver que sólo tiene sentido plantearse redes con dependencia espacial cuando la escala del evento es resuelta por la observación pero no por el modelo de circulación atmosférica.

7.3.2. *Predicciones Marginales y Conjuntas*

Normalmente se está interesado en calcular la probabilidad de cada localidad, condicionada al tipo de tiempo previsto $p(x_i|c)$ sin la necesidad de que las predic-

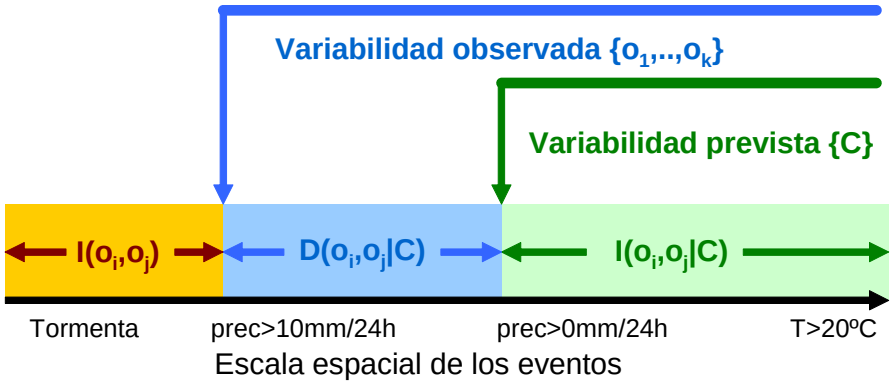

Figura 7.8: Para eventos cuya escala es inferior a la de la observación (típicamente por debajo de 10km), como por ejemplo una tormenta, se cumple la hipótesis *ingenua*: $I(o_i, o_j)$. Lo mismo ocurre para eventos cuya escala es superior a la que resuelve el predictor, ya que éste se puede considerar perfecto en esas escalas (típicamente por encima de 50km).

ciones de las distintas localidades sean consistentes. En este caso, como se verá más adelante, el modelo ingenuo es muy eficiente, aunque no se cumpla la hipótesis de independencia necesaria. Sin embargo, si se desea realizar una predicción conjunta de varias localidades $p(x_1, ..., x_k|c)$, entonces los modelos ingenuos son muy sensibles a la validez de la hipótesis de independencia, y solo en el caso de eventos cuya escala es resuelta por el modelo de circulación, resultan apropiados a tal efecto, mientras que para aquellos eventos cuya variabilidad es resuelta por la red de observación y no por el modelo de circulación, generan una predicción conjunta $p(\mathbf{x}|c)$ incoherente con la estructura espacial del evento. Por ejemplo, en la Fig. 7.5 se ilustra como el modelo ingenuo no es capaz de modelizar correctamente la distribución conjunta de la precipitación en dos localidades cercanas. Para evaluar esta sensibilidad, se ha tomado una pareja de estaciones muy próximas, como son *San Sebastián* (SS) y *Fuenterrabía* (FU), analizadas en un ejemplo anterior, y otra pareja de estaciones alejadas, como son *San Sebastian* (SS) y *Oviedo* (OV). Sobre ellas, se ha aplicado un modelo ingenuo suponiendo dos evidencias distintas: en el tipo de tiempo (paradigma de downscaling), y en el tipo de tiempo y la localidad más próxima (paradigma de predicción).

Los resultados obtenidos en el caso de *downscaling*, para el evento $prec > 10mm/24h$, se muestran en la Fig. 7.9, donde se puede ver que la diferencia entre el modelo ingenuo y el general es mayor cuanto más cercanas sean las estaciones; además, el modelo ingenuo sobreestima la ocurrencia de eventos contrarios y subestima las coincidencias. La magnitud de esta diferencia depende del tipo de tiempo;

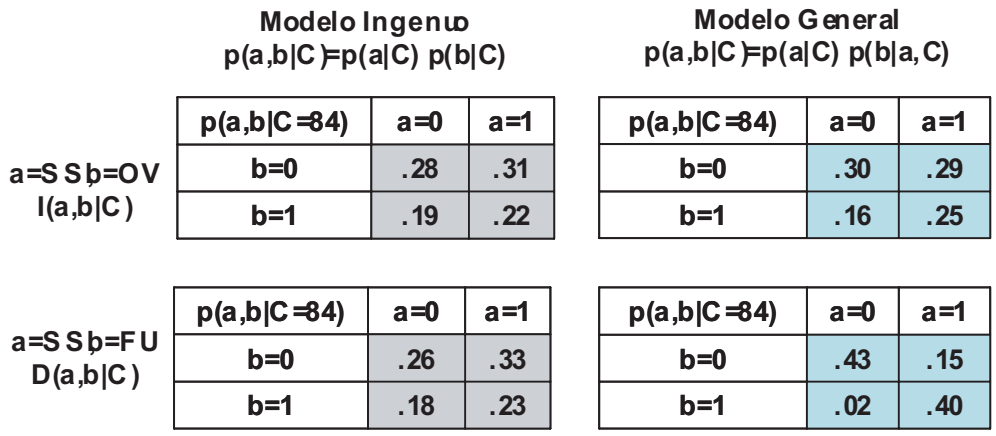

Figura 7.9: El error sistemático (bias) del modelo ingenuo frente al general, para el problema de downscaling. Ejemplo realizado para $c = 84$.

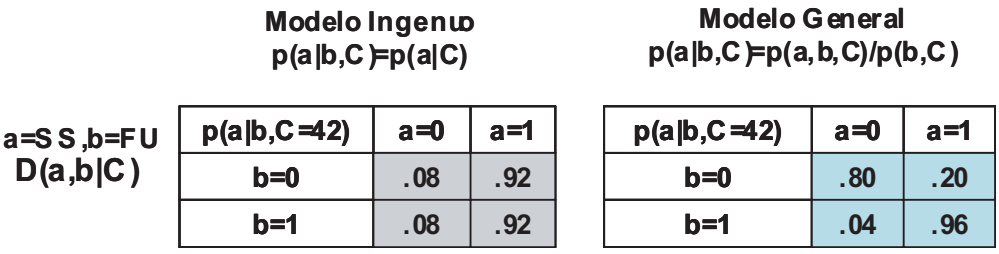

Figura 7.10: Error sistemático para el caso de predicción.

en el ejemplo se ha buscado uno de los casos donde las diferencias son mayores: $c = 84$.

La Fig. 7.10 muestra que, para el caso de predicción, las diferencias son mayores llegando a alcanzar sesgos relativos del 80 % para algunos tipos de tiempo. En este ejemplo se ha considerado $c = 42$.

Para concluir esta sección, se puede afirmar que los resultados presentados justifican la utilización de modelos más complejos que el ingenuo, ya que son numerosos los casos en los que claramente no se cumple la hipótesis de independencia requerida. Por otra parte, cuando se utiliza el modelo ingenuo para estimar probabilidades conjuntas, se comete un error sistemático que, en los ejemplos analizados, subestima las coincidencias; este error es muy sensible a la hipótesis de independencia.

7.3.3. Observaciones Disponibles y Sobreajuste

A pesar de trabajar con predictandos que representan variables binarias, el elevado número de variables de un problema práctico (cientos de localidades), obliga a los modelos a ajustar un elevado número de parámetros. Aunque los efectos del sobreajuste dependen de la longitud de la serie disponible para realizar el entrenamiento (el registro histórico disponible) y del número de parámetros, es muy útil tener una idea cuantitativa de este efecto. Para ello, se realizó un experimento donde se estudia el comportamiento del modelo ingenuo para diferentes longitudes del conjunto de entrenamiento considerando dos tipos de predictor: un predictor 'pobre' con 10 clases, C^{10}, y otro, de mayor resolución, con 100 clases, C^{100} (este tamaño se ha obtenido de forma empírica como el número de clases óptimo para maximizar la capacidad predictiva).

Los resultados de este experimento se muestran en la Fig. 7.11, donde se ha representado un *boxplot* de los RSA obtenidos por el modelo ingenuo para un conjunto de test de 5 años, considerando conjuntos de entrenamiento de distinta longitud, entre 1 y 18 años. El experimento se ha repetido 10 veces, para 10 conjuntos de entrenamiento y de test diferentes, considerando el evento $prec > 0mm/24h$, con los dos predictores anteriormente mencionados. Esta figura muestra que la capacidad predictiva del modelo con 100 tipos de tiempo es claramente superior a la del modelo con 10 estados, para casi todas las longitudes de entrenamiento; esto confirma que el número de clases del predictor óptimo crece con el conjunto de entrenamiento, hasta alcanzar la saturación. En segundo lugar, se constata la calidad del predictor decrece de forma esponencial para longitudes inferiores a 10 años; el predictor con menos estados tiene una curva de pérdida de calidad mucho menos acusada, ya que ajusta menor número de parámetros. Este experimento se ha realizado también para el evento $prec > 10mm/24h$, con resultados muy similares; solo cabe añadir que cuanto más raro es el evento, más largas han de ser las serires de entrenamiento necesarias para ajustar los parámetros. Finalmente, hay que indicar que estos resultados se han obtenido para métodos de predicción de precipitación diaria en la Cornisa Cantábrica. Resumiendo, en este ejemplo un modelo ingenuo debe disponer de registros históricos diarios de al menos 10 años para evitar el sobreajuste, y de 15 a 20 años para alcanzar la calidad óptima.

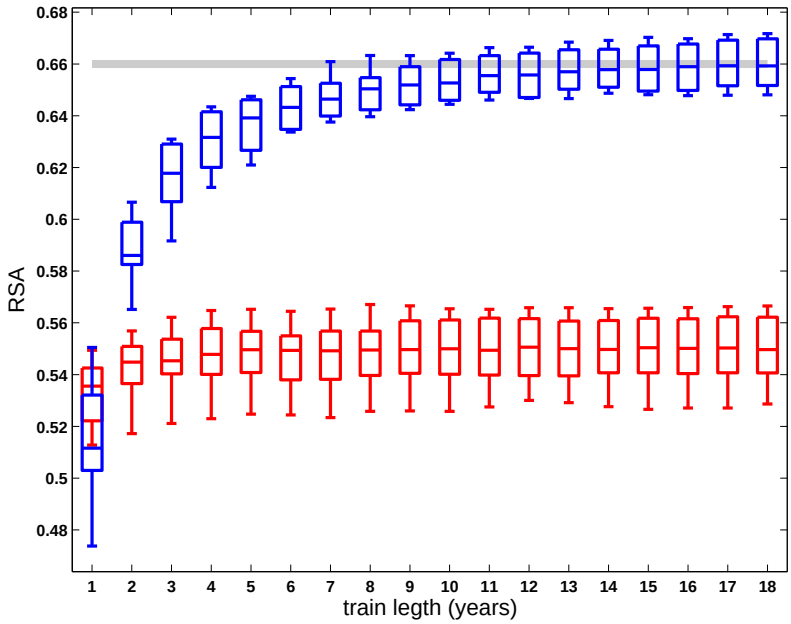

Figura 7.11: Evolución de la capacidad predictiva del modelo ingenuo para diferentes longitudes del conjunto de entrenamiento. La línea gris sirve como referencia para ilustrar la pérdida de pericia al reducir el tamaño del conjunto de entrenamiento, para el predictor C^{100}. El color azul corresponde al predictor C^{100} y el rojo al C^{10}. El boxplot se refiere a 10 realizaciones sobre diferentes conjuntos de entrenamiento-test, para la longitud dada.

CAPÍTULO 8

Algoritmos de Aprendizaje para la Predicción Local

En los ejemplos de modelos basados en Redes Bayesianas (RBs) mostrados en el Capítulo 6, se ha utilizado la metodología estándar de aprendizaje (descrita en la Sec. 3.5) para obtener las redes Bayesianas a partir de los datos; sin embargo, para analizar de forma óptima algunas de estas aplicaciones, como la predicción local de eventos de precipitación, se requieren métodos de aprendizaje específicos, que tengan en cuenta las particularidades del problema (Cano et al., 2004; Gámez and Puerta, 2005). Estos métodos trabajan con un elevado número de variables y tienen un conjunto de entrenamiento relativamente reducido, por lo que han de restringir el número de enlaces permitidos en el grafo, reduciendo simultáneamente el número de parámetros y el coste computacional. En este capítulo se proponen soluciones a este problema para el caso de algoritmos de búsqueda, que son los únicos que se consideran en esta Tesis.

La organización de contenidos de este capítulo es la siguiente. En primer lugar, en la Sec. 8.1, se analizan las posibilidades y limitaciones de los algoritmos de búsqueda estándar y se introducen soluciones en base a algoritmos de búsqueda local, como *localK2* y *localB*, que se utilizarán posteriormente en esta Tesis. Finalmente, en la Sec. 8.2 se trata el problema de la elección de una medida de calidad adecuada para la predicción local.

8.1. Algoritmos de Búsqueda en la Predicción Local

El principal problema del aprendizaje estructural basado en algoritmos de búsqueda se encuentra en el propio proceso de búsqueda ya que, en el caso general, se trata de un problema *NP-completo*. Las distintas simplificaciones que han sido propuestas para abordar este problema, como los algoritmos voraces $K2$ o B (ver Sec. 3.5), tienen complejidades cuadrática y cúbica en el número de nodos, respectivamente, por lo que reducen drásticamente la complejidad del problema, a costa de obtener una red de "buena calidad" renunciando a obtener la óptima. Como se vio en la Sec. 3.5.2 el problema del algoritmo $K2$ es la necesidad de imponer una pre-ordenación en las variables, mientras que el problema del algoritmo B es que resulta prohibitivo en casos con un elevado número de variables (cientos o miles), por lo que no es apropiado para el problema de predicción local que se aborda en esta Tesis.

En la Sec. 8.1.1 se analizan algunas de las deficiencias de los algoritmos de búsqueda con pre-ordenación y en la Sec. 8.1.2 se propone una modificación de los algoritmos de búsqueda que permite reducir un orden su complejidad, aprovechando las características particulares de la predicción local en meteorología, que permiten reducir el proceso de búsqueda a un vecindario (geográfico) de cada nodo.

8.1.1. Deficiencias de los Algoritmos con Pre-ordenación

Una de las características más importantes del algoritmo $K2$ es la pre-ordenación de las variables, de manera que sólo se permiten enlaces en un sentido de la ordenación. Esta restricción reduce el tiempo de búsqueda y también garantiza que no se formen ciclos en la red (ver Sec. 3.5.2). Sin embargo, en algunas aplicaciones, esta pre-ordenación es arbitraria y produce efectos secundarios que limitan considerablemente la utilidad de estos algoritmos. Este es el caso de la predicción local, donde la ordenación de las distintas localidades no tiene relación directa con el problema.

Estructuras Inverosímiles

En algunas aplicaciones prácticas existe un cierto grado de dependencia entre todas las variables del problema; por ejemplo, en el caso de la predicción local de precipitación, las localidades cercanas pueden tener un elevado grado de dependencia. En estos casos, dada una ordenación previa arbitraria de las variables, aquellas que eventualmente son colocadas en las primeras posiciones tienen más posibilidades de tener hijos que las demás, creando estructuras que además de resultar inverosímiles, tienen baja capacidad predictiva. Por ejemplo, en la Fig. 8.1 se mues-

tra el gran número de enlaces creados desde el nodo que ocupa eventualmente el primer lugar cuando se aplica el algoritmo *K2* al evento $prec > 0mm/24h$ sobre el dominio de la Península Ibérica descrito en la Sec. 5.1. Aunque inicialmente los enlaces, uno por uno, mejoran incrementalmente la calidad del grafo, muchos de ellos (los de largo alcance) serán prescindibles en una etapa posterior del algoritmo de búsqueda. Además, este hecho provoca que se obtengan grafos muy diferentes para distintas ordenaciones previas, lo que dificulta la interpretación del modelo resultante por parte del usuario (el meteorólogo en este caso).

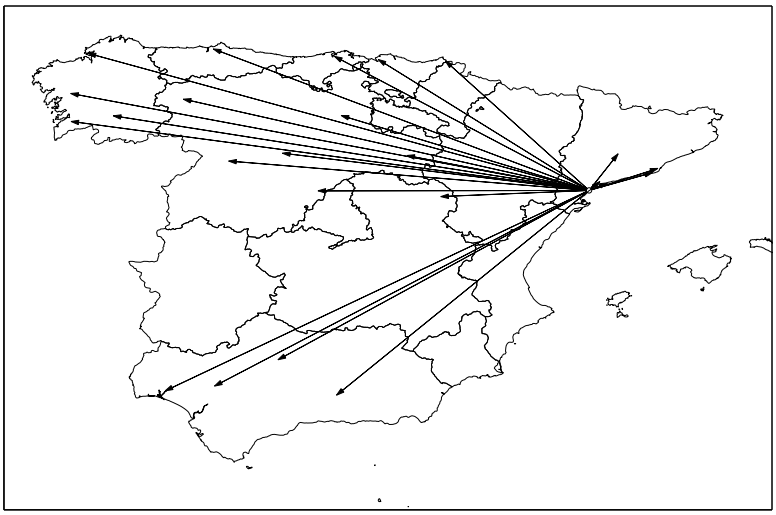

Figura 8.1: Ilustración del efecto de la pre-ordenación en la estructura espacial del grafo resultante. Algoritmo *K2* aplicado a un ordenamiento concreto de las variables: El primer nodo, en este caso Tarragona, es padre de 22 localidades.

Este problema podría ser solventado imponiendo como restricción adicional al algoritmo un número máximo de padres para cada nodo. De esta manera, los nodos ordenados en primeras posiciones no podrían enlazarse arbitrariamente con el resto de nodos del grafo. Sin embargo, como se verá más adelante, existen otras soluciones más eficientes para este problema.

Pérdida de Información

Otro efecto secundario de la pre-ordenación es el deterioro que puede sufrir la información cuando se propaga por caminos inadecuados. Aunque este no es un defecto exclusivo del algoritmo *K2*, la pre-ordenación lo magnifica al colocar desde el inicio enlaces poco realistas, como son los enlaces de largo alcance mostrados en la Fig. 8.1.

Este efecto se manifiesta con claridad en el "ejemplo de Parayas" descrito en el Ejemplo 3.2. Los grafos de las Fig. 3.6 y 3.7 muestran las estructuras aprendidas aplicando los algoritmos $K2$ y B, respectivamente, junto con las probabilidades condicionadas a la ocurrencia del evento *tormenta*. En el primero de los casos se tiene $p_{K2}(lluvia|tormenta) = 0.365$, mientras que el valor real (obtenido directamente de los datos) es 0.763, que coincide exactamente con el valor del segundo caso. La Tabla 8.1 muestra una comparación de las probabilidades obtenidas para distintos eventos y evidencias (esta tabla es una extensión de la Tabla 3.5.2); en esta tabla se han destacado en negrita los valores más discrepantes en las probabilidades dadas por el modelo $K2$, respecto al valor real. Esta discrepancia se debe a la pérdida de información en el grafo generado por *K2* ya que, debido exigencias de la ordenación, el flujo de información entre las variables tormenta y lluvia no es directo, como se puede comprobar en el grafo.

	Ll	Nv	Gr	Ro	Es
$p(x\|Tr)$	0.763	0.011	0.171	0.103	0.020
$p_B(x\|Tr)$	0.763	0.001	0.171	0.078	0.013
$p_{K2}(x\|Tr)$	**0.365**	0.029	0.169	0.117	0.020
$p(x\|Nb)$	0.401	0.003	0.015	0.299	0.019
$p_B(x\|Nb)$	0.401	0.003	0.013	0.299	0.020
$p_{K2}(x\|Nb)$	0.406	0.004	**0.001**	0.298	0.044
$p(x\|Tr, Nb)$	0.739	0.011	0.171	0.242	0.019
$p_B(x\|Tr, Nb)$	0.767	0.001	0.156	0.191	0.013
$p_{K2}(x\|Tr, Nb)$	**0.390**	0.028	**0.042**	0.286	0.044

Tabla 8.1: Probabilidades condicionadas para distintas evidencias y variables (Lluvia, Nieve, Granizo, Rocío, Escarcha) obtenidas a partir de los datos y de dos modelos distintos: $K2$ y B.

8.1.2. Algoritmos de Búsqueda Local

Una de las maneras más simples de reducir la complejidad de los algoritmos de búsqueda es limitar el número de hijos potenciales de cada nodo. En el algoritmo $K2$ todos los nodos con numeración posterior a uno dado pueden ser potencialmente hijos suyos, mientras que en el algoritmo B no existen restricciones. Para ello, es necesario que la dependencia entre las variables del problema tenga cierta estructura local, que permita definir el número de vecinos potenciales, k, del total de los n nodos que sea razonable considerar. Cuando $k \ll n$ entonces la complejidad de los

algoritmos anteriores se podrá reducir en un orden de magnitud, ya que la búsqueda de hijos se hará sobre un número constante de nodos. De esta manera se pueden resolver los problemas del algoritmo $K2$, haciéndole de paso más eficiente, y también se puede hacer que el algoritmo B resulte aplicable a problemas con un elevado número de variables (ahora sería cuadrático).

El algoritmo "Sparse Candidate"

El primer algoritmo en proponer esta idea fue el algoritmo *Sparse Candidate* de Friedman et al. (1999), que es una extensión del algoritmo de *Chow-Liu* para árboles (Chow and Liu, 1968). Se trata de un algoritmo de búsqueda local ya que establece la limitación del número de padres potenciales, que han de ser recalculados cada vez que se añade un enlace a la red, como se justifica en la Fig. 8.2, donde se puede ver que en el paso 1, el enlace que aporta mayor MI es el $C \rightarrow A$. En cambio, en el paso 2, aunque previamente la información mútua entre D y A es mayor: $MI(A, D) > MI(A, B)$, una vez establecido el enlace $C \rightarrow A$ resulta: $MI(A, B|C) > MI(A, D|C)$, por lo que el enlace que aporta mayor MI en el paso 2 es el $B \rightarrow A$.

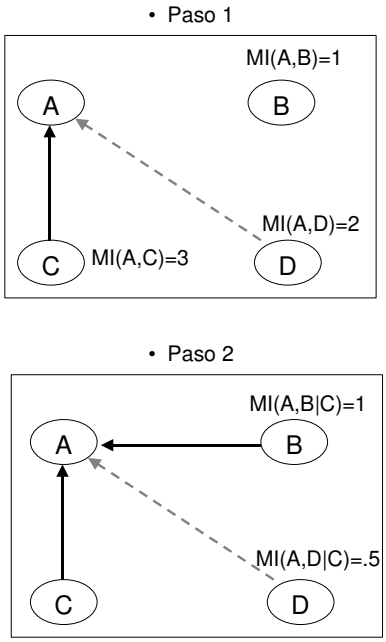

Figura 8.2: Ilustración de la necesidad de actualizar el conjunto de padres potenciales.

Este algoritmo está estructurado en dos partes: por un lado la *parte restrictiva*, donde se hace la selección de candidatos, y por otro la *parte de optimización*, que es donde se hace la búsqueda en ese espacio reducido. Utiliza una medida de calidad

descomponible, es decir que cumple la propiedad

$$Q(B) = \sum_i Q(x_i|\pi_i), \tag{8.1}$$

donde la medida de calidad de cada red $Q(B)$ se puede obtener como suma de las calidades parciales de cada nodo X_i dados sus padres Π_i en dicha red; de esta manera, cada vez que se añade un nuevo enlace solo es necesario recalcular los elementos del sumatorio cuyos padres cambian.

La medida de calidad empleada en este algoritmo se obtiene a partir de la información mutua entre cada pareja de variables:

$$MI(X,Y) = \sum_{x,y} \hat{P}(x,y) log \frac{\hat{P}(x,y)}{\hat{P}(x)\hat{P}(y)} = D_{kl}(\hat{P}(x,y), \hat{P}(x)\hat{P}(y)) \tag{8.2}$$

donde $\hat{P}$ son frecuencias observadas en la base de datos y D_{kl} es la *distancia de Kullback-Leibler* entre dos distribuciones R y S, que se define como:

$$D_{kl}(R(x), S(x)) = \sum_x R(x) log \frac{R(x)}{S(x)} \tag{8.3}$$

La información mútua puede ser interpretada como el error cometido al calcular la probabilidad conjunta suponiendo que X e Y son independientes. Así para cada red B, se puede definir una medida de discrepancia entre P_B y $\hat{P}$:

$$M_{disc}(X_i, X_j|B) = D_{kl}(\hat{P}(x_i, x_j), P_B(x_i, x_j)) \tag{8.4}$$

que es la utilizada como medida de calidad en este algoritmo, según la cual, la mejor red será aquella cuyo valor de M_{disc} sea más pequeño.

Cada vez que se crea un enlace, la medida de discrepancia debe ser actualizada, como se muestra en la Fig. 8.2. Además, esta medida requiere calcular $P_B(x_i, x_j)$ en cada paso de iteración para cada $X_j \neq X_i$ con $X_j \notin \Pi_i$. Este proceso puede llegar a ser muy costoso en el caso de grandes conjuntos de datos, por lo que en el algoritmo *Sparse Candidate*, para ganar eficiencia computacional, se aproximan los cálculos sobre una muestra reducida del total de realizaciones disponibles.

Algoritmo localK2 y localB

En algunos casos, el conjunto de candidatos puede ser pre-establecido a priori siguiendo algún criterio específico del problema que se trate, por lo que se podría evitar la tarea de recalcular los candidatos de cada nodo en cada paso del algoritmo

sparse candidate, que supone un alto coste computacional para el método. Por ejemplo, en el problema de la predicción local de precipitación, las variables corresponden a mediciones u ocurrencias en puntos geográficos distintos de una misma variable; por tanto, se podría considerar el criterio de similitud climatológica (o proximidad geográfica) para establecer el conjunto de padres candidatos de cada nodo, estableciendo una distancia máxima a partir de la cual los enlaces desde las localidades fuera de esta vecindad no aportarían capacidad predictiva al modelo, ni marginal ni condicionalmente. De esta manera no sería necesario re-calcular el conjunto de padres candidatos en cada paso del algoritmo de búsqueda. Este umbral de distancia podría ser establecido utilizando alguna medida de dependencia, como la correlación o la información mutua. También se podría aplicar algún criterio de vecindad local basada en algoritmos de agrupamiento, pero esta opción no se ha explorado en la Tesis.

Esta estrategia para mejorar la eficiencia del proceso de búsqueda, se aplica en los algoritmos denominados *localK2* y *localB* (Cano et al., 2004), donde el conjunto de candidatos es calculado una sola vez. Los algoritmos 3 y 4 muestran los detalles de ambas estrategias, devolviendo un grafo dirigido acíclico que representa las dependencias existentes en el conjunto de datos $\mathbf{D}$ considerando un número de padres potenciales k.

Algorithm 3 Algoritmo *localK2* para aprendizaje estructural de RBs.

Require: Una muestra de datos de un conjunto previamente ordenado de nodos $\mathbf{D} = \{\mathbf{X_1}, \ldots, \mathbf{X_n}\}$, una matriz con los k padres potenciales de cada nodo $\mathbf{V} = \{\mathbf{v_1}, \ldots, \mathbf{v_n}\}$ y un grafo disconexo B.

```
1: for i = 1 : n do
2:    S = score(D, B);
3:    for j = 1 : k do
4:       M = score(D, B ∪ {i → v_i(j)});
5:       if M > S then
            S = M; B(i, v_i(j)) = 1;
6:       end if
7:    end for
8: end for
```

Ensure: Un grafo dirigido acíclico B y su medida de calidad S.

El criterio de selección de padres potenciales (proximidad geográfica) ha sido desarrollado expresamente para resolver el problema de la predicción local de precipitación. En la Fig. 8.3 se muestra que la información mutua entre la precipitación de distintas localidades decrece rápidamente con la distancia entre éstas, siendo

Algorithm 4 Algoritmo *localB* para aprendizaje estructural de RBs.

Require: Una muestra de datos de un conjunto de nodos $\mathbf{D} = \{\mathbf{X_1}, \ldots, \mathbf{X_n}\}$, una matriz con los k-vecinos de cada nodo $\mathbf{V} = \{\mathbf{v_1}, \ldots, \mathbf{v_n}\}$ y un grafo disconexo B.

```
1: for i = 1 : n do
2:    S = score(D, B);
3:    for j = 1 : k do
4:       if acyclic(B ∪ {i → v_i(j)}) = 1 then
5:          M = score(D, B ∪ {i → v_i(j)});
6:          if M > S then
               S = M; B(i, v_i(j)) = 1;
7:          end if
8:       end if
9:    end for
10: end for
```

Ensure: Un grafo dirigido acíclico B y su medida de calidad S.

este comportamiento independiente de, por ejemplo, el estado de la atmósfera, o de otras localidades. Esta figura muestra que aunque el condicionamiento al estado de la atmósfera C altera las dependencias, no modifica el patrón de decrecimiento. Estos resultados sugieren que los *padres potenciales* pueden fijarse desde un principio en base a la proximidad geográfica, o a la información mutua marginal, sin afectar en gran medida al resultado del algoritmo.

La Fig. 8.4 compara el resultado del algoritmo estándar *K2* frente al local *localK2(10)*, con 10 padres potenciales para cada nodo, para modelizar las dependencias de la precipitación en el dominio de la Península Ibérica definido en la Sec. 5.1. Como ya se ha comentado, el algoritmo $K2$ genera un grafo diferente para cada ordenación y, además, genera enlaces de largo alcance que tienen difícil interpretación y escasa capacidad predictiva. En cambio, en el caso de $localK2(10)$ se ha constatado (aunque no se muestran figuras) que las diferentes ordenaciones generan grafos muy similares y, aunque su calidad es ligeramente inferior a la del grafo obtenido con *K2*, reúne cualidades, como una notable reducción del tiempo de cómputo y una mejor interpretación meteorológica del grafo, que justifican su aplicación.

Por otra parte, el algoritmo local reduce notablemente la complejidad del proceso de búsqueda pues, mientras que el algoritmo *K2* debe evaluar un promedio de $(n-1)/2$ candidatos por enlace, *localK2* debe evaluar tan sólo $(k-1)/2$, donde k es el número máximo de *padres potenciales* que se quieren considerar. En problemas con muchos nodos, $k << n$, por lo que la reducción en el tiempo de cómputo equivale a un orden de magnitud. De esta manera, el algoritmo *localK2* es de complejidad

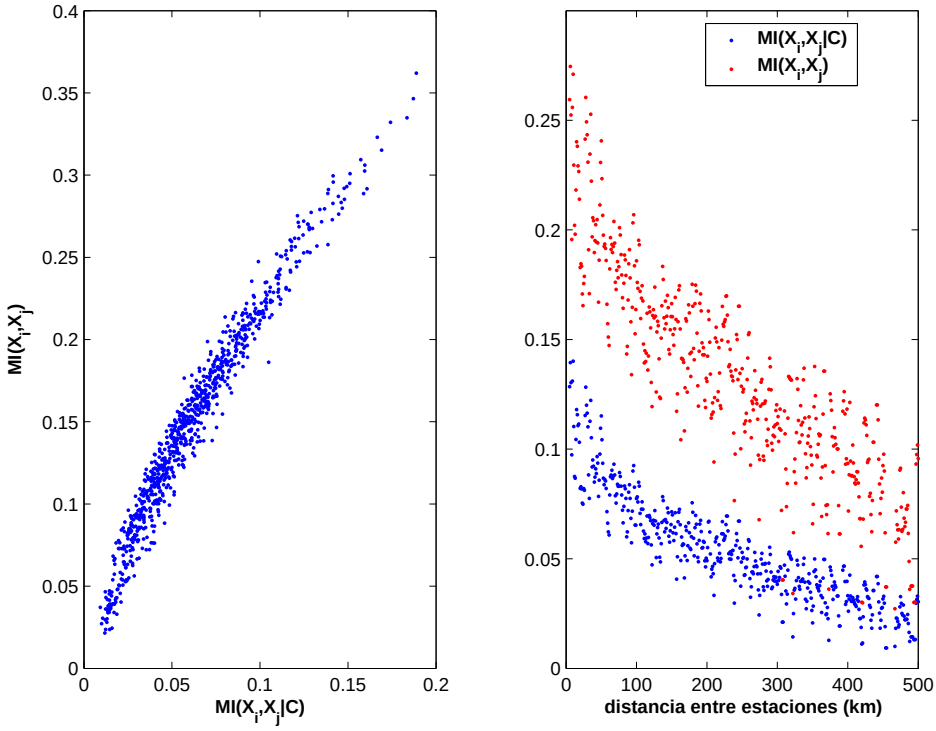

Figura 8.3: (izquierda) Información mutua de pares de localidades en función de su distancia para el evento binario $prec > 0mm/24h$ en las 42 estaciones de la Cuenca Norte (son $42 * 41/2$ pares); se muestran tanto los valores a priori, como los condicionados al estado de la atmósfera C^{100}. (derecha) Comparación entre los valores de información mutua a priori y la condicionada al estado de la atmósfera.

lineal, mientras que *localB* es de complejidad cuadrática, lo cual lo hace equivalente al *K2* estándar, pero sin los inconvenientes derivados de la pre-ordenación. En la Fig. 8.5 se muestra la evolución del tiempo de CPU del proceso de aprendizaje estructural en función del número de nodos, para tres tipos diferentes de algoritmos de aprendizaje: $K2$ estándar, $localK2(5)$ y $localK2(10)$, con el número de *padres potenciales* indicado entre paréntesis.

Por último, en la Figura 8.6 se compara la medida de calidad de *K2* con *localK2(k)* para diferente número de padres potenciales k, también para el dominio de la Península Ibérica. En esta figura se puede apreciar la convergencia exponencial de la calidad de *localK2(k)* hacia la de *K2* a medida que k crece. Esta figura indica que con un reducido número de *padres potenciales* se pueden obtener resultados muy similares al algoritmo *K2*; en este ejemplo concreto se ve que tomando 10 padres potenciales, se alcanza prácticamente la misma calidad que tomándolos todos.

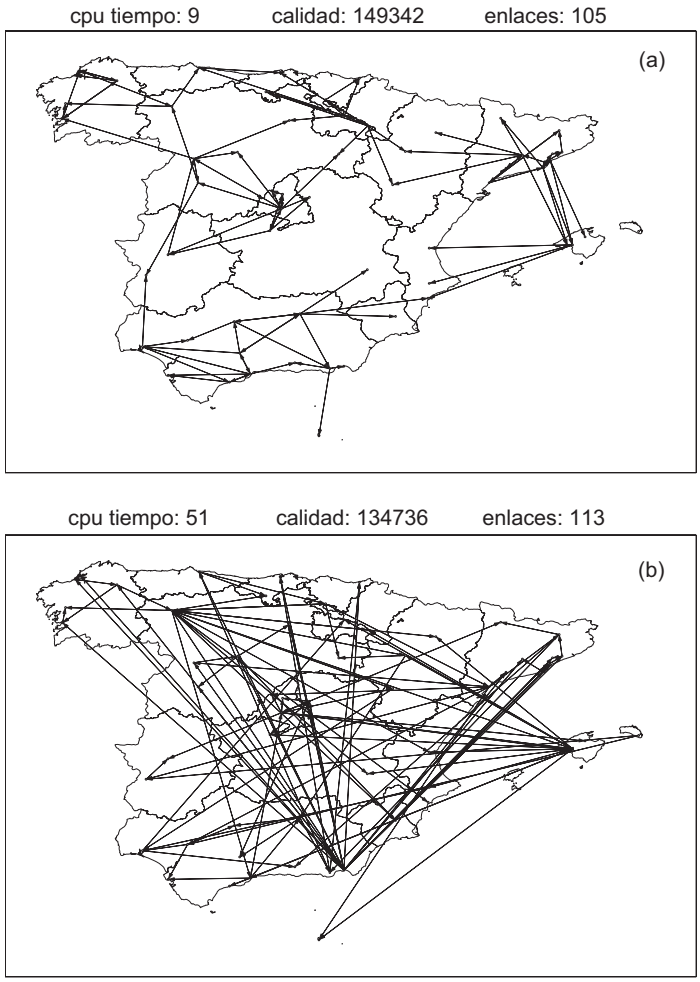

Figura 8.4: Grafos para precipitación obtenidos aplicando el algoritmo (a) *localK2(10)* y (b) *K2*; el máximo número de padres se ha fijado en dos. El tiempo de ejecución, la medida de calidad de cada grafo (MDL, cambiada de signo) y el número total de enlaces que contiene se muestran para cada grafo.

8.2. Medidas Específicas de Calidad

Los algoritmos de búsqueda local desarrollados en la sección anterior son eficientes para trabajar con un elevado número de variables, lo que facilita su aplicación en el problema de la predicción meteorológica local; sin embargo, para que estos algoritmos estén realmente adaptados al problema de la predicción local de precipitación, es necesario modificar las medidas de calidad estándar, pues éstas evalúan la capacidad global de la red para reproducir todos los datos observados, y no la capacidad predictiva de la misma; es decir, es necesario definir medidas de calidad específicas que tengan en cuenta las particularidades del problema que se quiera

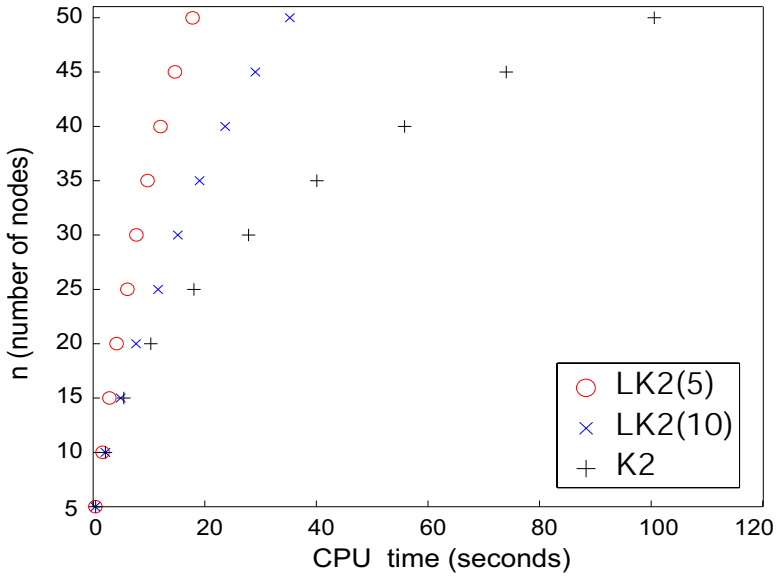

Figura 8.5: Número de nodos versus tiempo de CPU para tres algoritmos diferentes: *K2* estándar y *localK2* (*LK*2 en la figura) con 5 y 10 *padres potenciales*, respectivamente.

abordar. Por ejemplo, en Mach and Mráz (2005), se destaca lo complicado que es medir la calidad de un clasificador de forma objetiva utilizando métricas globales (como MDL) y se considera la necesidad de emplear métricas específicas para este problema; asimismo, Bradley (1997) utiliza el area bajo la curva ROC (AUC, ver Sec. 4.2) para evaluar el comportamiento de distintos métodos de clasificación, mostrando sus ventajas frente a otras medidas globales. Posteriormente Huang and Ling (2005) extienden estas ideas al caso de clasificadores ingenuos y extendidos TAN (ver Sec. 3.6), mostrando que medidas como AUC permiten discriminar mejor la capacidad clasificadora de los modelos.

En el caso de la predicción meteorológica local probabilística no existe una solución equivalente al caso de clasificación y, por tanto, el desarrollo de una medida de calidad específica es diferente. En la Sec. 8.2.3 se aborda este problema y se propone el uso de medidas específicas de calidad, alternativas a MDL y otras medidas de calidad global, como el RSA (análogo a AUC), en el proceso de aprendizaje de la estructura de la red. Es decir, se propone utilizar una medida distinta, que esté orientada exclusivamente a la capacidad predictiva del modelo. A continuación se analiza la medida de calidad global tanto para problemas de clasificación (ver Sec. 3.6) como para problemas de predicción.

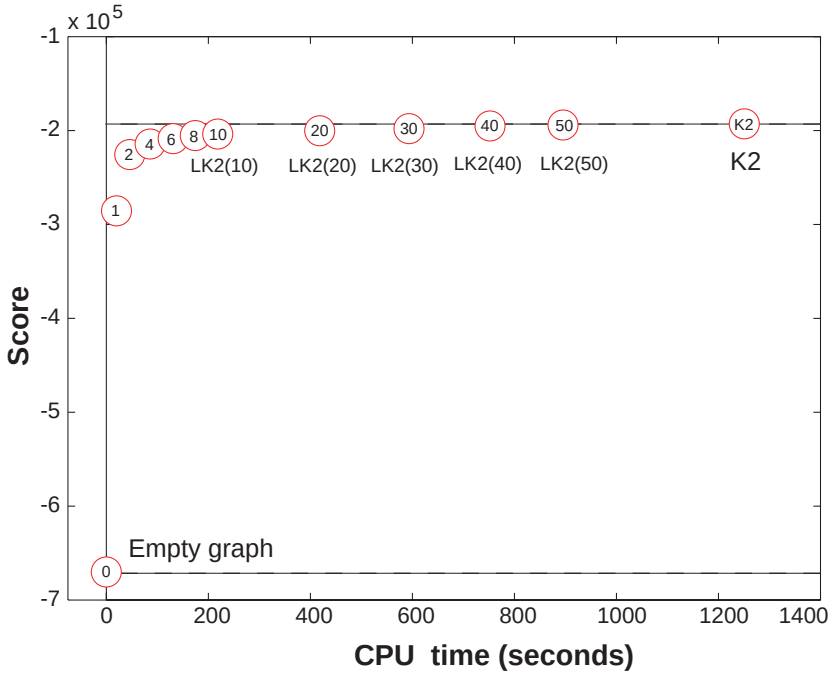

Figura 8.6: Comparación entre el algoritmo *K2* y varios *localK2(k)* para diferentes valores de k. La línea horizontal indica la medida de calidad obtenida con el algoritmo *K2*.

8.2.1. Efectos del Término de Penalización

Las métricas globales, como MDL, optimizan la calidad de la red para representar el conjunto de datos, lo cual incluye también al predictor (el estado de la atmósfera), que en nuestro caso es una variable con más estados que las demás (los predictandos son binarios, mientras que el predictor puede tener típicamente 100 estados, C^{100}, utilizando la notación establecida en 5.3). Esta inhomogeneidad en las variables tiene un efecto negativo, dado que los enlaces al predictor se penalizan en exceso a causa del primer término de la Ec. 3.11 (el término de penalización del número de parámetros), propiciando redes de muy baja capacidad predictiva con muy pocos o incluso ningún enlace al predictor.

Para ilustrar este problema, se ha realizado un experimento con el caso más simple, que se resume en la Fig. 8.7, donde se muestra el comportamiento de un algoritmo basado en MDL cuando compara dos enlaces como los mostrados en rojo en la figura. Este ejemplo corresponde a un problema de predicción local de precipitación ($precip > 0mm$) para $N = 9860$ días en dos localidades (A: San Sebastián, y B: Fuenterrabía), ambas con dos estados, considerando como predictor un estado de la atmósfera con 100 tipos de tiempo C^{100}, y utilizando la medida de calidad $MDL(B) = -r_B(C, a, b)log(N) - loglik_B(C, a, b)$.

El algoritmo elige el enlace que aporta el mayor valor de MDL, que resulta ser el del modelo B_1. Sin embargo, utilizando una medida estándar de validación, como RSA, se comprueba que la capacidad predictiva del modelo B_2 (en el nodo 'b') es claramente superior, debido a que aunque el enlace propuesto por el modelo B_2 aporta mayor verosimilitud, al mismo tiempo está excesivamente penalizado. Una forma de evitar este problema es eliminar el término de penalización de la ecuación 3.11, y limitar el número de padres por nodo para evitar el sobreajuste (Grossman and Domingos, 2004). Posteriormente, en la Sec. 9.1, se verán con algunos ejemplos las ventajas y defectos de esta consideración.

8.2.2. Medidas Específicas para Clasificación y Predicción

Como ya se ha mostrado en la Sec. 3.6, en un problema de clasificación, las variables se pueden dividir en clasificador (C) y atributos ($\mathbf{Y}$), por lo que la verosimilitud de cada realización i se puede descomponer como (Friedman et al., 1999):

$$\begin{aligned} loglik(D_i, B_{clas}) &= loglik(\mathbf{Y}_i, C_i, B_{clas}) \\ &= logP_{B_{clas}}(\mathbf{Y}_i) + logP_{B_{clas}}(C_i \mid \mathbf{Y}_i) \end{aligned} \tag{8.5}$$

por lo que la métrica de aprendizaje se puede expresar como:

$$\begin{aligned} MDL(D, B_{clas}) &= -r_{B_{clas}} * log(N/2) \\ -\sum_{i=1}^{N} logP_{B_{clas}}(\mathbf{Y}_i) &- \sum_{i=1}^{N} logP_{B_{clas}}(C_i \mid \mathbf{Y}_i). \end{aligned} \tag{8.6}$$

En la verificación del clasificador solamente interesa el último término, correspondiente al logaritmo de la verosimilitud condicional, de manera que es muy sencillo construir un MDL condicionado (CMDL) para unificar ambas métricas y así obtener una mejor validación:

$$CMDL(D, B_{clas}) = -r_{B_{clas}} * log(N/2) - \sum_{i=1}^{N} logP_{B_{clas}}(C_i \mid \mathbf{Y}_i), \tag{8.7}$$

donde se ha suprimido el término $\sum_{i=1}^{N} logP_{B_{clas}}(\mathbf{Y}_i)$, por ser independiente de la calidad del clasificador. Sin embargo, esta métrica tiene el problema de no ser descomponible, resultando su cálculo más costoso que el de MDL (Grossman and Domingos, 2004).

En cambio, para el caso de la predicción local (problema inverso al de clasificación) resulta:

$$\begin{aligned} loglik(D_i, B_{down}) &= loglik(\mathbf{Y}_i, C_i, B_{down}) \\ &= logP(C_i) + logP_{B_{down}}(\mathbf{Y}_i \mid C_i). \end{aligned} \tag{8.8}$$

donde el primer sumando de la ecuación anterior es constante respecto a la estructura de la red y por lo tanto, en este caso, el logaritmo de la verosimilitud es, salvo una constante, una medida de la capacidad predictiva conjunta.

8.2.3. *Capacidad Predictiva Marginal: Medida* RSA

El paradigma de predicción local meteorológica más utilizado es el de *downscaling*, en el cual la única evidencia es el estado de la atmósfera, C, que actúa como predictor y, además, es un nodo raíz del grafo (ver Cap. 7). Por tanto, interesa utilizar medidas de calidad que permitan a los predictandos recibir el máximo de información posible del predictor. Eliminando de MDL el término de penalización, queda el logaritmo de la verosimilitud *Loglik*. Esta métrica es, como ya hemos mencionado anteriomente, adecuada para evaluar la capacidad predictiva conjunta en el proceso de aprendizaje, $P(\mathbf{Y}|C)$; sin embargo, no resulta práctica para evaluar la capacidad predictiva marginal de cada localidad y, al no incluir término de penalización, requiere restricciones adicionales sobre el número máximo de padres por nodo para evitar el sobreajuste.

Como ya se expuso en el Cap. 4, en meteorología, existen numerosos índices de calidad que aportan información de la capacidad predictiva desde distintos puntos de vista. De todos ellos, el RSA (*Relative Operating Characteristics Skill Area*), es muy utilizado para evaluar la capacidad predictiva local o marginal; además, es una medida también conocida y utilizada en la evaluación de algoritmos de aprendizaje automático (Bradley, 1997). Su valor se obtiene a partir de la curva ROC, definida en la Sec. 4.2, como: *RSA=2ROCArea-1* y representa el área sobre la diagonal en la gráfica ROC. El valor máximo de este índice, para un sistema de predicción perfecto, sería $RSA = 1$, correspondiendo el valor cero a un sistema de predicción climatológico, mientras que sistemas cuya capacidad predictiva es superior (inferior) a la de la climatología, obtendrían valores positivos (negativos) de RSA. De esta manera, para un evento dado, se pueden comparar diferentes métodos, resultando mejor aquel que tenga mayor valor de RSA; para una descripción completa de estas y otras medidas de validación empleadas en meteorología y climatología, véase Richardson (2000).

A diferencia de MDL, RSA es una medida de calidad marginal, es decir, evalúa la capacidad predictiva de cada predictando por separado, y se puede expresar como:

$$RSA(Y_j, C, B) = f(P_B(Y_j \mid C)) \tag{8.9}$$

permitiendo estimar una medida global a partir de las individuales, por ejemplo:

$$RSA(\mathbf{Y}, C, B) = \frac{1}{n} \sum_{j=1}^{n} RSA(Y_j, C, B) \tag{8.10}$$

en caso de considerar la media.

En general, RSA es una medida más costosa de calcular que MDL, por no ser descomponible, aunque sería posible mejorar su eficiencia computacional considerando algoritmos de propagación orientada a un objetivo (Castillo et al., 1996), identificando el subconjunto de probabilidades en (8.10) que es necesario recalcular cada vez que se modifica un enlace en el algoritmo de aprendizaje. En esta Tesis no se ha abordado esta modificación, quedando pendiente como un trabajo futuro. Por otra parte, en el caso particular de problemas de *downscaling*, la complejidad computacional del algoritmo se reduce sustancialmente ya que, cada vez que se modifica un enlace, sólo es necesario recalcular la probabilidad del nodo hijo y de sus descendientes, con un coste similar al de MDL, como se deduce de las Ec. 9.1 y 9.2. Al no contar con un término de penalización, los algoritmos de aprendizaje basados en RSA, que evalúa directamente la calidad de la predicción probabilística, no controlan directamente el sobreajuste. En el caso de la predicción local de precipitación, lo que se hace es limitar la estructura del grafo imponiendo un número máximo de padres, y utilizar un conjunto de entrenamiento (por ejemplo, el 70 % de los datos disponibles) y otro de test (el 30 % restante), para evitar el sobreajuste. Este ha sido el criterio seguido en esta Tesis.

Para ilustrar el comportamiento de esta medida de calidad, comparada con una medida de calidad global, como MDL, se ha repetido el experimento mostrado en la Fig. 8.7, predicción probabilística del evento $prec > 0mm/24h$ usando como predictor C^{100}, considerando 66 parejas distintas de estaciones. La Fig. 8.8 muestra las medidas de calidad de los grafos de tipo B_1 y B_2 utilizando MDL y RSA, respectivamente. El diagrama de cajas del panel superior compara los valores de MDL, mientras que en el panel inferior se comparan los valores de RSA, en el nodo B. La medida MDL favorece preferentemente enlaces del tipo B_1, mientras que RSA favorece los del tipo B_2, que son los que tienen mayor capacidad predictiva en el modo *downscaling*.

El ejemplo anterior muestra la bondad de la medida RSA para el aprendizaje estructural, pero es necesario realizar un ejemplo real para probar su utilidad práctica. Para ello, se ha considerado el ejemplo de la Cornisa Cantábrica mostrado en la Fig. 5.3, utilizando un conjunto reducido de datos de entrenamiento (un año), para poner de manifiesto el problema del sobreajuste. La Fig. 8.9, compara los resultados del modelo ingenuo con los modelos obtenidos aplicando localB(mdl) y localB(rsa). En este ejemplo se trata de poner en evidencia tanto los efectos de la discrepancia en métricas de entrenamiento y test, como los del término de penalización. En la Fig. 8.9(a) se pone de manifiesto la necesidad de limitar el número de parámetros

ya que incluso un modelo tan sencillo como es el ingenuo se sobreajusta a los datos, como se deduce de la gran diferencia en la medida de entrenamiento y la de test. En la Fig. 8.9(b) se puede apreciar que el algoritmo basado en la medida MDL no consigue establecer ningún enlace al predictor (con 100 estados) debido al término de penalización, por lo que su capacidad predictiva es nula, pese a ser el que tiene mejor valor de MDL. La Fig. 8.9(c) muestra que el algoritmo basado en RSA mejora la capacidad predictiva del modelo ingenuo, utilizando un modelo con menos parámetros y más robusto desde el punto de vista predictivo. La Tabla 9.1 muestra los resultados, localidad por localidad, del valor de test de RSA para los modelos ingenuo y RSA.

Tabla 8.2: Valores de test de RSA correspondientes a los grafos *ingenuo* (panel(a)) y *localB(rsa)* (panel (c)) de la Fig. Ordenadas según la diferencia; el grupo encabezado por Espinama, son las 7 estaciones que están directamente enlazadas al estado de la atmósfera, ver fig. 8.9, por ello sus diferencias en test son casi nulas.

Localidad	*localB(rsa)*	*ingenua*	diferencia
ESPINAMA	0.31	0.33	-0.02
RIOSECO DE SOBRESCOBIO	0.53	0.55	-0.02
SAN SEBASTIAN 'IGUELDO'	0.48	0.48	0
ELDUAYEN	0.60	0.60	0
EL MERCADILLO DE LIERGANES	0.57	0.57	0
LASARTE-MICHELIN	0.50	0.50	0
SANTANDER 'CENTRO'	0.47	0.47	0
ESCOBEDO DE VILLAFUFRE	0.47	0.50	-0.03
GENESTAZA	0.41	0.42	-0.01
ECHEVARRIA	0.55	0.55	0
ZARDAIN	0.50	0.48	0.02
AMIEVA 'RESTAÑO'	0.60	0.58	0.02
VILLACARRIEDO	0.64	0.61	0.03
EIBAR-BANCO DE PRUEBAS	0.56	0.53	0.03
FUENTERRABIA 'AEROPUERTO'	0.62	0.58	0.04
CARRANZA	0.54	0.50	0.04
MIRONES	0.61	0.55	0.06
SOTO DE LOS INFANTES	0.32	0.25	0.07
GIJON	0.50	0.42	0.08
BEZANES	0.46	0.38	0.08
AMURRIO 'INSTITUTO'	0.55	0.47	0.08
DURANGO-VIVERO	0.55	0.47	0.08
BILBAO 'AEROPUERTO'	0.57	0.47	0.10
ROZADIO	0.58	0.48	0.10
CUEVAS DE ALTAMIRA	0.58	0.48	0.10
RANON 'AEROPUERTO'	0.55	0.45	0.10
ABADIANO 'MENDIOLA'	0.61	0.49	0.12
COTERILLO DE AMPUERO	0.67	0.55	0.12
OVIEDO 'EL CRISTO'	0.58	0.45	0.13
CELIS	0.61	0.46	0.15
PRESA DE LA BARCA	0.55	0.40	0.15
SAN SEBASTIAN 'ATEGORRIETA'	0.50	0.34	0.16
SOTO DE RIBERA	0.54	0.38	0.16
MERES DE SIERO	0.64	0.48	0.16
RIBADESELLA 'FARO'	0.60	0.43	0.17
GRADO	0.41	0.24	0.17
GURIEZO 'G.C.'	0.64	0.47	0.17
CAMIJANES	0.56	0.37	0.19
TORRELAVEGA 'SNIACE'	0.73	0.52	0.21
TAMA	0.55	0.33	0.22
SOTO DE LA BARCA	0.46	0.24	0.22
CANGAS DE ONIS	0.52	0.27	0.25

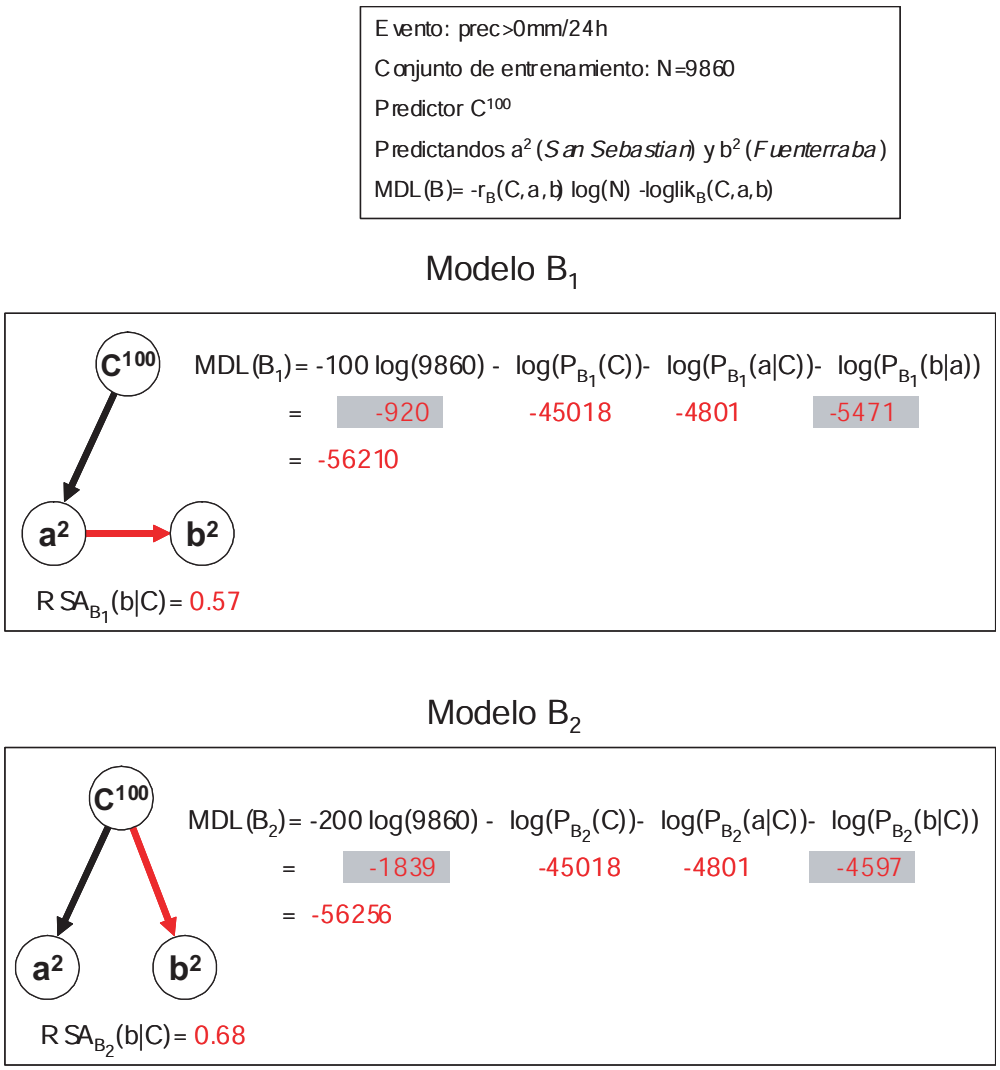

Figura 8.7: Proceso seguido por el algoritmo MDL para seleccionar (a) un enlace entre predictandos y (b) un enlace de predictando a predictor.

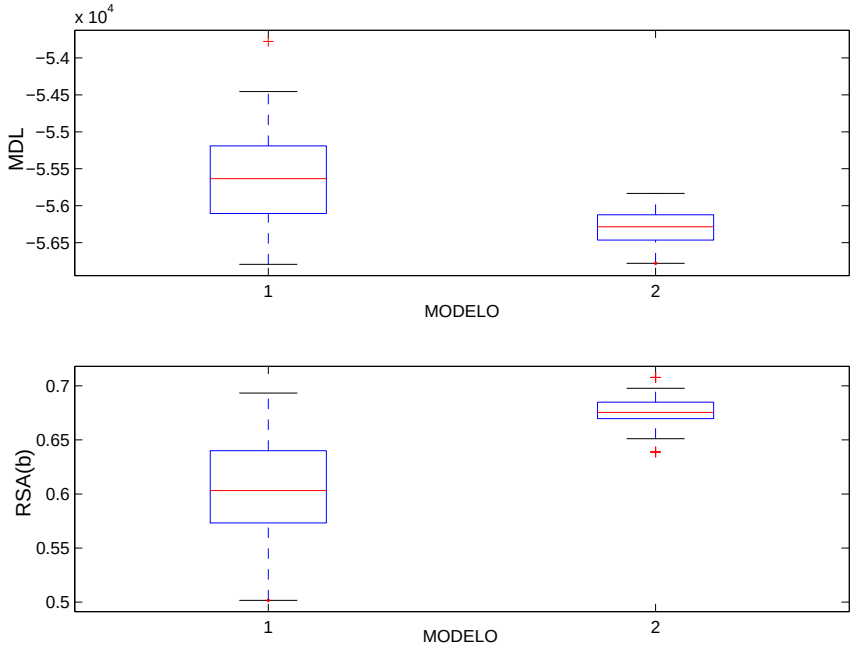

Figura 8.8: Comparación de las medidas de calidad MDL y RSA para varias repeticiones del ejemplo mostrado en la Fig. 8.7.

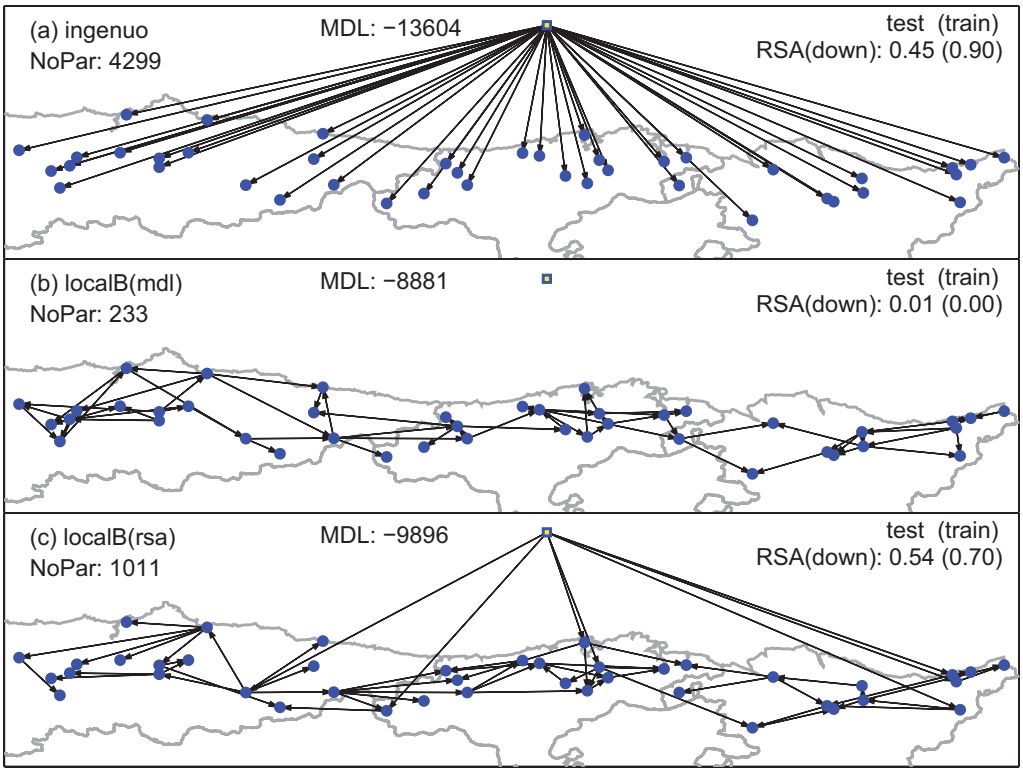

Figura 8.9: Comparación de (a) el modelo ingenuo, y los grafos obtenidos utilizando el algoritmo *localB* con las medidas de calidad (b) MDL y (c) RSA; en este último caso, los valores de capacidad predictiva conjunta se han obtenido como la media de las marginales (el valor entre paréntesis corresponde a los valores en el conjunto de datos de entrenamiento, mientras que los valores fuera del paréntesis son los de test.

CAPÍTULO 9

Predicción Local con Redes Bayesianas

En el Cap. 6 se presentaron distintas aplicaciones prácticas de las redes Bayesianas (RBs) en Meteorología, en especial su aplicación a la predicción probabilística local, tema que fué introducido con detalle en el Cap. 7, planteando tres paradigmas que cubren las diferentes situaciones prácticas que se puedan presentar, como son: *interpolación*, *downscaling*, y *predicción*. Posteriormente, en el Cap. 8 se analizaron algoritmos de aprendizaje específicos para este tipo de aplicaciones, que permiten evitar los problemas de coste computacional y de sobreajuste que se plantean al trabajar con RBs con cientos de variables.

Continuando con la predicción local, en este capítulo se describen una serie de experimentos, utilizando varios algoritmos de aprendizaje con diferentes métricas de calidad, tanto para RBs de tipo *extendido* (obtenidas a partir de la ingenua), como para RBs *genéricas* (ver Sec. 7.1). Estos experimentos se han realizado para cada uno de los tres paradigmas, con la idea de establecer de forma objetiva cual es el mejor modelo de RB en cada uno de ellos.

Este capítulo está dividido en dos partes. En la Sec. 9.1 se describen las características esenciales del diseño experimental utilizado para llevar a cabo la comparativa, analizando los aspectos más relevantes de cada uno de los paradigmas tratados. Finalmente, en la Sec. 9.2 se analizan los resultados tanto en su conjunto, como de forma separada para cada uno de los paradigmas.

9.1. Comparativa de Algoritmos

En esta comparación se ha utilizado el conjunto homogéneo de datos diarios en las 42 estaciones del dominio de alta resolución descrito en la Sec. 5.1, y se han aplicando diferentes versiones del algoritmo de aprendizaje *localB*, descrito en la Sec. 8.1.2. Las especificaciones de este experimento son las siguientes:

- Para simplificar los cálculos, se ha considerado una red formada por localidades que presentan series diarias de la misma longitud y sin lagunas; de esta manera, el aprendizaje de las RBs se realiza en condiciones homogéneas. Por otra parte, todos los nodos de la red (variables) son binarios, excepto el que representa el estado de la atmósfera, que puede tener 10 o 100 estados (tipos de tiempo).

- La calidad de una RB se medirá de dos formas. Por una parte, utilizando la métrica estándar MDL (3.6) y, por otra, con la métrica propuesta en esta Tesis, basada en evaluar la capacidad predictiva marginal de cada uno de los predictandos, estimando cada $p(x_i|c)$ a partir de la red, y obteniendo los correspondientes $RSA(X_i)$ (ver 8.2); finalmente, la capacidad predictiva global de la red se estima utilizando la media de los RSA individuales. Nótese que en el caso general se tendría que considerar la capacidad predictiva conjunta $p(\mathbf{y}|c)$, donde $\mathbf{Y} \subset \mathbf{X}$; sin embargo, ese estudio no se aborda en esta Tesis.

- Se utilizan diferentes versiones del algoritmo *localB*, correspondientes a las medidas de calidad MDL y RSA anteriores. Los algoritmos resultantes se denominan *localB(mdl)* y *localB(rsa)*, indicando entre paréntesis el tipo de medida de calidad utilizado.

- El experimento se ha realizado para cuatro supuestos, resultantes de combinar los eventos $prec > 0mm/24h$ y $prec > 10mm/24h$ con los predictores C^{10} y C^{100} (el superíndice indica el número de clases del estado de la atmósfera, ver Sec. 5.3 para más detalles).

- Para todos los casos se ha reservado el mismo conjunto de test, formado por el 30 % del total de datos. Sin embargo, para analizar el impacto de la longitud de entrenamiento, se han considerado dos situaciones, en el primer caso, se utiliza un subconjunto del 10 % de los datos de aprendizaje, mientras que en el segundo se emplea el 100 %; corresponden respectivamente a ejemplos donde se utiliza una serie corta de datos (alrededor de un año) y una serie larga (alrededor de diez años).

- Por último, todos los experimentos se han repetido 10 veces, para estimar la variabilidad introducida por la partición entrenamiento/test. En algunos casos se muestran los valores de los 10 casos y, en los restantes, su media.

En resumen, las características de este experimento son:

- 2 eventos binarios ($prec > 0mm/24h$ y $prec > 10mm/24h$), correspondientes a un evento normal y un evento raro (pequeña probabilidad de ocurrencia).
- 42 predictandos (observatorios de la AEMET)
- 2 tipos de predictor (C^{10} y C^{100} con 10 y 100 tipos de tiempo, respectivamente)
- 3 paradigmas: *downscaling*, *predicción* e *interpolación*
- 4 tipos de algoritmo: *ingenua*, *localB(rsa)*, *localB(mdl)* y *extLocalB(mdl)*
- 2 longitudes de entrenamiento (uno y diez años).
- 10 repeticiones aleatorias para diferentes conjuntos de entrenamiento y test.

Antes de analizar los resultados concretos, se describen algunas particularidades y consideraciones generales sobre cada uno de los tres paradigmas.

9.1.1. Modo Downscaling

Como se explicó en la Sec. 7.1, este modo de operación corresponde al caso en el que se está interesado en calcular $p(\mathbf{Y}|C)$, donde C es la única evidencia disponible, procedente del estado de la atmósfera (el tipo de tiempo), e $\mathbf{Y}$ es un subconjunto de predictandos. Por tanto, en los experimentos correspondientes a este paradigma, para cada día del conjunto de test se ha evidenciado la variable C (considerando el estado ocurrido) y se han calculado las predicciones probabilísticas para cada estación $p(y_i|c), \forall (y_i, c) \in \{Y_i, C\}$.

En este caso concreto, puede probarse que la capacidad predictiva local de las topologías *extendida* e *ingenua* son iguales, como se muestra a continuación:

$$p(a|c)_{ingenua} = \sum_{\mathbf{b}} \frac{p(c)p(a|c)p(\mathbf{b}|c)}{p(c)} = p(a|c) \tag{9.1}$$

$$p(a|c)_{extendida} = \sum_{\mathbf{b}} \frac{p(c)p(a|\mathbf{b},c)p(\mathbf{b}|c)}{p(c)} = \sum_{\mathbf{b}} p(a,\mathbf{b}|c) = p(a|c) \tag{9.2}$$

Esto significa que, en el modo *downscaling*, la consideración de la dependencia espacial entre observatorios no aporta capacidad predictiva individual, independientemente de la calidad del predictor y de la resolución espacial de la red de predictandos. Por tanto, los resultados del modelo *extLocalB(rsa)* no se han incluido en la comparación posterior, ya que este algoritmo no puede añadir ningún enlace al modelo *ingenuo*, dado que no puede mejorar su capacidad predictiva (nótese que el RSA de la red se calcula como promedio del RSA de cada una de las estaciones). En cambio, cuando se considera la predicción conjunta de varias estaciones, ambos modelos dan resultados diferentes (como se constata en los resultados del cuadro 7.5 de la Sec. 7.2.2, ya que $p(\mathbf{X}|C)_{ingenua} \neq p(\mathbf{X}|C)_{genérica}$; es en este caso donde los modelos de red *genérica* son más adecuados, ya que los modelos de red *ingenua* no son capaces de compatibilizar espacialmente predicciones colectivas, como las que necesita por ejemplo, un modelo hidrológico de una cuenca fluvial. Este estudio comparativo será uno de los trabajos futuros de esta Tesis.

9.1.2. Modo Predicción

Como se explicó en la Sec. 7.1, este modo de operación corresponde al caso más general, donde además de tener evidencia en el estado de la atmósfera, es posible que haya otro tipo de evidencias en algún predictando (por ejemplo, en estaciones automáticas que reportan las observaciones en tiempo real). Por tanto, se trata de calcular $p(\mathbf{Y}|C, \mathbf{Z})$, donde C es el estado de la atmósfera, $\mathbf{Y}$ son los predictandos para los cuales se quiere calcular una predicción, mientras que $\mathbf{Z}$ son los predictandos evidenciales.

En los experimentos correspondientes a este paradigma, para cada día del conjunto de test, se han evidenciado tanto la variable C como los nodos del Markov Blanket correspondiente a cada nodo y_i (que es el subconjunto de nodos formado por sus padres, sus hijos y los padres de sus hijos) $\mathbf{W}_i = MB(Y_i)$. Por tanto, la predicción probabilística de cada estación se ha calculado como $p(y_i|c, \mathbf{w}_i), \forall(y_i, \mathbf{w}_i, c) \in \{Y_i, \mathbf{W}_i, C\}$. De esta manera, se obtiene el umbral superior de la capacidad predictiva en este paradigma, pues toda la información adicional fuera del Markov Blanket es irrelevante para el cálculo de la probabilidad.

En este caso, la topología *ingenua* no tiene utilidad ya que no distingue entre *predicción* y *downscaling*, como se demuestra a continuación:

$$p(a|c, \mathbf{b})_{ingenua} = \frac{p(c)p(a|c)p(\mathbf{b}|c)}{p(c, \mathbf{b})} = p(a|c) \tag{9.3}$$

9.1.3. Modo Interpolación

Como ya se explicó en la sección 7.1, en el paradigma de *diagnóstico*, únicamente consideraremos el modo de *interpolación*, correspondiente al caso $p(\mathbf{Y}|\mathbf{Z})$, donde $\mathbf{Z}$ es el subconjunto de predictandos cuyo valor es conocido, e $\mathbf{Y}$ es el subconjunto de predictandos para los que se quiere estimar un valor.

Por tanto, en los experimentos correspondientes a este paradigma, para cada día del conjunto de test se han evidenciado únicamente aquellas variables de $MB(Y_i)$ que son predictandos: $\mathbf{W}_i = \{MB(Y_i) \backslash C\}$ (es decir, no se da evidencia en C para distinguirlo del paradigma de *predicción*) y se han calculado las predicciones probabilísticas para cada estación $p(y_i|\mathbf{w}_i), \forall (y_i, \mathbf{w}_i) \in \{Y_i, \mathbf{W}_i\}$. De esta manera no solo se evalúa el potencial de esta metodología (dando el máximo de información posible), sino que además se pueden comparar los resultados con el caso de predicción, de forma que se puede analizar el efecto de ambas componentes (tipo de tiempo previsto frente a ocurrencia en estación cercana).

9.2. Resultados

Como resultado de los experimentos citados se han obtenido 320 grafos diferentes, aparte del *ingenuo*, y 50400 valores individuales de RSA, que necesariamente han sido resumidos en varias figuras que ilustran distintos aspectos del análisis, como se explica a continuación.

En primer lugar, las Figs. 9.1 y 9.2 muestran resultados comparativos globales utilizando diagramas de cajas (*boxplot*), que indican con el rectángulo los cuartiles de la muestra, y con las líneas el rango y extremos de la misma. En la Fig. 9.1, se analiza el RSA promedio de la red (la media de las 42 estaciones) en las 10 repeticiones de entrenamiento/aprendizaje realizadas. Para ello, se muestran diagramas de cajas para cada uno de los algoritmos considerados: *ingenuo*, *localB(rsa)*, *localB(mdl)* y *extLocalB(mdl)*. Cada fila corresponde a una de las cuatro combinaciones de los eventos ($Precip > 0mm$ y $Precip > 10mm$) con los tamaños de muestra de entrenamiento (10 % y 100 % del total disponible), mientras que cada columna muestra uno de los paradigmas: (*downscaling*, *predicción* e *interpolación*). Así, por ejemplo, en el gráfico superior izquierdo se muestran los resultados el evento $prec > 0mm/24h$ para la muestra larga de entrenamiento. Además, el color rojo corresponde a los resultados para el predictor C^{100} (100 tipos de tiempo) y las azules al C^{10} (10 tipos de tiempo). Finalmente, los círculos indican el valor medio de RSA evaluado en el conjunto de entrenamiento, de forma que se pueda valorar el posible sobreajuste del

modelo (nótese que en los casos de muestra corta de entrenamiento, los parámetros de la red se ajustan con unos 365 datos y, por tanto, el modelo es más susceptible de sobreajustarse a los datos).

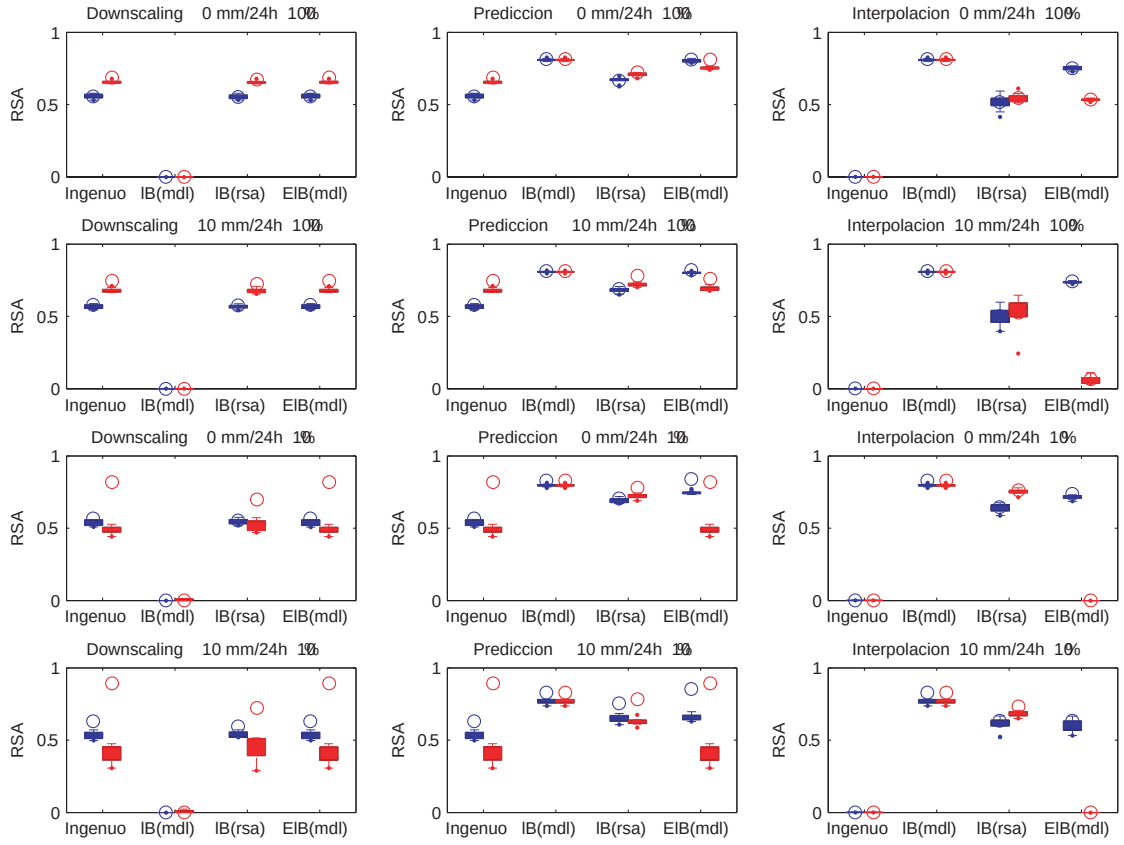

Figura 9.1: Resumen comparativo de los diferentes algoritmos para distintos eventos y tamaños de la muestra de entrenamiento. En los gráficos, cada caja representa la dispersión de los valores medios para las 10 repeticiones efectuadas (atribuible a la sección del conjunto de test). En azul se muestra el caso C^{10} y en rojo C^{100}; los círculos representan la media para el conjunto de entrenamiento. En cada gráfico, el eje de abscisas indica el tipo de algoritmo utilizado y el eje de ordenadas, la calidad obtenida (RSA promedio).

La Fig. 9.2, es similar, aunque en este caso se realiza la comparativa considerando los resultados promedio de las 42 estaciones (tomadas sobre las 10 repeticiones).

Las dos figuras anteriores muestran que la variabilidad procedente de la selección arbitraria entrenamiento/test (10 repeticiones) es inferior a la variabilidad espacial (42 estaciones). Asimismo, tambien se puede observar que en los experimentos realizados con la muestra de entrenamiento corta para el predictor C^{100} y el evento $prec > 10mm/24h$, los círculos de color rojo están muy por encima de las cajas, evidenciando el efecto del sobreajuste producido en aquellas RBs que tienen muchos

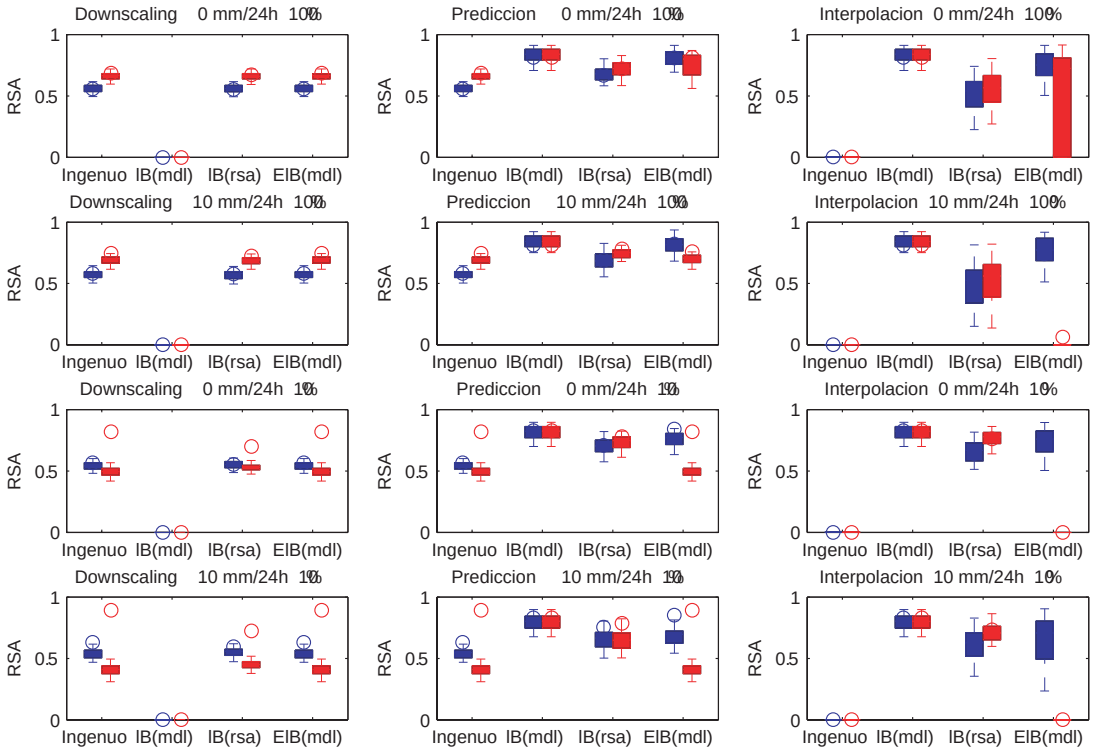

Figura 9.2: Comparativa de los diferentes algoritmos. En este caso, los diagramas reflejan la variabilidad entre estaciones.

enlaces con C, como son: *ingenua* y *extLocalB(mdl)*.

Por otra parte, en la Fig. 9.3 se muestran muestra los 4 grafos resultantes en la primera de las 10 realizaciones. En cada grafo, a la izquierda, aparece el tipo de algoritmo utilizado y el número de parámetros que requiere la RB correspondiente ($NoPar$). En el centro aparece el valor de MDL, mientras que las cantidades que aparecen a la derecha corresponden a los valores medios de RSA en el conjunto de test y de entrenamiento (este último entre paréntesis), para cada uno de los tres paradigmas. Como en la descripción del experimento, se presentan los resultados para los distintos paradigmas de predicción. Esta figura pone de manifiesto que el algoritmo *localB* con la medida de calidad RSA presenta los mejores resultados globales, con menor sobreajuste que el resto de algoritmos (Fig. 9.3).

En las siguientes secciones se analizan los resultados en más detalle para cada uno de los tres paradigmas analizados. Las figuras 9.4-9.10 muestran los grafos obtenidos con los cuatro modelos de aprendizaje para cada uno de los 8 casos propuestos: los eventos normal y raro ($prec > 0mm/24h$ y $prec > 10mm/24h$) con longitud de entrenamiento corta y larga (uno y diez años)

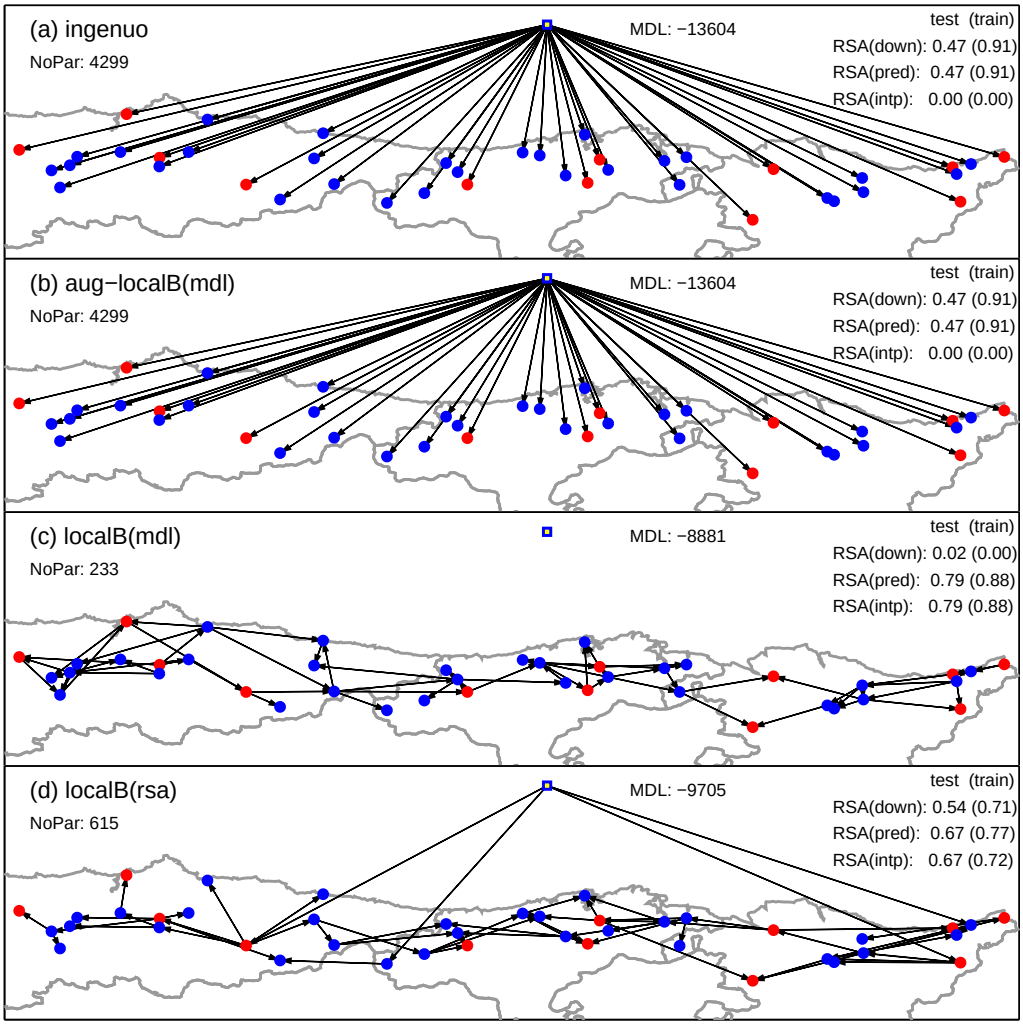

Figura 9.3: Grafos obtenidos con los distintos métodos de aprendizaje en el caso de mayor sobreajuste: longitud corta de entrenamiento para el evento $prec > 10mm/24h$ con la variable atmosférica C^{100} ($NoPar$ es el número de parámetros del grafo).

9.2.1. Modo Downscaling

En primer lugar, analizando la primera columna de las figuras resumen 9.1 y 9.2, que corresponde al modo *downscaling*, se observa lo siguiente:

1. Las redes obtenidas con el algoritmo *localB(mdl)* presentan un RSA nulo, salvo algún caso aislado, debido a que el término de penalización de este algoritmo no permite establecer enlaces al predictor C, especialmente cuando hay pocos datos de entrenamiento. Los enlaces se establecen entre las estaciones, porque tienen una penalización mucho menor debido a su carácter binario.

2. El modelo *ingenuo* es el modelo óptimo para *downscaling* (cuando se considera la validación marginal, estación a estación); sin embargo, puede ser mejorado con modelos más sencillos (como los obtenidos mediante *localB(rsa)*), cuando hay sobreajuste.

3. Los modelos *ingenuo* y el *extendido* tienen distribuciones de RSA idénticas, de acuerdo con lo establecido por las ecuaciones 9.1 y 9.2.

4. En casos de posible sobreajuste, cuando se dispone de pocos datos de entrenamiento, el predictor C^{10} es mejor que C^{100} para todos los modelos posibles, pues involucra menos parámetros.

Algunas de estas conclusiones se pueden comprobar en más detalle en los grafos de la Fig. 9.3 que, como ya se ha mencionado, corresponden al caso de mayor sobreajuste potencial; en la misma cabe destacar lo siguiente:

- El panel (c) muestra que, a veces, el algoritmo *localB(mdl)* no consigue establecer ningún enlace al predictor debido al término de penalización, por lo que su capacidad predictiva en el modo *downscaling* es nula, a pesar de tener la mayor calidad en términos de MDL.

- En el panel(d) se observa que el algoritmo *localB(rsa)* obtiene el mejor modelo con sólamente 4 enlaces directos al predictor (aparecen en negrita en la Tabla 9.1), demostrando que en casos de sobreajuste es más robusto, en general, utilizar enlaces entre variables binarias, que corresponden a la dependencia espacial, para recibir la información del predictor (tipo de tiempo). La tabla 9.1 muestra los valores de RSA obtenidas con los grafos mostrado en el panel (d) para todas las estaciones (las estaciones con enlace directo a la atmósfera se muestran en negrita en esta tabla).

Considerando que las redes *ingenua* y *extendida* son marginalmente equivalentes en problemas de *downscaling*, las mejoras respecto a la red *ingenua*, obtenidas por *localB(rsa)* sólo pueden proceder de estaciones que no tienen conexión directa con C, como se constata en los resultados de la tabla 9.1, correspondientes a la figura 9.3(d). Para interpretar correctamente estos resultados hay que recordar que el algoritmo *localB(rsa)* establece enlaces sólo cuando se supera localmente la capacidad predictiva del modelo *ingenuo* en el conjunto de entrenamiento (esto puede no ocurrir en el test, como demuestran algunos de los valores que aparecen en la tabla). Las estaciones que mantienen su estatus de modelo *ingenuo* aparecen en negrita en la tabla, y el resto son estaciones que no tienen enlace directo con el estado de la atmósfera. La mayoría de las estaciones del segundo grupo muestran mejoras respecto al modelo ingenuo, confirmando en el test lo aprendido en el entrenamiento. En algunos casos, las diferencias son muy significativas como, por ejemplo, en *Soto de la Barca* y *Cangas de Onís*. También es cierto que en otras estaciones ocurre lo contrario, como ocurre en *Tama*, aunque son las menos. Posteriormente, en el Cap. 10, se analizará cómo se pueden mejorar estos resultados modificando los algoritmos de aprendizaje para que operen de forma jerárquica.

9.2.2. Modo Predicción

En este caso, la capacidad predictiva local queda compartida entre el predictor atmosférico y las observaciones locales entorno a cada estación analizada (*Markov Blanket*); por ello los resultados son menos sensibles al número de estados de C (tipos de tiempo), pues siempre se dispone de la evidencia en el vecindario. Los resultados para este tipo de predicciones se muestran en la segunda columna de las Figuras 9.1 y 9.2, a partir de las cuales se puede concluir que:

1. Comparando con los resultados en modo *downscaling* (primera columna), se puede ver que el modelo *ingenuo* no distingue entre *downscaling* y *predicción*, conforme a lo establecido en la Ecuación 9.3.

2. Al contar con la capacidad predictiva que aporta el *Markov Blanquet*, el predictor C^{10} es igual o mejor que C^{100}.

3. *localB(mdl)* obtiene el mejor modelo independientemente del sobreajuste y de la calidad del predictor.

A partir de los ejemplos concretos mostrados en las Figuras 9.3-9.10 también se puede concluir que el método *localB(mdl)* es el que mejor representa la dependencia

espacial, dejando en un segundo plano el estado de la atmósfera, que en este caso es poco relevante. Por tanto, este modelo tiene la menor tasa de sobreajuste junto con los mejores resultados. Por contra, la Fig. 9.3(b) muestra que *extLocalB(mdl)* no permite establecer ningún enlace entre predictandos, obteniendo resultados muy malos cuando el conjunto de entrenamiento es pequeño.

9.2.3. Modo Interpolación

Finalmente, la tercera columna de las figuras 9.1 y 9.2 muestra los resultados para el modo de *interpolación*:

1. En este caso $\{\mathbf{W}_i = \emptyset \forall Y_i \in \mathbf{Y}\}$, por lo que resulta que la calidad del modelo *ingenuo* es nula (coincide con la climatología) cuando se interpola utilizando el *Markov Blanket* excluyendo C.

2. Ahora la capacidad predictiva es aportada exclusivamente por la ocurrencia en las estaciones vecinas, de manera que la calidad del predictor pierde trascendencia, por lo que los resultados son similares para C^{10} y C^{100}.

3. *localB(mdl)* obtiene el mejor modelo independientemente del sobreajuste y de la calidad del predictor, además sus resultados son similares a los obtenidos en el modo *predicción*, indicando que la información del predictor se hace redundante cuando ya se dispone de la información procedente del vecindario de la estación a predecir.

4. Las figuras 9.3-9.10 muestran de nuevo que el modelo *localB(mdl)* es el que mejor representa la dependencia espacial.

Tabla 9.1: Valores locales de RSA correspondientes a los grafos *ingenuo* y *localB(rsa)* de las Figs. 9.3(a) y (d), respectivamente.

Localidad	*localB(rsa)*	*ingenua*	diferencia
FUENTERRABIA 'AEROPUERTO'	0.57	0.58	-0.01
SAN SEBASTIAN 'IGUELDO'	0.44	0.48	-0.03
ELDUAYEN	0.60	0.60	0.00
AMURRIO 'INSTITUTO'	0.52	0.47	0.05
BILBAO 'AEROPUERTO'	0.57	0.47	0.10
EL MERCADILLO DE LIERGANES	0.60	0.58	0.03
VILLACARRIEDO	0.59	0.61	-0.02
ROZADIO	0.55	0.48	0.07
RANON 'AEROPUERTO DE ASTURIAS'	0.53	0.45	0.08
RIOSECO DE SOBRESCOBIO	0.55	0.55	0.00
OVIEDO 'EL CRISTO'	0.53	0.45	0.08
ZARDAIN	0.53	0.48	0.05
SAN SEBASTIAN 'ATEGORRIETA'	0.34	0.34	0.00
LASARTE-MICHELIN	0.50	0.50	0.01
EIBAR-BANCO DE PRUEBAS	0.55	0.53	0.03
ECHEVARRIA	0.55	0.54	0.01
ABADIANO 'MENDIOLA'	0.59	0.49	0.10
DURANGO-VIVERO	0.53	0.46	0.07
GURIEZO 'G.C.'	0.62	0.47	0.15
CARRANZA	0.59	0.50	0.09
COTERILLO DE AMPUERO	0.63	0.54	0.08
MIRONES	0.54	0.55	-0.01
SANTANDER 'CENTRO'	0.56	0.47	0.08
ESCOBEDO DE VILLAFUFRE	0.45	0.49	-0.04
TORRELAVEGA 'SNIACE'	0.55	0.53	0.03
CELIS	0.59	0.46	0.13
CAMIJANES	0.56	0.37	0.19
ESPINAMA	0.32	0.32	0.00
TAMA	0.23	0.33	-0.10
AMIEVA 'RESTAÑO'	0.58	0.58	-0.00
CANGAS DE ONIS	0.49	0.27	0.22
RIBADESELLA 'FARO'	0.59	0.43	0.16
GIJON	0.46	0.42	0.04
BEZANES	0.41	0.38	0.03
SOTO DE RIBERA	0.57	0.38	0.19
MERES DE SIERO	0.62	0.48	0.14
GRADO	0.41	0.24	0.17
GENESTAZA	0.46	0.42	0.04
SOTO DE LA BARCA	0.46	0.24	0.22
PRESA DE LA BARCA	0.53	0.40	0.13
SOTO DE LOS INFANTES	0.28	0.25	0.03
CUEVAS DE ALTAMIRA	0.51	0.48	0.03

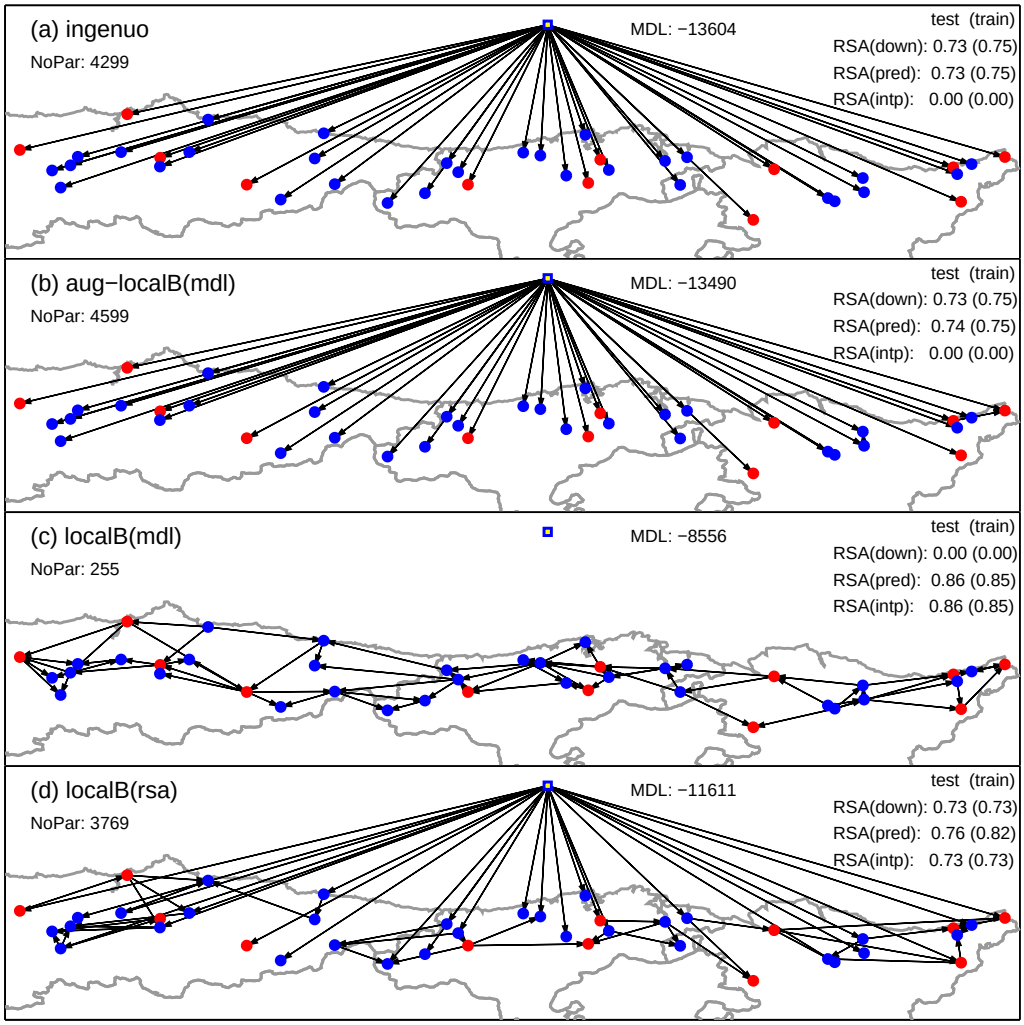

Figura 9.4: Grafos obtenidos con una longitud de entrenamiento del 100 % para el evento $prec > 10mm/24h$ con el clasificador C^{100}.

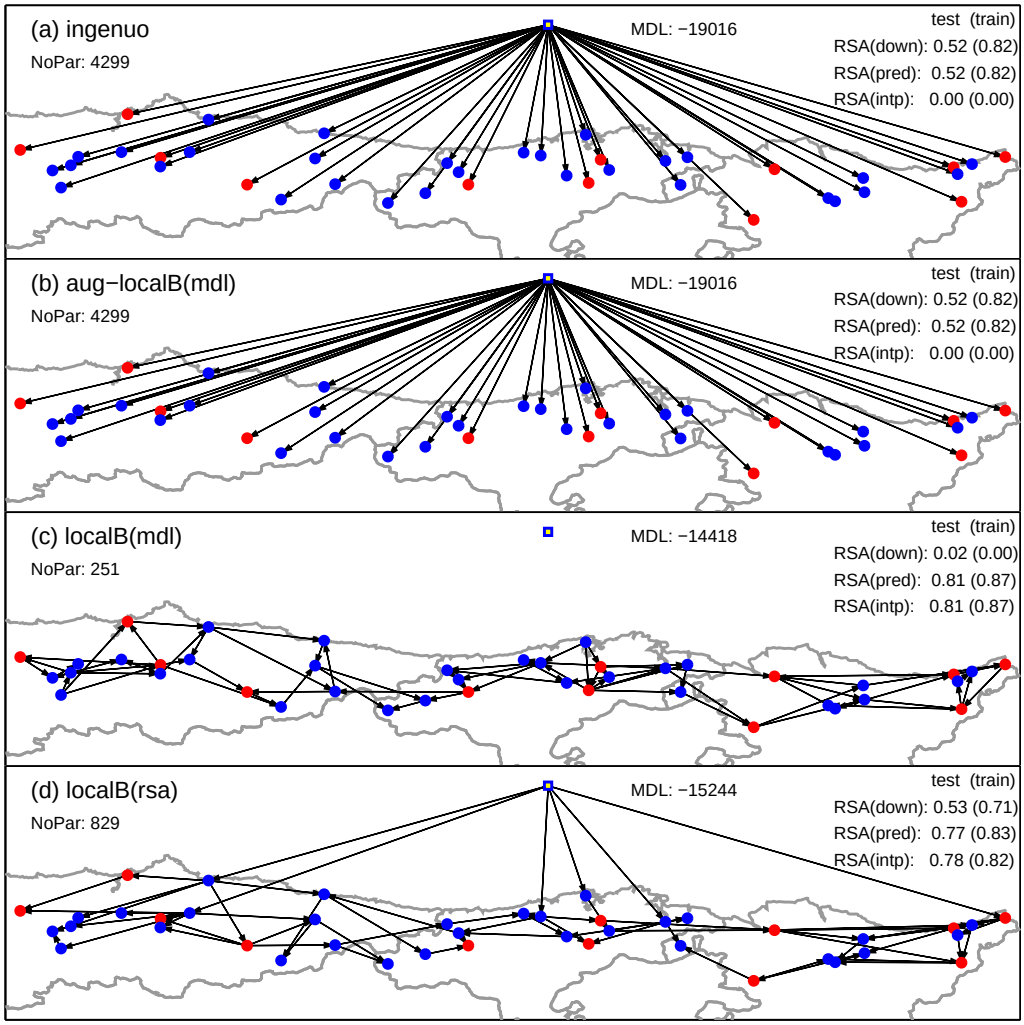

Figura 9.5: Grafos obtenidos con una longitud de entrenamiento del 10 % para el evento $prec > 0mm/24h$ con el clasificador C^{100}.

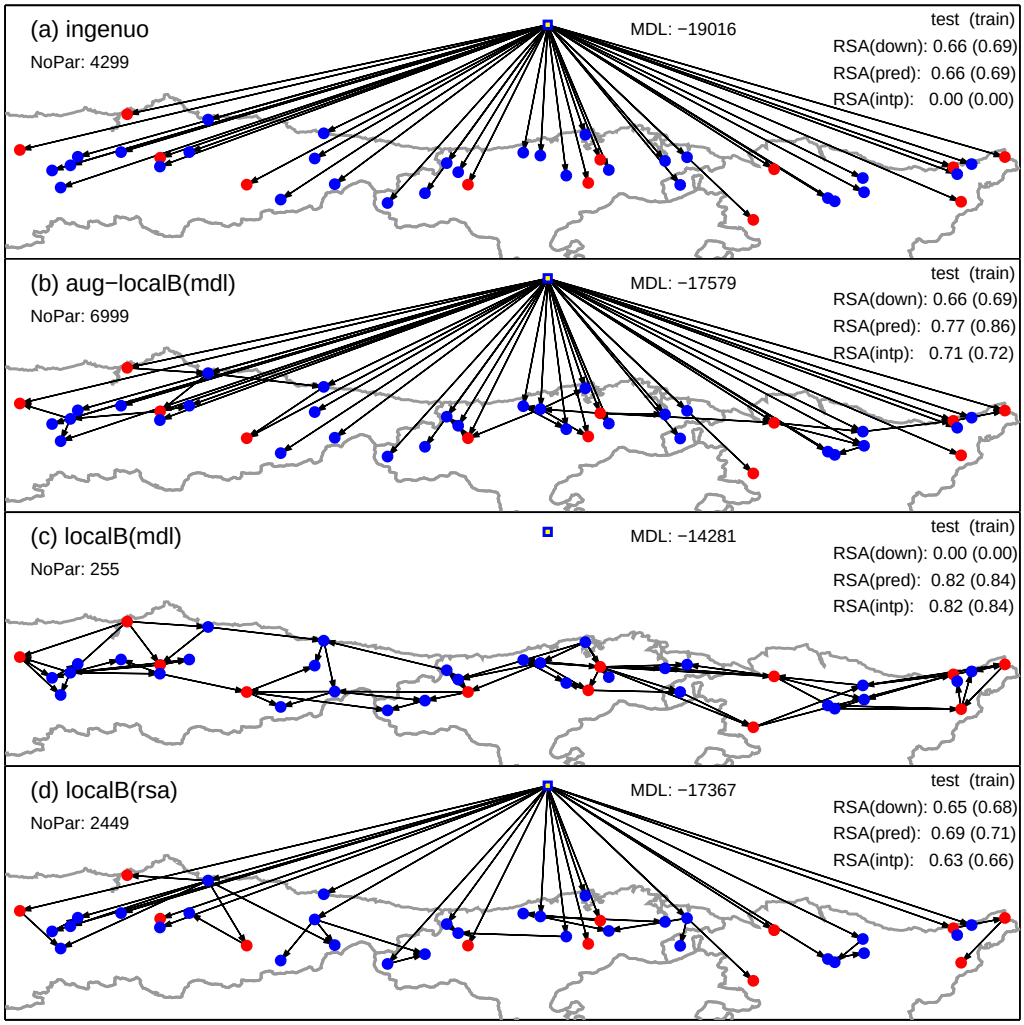

Figura 9.6: Grafos obtenidos con una longitud de entrenamiento del 100 % para el evento $prec > 0mm/24h$ con el clasificador C^{100}.

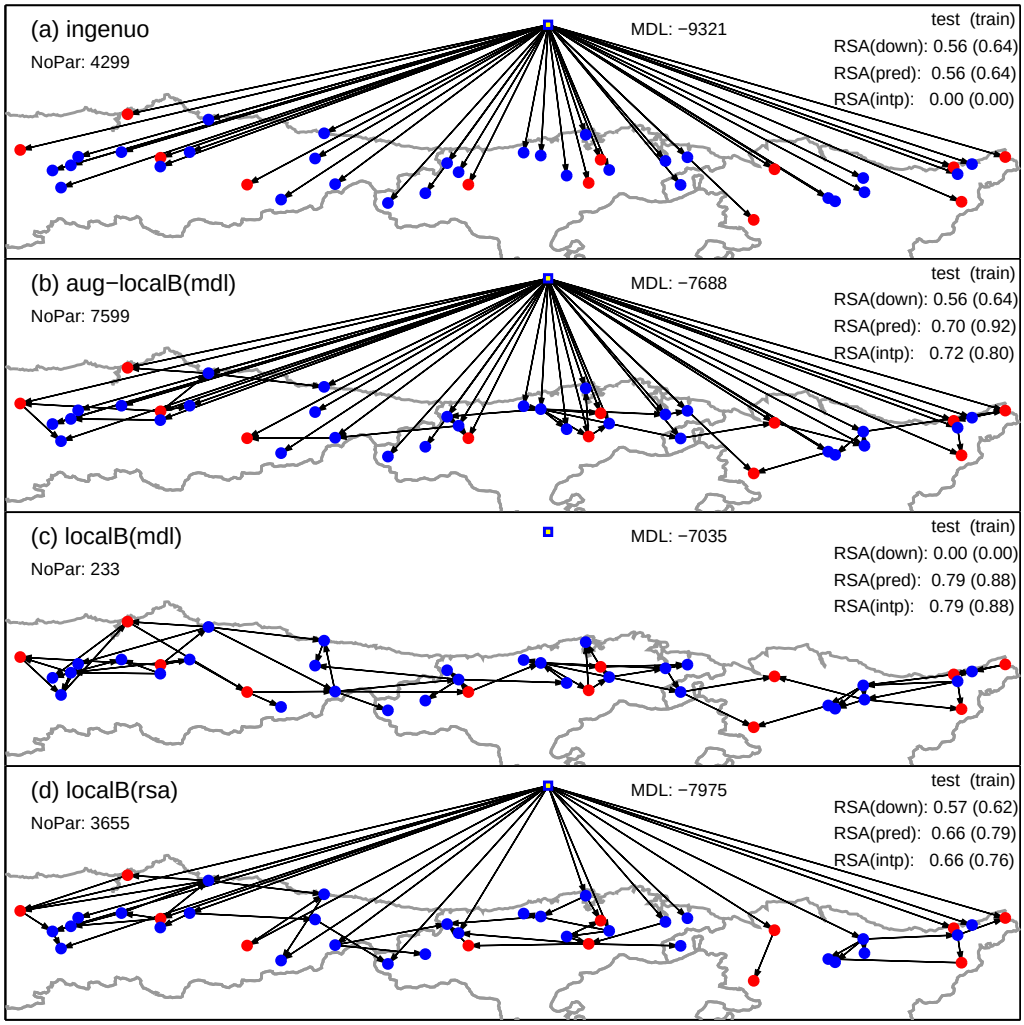

Figura 9.7: Grafos obtenidos con una longitud de entrenamiento del 10 % para el evento $prec > 10mm/24h$ con el clasificador C^{10}.

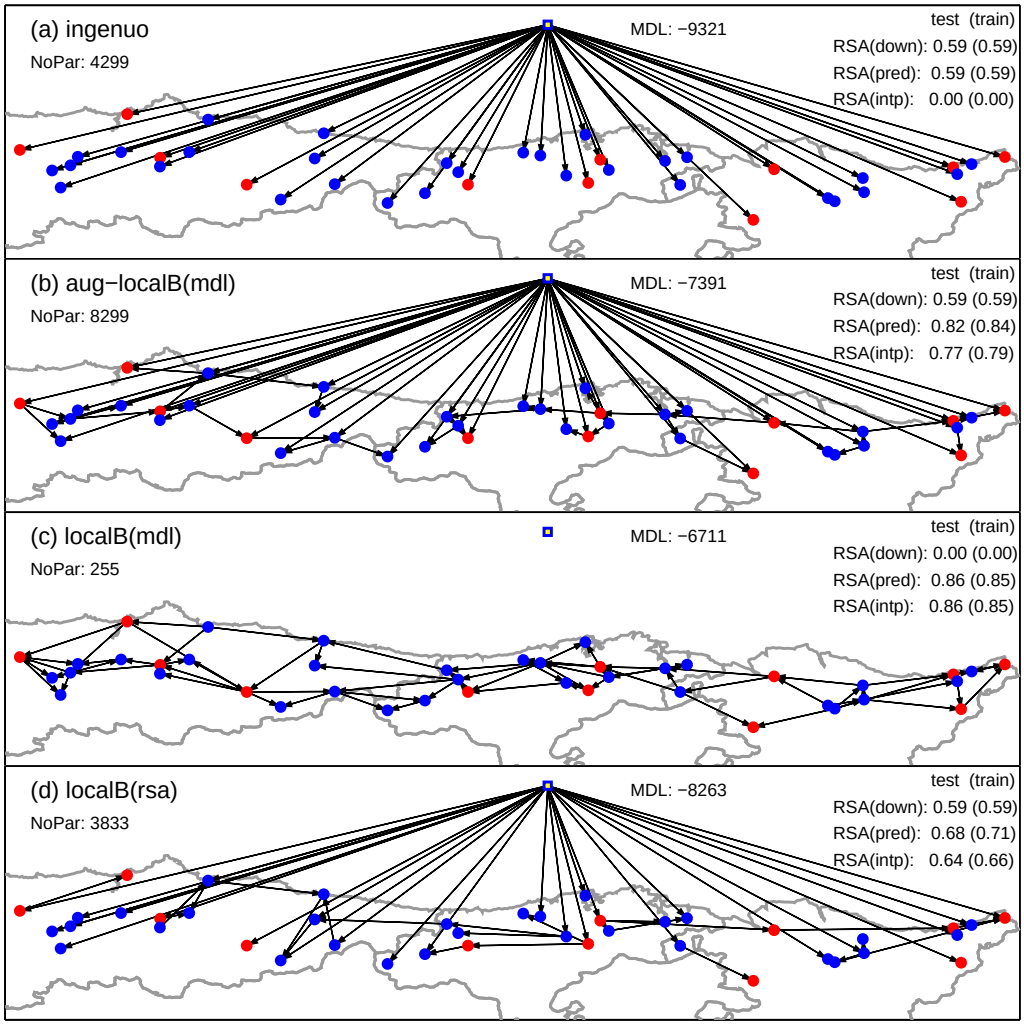

Figura 9.8: Grafos obtenidos con una longitud de entrenamiento del 100 % para el evento $prec > 10mm/24h$ con el clasificador C^{10}.

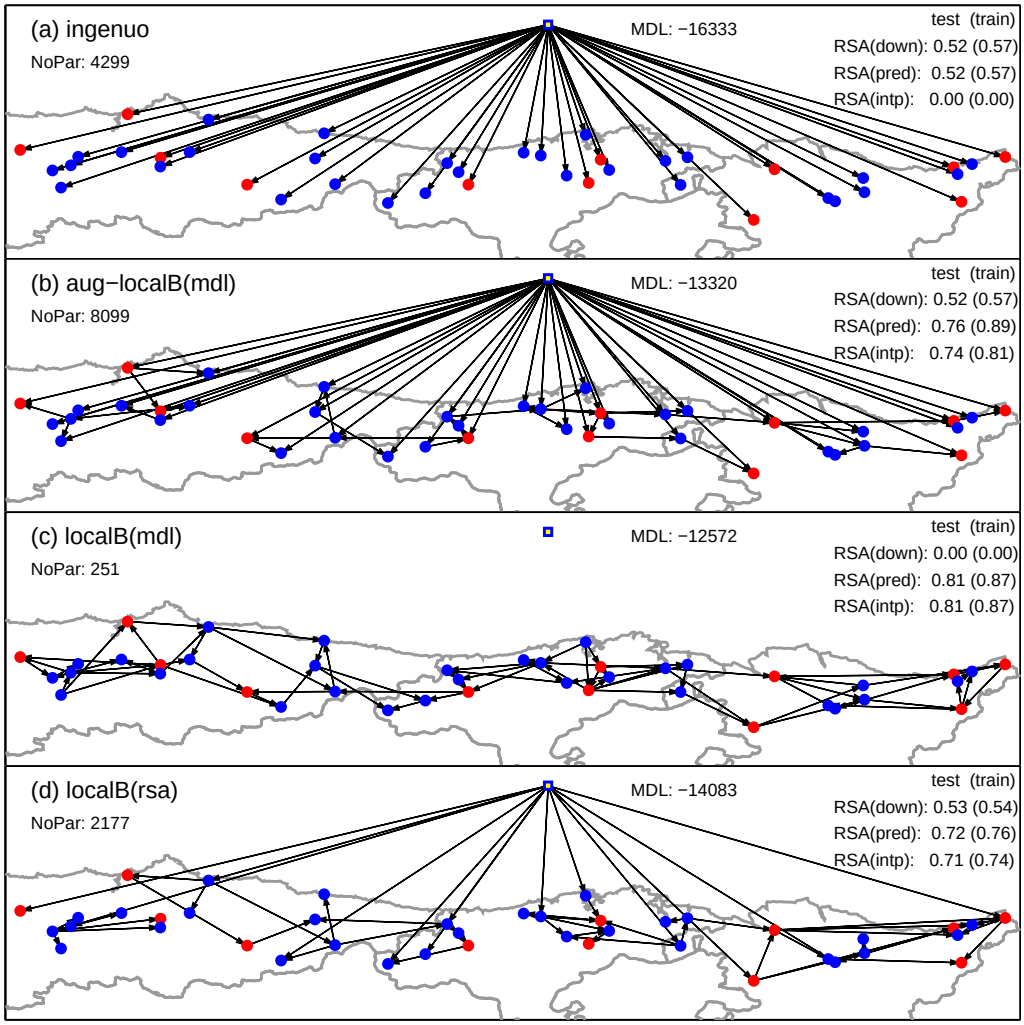

Figura 9.9: Grafos obtenidos con una longitud de entrenamiento del 10 % para el evento $prec > 0mm/24h$ con el clasificador C^{10}.

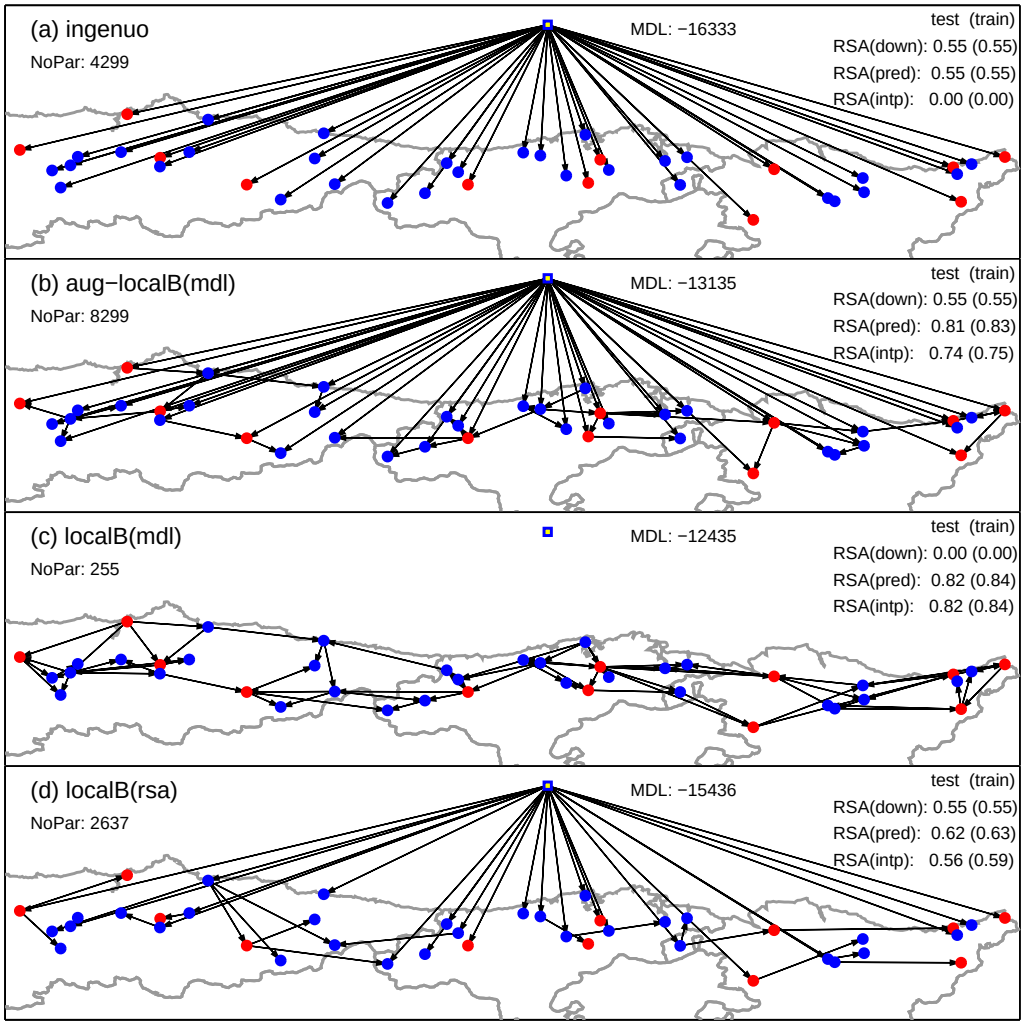

Figura 9.10: Grafos obtenidos en el caso de menor sobreajuste: con una longitud de entrenamiento del 100 % para el evento $prec > 0mm/24h$ con el clasificador C^{10}.

CAPÍTULO 10

Downscaling en Redes Heterogéneas

En el capítulo anterior hemos visto que las redes Bayesianas (RBs) generadas por *localB(mdl)* obtienen los mejores modelos en *predicción* e *interpolación* independientemente de la longitud de entrenamiento, de la calidad del predictor y del evento; mientras que para *downscaling* marginal el mejor modelo es el *ingenuo*. Sin embargo, en este capítulo veremos que cuando se produce sobreajuste, existen modelos más sencillos, como los generados por *localB(rsa)*, que obtienen mejores resultados para el *downscaling* marginal; además, las condiciones bajo las que se produce sobreajuste, en este tipo de modelos, son relativamente habituales, como se comprenderá enseguida.

Los sistemas de predicción estadística local igualan, e incluso mejoran, la resolución de la red de observación, lo cual es de gran utilidad para responder a la creciente demanda de predicciones detalladas sobre fenómenos de gran variabilidad como la precipitación o el viento, siempre y cuando se disponga de redes de observación capaces de resolver dicha variabilidad. Afortunadamente, cada vez existen más y mejores sistemas de teledetección: satélites, radares, perfiladores de viento, detectores de descargas, etc, así como más y mejores sistemas de observación terrestre, gracias a la proliferación de redes distribuidas estratégicamente en función de los intereses regionales y locales, y a la importante aportación de las estaciones privadas, cada vez más numerosas y fiables (Zepeda-Arce et al. (2000) y Janis et al. (2004)). Lamentablemente, las series de observaciones han de contener un número de registros suficiente como para estimar sus parámetros de forma robusta, lo cual

se traduce en que, por ejemplo, en un sistema de predicción de precipitación a escala diaria como el modelo *ingenuo*, han de transcurrir de 5 a 10 años para evitar el sobreajuste y de 10 a 15 años hasta alcanzar su máxima capacidad predictiva (Jagannathan and Arlqy, 1967) (ver Sec. 7.3.3).

En este capítulo mostraremos, con un ejemplo lo más realista posible, cómo las RBs son capaces de incluir en un mismo modelo redes de observación de longitudes diferentes, utilizando un algoritmo de aprendizaje específico denominado *H2*, que además consigue reducir significativamente la longitud mínima del conjunto de entrenamiento necesaria para evitar el sobreajuste.

La idea de utilizar RBs para explotar anticipadamente la capacidad predictiva marginal, aunque es nueva, comparte características tanto de los algoritmos para clasificadores ingenuos extendidos como de los jerárquicos (ver Sec. 3.6). Su principal diferencia es que han de optimizar el modo *downscaling*, en lugar del modo *clasificación*. Además, en la versión más simple, que es la que desarrollamos en esta Tesis, solamente se consideran dos tipos de estaciones que denominaremos *antiguas* y *nuevas*, en relación a la cantidad de datos históricos disponibles.

10.1. Redes Heterogéneas. El Algoritmo H2

La mayoría de los sistemas de predicción estadística, están orientados a redes de observación de carácterísticas homogéneas en cuanto a su naturaleza y longitud temporal, contemplando únicamente la heterogeneidad espacial. En esste sentido, el algoritmo *H2* que se explica a continuación, combina redes heterogéneas en el espacio y el tiempo, habilitando la incorporación de nuevas redes de observación a la predicción operativa; en realidad *H2* realiza una doble función ya que además de tratar con redes heterogéneas, veremos que tambien consigue acelerar significativamente la incorporación de nuevas redes de observación a la predicción operativa.

La fundamentación del Algoritmo *H2* es relativamente sencilla y se justifica en base a dos importantes conclusiones del experimento que se realizó en el Cap. 9. Por una parte, siempre que el registro histórico sea suficientemente largo, la red *ingenua* es el modelo de downscaling óptimo en términos de capacidad predictiva local (marginal), independientemente de la calidad del predictor y de la resolución espacial de los predictandos. Por tanto, como se verá más adelante, en el algoritmo *H2* se impone este tipo de estructura para las series *antiguas*, o largas (de forma similar a como actúan los *superpadres* de SP (Keogh and Pazzani, 1999), o las variables latentes de HBN (Langseth and Nielsen, 2002)). Por otra parte, las redes *ingenuas* tienen el problema del sobreajuste de las estaciones *nuevas*, debido al número rela-

tivamente elevado de estados del predictor (en aplicaciones operativas es necesario considerar hasta 100 tipos de tiempo para que el modelo operativo resulte competitivo). Sin embargo, como se ha constatado en la Sec. 9.1, este tipo de redes pueden mejorar su capacidad predictiva si pierden su enlace al predictor y se conectan a otro predictando (estos son binarios), ya que así se reduce notablemente el número de parámetros.

El algoritmo *H2* tiene en cuenta esos factores y facilita que las series *nuevas* puedan conectarse a otras estaciones para obtener indirectamente la información del predictor. Así, el *H2* se inicia con una red *ingenua* restringida (red *ingenua* en la cual un subconjunto de los nodos están aislados), evalúa todos los enlaces legales, añadiendo el que agrega mayor capacidad predictiva local. De esta manera la estación ganadora queda habilitado para ser padre de los restantes estaciones *nuevas*. El proceso continúa hasta que no se consigue la mejora deseada, o bien se alcanza el límite de enlaces. En este sentido, la limitación del número máximo de padres por nodo es apropiada ya que se trata de evitar el sobreajuste utilizando conjuntos de entrenamiento pequeños; en los experimentos realizados se han permitido dos padres como máximo, obteniendo buenos resultados.

Su esquema de funcionamiento es el siguiente (ver Algoritmo 5):

1. Se consideran tres tipos de nodos, el predictor, los predictandos *antiguos* y los predictandos *nuevos*.

2. La red inicial es un modelo *ingenuo* restringido, donde las series *antiguas* tienen como padre al predictor y las series *nuevas* están desconectadas.

3. Sólamente se admiten y evalúan enlaces desde los nodos con padre a nodos *nuevos* sin hijos y con menos de dos padres.

4. Una vez evaluada la capacidad predictiva de todos los enlaces permitidos, se establece el mejor y se vuelve al punto 3, mientras no se cumpla el criterio de parada.

Estas características hacen de *H2* un algoritmo relativamente eficiente, puesto que para elegir cada enlace, hay que evaluar la capacidad predictiva de $m \times n$ enlaces, donde m son los predictandos con padre y n son los nodos *nuevos*, frente a las n^2 evaluaciones por enlace que habría que realizar en el caso del algoritmo B, siendo n el número de nodos. La clave de su eficiencia se debe a la regla 3, que no permite enlaces a nodos con hijos, para no modificar la capacidad predictiva del nodo receptor y de todos sus descendientes, lo cual crearía un algoritmo innecesariamente complejo (ver

Algorithm 5 Algoritmo *H2* para aprendizaje estructural de RBs.

Require: Una muestra de datos de un conjunto de nodos separados en tres grupos $\mathbf{D} = \{C, \mathbf{X_\Omega}, \mathbf{X_\Phi}\}$, donde $\Omega = \{2, ..., m\}$ hace referencia a los padres potenciales y $\Phi = \{m+1, ..., n\}$, a los hijos potenciales, siendo C el estado de la atmósfera. El grafo inicial esta definido por una matriz de adyacencias B: $B(1, j) = 1$ $\forall j \in \Omega$, y el número máximo de padres por nodo es r.

while $\Phi \neq \emptyset$ **do**
  for $i \in \Phi$ **do**
    $S = score(\mathbf{D}, B); k = \emptyset;$
    for $j \in \Omega$ **do**
      if $acyclic(B \cup \{j \rightarrow i\}) = 1$ **then**
        $M = score(\mathbf{D}, B \cup \{j \rightarrow i\});$
        if $M > S$; **then** $S = M; k = j;$ **endif**
      end if
    end for
    if $k \neq \emptyset$ **then**
      $B(k, i) = 1;$
      $\Omega = \{j \in [2, n] : \pi_j \neq \emptyset\};$
      $\Phi = \{i \in [m+1, n] : children(i) = \emptyset \wedge \sharp(\pi_i) \leq r\};$
    else
      $\Phi = \{\Phi \backslash \{i\}\};$
    end if
  end for
end while

Ensure: Un grafo dirigido acíclico B y su medida de calidad S.

figura 10.1). Al mismo tiempo, la regla 3 garantiza un cierto incremento de calidad pues si un nodo tiene hijos, significa que ya ha sido mejorado al menos una vez. Esta regla, tampoco permite enlaces entre los nodos *antiguos*, ni enlaces entre nodos sin padre, ya que de acuerdo con los resultados del experimento 9.1, no añadirían capacidad predictiva individual.

Finalmente, para agilizar la computación, dependiendo del número de nodos, se puede plantear una búsqueda local en los mismos términos que los establecidos en el algoritmo *localk2*, como se describe en la Sec. 8.1.2.

10.2. Ejemplo de Aplicación

Para mostrar el funcionamiento de *H2* sobre *redes heterogéneas*, utilizaremos las 42 estaciones del dominio de alta resolución, descrito en 5.1, divididas en dos grupos:

- *Antiguas*: Grupo formado por 12 estaciones, cuyo conjunto de entrenamiento

consta de $9860 - 1445 = 8415$ realizaciones, suficientemente largo como para evitar el sobreajuste; son las 12 primeras de la tabla 5.1.

- *Nuevas*: Grupo formado por las 30 restantes, cuyo conjunto de entrenamiento es de $4816 - 1445 = 3371$ realizaciones, aunque para simular un caso real y poner de manifiesto los efectos del sobreajuste, es deliberadamente reducido y solo se empleará un 10 %, es decir 337 realizaciones.

En todas ellas se han reservado 1445 realizaciones para test. En la Tabla 5.1, se ofrecen algunos datos geográficos y climatológicos básicos sobre las 42 estaciones seleccionadas. Además, se utilizará el predictor de 100 clases, ya que aunque el de 10 clases se sobreajusta menos, no aporta suficiente capacidad predictiva marginal como para ser operativo, de acuerdo con lo que ya se explicó en la Sec. 7.1.

En la Fig. 10.1, se ofrece una explicación gráfica del algoritmo en la que hemos utilizando colores para indicar las reglas del mismo. Los tipos de nodos son: *predictor*, con la etiqueta C^{100}_{8415}, en negro, predictando *antiguo*, con la etiqueta A^{2}_{8415}, con el borde negro y fondo gris, y predictando *nuevo*: A^{2}_{3371}, con el borde rojo y fondo inicialmente blanco. El superíndice indica la cardinalidad y el subíndice indica la máxima longitud de entrenamiento disponible, para diferenciar las series largas de las cortas. Los enlaces son de dos tipos: en negro, enlaces con 100 parámetros y en rojo: enlaces con 2 parámetros. Si un nodo tiene el fondo gris puede ser padre potencial, mientras que un nodo con el fondo blanco no, y un nodo con el borde negro no puede recibir enlaces, mientras que un nodo con el borde rojo si; de esta manera, los nodos cortos pasan a tener borde negro cuando tienen hijos.

10.2.1. Resultados

Como en el experimento 9.1, el algoritmo ha sido puesto a prueba para 10 repeticiones, realizadas cambiando los conjuntos de entrenamiento y de test aleatoriamente, obteniendo igualmente una gran cantidad de resultados, que se resumen en los gráficos y tablas que se muestran a continuación.

La Tabla 10.2.1, contiene los resultados de tomar la media de las capacidades predictivas locales (para cada estructura B, $RSA_B(a) = f(p_B(a|c))$) de las 10 repeticiones aleatorias, comparando la red *ingenua* con la generada por *H2*. Evidentemente en las 12 estaciones *antiguas* no se produce ningún cambio; sin embargo, para las restantes estaciones, como resultado de la sustitución del enlace al predictor (con 100 parámetros) por enlaces a los predictandos vecinos (con 2 parámetros), se obtienen capacidades predictivas marginales (valores de RSA en el conjunto de test)

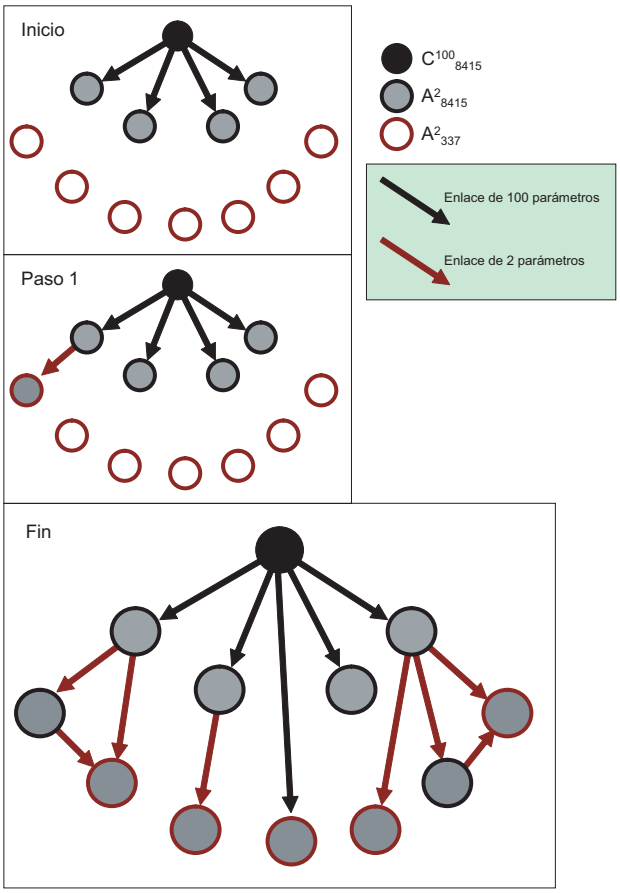

Figura 10.1: Explicación gráfica del algoritmo **H2** para *downscaling*.

claramente superiores en la mayoría de los casos. En esta tabla, hay además una tercera columna con la etiqueta *N max*, que estima el valor de RSA que potencialmente alcanzaría una red *ingenua* sin sobreajuste, es decir, suficientemente larga, y que en este experimento se estima utilizando la máxima longitud de entrenamiento (3371 datos, aproximadamente 10 años), cercana a la saturación, de acuerdo con el experimento que se describió en la figura 7.11.

Los resultados obtenidos se resumen tambien en las Figuras 10.3 y 10.4, constatando la capacidad de *H2* para adelantar la incorporación de nuevas redes de observación, y por tanto mejorando la resolución de la predicción local de precipitación, con calidades que dependen de la probabilidad del evento y, por supuesto, de la disponibilidad de estaciones antiguas en la vecindad. Por ejemplo, *Santander 'Centro'* es, desde 1997, el nuevo observatorio meteorológico principal de la Delegación de AEMET en Cantabria, y por ello dispone de un registro histórico que podemos catalogar de tipo *corto*, con un valor de $RSA = 0.48$ para el evento $prec > 0mm/24h$

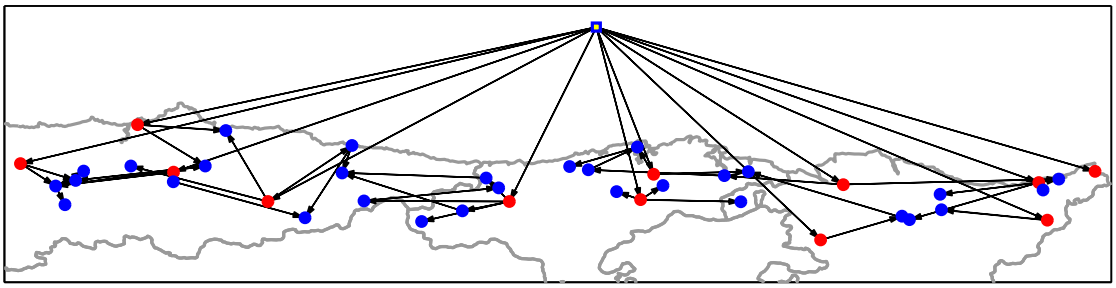

Figura 10.2: Ejemplo de DAG realizado con el algoritmo *H2* para el evento $prec > 10mm/24h$. Las series *antiguas* aparecen en rojo y, como se puede ver, son las únicas que tienen enlaces directos a la evidencia.

cuando se utilizar una red ingenua. Sin embargo, cuando se aplica *H2*, *Santander 'Centro'* se desconecta del estado de la atmósfera y se conecta a *El Mercadillo de Liérganes* (ver Fig. 10.2) mejorando su capacidad predictiva local hasta el valor $RSA = 0.67$, para el mismo evento y conjunto de entrenamiento.

Para terminar, se ha comparado *H2* con otra alternativa sencilla, con pocos parámetros, para estimar las predicciones en los nodos *cortos*; esta alternativa consiste en interpolar (o copiar) las predicciones obtenidas en los observatorios *largos* más cercanos. En la Figura 10.6 se muestran los resultados de dicha comparativa, donde se puede observar que, atendiendo a su sencillez, el método de *interpolación* es competitivo en áreas homogéneas, sin embargo, en áreas heterogéneas, donde las estaciones *cortas* no disponen de vecinos *largos*, resulta sensiblemente inferior a *H2*.

Los resultados experimentales obtenidos demuestran que, en casos de sobreajuste, donde solo tienen cabida métodos sencillos como la *interpolación* o el *clasificador ingenuo*, los resultados obtenidos por $H2$ son significativamente superiores a los citados métodos, aportando un aumento real de resolución en los lugares donde existen heterogeneidades. Por todo ello, la conclusión final es que la metodología propuesta permite incorporar redes relativamente nuevas (un año de funcionamiento) en los sistemas de predicción local, mejorando la resolución de las predicciones y logrando un avance en la generación de predicciones sobre fenómenos de variabilidad elevada. Por otra parte, este algoritmo se puede aplicar a otros eventos, teniendo en cuenta que su capacidad de adelanto respecto al modelo ingenuo, depende de la probabilidad del evento.

Tabla de medias de RSA para los eventos 0 y 10 $mm/24h$ (aparecen indicados entre paréntesis). Se comparan los resultados de la red *ingenua*, I, con los de *H2*. También se pone el valor de saturación que alcanzaría la red *ingenua*, IS, suficientemente larga (que para las series largas, es el mismo en los tres casos).

Localidad	$H2(0)$	$I(0)$	$IS(0)$	$H2(10)$	$I(10)$	$IS(10)$
FUENTERRABIA 'AEROPUERTO'	0.68	0.68	0.68	0.73	0.73	0.73
SAN SEBASTIAN 'IGUELDO'	0.68	0.68	0.68	0.72	0.72	0.72
ELDUAYEN	0.62	0.62	0.62	0.69	0.69	0.69
AMURRIO 'INSTITUTO'	0.66	0.66	0.66	0.67	0.67	0.67
BILBAO 'AEROPUERTO'	0.68	0.68	0.68	0.70	0.70	0.70
EL MERCADILLO DE LIERGANES	0.66	0.66	0.66	0.68	0.68	0.68
VILLACARRIEDO	0.66	0.66	0.66	0.70	0.70	0.70
ROZADIO	0.63	0.63	0.63	0.72	0.72	0.72
RANON 'AEROPUERTO DE ASTURIAS'	0.65	0.65	0.65	0.64	0.64	0.64
RIOSECO DE SOBRESCOBIO	0.68	0.68	0.68	0.68	0.68	0.68
OVIEDO 'EL CRISTO'	0.67	0.67	0.67	0.67	0.67	0.67
ZARDAIN	0.71	0.71	0.71	0.69	0.69	0.69
SAN SEBASTIAN 'ATEGORRIETA'	0.60	0.40	0.67	0.57	0.32	0.67
LASARTE-MICHELIN	0.61	0.41	0.66	0.62	0.29	0.69
EIBAR-BANCO DE PRUEBAS	0.66	0.45	0.70	0.71	0.42	0.75
ECHEVARRIA	0.65	0.45	0.68	0.68	0.37	0.73
ABADIANO 'MENDIOLA'	0.62	0.41	0.66	0.67	0.35	0.74
DURANGO-VIVERO	0.63	0.41	0.66	0.63	0.32	0.69
GURIEZO 'G.C.'	0.60	0.38	0.65	0.68	0.41	0.74
CARRANZA	0.64	0.44	0.68	0.71	0.38	0.76
COTERILLO DE AMPUERO	0.64	0.42	0.67	0.68	0.37	0.73
MIRONES	0.65	0.45	0.68	0.66	0.36	0.72
SANTANDER 'CENTRO'	0.67	0.47	0.71	0.68	0.41	0.74
ESCOBEDO DE VILLAFUFRE	0.61	0.38	0.64	0.63	0.31	0.69
TORRELAVEGA 'SNIACE'	0.68	0.45	0.70	0.72	0.38	0.76
CELIS	0.64	0.43	0.67	0.69	0.34	0.75
CAMIJANES	0.62	0.41	0.66	0.67	0.38	0.75
ESPINAMA	0.54	0.30	0.60	0.37	0.17	0.64
TAMA	0.62	0.41	0.67	0.59	0.34	0.78
AMIEVA 'RESTAÑO'	0.64	0.45	0.68	0.66	0.39	0.74
CANGAS DE ONIS	0.60	0.36	0.63	0.61	0.25	0.71
RIBADESELLA 'FARO'	0.63	0.41	0.68	0.64	0.32	0.74
GIJON	0.65	0.43	0.69	0.63	0.30	0.74
BEZANES	0.61	0.34	0.65	0.53	0.29	0.66
SOTO DE RIBERA	0.62	0.41	0.66	0.62	0.31	0.73
MERES DE SIERO	0.63	0.42	0.67	0.62	0.26	0.73
GRADO	0.61	0.37	0.66	0.57	0.28	0.71
GENESTAZA	0.67	0.45	0.69	0.61	0.25	0.68
SOTO DE LA BARCA	0.68	0.48	0.73	0.63	0.26	0.72
PRESA DE LA BARCA	0.68	0.46	0.72	0.60	0.27	0.73
SOTO DE LOS INFANTES	0.41	0.23	0.49	0.47	0.22	0.61
CUEVAS DE ALTAMIRA	0.64	0.42	0.67	0.63	0.30	0.71

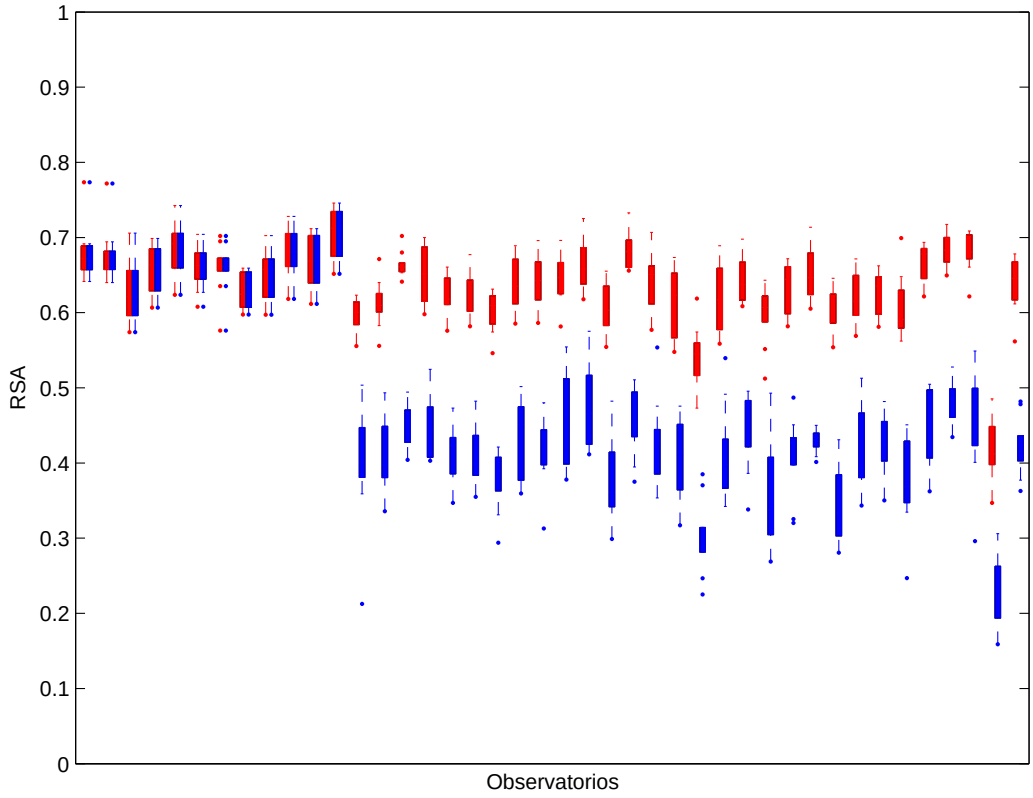

Figura 10.3: Diagrama de cajas con las mejoras obtenidas por *H2* para el evento $prec > 0mm/24h$ (en rojo) frente al clasificador ingenuo (en azul). Las 12 primeras son las series *antiguas*, para las cuales no se produce ningún cambio.

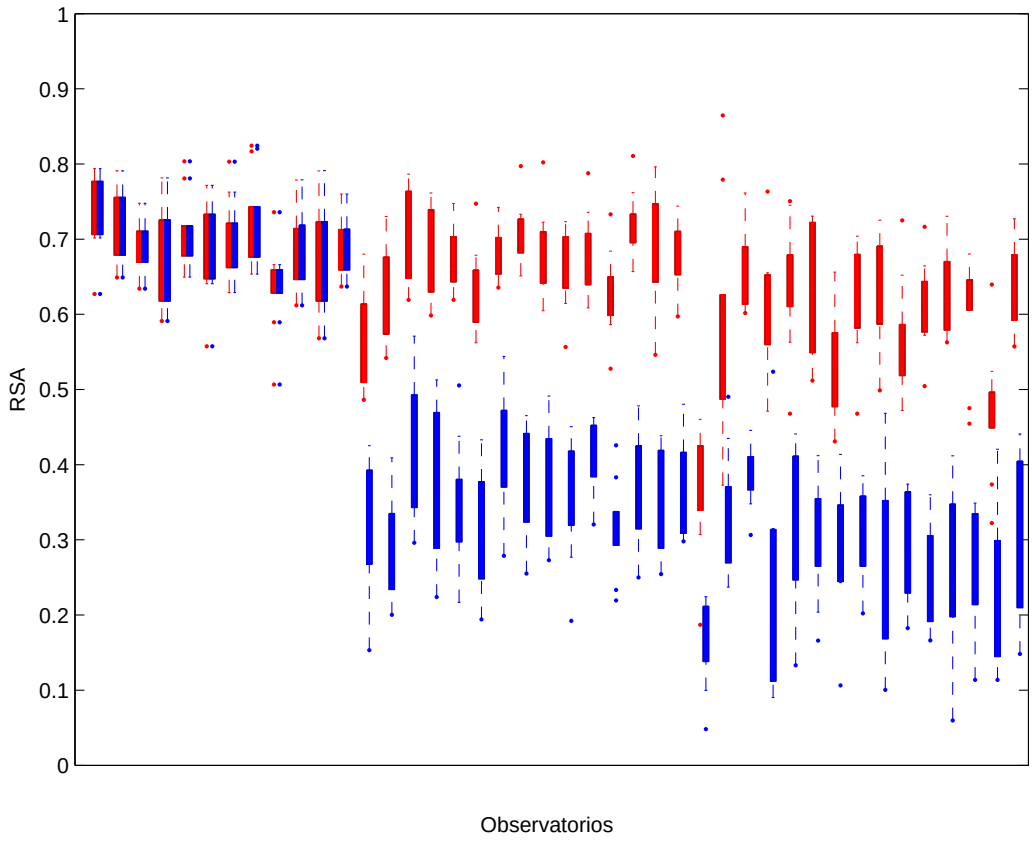

Figura 10.4: Diagrama de cajas con las mejoras obtenidas por *H2* (en rojo) frente al clasificador ingenuo (en azul), para el evento $prec > 10mm/24h$; las 12 primeras son las series *antiguas*, para las cuales no se produce ningún cambio.

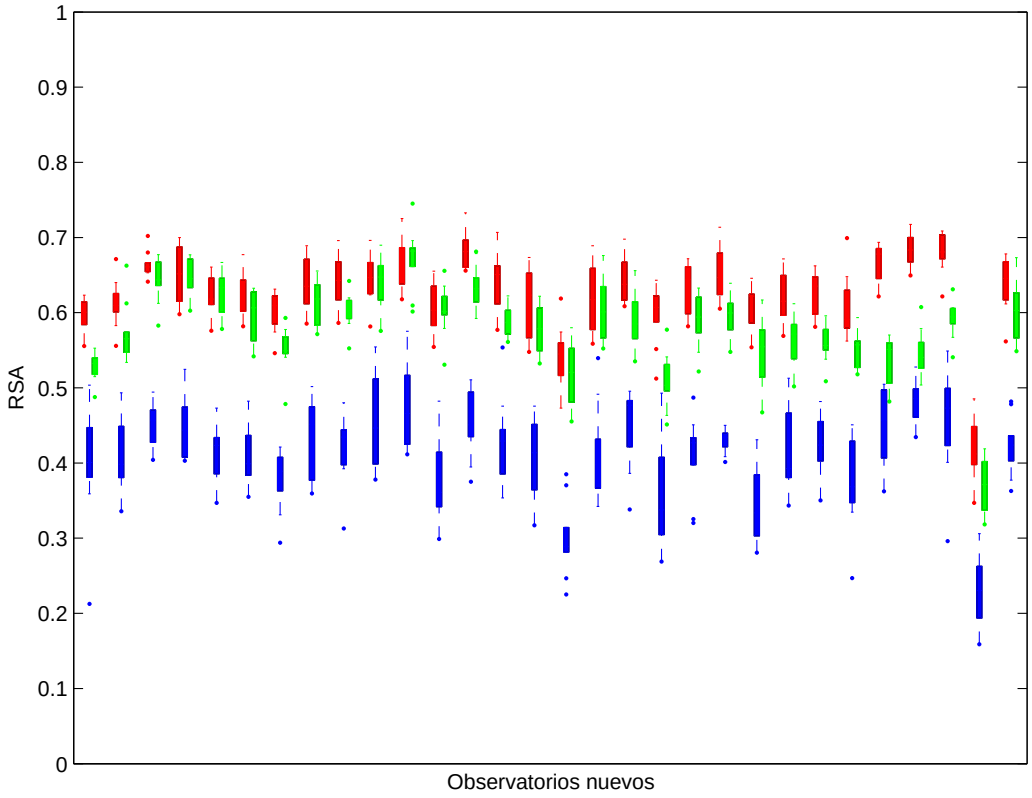

Figura 10.5: Diagrama de cajas comparativo entre *H2* y el método *interpolación* para el evento $prec > 0mm/24h$. Los resultados de *H2* aparecen en rojo, frente al clasificador ingenuo, en azul y los de *interpolación* en verde. Solamente se muestran las series *nuevas*.

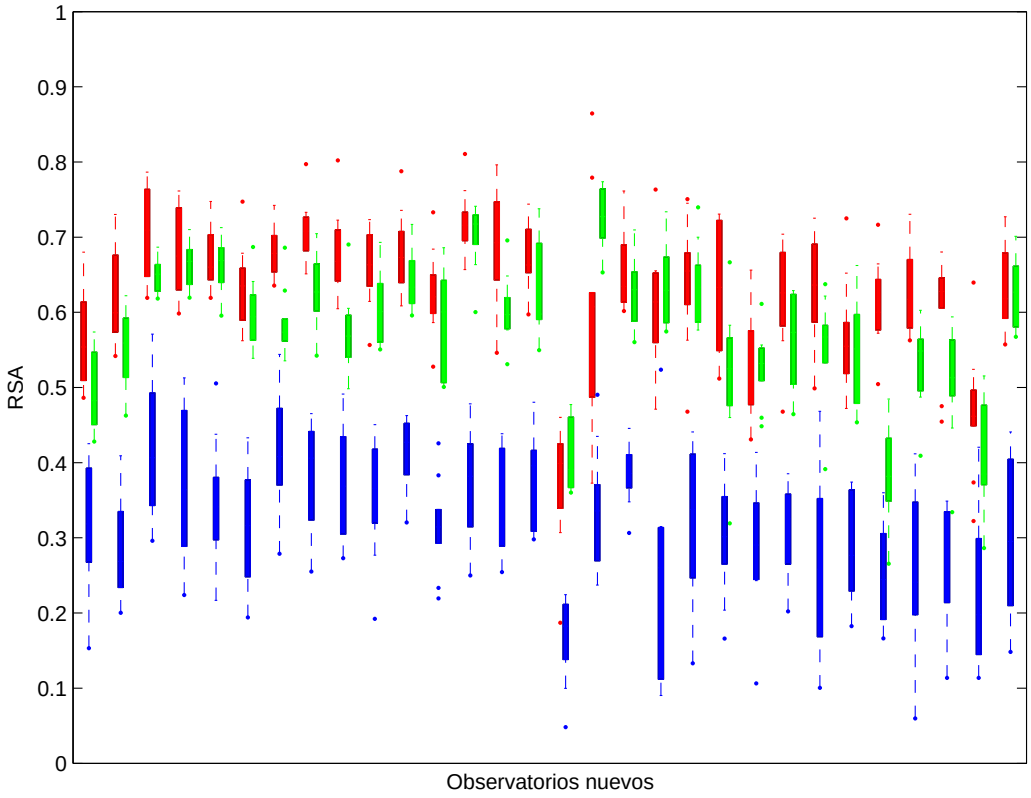

Figura 10.6: Diagrama de cajas comparativo entre *H2* y el método *interpolación* para el evento $prec > 10mm/24h$

.

CAPÍTULO 11

Conclusiones y Trabajo Futuro

Esta Tesis supone el primer estudio general de las aplicaciones de las redes probabilísticas, en particular de las redes Bayesianas (RBs), al servicio de la predicción meteorológica. En particular, se han analizado aplicaciones en el ámbito de los *Generadores de Tiempo* y la *Predicción Local*, poniendo de manifiesto que las RBs son una metodología capaz de modelar el acoplamiento espacial de los eventos meteorológicos mediante un modelo probabilístico global. Posteriormente, se ha realizado un estudio más profundo de la *Predicción Local de Precipitación*, para lo cual, ha sido necesario adaptar métodos específicos de aprendizaje automático, desarrollando algoritmos de búsqueda local, como *localK2* y *localB* (ver sec. 8.1.2). Además, a partir de *localB*, se ha desarrollado un algoritmo jerárquico para la predicción de eventos binarios de precipitación de alta resolución, denominado H2 (ver sec. 10.1), que permite incorporar nuevas redes de observación a la cadena operativa de predicción contando con el apoyo de estaciones de referencia, con series más largas. Aparte de las aplicaciones en sí, también se ha realizado un estudio comparativo mediante varios experimentos de predicción local para tres topologías básicas de RB (siempre multinomial y estática): *ingenua*, *extendida* y *genérica*, comparando medidas globales (como MDL), con medidas locales (como AUC), con el fin último de evaluar las posibilidades y limitaciones de los diferentes modelos. Concretamente, la topología de red ingenua se corresponde con un método estándar de predicción local (el método de análogos), utilizado en numerosas aplicaciones prácticas. Por tanto, además de mostrar que las RBs son capaces de generalizar esta metodología estándar, también hemos probado que generalizan en un único modelo tres de los

principales modos de operación en meteorología como son: *downscaling*, *predicción* y *diagnóstico*.

A continuación se resumen las principales conclusiones que aportan alguna mejora al estado del conocimiento.

11.1. Conclusiones

1. Por primera vez se constata la utilidad de las RBs en problemas relacionados con la predicción local y especialmente en el (*downscaling* estadístico).

2. Los algoritmos estándar de aprendizaje automático que resultan eficientes en este problema para aprender los modelos a partir de los datos presentan ciertas deficiencias. Por ejemplo, el algoritmo *K2* sufre una notable degradación de la información debido a la ordenación previa (ver figura 8.1 y la sección 8.1.1). Por otra parte, otros algoritmos más complejos, como *B*, resultan ineficientes para este tipo de problemas de elevada dimensión. Por tanto, para tener en cuenta las características particulares de este problema (elevada dimensión, alta correlación espacial, etc.), ha sido necesario desarrollar algoritmos específicos de aprendizaje automático (ver Sec. 8.1).

3. Asimismo, se recomienda utilizar como medida de calidad el AUC (área bajo la curva ROC, RSA), ya que se ha constatado que la medida de calidad MDL (longitud de descripción mínima) no es representativa de la capacidad predictiva local en este tipo de sistemas (ver Sec. 8.2).

4. En consecuencia, se han desarrollando nuevos algoritmos de búsqueda local (*localK2* y *localB*) específicos para series de precipitación, adaptando algunos métodos estándar de aprendizaje estructural para su aplicación en la predicción local (ver sección 8.1.2); también se han realizado pruebas sobre otras variables con resultados prometedores.

5. Los modelos basados en RBs son útiles para explicar el comportamiento, tanto individual como colectivo, de predictandos cuya variabilidad observada no es capturada por los modelos de circulación. De esta manera, su campo óptimo de operación son los sistemas formados por modelos globales de circulación y redes de observación de alta resolución, resultando menos útiles en el *downscaling* individual a partir de LAMs sobre series largas de baja resolución, ver Sec. 7.3.1.

6. En general, para conjuntos de entrenaniento suficientemente grandes, la capacidad predictiva de las RBs es superior a la del *modelo Ingenuo*, ver fig. 7.8; ya que, cuando se utiliza el *modelo Ingenuo* para estimar probabilidades conjuntas, se comete un error sistemático (que en en el caso de la precipitación subestima las coincidencias), que resulta ser tanto mayor cuanto más lejos estemos de la hipótesis *ingenua*, ver sec. 7.3.1. Sin embargo, cuando el predictor explica la variabilidad observada del predictando o cuando la variabilidad del predictando no es resuelta por la red de observación, se cumple la hipótesis de independencia condicional entre los predictandos y por lo tanto el mejor modelo es el *Ingenuo* en cualquiera de los paradigmas, tanto para el comportamiento individual, como para el colectivo de los predictandos (ver Sec. 7.3). Incluso aunque no se cumpla dicha hipótesis, el *modelo Ingenuo* es el que posee mayor capacidad predictiva individual en modo *downscaling*, independientemente de la calidad del predictor y de la resolución de la red de predictandos (ver Sec. 9.1).

7. Las conclusiones son diferentes cuando se trabaja con conjuntos de entrenamiento pequeños, concretamente cuando se trabaja en modo *downscaling*, se ha constatado que el *modelo Ingenuo* es más sensible al sobreajuste (pierde su capacidad predictiva individual más rápidamente) que las redes generadas con el algoritmo *localB(rsa)*. Esto ocurre debido a que *localB(rsa)* es capaz de sustituir enlaces directos al estado de la atmósfera por otros indirectos a través de nodos vecinos, reduciendo el número de parámetros. Por lo tanto, cuando la longitud de entrenamiento no permita un adecuado ajuste de parámetros (como es el caso de la mayoría de las nuevas redes de observación de alta resolución), incluso el *modelo Ingenuo* resulta demasiado complejo y debe ser sustituido por otro con menos parámetros, ver sec. 7.3.3. Para ello, se ha desarrollado un nuevo algoritmo, denominado *H2*, que permite integrar en un mismo modelo diferentes redes de observación, y que además, allí donde existan estaciones antiguas en la vecindad, es capaz de habilitar estaciones recientes a la predicción operativa, necesitando un periodo de aprendizaje significativamente inferior al requerido por el *Clasificador Ingenuo* para evitar el sobreajuste.

8. La metodología propuesta puede ser aplicada a la predicción local de diferentes fenómenos meteorológicos en diversas escalas.

11.2. Trabajo Futuro

EL trabajo realizado en esta Tesis supone un estudio en "modo perfecto", es decir, utilizando el .estado real"de la atmósfera (en realidad el estado dado por los proyectos de reanálisis, que asimilan las observaciones desde el espacio real al espacio del modelo). Las posibilidades abiertas en el campo del análisis, la diagnosis y la predicción, tanto meteorológica como climatológica, son enormes, abriendo muy prometedoras líneas de trabajo futuro. Entre las más inmediatas cabe citar:

- Se adapta naturalmente a las nuevas tendencias de predicción por conjuntos y multimodelo actualmente operativas en el medio plazo, estacional y climático, por lo que su migración a dichos alcances es relativamente sencilla.

- Ya se han realizado estudios preliminares para el tratamiento conjunto de series de variables contínuas -utilizando redes *gaussianas*, definidas en la sección 3.4.1- para un mejor control de calidad, homogeneización y relleno de lagunas, publicados en Cano and Gutiérrez (2004); por lo que se ampliará su aplicación a variables *gaussianas* utilizando esta metodología.

- Generación de Modelos Conceptuales: Los patrones atmosféricos empleados en esta Tesis, son genéricos, por lo que queda pendiente la investigación en la selección de patrones específicos, obteniendo simultáneamente un mejor conocimiento y una mejor predicción de los sistemas que dan lugar a fenómenos adversos. En este aspecto, ya se han obtenido algunos resultados, por lo que se puede adelantar que la mejora será significativa en la mayoría de los predictandos.

- Actualmente el sistema sólo se activa con la entrada de nueva información prevista por los modelos numéricos, es decir sólo está desarrollado operativamente el *software* en modo *downscaling*, quedando interesantes opciones por desarrollar, como son los modos de *predicción* y *diagnóstico*; este último en sus versiones de *clasificación* e *interpolación*. La opción de *predicción*, por ejemplo, permitiría operar al sistema en el rango de la vigilancia y el muy corto plazo, para lo cual basta con activar y actualizar la predicción cada vez que se recibe nueva información observada. Al tratarse de un sistema computacionalmente muy eficiente, esta operación se podría realizar y monitorizar en tiempo real.

- Los modelos probabilísticos basados en RBs tambien pueden ser utilizados en la simulación y análisis de escenarios atmosféricos de interés. Uno de los ejemplos más interesantes de esta alternativa es el estudio del cambio climático,

donde, desde esta nueva perspectiva, se pueden estimar las heterogeneidades en las anomalías de la fenomenología local asociada a los diferentes escenarios atmosféricos previstos por los modelos climáticos. Esta es una de las tareas más inmediatas a desarrollar, ya que una vez constatado el calentamiento global, queda mucho trabajo pendiente, pues aunque se trate de un fenómeno global, no lo hace con la misma intensidad en todas partes; motivo por el cual, los agentes sociales tienen la apremiante necesidad de conocer en detalle cuales son las áreas más sensibles al cambio, para así poder poner en marcha procedimientos de evaluación de impacto, vulnerabilidad, mitigación y, en su caso, de adaptación. En este sentido, quizás este sea el aspecto con mayor prioridad a la hora de promover su continuidad en un futuro próximo.

- Aunque en esta Tesis se ha estudiado principalmente el paradigma de clasificación inverso o *downscaling*, la clasificación es una línea de trabajo con mucho por desarrollar en meteorología, por ejemplo, en áreas como la climatología descriptiva o en el diagnóstico probabilístico de sistemas multimodelo, donde sería interesante aplicar el concepto de comités de clasificadores en la línea que se describe en: (Sierra et al., 2001).

- Una línea futura de investigación, ya comprometida, es la adaptación de los métodos y herramientas desarrollados a diferentes alcances temporales, variables y características, ya que los resultados de esta Tesis tienen una implicación directa con parte de la actividad investigadora propia de la AEMET. Contando con la participación del autor de la Tesis en diferentes programas y proyectos, concretamente, participa de forma directa en el *Programa Coordinado para la Generacion de Escenarios Climáticos* (coordinado por la AEMET, como parte del Plan Nacional de Adaptación al Cambio Climático); asimismo participa en otro proyecto del Gobierno de Cantabria titulado *Escenarios Regionales de Cambio Climatico en Cantabria*, siendo previsible también su participación en otro proyecto tambien del Gobierno de Cantabria titulado: *Estrategia de Accion Frente al Cambio Climático en Cantabria*. Además, como parte del grupo *Meteorology & Data Mining*, se mantiene una intensa cooperación con la UNICAN en otros proyectos como ENSEMBLES del VI Programa Marco (2004-2008) y el citado *Programa Coordinado*.

- Finalmente, la AEMET ha dispuesto otorgar, bajo la tutoría del autor de esta Tesis una beca de dos años de duración titulada: *Adaptación de métodos de downscaling estadístico a escalas de tiempo climático en la Cornisa Cantábrica*,

que se viene desarrollando desde Enero de 2009.

Parte III

Bibliografía

Bibliografía

Abramson, B., Brown, J., Edwards, J., Murphy, A., Winkler, R., and Robert, L. (1996). Hailfinder: A bayesian system for forecasting severe weather. *International Journal of Forecasting*, 12(1):57–71.

Acid, S., Campos, L. M., and Castellano, J. G. (2005). Learning bayesian network classifiers: Searching in a space of partially directed acyclic graphs. *Mach. Learn.*, 59(3):213–235.

Alexandersson, H. (1986). Homogeneity test applied to precipitation data. *Journal of Climatology*, 6:661–675.

Ancell, R., Cofiño, A. S., Sordo, C., San-Martín, D., and Gutiérrez, J. M. (2009). Multi-site probabilistic downscaling using bayesian networks. a new weather-typing approach. Submitted to Climate Dynamics.

Ancell, R. and Gutiérrez, J. M. (2009). Bayesian networks for high resolution probabilistic downscaling. Submitted to Journal of Hydrology.

Anderberg, M. R. (1973). *Cluster Analysis for Applications*. Academic Press, New York.

Ayuso, J. J. (1994). *Predicción Estadística Operativa en el INM*, vol. B-34 of *Monografías del Instituto Nacional de Meteorología*. Ministerio de Medio Ambiente, Madrid.

Barnett and Preisendorfer (1987). Origins and levels of monthly and seasonal forecast skill for united states surface air temperatures determined by canonical correlation analysis. *Monthly Weather Review.*, 115:1825–1850.

Barry, R. and Chorley, R. (1998). *Atmosphere, Weather and Climate*. Routledge, London, 7th ed.

Beersma, J. and Buishand, T. (2003). Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation. *Climate Research*, 25:121–133.

Bellone, E., Hughes, J., and Guttorp, P. (2000). A hidden markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Climate research*, 15(1):1–.

Bertacchi Uvo, C., Olsson, J., Morita, O., Jinno, K., Kawamura, A., Nishiyama, K., Koreeda, N., and Nakashima, T. (2001). Statistical atmospheric downscaling for rainfall estimation in kyushu island, japan. *Hydrology and Earth System Sciences*, 5:259–271.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer.

Blanco, R., Inza, I., and Larrañaga, P. (2003). Learning bayesian networks in the space of structures. *International Journal of Intelligent Systems*, 18:202–220.

Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. (1996). Context-specific independence in Bayesian networks. In *Procedings of the 12th Conference on Uncertainty in Artificial Intelligence*, pp. 115–123.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3.

Buizza, R. and Palmer, T. N. (1995). Singular vector structure of the atmospheric global circulation. *Journal of Atmospheric Sciences*, 52:1434.

Buntine, W. (1991). Theory refinement on bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 52–60. Morgan Kaufmann Publishers.

Buntine, W. (1996). A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8(2):195–210.

Burguer, G. (1996). Expanded downscaling for generating local weather scenarios. *Climate Research*, 7:111–128.

Cannon, A. J. (2007). 2007 special issue: Nonlinear analog predictor analysis: A coupled neural network/analog model for climate downscaling. *Neural Netw.*, 20(4):444–453.

Cano, R., Cofiño, A., López, F., Gutiérrez, J., and Rodríguez, M. (2001). Aplicación de métodos de clasificación al downscaling estadístico. In *Actas del V Simposio Nacional de Predicción*, pp. 49–50. Instituto Nacional de Meteorología, Madrid.

Cano, R. and Gutiérrez, J. (2004). Relleno de lagunas y homogeneización de series de precipitación en redes densas a escala diaria. In *El Clima, entre el Mar y la Montaña*, pp. 441–450. AEC, 2004, Serie A, nº4. URL `www.aeclim.org/publicac.html`.

Cano, R., Sordo, C., and Gutiérrez, J. M. (2004). Bayesian networks in meteorology. In J. A. Gámez, S. Moral, and A. Salmerón, eds., *Advances in Bayesian Networks*, pp. 309–327. Springer Verlag.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1996). Goal oriented symbolic propagation in bayesian networks. In *AAAI/IAAI, Vol. 2*, pp. 1263–1268.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1997). *Expert Systems and Probabilistic Network Models*. Springer-Verlag. Free Spanish version `http://personales.unican.es/gutierjm`.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1998). Modelling probabilistic networks of discrete and continuous variables. *Journal of Multivariate Analysis*, 64:48–65.

Cavazos, T. (1997). Downscaling large-scale circulation to local winter rainfall in north-eastern Mexico. *International Journal of Climatology*, 17:1069–1082.

Cavazos, T. (2000). Using Self-Organizing Maps to investigate extreme climate event: An application to wintertime precipitation in the Balkans. *Journal of Climate*, 13:1718–1732.

Chen, M., Xie, P., Janowiak, J. E., and Arkin, P. A. (2002). Precipiation: A 50-yr monthly analysis based on gauge observations. *Journal of Hydrometeorology*, 3:249–266.

Cheng, J., Bell, D. A., and Liu, W. (1997). Learning belief networks from data: An information theory based approach. In *In Proceedings of the Sixth ACM International Conference on Information and Knowledge Management*, pp. 325–331.

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees.

Cios, K., Pedrycz, W., Swiniarski, R., and Kurgan, L. (2007). *Data Mining: A Knowledge Discovery Approach (Hardcover).*. Springer. ISBN 978-0-387-33333-5.

Cofiño, A. S. (2004). *Técnicas Estadísticas y Neuronales de Agrupamiento Adaptativo para la Predicción Probabilística de Fenómenos Meteorológicos Locales. Aplicación en el Corto Plazo y en la Predicción Estacional*. Ph.D. thesis, Universidad de Cantabria.

Cofiño, A. S., Cano, R., Sordo, C., and Gutiérrez, J. M. (2002). Bayesian networks for probabilistic weather prediction. In *Proceedings of the 15th European Conference on Artificial Intelligence*, pp. 695–700.

Cooper, G. F. and Herskovitz, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

de Campos, L. (2006). A scoring function for learning bayesian networks based on mutual information and conditional independence tests. *J. Mach. Learn. Res.*, 7:2149–2187.

de Campos, L. and Castellano, J. (2007). Bayesian network learning algorithms using structural restrictions. *International Journal on Approximate Reasoning*, 45(2):233–254.

Duda, R., Hart, P., and Stork, D. (2000). *Pattern Classification*. Wiley-Interscience, 2nd ed.

Eckert, P., Cattani, D., and Ambühl, J. (1996). Classification of ensemble forecasts by means of an artificial neural network. *Meteor. Appl.*, 3:169–178.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, 29(2–3):131–163.

FRIEDMAN, N. and GOLDSZMIDT, M. (1996). Building classifiers using bayesian networks. In *Procedings of the National Conference on Artificial Intelligence*, pp. 1277–1284.

FRIEDMAN, N., LINIAL, M., NACHMAN, I., and PEÈR, D. (2000). Using bayesian networks to analyze expression data. *J Comp Biol*, 7:601–620.

FRIEDMAN, N., NACHMAN, I., and PEÈR, D. (1999). Learning bayesian network structure from massive datasets: The 'sparse candidate' algorithm. In *Procedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pp. 206–215.

GABRIEL, K. and NEUMANN, J. (1962). A markov chain model for daily rainfall occurrence at tel aviv. *Quarterly Journal of the Royal Meteorological Society*, pp. 90–95.

GALEA, M., SHEN, Q., and LEVINE, J. (2004). Evolutionary approaches to fuzzy modelling for classification. *Knowl. Eng. Rev.*, 19(1):27–59.

GARDNER, M. W. and DORLING, S. R. (1998). Artificial neural networks (the multilayer perceptron). a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32:2627–2636.

GEIGER, D. and HECKERMAN, D. (1994). Learning gaussian networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pp. 235–243. Morgan Kaufmann Publishers.

GLAHN, H. (1974). Forecasting local weather by means of model output statistics. *Bulletin of the American Meteorological Society*, 55:1217–1227.

GÁMEZ, J. and PUERTA, J. (2005). Constrained score+(local)search methods for learning bayesian networks. *Lecture Notes in Computer Science*, 3571:161–173.

GRASSBERGER, P. and PROCACCIA, I. (1983). Characterization of strange attractors. *Physical Review Letters*, 50:346–349.

GROSSMAN, D. and DOMINGOS, P. (2004). Learning bayesian network classifiers by maximizing conditional likelihood. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, p. 46. ACM, New York, NY, USA. ISBN 1-58113-828-5.

Gutiérrez, J. M., Primo, C., Rodríguez, M. A., and Fernández, J. (2008). Spatiotemporal characterization of ensemble prediction systems - the mean-variance of the logarithms (MVL) diagram. *Nonlin. Processes Geophys.*, 15(1):109–114. URL `http://www.nonlin-processes-geophys.net/15/109/2008`.

Gutiérrez, J. M., Cano, R., Cofiño, A. S., and Rodríguez, M. A. (1999). Redes neuronales y patrones de analogías aplicados al downscaling climático. In *La Climatología Española en los Albores del Siglo XXI*, pp. 113–121. Asociación Española de Climatología.

Gutiérrez, J. M., Cano, R., Cofiño, A. S., and Rodríguez, M. A. (2004a). Clustering methods for statistical downscaling in short-range weather forecast. *Monthly Weather Review*, 132(9):2169–2183.

Gutiérrez, J. M., Cano, R., Cofiño, A. S., and Sordo, C. M. (2004b). *Redes Probabilísticas y Neuronales en las Ciencias Atmosféricas*. Monografías Ministerio de Medio Ambiente.

Gutiérrez, J. M., Cofiño, A. S., and Cano, R. (2005). Analysis and downscaling multi-model seasonal forecasts in perú using self-organizing maps. *Tellus A*, 57:435–447.

Gutiérrez, J. M., Cofiño, A. S., Cano, R., and Sordo, C. (2002). A generalization of analogue downscaling methods by bayesian networks. In *International Conference on Quatitative Precipitation Forecasting*, pp. 87–87. The World Weather Research Programme's WWRP.

Harris, D., Foufoula-Georgiou, K., Droegemeier, K., and Levit, J. (2001). Multiscale statistical properties of a high-resolution precipitation forecast. *Journal of Hydrometeory*, 2:406–418.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

Heckerman, D. (1996). A tutorial on learning with bayesian networks. Tech. Rep. MSR-TR-95-06, Microsoft Research.

Henrion, M. (1988). Propagation of uncertainty by logic sampling in bayes networks. In J. F. Lemmer and L. N.. Kanal, eds., *Proceedings of the 2nd Annual Conference on Uncertainty in Artificial Intelligence*, pp. 149–164.

Herrera, S., Pons, M. R., Cofiño, A. S., and Gutiérrez, J. M. (2007). Development and analysis of a daily high resolution grid over spain (1950-2000). In *8th European Conference on Applications of Meteorology*. EMS.

Holton, J. (1992). *An Introduction to Dynamic Meteorology*. Academic Press.

Hruschka, E. and Ebecken, N. (2005). Applying bayesian networks for meteorological data mining. In *Applications and Innovations in Intelligent Systems XIII*, pp. 122–133. Springer-Verlag.

Hsieh, W. W. (2000). Nonlinear canonical correlation analysis by neural networks. *Neural Networks*, 13:1095–1105.

Huang, J. and Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Trans. on Knowl. and Data Eng.*, 17(3):299–310.

Hughes, J., Guttorp, P., and Charles, S. (1999). A non-homogeneous hidden markov model for precipitation occurrence. *Applied statistics*, 48(1):15–.

Hughes, J., Lettemaier, D. P., and Guttorp, P. (1993). A stochastic approach for assessing the effect of changes in regional circulation patterns on local precipitation. *Water Resources Research*, 29:3303–3315.

Jagannathan, P. and Arlqy, R. (1967). A note on climatological normals. Tech. Rep. 84-208, W.M.O.

Jangmin, O., Lee, J., Park, S., and Zhang, B. (2004). Stock trading by modelling price trend with dynamic bayesian networks. *Lectures Notes in Computer Science*, 3177:794–799.

Janis, M., Hubbard, K., and Redmond, K. (2004). Station density strategy for monitoring long-term climatic change in the contiguous united states. *J. Climate*, 17:151–162.

Jansá, A. (1990). *Notas sobre análisis meteorológico mesoescalar en niveles bajos.*. Instituto Nacional de Meteorología, Madrid.

Jing, Y., Pavlović, V., and Rehg, J. M. (2005). Efficient discriminative learning of bayesian network classifier via boosted augmented naive bayes. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pp. 369–376. ACM, New York, NY, USA. ISBN 1-59593-180-5.

Jolliffe, I. and Stephenson, D. (2003). *Forecast Verification*. John Wiley and Sons.

Jordan, M. (1998a). *Learning in Graphical Models*. MIT Press.

Jordan, M. I. (1998b). *Learning in Graphical Models*. MIT Press.

Kalnay, E. (2003). *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, Cambridge.

Kalnay, E., Kanamitsu, M., Kistle, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., and Joseph, D. (1996). The NMC/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77:437–471.

Kennett, R., Korb, K., and Nicholson, A. (2001). Seabreeze prediction using bayesian networks. *Lecture Notes in Computer Science*, 2035:1611–3349.

Keogh, E. and Pazzani, M. (1999). Learning augmented bayesian classifiers: A comparison of distribution-based and classification-based approaches. In *In Proceedings of the Seventh International Workshop on Artificial Intelligence Statistics*, pp. 225–230.

Kovalerchuk, B. and Vityaev, E. (2003). *Data Mining in Finance: Advances in Relational and Hybrid Methods.*. Springer, Kluwer Academic. ISBN 0792378040.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *Neural Computation*, 9(7):1493–1516.

Krause, P. (1999). Learning probabilistic networks. *Knowl. Eng. Rev.*, 13(4):321–351.

Lam, W. and Bacchus, F. (1994). Learning bayesian beliefe networks: An approach based on the mdl principle. In *Computational Intelligence*, pp. 269–293.

Langseth, H. and Nielsen, T. D. (2002). Classification using hierarchical naive bayes models. URL `citeseer.ist.psu.edu/langseth02classification.html`.

Langseth, H. and Nielsen, T. D. (2006). Classification using hierarchical naïve bayes models. *Mach. Learn.*, 63(2):135–159.

Larrañaga, P., Kuijpers, C., Murga, R., and Yurramendi, Y. (1996a). Learning bayesian network structures by searching for the best ordering with genetic algorithms. *IEEE Transaction on Systems, Man and Cybernetics - Part A*, 26:487–493.

Larrañaga, P., Lozano, J., Peña, J., and Inza, I. (2003). Workshop on probabilistic graphical models for classification. In *Proceedings of the Workshop on Probabilistic Graphical Models for Classification. ECML/PKDD, 71-82, 2003*. ECML/PKDD.

Larrañaga, P., Lozano, J., Peña, J., and Inza, I. (2005). Editorial. *Mach. Learn.*, 59(3):211–212.

Larrañaga, P., Poza, M., Yurramendi, Y., Murga, R. H., and Kuijpers, C. M. H. (1996b). Structure learning of bayesian networks by genetic algorithms: A performance analysis of control parameters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(9):912–926.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of Atmospheric Science*, 20:130–141.

Lorenz, E. N. (1969). Atmospheric predictability as revealed by naturally occuring analogues. *Journal of the Atmospheric Sciences*, 26:636–646.

Mach, M. and Mráz, M. (2005). Quality measures for bayesian network classifier design. http://www.bmf.hu/conferences/SAMI2005/mach.pdf.

Mandelbrot, B. (2004). *Fractals and Chaos*. Springer Verlag.

McCann, R., Marcot, B., and Ellis, R. (2006). Bayesian belief networks: applications in natural resource management. *Canadian Journal of Forest Research*, 36:3053–3062.

Mediero Orduña, L. (2007). *Pronóstico probabilístico de caudales de avenida mediante redes bayesianas aplicadas sobre un modelo hidrológico distribuido.*. Ph.D. thesis, Ingeniería Civil: Hidráulica y Energética / E.T.S.I. Caminos, Canales y Puertos (UPM).

Michaelides, S. C., S., P. C., and Kleovoulou, G. (2001). Classification of rainfall variability by using artificial neural networks. *International Journal of Climatology*, 21:1401–1414.

Middendorf, M., Reischle, F., and Schmeck, H. (2002). Multi colony ant algorithms. *Journal of Heuristics*, 8(3):305–320.

Mittal, A. and Kassim, A. (2007). *Bayesian Network Technologies: Applications and Graphical Models*. IGI Publishing, 2007.

Mohamadi, H., Habibi, J., Abadeh, M. S., and Saadi, H. (2008). Data mining with a simulated annealing based fuzzy classification system. *Pattern Recogn.*, 41(5):1824–1833.

Murphy, A. H. (1973). A new vector partition of probability score. *Journal of Applied Meteorology*, 12:595–600.

Murphy, A. H. (1993). What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, 8:281–293.

Murphy, A. H. and Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, 115:1330–1338.

Murphy, J. (1999). An evaluation of statistical and dynamical techniques for downscaling local climate. 12:2256–2284.

Murphy, K. P. (2001). The bayes net toolbox for matlab. *Computing Science and Statistics*, 33. Http://www.cs.berkeley.edu/ murphyk/Bayes/bnt.html.

Nadi, F., Agogino, A., and Hodges, D. (1991). Use of influence diagrams and neural networks in modeling semiconductor manufacturing processes. *IEEE Transactions on Semiconductor Manufacturing*, 4:52–58.

Neapolitan, R. (2003). *Learning Bayesian Networks*. Prentice-Hall, Inc.

Neil, M., Fenton, N., and Nielson, L. (2000). Building large-scale bayesian networks. *The Knowledge Engineering Review*, 15:257–284.

Nikovski, D. (2000). Constructing bayesian networks for medical diagnosis from incomplete and partially correct statistics. *EEE Transactions on Knowledge and Data Engineering*, 12(4):509 – 516. Special issue on constructing Bayesian networks.

Orlanski, I. (1975). A rational subdivision of scales for atmospheric processes. *Bulletin of the American Meteorological Society*, 56(5):527–530.

P. Larra n., Kuijpers, C. M. H., Murga, R. H., Inza, I., and Dizdarevic, S. (1999). Genetic algorithms for the travelling salesman problem: A review of representations and operators. *Artif. Intell. Rev.*, 13(2):129–170.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Delecluse, P., Deque, M., Diez, E., Doblas-Reyes, F. J., Feddersen, H., Graham, R., Gualdi, S., Gueremy, J. F., Hagedorn, R., Hoshen, M., Keenlyside, N., Latif, M., Lazar, A., Maisonnave, E., Marletto, V., Morse, A. P., Orfila, B., Rogel, P., Terres, J. M., and Thomson, M. C. (2004). Development of a european ensemble system for seasonal to interannual prediction (DEMETER). *Bulletin of the American Meteorological Society*, 85(6):853–872. URL `http://www.ecmwf.int/research/demeter/`.

Peña, J. M., Lozano, J. A., and Larrañaga, P. (1999). An empirical comparison of four initialization methods for the k-medias algorithm. *Pattern Recognition Letters*, 20:1027–1040.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Philander, S. G. (1990). *El Niño, La Niña, and the Southern Oscillation*. Academic Press, San Diego.

Pourret, O., Naim, P., and Marcot, B. (2008). *Bayesian Belief Networks: A Practical Guide to Applications*. John Wiley.

Pérez-Muñuzuri, V. and Gelpi, I. R. (2000). Application of nonlinear forecasting techniques for meteorological modeling. *Annales Geohysicae*, 18:1349–1359.

Primo, C., Szendro, I., Rodríguez, M. A., and Gutiérrez, J. M. (2007). Error growth analysis in systems with spatial chaos: Coupled map lattices and global weather models. *Physical Review Letters*, 98:108501.

Pu, Z.-X., Kalnay, E., Parrish, D., Wu, W., and Toth, Z. (1997). The use of the bred vectors in the ncep operational 3-dimensional variational system. *Weather and Forecasting*, 12:689–695.

Rayner, N. A., Horton, E. B., Parker, D. E., Folland, C. K., and Hackett, R. B. (1996). Version 2.2 of the global sea-ice and sea surface temperature data set, 1903-1994. report crtn 74. Hadley Centre for Climate Prediction and Research.

Richardson, C. W. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17:182–190.

Richardson, D. S. (2000). Skill and economic value of the ECMWF ensemble prediction system. *Quaterly Journal of the Royal Meteorological Society*, 126:649–668.

Sanchez, E., J.L.and Garcia Ortega and Marcos, J. (2001). Construction and assessment of a logistic regression model applied to short-term forecasting of thunderstorms in leon (spain). 15:57–71.

Sanguesa, R. and Cortes, U. (1997). Learning causal networks from data: a survey and a new algorithm for recovering possibilistic causal networks. *AI Commun.*, 10(1):31–61.

Sansom, J. (1999). Large-scale spatial variability of rainfall through hidden semi-markov models of breakpoint data. *Journal of geophysical research*, 104(24):31361.

Sierra, B., Serrano, N., Larranaga, P., Plasencia, E. J., Inza, I., Jimenez, J. J., Revuelta, P., and Mora, M. L. (2001). Using bayesian networks in the construction of a bi-level multi-classifier. a case study using intensive care unit patients data. *Artificial Intelligence in Medicine*, 22(3):233–248. URL `citeseer.ist.psu.edu/article/sierra00using.html`.

Smith, T. M., Reynolds, R. W., Livezey, R. E., and Stokes, D. C. (1996). Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *Journal of Climate*, 9:1403–1420.

Strogatz, S. (2001). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*. Perseus Book Group, 1st ed.

Su, J. and Zhang, H. (2006). Full bayesian network classifiers. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pp. 897–904. ACM, New York, NY, USA. ISBN 1-59593-383-2.

Talagrand, O. (1997). Evaluation of probabilistic prediction systems. In *Workshop on predictability*, pp. 1–25. ECMWF.

Tsagalidis, E., Tsitouridis, K., Evangelidis, G., and Dervos, D. (2008). Hail size estimation and prediction using data mining techniques. In *ERAD 2008. The Fifth European con-*

ference on Rafa in Meteorology and Hydrology, p. 072. URL `http://erad2008.fmi.fi/proceedings/extended/erad2008-0072-extended.pdf`.

VAN DEN DOOL, H. M. (1989). A new look at weather forecasting though analogs. *Montly Weather Review*, 117:2230–2247.

VON STORCH, H. (1999). On the use of "inflation"in statistical downscaling. *Journal of Climate*, 12:3505–3506.

WEBB, G. I., BOUGHTON, J. R., and WANG, Z. (2005). Not so naive bayes: Aggregating one-dependence estimators. *Mach. Learn.*, 58(1):5–24.

WILBY, R., TOMLINSON, O., and DAWSON, C. (2003). Multi-site simulation of precipitation by conditional resampling. *Climate Research*, 23:183–194.

WILBY, R. L. and WIGLEY, T. M. L. (1997). Downscaling general circulation model output. A review of methods and limitations. *Progress in Physical Geography*, 21:530–548.

WILBY, R. L. and WILKS, D. S. (1999). The weather generation game. A review of stochastic weather models. *Progress in Physical Geography*, 23:329–357.

WILKS, D. S. (1995). *Statistical Methods in the Atmospheric Sciences. An Introduction.*. Academic Press.

WILKS, D. S. (1999a). Multisite downscaling of daily precipitation with a stochastic weather generator. *Climate Research*, 11:125–136.

WILKS, D. S. (1999b). Multisite downscaling of daily precipitation with a stochastic weather generator. *Climate Research*, 11:125–136.

WILSON, L. J. (2001). Verification of precipitation forecasts: A survey of methodology. Tech. rep., Environment Canada. URL `www.chmi.cz/meteo/ov/wmo/src/wilson_paper_prague.pdf`.

WITTEN, I. and FRANK, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*.. Morgan Kaufmann. ISBN 0-12-088407-0.

WONG, S. and LI, C. (2006). *LIFE SCIENCE DATA MINING (Science, Engineering, and Biology Informatics)*.. World Scientific Publishing; 1 edition (December 29, 2006). ISBN 978-9812700650.

Wood, A. (2004). Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, 62:1573–1480.

Yu, P., Chen, S., and Chang, I. (2006). Support vector regression for real-time flood stage forecasting. 328:704–716.

Zepeda-Arce, J., Foufoula-Georgiou, K., and Droegemeier, K. (2000). Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *J. Geophys. Res.*, 10:129–146.

Zhang, N. (2002). Hierarchical latent class models for cluster analysis. URL `citeseer.ist.psu.edu/zhang02hierarchical.html`.

Zorita, E., Hughes, J. P., Lettemaier, D. P., and von Storch, H. (1995). Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. *Journal of Climate*, 8:1023–1042.

Zorita, E. and von Storch, H. (1999). The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *Journal of Climate*, 12:2474–2489.

Zwiers, F. and von Storch, H. (1990). Regime dependent auto-regressive time series modeling of the southern oscillation. *Journal of Climate*, pp. 1347–1360.