VNIVERSITAT DE VALÈNCIA

Facultat de Ciències Matemàtiques

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Tratamiento bayesiano de valores ausentes en datos espacio-temporales

Tesis presentada por:
Carlos Abellán de Andrés
y dirigida por:
Antonio López Quílez
dentro del programa de doctorado:
Estadística y Optimización

Valencia, Septiembre de 2015

Antonio López Quílez, profesor titular del Departamento de Estadística e Investigación Operativa de la Universitat de València,

CERTIFICA que la presente Tesis

Tratamiento bayesiano de valores ausentes en datos espacio-temporales

ha sido realizada bajo su dirección por Carlos Abellán de Andrés en el Departamento de Estadística e Investigación Operativa.

Y para que así conste, firma el presente certificado.

Burjassot, de Mayo de 2015
Antonio López Quílez

A Pau, Martina y, por supuesto, a Rut

Prólogo

La historia de esta tesis es bien larga. Tanto, que sería muy difícil detallarla a través de estas líneas que, dicho sea de paso, son las que mayor placer me proporciona escribir.

Desde el 2001, año en el que comencé mi andadura en el departamento de Estadística e Investigación Operativa de la Universitat de València, hasta el presente 2015 en el que presento esta tesis siendo profesor de matemáticas en enseñanza secundaria, no sólo han pasado 14 años, también han pasado infinidad de experiencias vitales.

Queramos o no, las decisiones que tomamos a lo largo de nuestra vida, por insignificantes que puedan ser, van desentrañando la multitud de sendas y caminos que forman el mapa de nuestro destino y que, presuntamente, se cruzan y se separan caprichosamente o, como diríamos en nuestro argot, de forma aleatoria. Sin embargo, hay momentos en la vida, en que todo fluye suavemente y dicho entramado de rutas parece converger hacia un mismo punto, más elevado que el resto, en el que se divisa con total claridad el camino recorrido y, mejor aún, el que queda por recorrer. Tengo la certeza de que este momento, es uno de ellos.

Desde la paz y la calma que se disfruta en este punto del recorrido, se aprecia con mayor nitidez que más importante aún que el propio camino andado son las personas con las que he compartido la aventura. Miro a mí alrededor, y no encuentro sino palabras de agradecimiento a aquellos que, en mayor o menor media, han sido protagonistas de esa infinidad de experiencias vitales. Famila,

amigos, compañeros de trabajo, profesores, tutores... el trayecto no ha sido un esfuerzo individual sino el resultado de la suma de un esfuerzo colectivo. Y a todos les estaré eternamente agradecido.

Digno de destacar es el verdadero esfuerzo que ha tenido que hacer Antonio López Quílez, director de este trabajo, para no perder la paciencia con un alumno tan irregular como yo. En varias ocasiones he dejado de lado el desarrollo de esta tesis, en alguna de ellas con visos de un abandono definitivo, para posteriormente ponerme en contacto con él con el fin de retomarla. Antonio, lejos de ejercer el merecido derecho al desistimiento de ayudar a tan informal estudiante, siempre me ha acogido con los brazos abiertos y la mejor de sus sonrisas. Más aun, sin gozar de las mejores condiciones personales para soportar la tutoría por partes que esta tesis ha supuesto, Antonio ha ejercido su labor de dirección con la sabiduría que acostumbra y un ánimo infranqueable a la vez que contagioso. Proporcionándome siempre tanto esa explicación que ilustra con claridad lo que tantas dudas me generaba en un momento determinado, como el aliento necesario en esos otros momentos en los que el peso de la investigación, del trabajo o incluso de la vida misma estaba haciendo mella en mi esfuerzo. Muchas gracias Antonio.

Otro esfuerzo encomiable ha sido el de Juanjo Abellán Andrés, y no sólo el relativo a mostrar interés por leerse el presente trabajo con el espíritu constructivo adecuado como para aportar valor añadido, sino por el esfuerzo que ha supuesto ser mi hermano mayor, preocuparse por mi educación, por mi bienestar personal y profesional. Nuestro parentesco me sitúa en la situación privilegiada de haber compartido con él muchísimos de los momentos vitales que he experimentado, de disfrutar de su compañía, tanto para el ocio como para lo profesional, de mantener profundos debates sobre cualquier cosa, humana o divina, y de compartir una de nuestras grandes pasiones en común, la estadística. A este respecto, lo que para él supone una trivial conversación, para mí son lecciones magistrales que siempre me han ayudado a aprender y a crecer en

seguridad y motivación. Gràcies de tot cor Germanot.

Pero sin duda, la persona que mayor esfuerzo ha hecho para que este proyecto salga adelante es Rut. Mi ángel de la guarda. Mi compañera de fatigas. Ella se ha preocupado de todo lo necesario para proporcionarme las condiciones óptimas para la realización de este trabajo. Gracias a ella he encontrado el ánimo y, sobre todo, el tiempo para finalizarlo. Sin su ayuda estoy seguro de que esta tesis habría acabado en el tintero, y sin su presencia ni siquiera hubiera tenido sentido acabarla. La mayor de mis suertes es poder poner el punto y final a este trabajo a su lado. Y además tener la satisfacción de celebrarlo no sólo con ella, sino con nuestros dos amores, Pau y Martina. Que, lejos de suponer una rémora en el tiempo necesario para llevar adelante esta tesis, han supuesto el oasis de relajación y desconexión, tan útil como necesario, para retomar con mayor fuerza la faena pendiente. Y cuya opinión ha sido crucial a la hora de tomar decisiones nada triviales como el color de las gráficas.

Se dice que para embarcarse en una empresa del calibre de una tesis hace falta mucho ánimo, fuerza de voluntad y tiempo. En mi caso, todo se reduce a contar con la más importante y la mejor de mis experiencias vitales: los Abellán-López. Os quiero infinito. Gracias por todo vuestro esfuerzo.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

La estadística es la ciencia que utiliza conjuntos de datos para obtener inferencias basándose en el cálculo de probabilidades. Sin embargo, con frecuencia se tiene que hacer frente a la presencia de valores ausentes en dicho conjunto de datos sobre los que se desea realizar un análisis estadístico.

La existencia de valores ausentes representa un auténtico quebradero de cabeza a la hora de implementar un estudio. No sólo implica una mayor complejidad asociada al análisis sino que también introduce mayor incertidumbre en el estudio. Este hecho puede derivar en posibles sesgos en los resultados finales y, por tanto, cuestionar la fiabilidad en las conclusiones que de estos se deriven.

No es de extrañar entonces, la innumerable cantidad de publicaciones existentes asociadas al tema. Estas abarcan casi la totalidad de campos de investigación en los que se requiere del análisis estadístico.

Sin embargo, salvo excepciones, el tratamiento que se realiza de la existencia de valores ausentes no es el adecuado. La mayor parte de estas publicaciones, o bien hacen caso omiso de la no completitud de sus datos, procediendo a analizarlos mediante técnicas estadísticas para datos completos; o bien hacen referencia a cómo abordar el análisis de este tipo de datos incompletos

únicamente desde el punto de vista de la imputación de los valores ausentes. Es decir, plantean completar los datos no observados con metodologías que varían en complejidad pero no llegan a estudiar en profundidad el origen de las no-respuestas. Ni tampoco plantean ningún tipo de estudio de sensibilidad asociado a las metodologías utilizadas para la imputación.

Basta ver, a modo de ejemplo, el estudio que realizaron Wood *et al.* (2004) en el que revisaron las publicaciones entre Julio y Diciembre de 2001 aparecidas en las revistas médicas BMJ, JAMA y Lancet and New England Journal of Medicine. En esta revisión, observaron que de un total de 71 ensayos clínicos el 89 % presentaba valores ausentes. Este hecho demuestra lo habitual que resulta este problema en el análisis de datos. No obstante, lo más inquietante fue que en el 92 % de los 26 ensayos clínicos univariantes y en el 46 % de los 37 ensayos clínicos con medidas repetidas, se analizaron los datos con técnicas estadísticas diseñadas para bancos de datos completos (en algunos casos se eliminaban las observaciones con algún dato ausente y en otros se incluían igualmente en el estudio). En el resto se llevó a cabo una imputación de los valores ausentes haciendo uso de técnicas rudimentarias. Y únicamente en el 21 % de estas imputaciones se implementó un análisis de sensibilidad respecto a la estimación de los datos no observados.

Si esto ocurre en apenas 6 meses y únicamente en el campo de investigación de ensayos clínicos, este ejemplo muestra que quizás el tratamiento habitual que se haga de la información faltante en un análisis de datos no sea el adecuado.

Cuando el conjunto de valores se analiza como si fuera un banco de datos completamente observado, la fiabilidad de los resultados podría quedar comprometida. La información faltante podría ser importante a la hora de analizar los datos, pudiendo incluso modificar los resultados finales en caso de que se conociera. Por otro lado, si se eliminan las observaciones con algún valor ausente la pérdida de información es evidente. Disponer de menos observaciones conduce a una pérdida de potencia estadística y, en

consecuencia, de nuevo la fiabilidad de los resultados quedaría en entredicho.

Es necesario, por tanto, hacer frente con rigurosidad al problema de la existencia de valores ausentes. Se debe plantear el análisis de datos de forma global, no sólo diseñando el estudio estadístico a realizar sino también planteando un tratamiento adecuado de la información faltante.

Para este propósito dos factores asociados a la existencia de valores ausentes adquieren relevancia. En primer lugar identificar el mecanismo de aparición de los valores ausentes. Su modelización es clave, pues dicho mecanismo influirá en el posterior tratamiento de los datos no observados. En segundo lugar, una vez planteado un modelo para el mecanismo de aparición de valores ausentes, se debe diseñar un modelo de imputación de datos y se debe decidir si completar la base de datos mediante imputación simple, imputación múltiple o sencillamente incluir la modelización de los valores ausentes dentro del estudio principal.

En el primer caso, identificar el mecanismo de generación de los valores ausentes conlleva la detección de una posible dependencia en la aparición de estos, o entre la aparición de valores ausentes y los valores observados. Desafortunadamente, es muy difícil captar este patrón de aparición de valores ausentes a partir de los propios datos. No obstante, la experiencia que manifiestan las publicaciones que analizan este aspecto dictamina que la hipótesis de independencia en la aparición de datos no observados es la más frecuente. Este hecho es importante pues simplifica el proceso de inferencia, permitiendo así centrar la atención únicamente en la estructura de los datos incompletos con el fin de modelizar una imputación de los valores ausentes.

En casos infrecuentes de sospecha de no independencia en la generación de valores ausentes, ampliar el banco de datos observados mediante la incorporación de covariables o realizar un análisis de sensibilidad respecto a mecanismos que modelicen la aparición de los valores ausentes, serían las estrategias más

adecuadas para tratar el problema.

Respecto a los tipos de imputación, si se opta por la imputación simple se proporcionará un valor estimado para cada uno de los valores ausentes. De este modo el conjunto de datos pasaría a estar completamente observado. En el caso de optar por imputación múltiple, se asignará una muestra de m estimaciones para cada valor ausente. Así, se dispondría de m conjuntos de datos completamente observados implicando la realización de m análisis estadísticos. Mediante el proceso conocido como *Reglas de Rubin* se resumirían los resultados de estos m análisis para integrarlos en un único resultado final. Por último, si lo que realmente interesa es analizar los datos sin necesidad de reparar en la imputación de los valores ausentes, sería conveniente incluir el modelo de imputación de datos dentro del estudio principal.

Si la existencia de valores ausentes es habitual en bases de datos, no lo es menos en aquellos conjuntos de datos caracterizados por la presencia de correlación espacio-temporal. El tratamiento y análisis de datos con dicha estructura forma parte del campo de la denominada *estadística espacio-temporal*. Esta parte del análisis estadístico de datos ha experimentado un importante auge en las últimas décadas coincidiendo con el rápido desarrollo de nuevas tecnologías computacionales. No obstante, y pese a la gran cantidad de publicaciones científicas existentes alrededor de la estadística espacio-temporal, no existe ningún estudio en este campo propiamente dedicado al tratamiento de valores ausentes.

Así pues, con el ánimo de hacer frente a bases de datos provistas de estructura espacio-temporal e incompletas debido a la existencia de valores ausentes, en esta tesis nos planteamos los siguientes objetivos:

1. Plantear una comparativa de tres posibles modelos de imputación, diseñados para adaptarse a la realidad espacio-temporal de los datos a los que nos enfrentamos, con el fin de llevar a cabo la imputación de la información faltante asumiendo un proceso de aparición de valores ausentes

aleatorio.

El planteamiento de las modelizaciones presentadas en esta tesis parte de la recurrencia con la que aparecen en la literatura específica de la estadística espacio-temporal. En este trabajo adaptamos las ideas que subyacen en cada uno de los modelos al campo de la imputación de valores ausentes. Y estudiamos no sólo si las imputaciones que proporcionan estos modelos son adecuadas sino también se analiza cómo cada modelo se adapta y ajusta a la correlación espacial y temporal de los datos.

Cabe destacar que, a pesar de que las modelizaciones estudiadas en esta tesis son muy citadas en la literatura espacial y temporal, el número de referencias en las que se plantea una comparativa de este tipo es más bien escaso. Por tanto, esta tesis pretende también ser el comienzo de una línea de investigación que estudie la mejor forma de tratar el frecuente problema de la existencia de valores ausentes en el campo de la estadística espacio-temporal, aportando al mismo tiempo una comparativa entre modelizaciones espacio-temporales poco frecuente hasta la fecha.

2. Llevar a cabo dicha comparativa de modelizaciones mediante su aplicación en la base de datos de calidad del agua potable en la Comunitat Valenciana. Esta información, al estar recogida a nivel municipal entre los años 1991-2000, se considera susceptible de ser tratada atendiendo a una posible correlación espacio-temporal. En particular, el estudio de las modelizaciones de imputación lo centramos sobre las concentraciones de magnesio y de nitratos existentes en el agua potable. Siendo estos últimos valores de concentraciones de nitratos los que, debido a su estructura, nos permitirán analizar con mayor detalle el comportamiento de los modelos.

3. Realizar un estudio de sensibilidad en el que llevaremos a cabo una simulación de bancos de datos dotados con estructura

espacio-temporal. Procediendo después de nuevo a comparar las tres modelizaciones de imputación sobre dichos datos simulados. De esta forma, el hecho de conocer los verdaderos valores asociados a los datos simulados como ausentes, nos aportará más claridad en el estudio de cómo se comporta cada uno de los modelos de imputación propuestos.

Dicho estudio simulado se planteará con la intención de apoyar el proceso comparativo planteado en el primer objetivo. Nuestra intención no es establecer un modelo de referencia para la imputación de valores ausentes en bases de datos con estructura espacio-temporal sino ilustrar un modo de afrontar el problema de la existencia de dichos valores ausentes en este contexto.

Todo el proceso descrito de estudio y comparación de modelizaciones espacio-temporales, tanto para el caso del banco de datos de calidad del agua potable como para los bancos de datos simulados, lo llevaremos a cabo bajo la perspectiva bayesiana.

Esta tesis está estructurada en ocho capítulos. Partiendo de la breve descripción realizada en el presente capítulo se pasa seguidamente a la revisión de la imputación de datos en el capítulo 2. En este se hará un pequeño resumen histórico y se explicará con más detalle el mecanismo de generación de valores ausentes. Se expondrá analíticamente el tratamiento estadístico asociado a un conjunto de datos incompleto por la presencia de valores no observados y, por último, se describirán los tipos habituales de imputación de estos.

En el capítulo 3 se hará un modesta revisión global de los diferentes procesos que conllevan el tratamiento de datos espaciales y espacio-temporales, así como de las técnicas aplicadas con mayor frecuencia para su análisis. Haremos hincapié en la cartografía de enfermedades, motor de la investigación en redes de localizaciones, con el fin de contextualizar nuestro trabajo. Finalmente presentaremos los datos de calidad del agua potable en la Comunitat Valenciana.

Posteriormente, comenzaremos a analizar dichos datos con el objetivo de estudiar y comparar modelos espacio-temporales de imputación. En primer lugar se justifica la necesidad de utilizar tales modelos a través del estudio exploratorio previo realizado en el capítulo 4. Para ello se plantea una comparativa entre modelos que aumentan progresivamente en complejidad. Estos varían desde un sencillo modelo de heterogeneidad hasta un modelo completo que incorpora componentes capaces de recoger la posible correlación espacio-temporal de los datos.

Seguidamente, en el capítulo 5, procederemos a la presentación y explicación detallada de los modelos espacio-temporales a comparar en la imputación de las concentraciones de nitratos. Expondremos la motivación que nos ha llevado a la selección de dichos modelos basándonos en su estructura y en la forma en que esta es capaz de captar la naturaleza espacio-temporal de los datos.

Los resultados de la comparación de los modelos se presentan y discuten en el capítulo 6. Se mostrarán los mapas de las concentraciones de nitratos completadas mediante imputación con cada uno de los modelos. Se expondrá el criterio tenido en cuenta para la comparación de las modelizaciones estudiadas. Los resultados mostrados incluirán una medida de la calidad de las imputaciones proporcionadas por los modelos. Asimismo proporcionaremos también una medida de cómo ajusta cada modelo la naturaleza espacio-temporal de los datos imputados. Finalmente trataremos de clarificar con ejemplos puntuales el comportamiento de cada uno de ellos en el proceso de imputación.

El análisis y comparación de los modelos termina en el capítulo 7. En este, se aborda el estudio del comportamiento de estos en la imputación de datos simulados. Cuantificaremos de nuevo la calidad de las imputaciones proporcionadas por cada modelo, esta vez medida sobre tres bancos de datos espacio-temporales diferentes.

Finalizaremos con el capítulo 8 en el que destacaremos las conclusiones más importantes de este trabajo así como se propondrán nuevas líneas de investigación que, a nuestro entender,

podrían ser de gran ayuda en el tratamiento del problema de la existencia de valores ausentes.

Capítulo 2

Imputación de datos

El problema de datos ausentes es tan antiguo como la propia estadística. Tener información faltante ha ido desde siempre ligado a la gran mayoría de análisis de datos en cualquier campo de aplicación de la estadística. Un banco de datos incompleto complica cualquier tipo de investigación asociada a dichos datos, introduciendo mayor incertidumbre en el estudio y llegando a provocar, si se ignora el problema o se trata de una forma errónea, resultados no fiables.

Esta omnipresencia del problema de datos ausentes en el análisis estadístico ha generado innumerables publicaciones científicas asociadas al problema. El número de referencias crece exponencialmente tal y como puede observarse en la figura 2.1. Tal cantidad de publicaciones abarcan tanto el aspecto metodológico como el de la aplicación en una amplia variedad de disciplinas científicas.

No obstante, desde la perspectiva histórica, puede considerarse a Donald B. Rubin como el pionero en abordar formalmente el problema de datos ausentes dando origen a lo que desde entonces se conoce como *imputación de datos*. Suyas son gran parte de las contribuciones científicas asociadas a la imputación de datos, comenzando por las más relevantes que ayudaron a desarrollar el problema en las décadas de los 70 y 80, y continuando hasta la

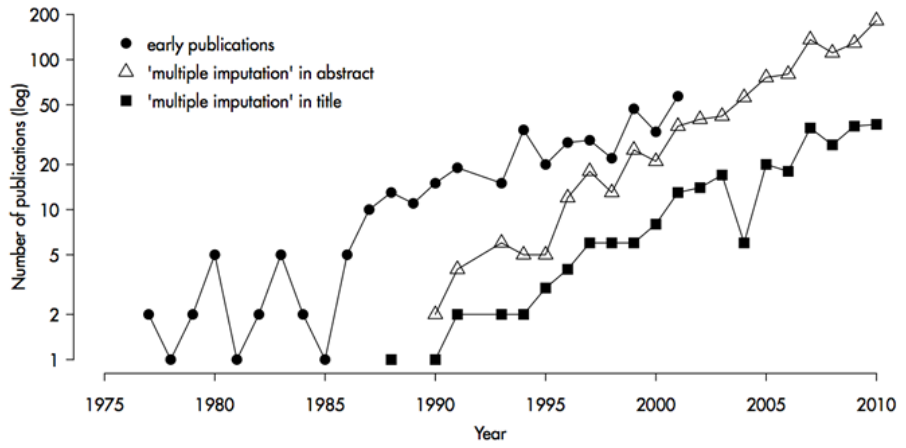


Figura 2.1: Número de publicaciones a escala logarítmica asociadas a imputación múltiple durante el periodo 1977-2010. Fuente: www.stefvanbuuren.nl

actualidad con la participación en libros y aplicaciones en diferentes campos de investigación.

Estos trabajos han supuesto no sólo la oficialización del problema de datos ausentes, casi inexistente en la literatura hasta entonces (Anderson, 1957; Buck, 1960; Afifi y Elashoff, 1966; Hartley y Hocking, 1971; Orchard y Woodbury, 1972), sino también una gran ayuda al desarrollo de posibles soluciones mediante técnicas de imputación de datos.

Al principio de los años 70 Rubin se encontró, trabajando en el Departamento de Evaluación de la Educación, en Princeton, con el problema de la existencia de valores ausentes en los estudios poblacionales que llevaban a cabo. Para solventar las dificultades que conllevaba trabajar con bancos de datos (de dimensiones importantes) incompletos, Rubin resolvió completar la información faltante dando comienzo a la hoy tan extendida imputación de datos ausentes. A partir de ahí surgieron gran cantidad de trabajos, véase por ejemplo Dempster *et al.* (1977); Rubin (1976, 1978, 1979), en los que se desarrollaron los conceptos asociados a la existencia de

valores ausentes así como las técnicas de imputación utilizadas para dar solución al problema.

Una década más tarde, publicó el libro que se conoce como el origen de la imputación múltiple (Rubin, 1987), la técnica de imputación más utilizada en la actualidad.

Posteriormente, en Rubin (1996), después de más de 18 años trabajando en el tema, Rubin realiza una completa revisión del problema de datos ausentes, de los tipos y técnicas de imputación así como de la literatura publicada hasta la fecha. Y nuevamente una década después, Rubin publicaría una nueva y completa revisión de obligada lectura, Little y Rubin (2002).

De forma coetánea a esta última revisión de la imputación de valores ausentes, empiezan a aparecer publicaciones de otros investigadores que también han contribuido favorablemente al tópico. Así, Schafer (1999); Schafer y Graham (2002) son dos interesantes referencias a tener en cuenta por la sencillez y claridad con la que explican y revisan la problemática de la información faltante, mientras que en Ibrahim *et al.* (2005) se vuelve a realizar una importante y completa revisión del problema para modelos lineales generalizados.

Con el transcurrir de los años, la investigación y desarrollo de la imputación de valores ausentes crece a un ritmo vertiginoso a la vez que se extiende a casi la totalidad de las ciencias. En las primeras décadas dicho desarrollo se llevaba a cabo básicamente desde la perspectiva frecuentista. Más o menos a partir de la década de los 90, gracias a los avances en computación, el problema también comienza a ser tratado bajo la óptica bayesiana. Este hecho provoca que el número de trabajos publicados alcance tal magnitud que dificulta sobremanera su revisión y seguimiento.

Debido a esto, cabe destacar el papel que juegan las obras más actuales, Molenberghs y Kenward (2007); Daniels y Hogan (2008); VanBuuren (2012), Carpenter y Kenward (2013) o Gelman *et al.* (2013), pues se han convertido en guías de la investigación y el tratamiento de valores ausentes en la actualidad.

El propósito de este capítulo es describir los aspectos más relevantes del tratamiento de valores ausentes. Comenzaremos por la descripción del mecanismo de generación de valores ausentes para posteriormente exponer el tratamiento analítico del problema. Finalizaremos con la explicación de los tipos de imputación de los datos no observados más importantes, imputación simple, imputación múltiple e imputación bayesiana.

2.1. Mecanismo de generación de valores ausentes

Un aspecto muy importante a tener en cuenta en el tratamiento de valores ausentes es el mecanismo que genera la aparición de dichos datos no observados. Esto es, ¿los datos ausentes siguen algún patrón detectable, o por el contrario la aparición de valores ausentes es totalmente aleatoria?

Detectar correctamente este mecanismo de generación de valores ausentes es crucial a la hora de realizar la imputación y, por tanto, a la hora de realizar inferencias. Si no se tiene en cuenta o si se detecta erróneamente, las imputaciones realizadas pueden estar sesgadas y por tanto, las conclusiones extraídas del estudio pueden ser incorrectas.

En muchas ocasiones, dada la complejidad de los datos, es muy difícil determinar correctamente dicho mecanismo de generación de valores ausentes. El procedimiento que se suele seguir en estos casos es realizar un análisis de sensibilidad respecto a los supuestos de mecanismos de generación de valores ausentes, véase por ejemplo Lee y Carlin (2010); Resseguier *et al.* (2011) o Mason *et al.* (2012b). Es decir, realizar una imputación diferente para cada uno de los supuestos de mecanismos de generación de valores ausentes y, para cada imputación, realizar la inferencia deseada. Así se estará en condiciones de analizar la robustez y firmeza de las conclusiones del estudio frente a la elección del mecanismo de generación de

datos ausentes.

Rubin (1976) sentó las bases de la inferencia con datos incompletos debido a la existencia de valores ausentes partiendo para ello de la clasificación del mecanismo de generación de valores ausentes. De forma resumida, según Rubin existen tres tipos de mecanismos: valores ausentes completamente aleatorios (MCAR, missing completely at random), valores ausentes aleatorios (MAR, missing at random) y valores ausentes no aleatorios (MNAR, missing not at random).

MCAR ocurre cuando la aparición de valores ausentes no depende de ninguna de las variables que intervienen en el análisis. No depende ni de los valores observados ni del resto de valores ausentes, es decir, no existe ningún patrón detectable de generación de valores ausentes o, en caso de existir, este no influye en el análisis. Por ejemplo, supongamos el estudio de una enfermedad sobre la que se está interesado analizar la posible influencia que una determinada covariable pueda tener. MCAR sería el caso en el que la causa de aparición de valores ausentes en dicha covariable no guarda relación alguna ni con la enfermedad a estudio ni con la parte observada de la covariable.

Sin embargo, un mecanismo clasificado como MAR es aquel en el que se observa dependencia entre la aparición de valores ausentes y los valores que sí se han observado. En el ejemplo anterior, se consideraría MAR el caso en el que la aparición de los valores ausentes en la covariable tiene algún tipo de relación con los valores que se han observado de la propia covariable.

Cuando un mecanismo no se puede considerar MCAR ni MAR, estamos ante un MNAR.

2.2. Tratamiento estadístico

Formalmente, si llamamos $\boldsymbol{y}$ al conjunto de datos (matriz de n registros con r items) del que se dispone para realizar un determinado estudio, introducimos la notación $\boldsymbol{y} = (\boldsymbol{y_{obs}}, \boldsymbol{y_{aus}})$, donde $\boldsymbol{y_{obs}}$ son los valores observados del conjunto de datos e $\boldsymbol{y_{aus}}$ son los valores ausentes.

Se define el indicador $\boldsymbol{I} = \{I_{ij}\}$ con $i = 1, ..., n$ y $j = 1, ...r$, donde I_{ij} será 1 si la correspondiente componente de $\boldsymbol{y}$ es observada y 0 si es ausente. De esta forma $\boldsymbol{I}$ está completamente observado.

Así, se tiene que la distribución conjunta de los datos, $\boldsymbol{y}$, y del indicador $\boldsymbol{I}$ es:

$$p(\boldsymbol{y}, \boldsymbol{I}|\boldsymbol{\phi}, \boldsymbol{\theta}) = p(\boldsymbol{y_{obs}}, \boldsymbol{y_{aus}}, \boldsymbol{I}|\boldsymbol{\phi}, \boldsymbol{\theta}), \tag{2.1}$$

donde $\boldsymbol{\phi}$ es el vector de parámetros de la distribución de los datos y $\boldsymbol{\theta}$ representa el vector de parámetros desconocidos que gobiernan la generación de valores ausentes.

De esta forma, se observa que aplicar métodos estándar para datos completos cuando el banco de datos no lo está debido a la existencia de valores ausentes, no sería adecuado ya que la verosimilitud (2.1) es tratada como $p(\boldsymbol{y}|\boldsymbol{\phi})$, cuando dicha verosimilitud depende de los valores ausentes también, es decir, se estaría obviando indebidamente $\boldsymbol{\theta}$.

Existen dos posibles factorizaciones de esta distribución conjunta (2.1) que dan pie a sendas teorías sobre el tratamiento de los valores ausentes. Por un lado si se considera

$$p(\boldsymbol{y}, \boldsymbol{I}|\boldsymbol{\phi}, \boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{I}, \boldsymbol{\phi})p(\boldsymbol{I}|\boldsymbol{\theta})$$

se estaría hablando de lo que se conoce en la actualidad como modelos de mixtura de patrones, véase Molenberghs y Kenward (2007) o Daniels y Hogan (2008).

Por otro lado, si se considera

$$p(\boldsymbol{y}, \boldsymbol{I}|\boldsymbol{\phi}, \boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\phi}, \boldsymbol{\theta})p(\boldsymbol{I}|\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\theta}) \tag{2.2}$$

se estaría hablando entonces de la otra perspectiva para el tratamiento de los valores ausentes conocida como modelos de selección.

El tratamiento que de los valores ausentes se realiza en esta tesis se lleva a cabo bajo el amparo de esta última factorización. Y es en (2.2) donde aparece el mecanismo de generación de valores ausentes definido como la distribución condicional de $\boldsymbol{I}$ dado $\boldsymbol{y}$, es decir $p(\boldsymbol{I}|\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\theta})$, que como en (2.1) reescribimos como $p(\boldsymbol{I}|\boldsymbol{y_{obs}}, \boldsymbol{y_{aus}}, \boldsymbol{\phi}, \boldsymbol{\theta})$

Así, si estamos ante un mecanismo considerado MCAR

$$p(\boldsymbol{I}|\boldsymbol{y_{obs}}, \boldsymbol{y_{aus}}, \boldsymbol{\theta}) = p(\boldsymbol{I}|\boldsymbol{\theta}).$$

de esta forma, el mecanismo de generación de valores ausentes no depende de $\boldsymbol{y}$ en ningún caso.

Mientras que si el mecanismo es considerado MAR

$$p(\boldsymbol{I}|\boldsymbol{y_{obs}}, \boldsymbol{y_{aus}}, \boldsymbol{\theta}) = p(\boldsymbol{I}|\boldsymbol{y_{obs}}, \boldsymbol{\theta}).$$

Por último, si el mecanismo de generación de valores ausentes es MNAR, no hay simplificación posible de la distribución condicional de $\boldsymbol{I}$.

Esta clasificación del mecanismo de generación de valores ausentes influye en el proceso de inferencia con datos incompletos, pues para poder llevarla a cabo se debe tener en cuenta que la única información disponible en el problema es $(\boldsymbol{y_{obs}}, \boldsymbol{I})$. Y que su distribución conjunta se obtiene de la siguiente forma

$$p(\boldsymbol{y_{obs}}, \boldsymbol{I}|\boldsymbol{\phi}, \boldsymbol{\theta}) = \int p(\boldsymbol{y_{obs}}, \boldsymbol{y_{aus}}, \boldsymbol{I}|\boldsymbol{\phi}, \boldsymbol{\theta}) dy_{aus}. \tag{2.3}$$

Factorizando la distribución conjunta de los datos y el indicador de acuerdo a la perspectiva de modelos de selección, ecuación (2.2), se tiene

$$p(\boldsymbol{y_{obs}}, \boldsymbol{I}|\boldsymbol{\phi}, \boldsymbol{\theta}) = \int p(\boldsymbol{y_{obs}}, \boldsymbol{y_{aus}}|\boldsymbol{\phi}, \boldsymbol{\theta}) p(\boldsymbol{I}|\boldsymbol{y_{obs}}, \boldsymbol{y_{aus}}, \boldsymbol{\phi}, \boldsymbol{\theta}). \tag{2.4}$$

Asumiendo que $\boldsymbol{I}|\boldsymbol{y},\boldsymbol{\theta}$ es condicionalmente independiente de $\boldsymbol{\phi}$ y que $\boldsymbol{y}|\boldsymbol{\phi}$ es condicionalmente independiente de $\boldsymbol{\theta}$, la expresión $p(\boldsymbol{y_{obs}},\boldsymbol{y_{aus}}|\boldsymbol{\phi},\boldsymbol{\theta})p(\boldsymbol{I}|\boldsymbol{y_{obs}},\boldsymbol{y_{aus}},\boldsymbol{\phi},\boldsymbol{\theta})$ puede simplificarse a $p(\boldsymbol{y_{obs}},\boldsymbol{y_{aus}}|\boldsymbol{\phi})p(\boldsymbol{I}|\boldsymbol{y_{obs}},\boldsymbol{y_{aus}},\boldsymbol{\theta})$. Es decir, la distribución conjunta vista en (2.2) quedaría

$$p(\boldsymbol{y_{obs}},\boldsymbol{y_{aus}},\boldsymbol{I}|\boldsymbol{\phi},\boldsymbol{\theta}) = p(\boldsymbol{y_{obs}},\boldsymbol{y_{aus}}|\boldsymbol{\phi})p(\boldsymbol{I}|\boldsymbol{y_{obs}},\boldsymbol{y_{aus}},\boldsymbol{\theta}) \quad (2.5)$$

y por tanto, la expresión (2.4) queda

$$p(\boldsymbol{y_{obs}},\boldsymbol{I}|\boldsymbol{\phi},\boldsymbol{\theta}) = \int p(\boldsymbol{y_{obs}},\boldsymbol{y_{aus}}|\boldsymbol{\phi})p(\boldsymbol{I}|\boldsymbol{y_{obs}},\boldsymbol{y_{aus}},\boldsymbol{\theta})dy_{aus}. \quad (2.6)$$

Y aquí es donde se observa que la clasificación del mecanismo de generación de valores ausentes influye en el proceso de inferencia pues, bajo el supuesto de mecanismo de generación de valores ausentes MAR, la expresión anterior queda

$$p(\boldsymbol{y_{obs}},\boldsymbol{I}|\boldsymbol{\phi},\boldsymbol{\theta}) = p(\boldsymbol{y_{obs}}|\boldsymbol{\phi})p(\boldsymbol{I}|\boldsymbol{y_{obs}},\boldsymbol{\theta}), \quad (2.7)$$

mientras que bajo MCAR se simplifica aún más, quedando

$$p(\boldsymbol{y_{obs}},\boldsymbol{I}|\boldsymbol{\phi},\boldsymbol{\theta}) = p(\boldsymbol{y_{obs}}|\boldsymbol{\phi})p(\boldsymbol{I}|\boldsymbol{\theta}). \quad (2.8)$$

Además, si los parámetros que gobiernan la distribución del mecanismo de generación de valores ausentes, $\boldsymbol{\theta}$, y la distribución de los datos, $\boldsymbol{\phi}$, son independientes a priori, entonces, desde el punto de vista bayesiano, se puede inferir sobre los parámetros $\boldsymbol{\phi}$ únicamente a partir de la verosimilitud de los valores observados $p(\boldsymbol{y_{obs}}|\boldsymbol{\phi})$ (Gelman *et al.*, 2013; Little y Rubin, 2002). En este caso se dice que el mecanismo de valores ausentes es *ignorable*.

La asunción de ignorabilidad simplifica mucho el problema, pues desde el punto de vista bayesiano los valores ausentes, al ser cantidades desconocidas, son tratados como parámetros y por tanto, asumiendo ignorabilidad no sólo podemos inferir sobre $\boldsymbol{\phi}$

sino también sobre los valores ausentes. Así, podemos obtener una distribución posterior predictiva sobre cada uno de los valores no observados del banco de datos, $\boldsymbol{y_{aus}}$.

Las hipótesis de MCAR o MAR, y por tanto de ignorabilidad, se presentan con bastante frecuencia en problemas con datos ausentes, lo que implica que el desarrollo descrito se ajusta a gran parte de los problemas donde se requiere imputar datos. En aquellos casos en los que no se pueda asumir ignorabilidad en el mecanismo de valores ausentes o la aleatoriedad en la presencia de éstos, es necesario un desarrollo más complejo que nos permita llegar a obtener la distribución posterior de los datos ausentes. Por ejemplo, desde el punto de vista bayesiano se requeriría asignar una distibución a $p(\boldsymbol{I}|\boldsymbol{y}, \boldsymbol{\theta})$, siendo distribuciones Bernoulli independientes la opción más utilizada en la literatura.

No obstante, siguiendo las afirmaciones de Rubin en Gelman *et al.* (2013), capítulo 18, es muy raro encontrar situaciones donde no se pueda asumir ignorabilidad. Además, respecto a la hipótesis de aleatoriedad, siempre se podrá asumir mediante la introducción de tantas covariables, completas y que aporten información, como sean necesarias (en la medida de lo posible). Así, aumentando el banco de datos observados se reduce la dependencia del mecanismo de generación de valores ausentes, dados los observados, de los propios valores ausentes.

2.3. Tipos de Imputación

2.3.1. Imputación simple

La imputación simple consiste en reemplazar cada valor ausente por una estimación de dicho valor, es decir imputar cada dato ausente con un único valor. De esta forma se consigue completar el banco de datos y trabajar con él como si fuera un banco de datos completo.

Esta técnica de imputación tiene varias ventajas. En primer

lugar, al completar el banco de datos se puede hacer uso de las técnicas estadísticas estándar para datos completos (implementadas en los softwares básicos de estadística) y por tanto no es necesario que dicho banco de datos sea tratado por usuarios con avanzados conocimientos de estadística.

En segundo lugar, el esfuerzo de realizar la imputación se lleva a cabo una única vez. Esto representa una ventaja a nivel práctico sobretodo en bancos de datos públicos o que se vayan a utilizar para estudios con diferentes finalidades. En estos casos se realiza la imputación simple completando la base de datos y distribuyéndola posteriormente, en lugar de distribuir los datos incompletos y cada usuario verse obligado a realizar una imputación según su propósito de estudio.

Otra ventaja asociada a la imputación simple y relacionada con bancos de datos destinados a ser distribuidos, es que la imputación la realiza la persona, o personas, que generalmente se encargan de su recopilación. Es decir, que trabajan a diario con ese tipo de datos y por tanto tienen un nivel de conocimiento de la naturaleza de los datos superior al de los futuros usuarios, en consecuencia la imputación resulta ser más adecuada a la estructura de los datos.

No obstante, también existen varias desventajas de imputar un único valor por cada dato ausente.

Sin duda, la principal desventaja es que al asignar un único valor por cada dato ausente no se está teniendo en cuenta la variabilidad asociada a la incertidumbre de los valores ausentes. Esto puede derivar en sesgos importantes al realizar estudios estadísticos sobre datos completados mediante imputación simple.

Con el fin de solventar este importante inconveniente, Rubin propone la imputación multiple de datos.

2.3.2. Imputación múltiple

La imputación múltiple consiste en proporcionar varios valores para cada dato ausente, es decir, asignar a cada valor ausente no

una única estimación sino una muestra de tamaño m. De esta forma se crean m bancos de datos completos cada uno de los cuales puede ser analizado mediante el empleo de técnicas estadísticas estándares para datos completos.

La idea de la imputación múltiple es dar solución a la principal desventaja de la imputación simple sin desaprovechar las ventajas prácticas de ésta. Así al proporcionar varias estimaciones de los valores ausentes se consigue tener en cuenta la variabilidad de los datos ausentes y por tanto evitar posibles sesgos en futuras inferencias que se realicen sobre los datos.

La forma de proceder en un estudio estadístico donde intervengan datos incompletos utilizando la imputación múltiple como técnica de imputación es la siguiente: como ya se ha comentado, una vez realizada la imputación múltiple de los valores ausentes lo que se tiene son m bancos de datos completos. Aquellos valores que no fueran ausentes se repetirán m veces, una por cada banco de datos. Sin embargo habrá un valor imputado diferente por dato ausente en cada uno de los m bancos de datos. Posteriormente se realiza el análisis estadístico sobre cada uno de los m bancos de datos, obteniéndose m estimaciones del parámetro o parámetros de interés del estudio. De esta forma la incertidumbre asociada a los valores ausentes es tenida en cuenta.

Finalmente se resumen esas m estimaciones siguiendo lo que se conoce como Reglas de Rubin (Rubin, 1987) para extraer conclusiones de dicho de parámetro o parámetros de interés.

La imputación múltiple presenta una serie de inconvenientes a nivel práctico (en comparación con la imputación simple), los cuales resumimos brevemente. El primero es que el esfuerzo necesario para realizar la imputación de datos se incrementa considerablemente respecto a la imputación simple. El segundo inconveniente es que el proceso de inferencia también requiere de un mayor tiempo y esfuerzo pues es necesario analizar m bancos de datos. Por último, al imputar con varios valores cada valor ausente y pasar a tener m bancos de datos, se suelen presentar problemas de almacenamiento

ya que se tiene un banco de datos *tridimensional* (n registros, r variables y m imputaciones) y suele ser necesario implementar rutinas especiales para tratar estadísticamente un banco de datos de estas características. Este problema puede ser definitivo para descartar la imputación múltiple si estamos trabajando con un gran volumen de datos y además se emplea un software estadístico incapaz de explotar un banco de datos tridimensional.

Sin embargo, si no se tiene problemas en almacenar un banco de datos tridimensional ni se tiene problemas para explotar dicho banco de datos, la imputación múltiple resulta ser una técnica de imputación óptima para dar solución al problema de trabajar con valores ausentes. Esto es debido a que la incertidumbre debida a la existencia de valores ausentes queda reflejada en el análisis al mismo tiempo que se evita el posible sesgo debido a la información faltante. Además, es una técnica práctica, sencilla a nivel metodológico y con un esfuerzo requerido para implementarla y realizar posteriormente inferencias válidas no demasiado alto teniendo en cuenta el problema al que se está dando solución.

2.3.3. Imputación bayesiana

Como en cualquier análisis de datos, el tratamiento de valores ausentes se puede abordar tanto desde la perspectiva clásica como desde la perspectiva bayesiana.

Obtener estimaciones de los valores ausentes desde un enfoque frecuentista, sobre todo en aquellos estudios donde la verosimilitud de la que deseamos extraer estimaciones es compleja, significa recurrir al algoritmo EM (Dempster et al., 1977), o a procesos iterativos como el descrito en Carpenter y Kenward (2013) que, aún siendo frecuentistas, están basados en muestreo Gibbs, descrito más adelante en esta sección.

El algoritmo EM proporciona estimaciones válidas de las cantidades de interés mediante un proceso iterativo en el que intervienen dos etapas: Esperanza y Maximización.

Así, partiendo de un punto inicial para las cantidades de interés, sea $\boldsymbol{\theta}^0$, en el primer paso del algoritmo se obtiene la esperanza del logaritmo de la verosimilitud estudiada, que será una función de $\boldsymbol{\theta}$, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^0)$. En el siguiente paso se maximiza dicha función Q obteniendo una nueva estimación (máximo verosímil) para $\boldsymbol{\theta}$ y volviendo de nuevo al primer paso. El algoritmo se detendrá cuando la diferencia entre las estimaciones proporcionadas por dos pasos sucesivos sea menor que una determinada tolerancia.

No obstante, el algoritmo EM no siempre se puede aplicar debido a que la verosimilitud, por muy compleja que sea, debe ser conocida. Si además el mecanismo de generación de valores ausentes no pudiera ser asumido *ignorable*, obtener estimaciones de los valores ausentes se convertiría en un problema de difícil solución desde el punto de vista clásico.

Por el contrario, si se enfoca el problema de la imputación desde el punto de vista bayesiano, con mayor o menor dificultad siempre seremos capaces de obtener estimaciones de los valores ausentes.

Haciendo uso de una modelización jerárquica se pueden ir introduciendo de forma natural en sucesivas capas todos los elementos intervinientes en el proceso, pudiendo añadir covariables *informativas* en caso de ser necesario e incluso la modelización del mecanismo de generación de valores ausentes si estuviéramos en un escenario MNAR. En la figura 2.2 se puede observar de forma esquemática el proceso de imputación de valores ausentes.

La gran ventaja del tratamiento de valores ausentes desde el punto de vista bayesiano es que, al obtener una muestra de la distribución posterior de los valores ausentes, podemos, o bien hacer una imputación simple (media, mediana) o, si se prefiere, una imputación múltiple. Es más, si el fin de la imputación de los valores ausentes es que el banco de datos sirva como covariable en un determinado estudio, el esquema de imputación mostrado en la figura 2.2 puede ser incluido dentro de la modelización de dicho estudio como una capa más. Así, la incertidumbre asociada a los valores ausentes es tenida en cuenta en la estimación de los

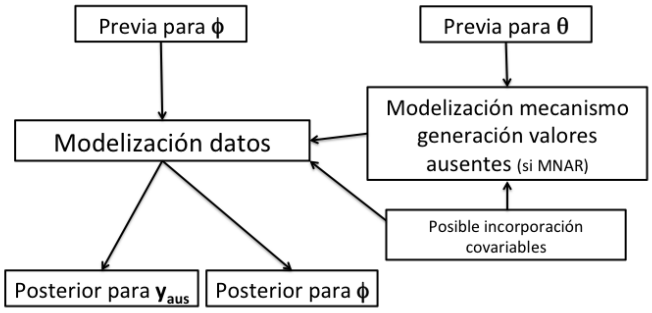


Figura 2.2: Estrategia para la imputación de valores ausentes desde el punto de vista bayesiano

parámetros de interés del estudio.

La dificultad del proceso bayesiano radica en la obtención de la distribución posterior de las cantidades de interés (en nuestro caso los valores ausentes). Aunque con el auge en los últimos años de los métodos MCMC en paralelo con avances en computación, esta dificultad ha dejado de ser un handicap a la hora de decantarse por abordar un análisis desde la perspectiva bayesiana.

Las técnicas numéricas basadas en cadenas de Markov, MCMC (Green 2001) nos permiten inferir sobre expresiones imposibles de hallar a nivel analítico. El objetivo de estas técnicas es la obtención de una muestra de la distribución posterior de las cantidades de interés. Para ello, a partir de un valor inicial, se va generando una cadena aleatoria de posibles valores de los parámetros o cantidades de interés. Bajo condiciones de equilibrio, dicha cadena contiene valores de la distribución posterior buscada y por tanto se puede considerar una muestra de la distribución de las cantidades de interés.

Existen diversas técnicas MCMC, donde el proceso de generación de las cadenas puede ser distinto según la técnica empleada. Las más comunes son el muestreo Gibbs (Geman y Geman, 1984) y el Metrópolis-Hastings (Hastings, 1970).

El algoritmo de muestreo Gibbs está basado en las distribuciones condicionales de los parámetros de interés, las cuales son conocidas y por tanto se pueden obtener valores de ellas. Partiendo de un punto inicial $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0, ..., \theta_n^0)$ en cada paso se obtiene un nuevo valor para cada componente, generando una observación aleatoria de su distribución condicionada al resto de componentes, esto es

$$\begin{aligned}
&\theta_1^1 \sim p(\theta_1|(\theta_2^0, \theta_3^0, ..., \theta_n^0)) \\
&\theta_2^1 \sim p(\theta_2|(\theta_1^1, \theta_3^0, ..., \theta_n^0)) \\
&... \\
&\theta_n^1 \sim p(\theta_n|(\theta_1^1, \theta_2^1, ..., \theta_{n-1}^1))
\end{aligned}$$

repitiendo el proceso a lo largo de N iteraciones, N lo suficientemente grande para asegurar la convergencia de la cadena, se llega a la obtención de una muestra de la distribución posterior para $\boldsymbol{\theta}$.

De una forma diferente funciona el algoritmo de Metrópolis-Hastings. La cadena aleatoria se obtiene en este caso generando valores para $\boldsymbol{\theta}$ a partir de una función denominada *instrumental*. Dichos valores son incluidos o no en la cadena a través de un mecanismo de aceptación-rechazo basado en la probabilidad, en la distribución posterior buscada, asociada a los valores generados. Generando un número suficiente de iteraciones para asegurar la convergencia de la cadena generada, se obtiene una muestra de la distribución posterior de $\boldsymbol{\theta}$.

Los algoritmos que emplean técnicas MCMC siguen siendo los más utilizados para la obtención de muestras de las distribuciones posteriores de las cantidades de interés. Pueden ser o bien implementados por el investigador encargado de la resolución del problema, o bien, se puede hacer uso de software específico que ya los lleva implementados. Un par de ejemplos de este tipo de software

son WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/) y JAGS (http://mcmc-jags.sourceforge.net/) ambos de libre distribución y que utilizan el muestreo Gibbs para la resolución de modelos bayesianos.

Sin embargo, los métodos MCMC pese a estar ampliamente extendidos tienen un gran inconveniente: el tiempo de computación. Para problemas de una determinada complejidad pueden ser necesarias horas e incluso días para asegurarse la convergencia del algoritmo y estar seguros que estamos muestreando de la distribución posterior de las cantidades de interés.

En los últimos años otra técnica diferente está en auge: INLA (Integrated Nested Laplace Approximations). Es una técnica de aproximación para la estimación de parámetros en modelos con estructura latente Gaussiana. Fue propuesta por Havard Rue (Rue *et al.*, 2009) y desde su aparición son cada vez más numerosos en la literatura los estudios bayesianos que recurren a esta técnica en lugar de utilizar métodos MCMC para la obtención de muestras de la distribución posterior.

La principal ventaja de INLA es que la aproximación a la distribución posterior de una determinada cantidad de interés tarda apenas unos minutos mientras que mediante técnicas MCMC puede tardar horas o incluso días. Esto la convierte en una técnica muy atractiva y con los años será el método básico de obtención de estimaciones de las cantidades de interés en cualquier proceso bayesiano.

Por el momento, al ser una técnica que está dando sus primeros pasos, existen todavía modelizaciones con estructura latente Gaussiana cuyas distribuciones posteriores no han podido ser aproximadas mediante INLA. Este es el caso de determinadas modelizaciones espacio temporales, alguna de las cuales forma parte de las tratadas en esta tesis. Por ello en este trabajo hemos seguido haciendo uso de técnicas MCMC.

En cualquier caso, gracias a los métodos MCMC e INLA, junto con los continuos avances en técnicas de computación,

en la actualidad se pueden abordar estudios desde el punto de vista bayesiano en los que intervienen modelizaciones avanzadas. Estudios que en anteriores décadas los investigadores no podían siquiera plantearse su resolución debido a la imposibilidad de hallar una expresión analítica para las distribuciones posteriores. Hoy, sin embargo, pueden ser resueltos de forma relativamente sencilla.

En nuestro caso, la complejidad de la modelización que se plantea en esta tesis hace inviable su tratamiento desde el punto de vista clásico. Sin embargo, abordar el problema de imputar valores ausentes desde el punto de vista bayesiano resulta ser un proceso relativamente sencillo que permite ir adaptando el modelo a la estructura y naturaleza de los datos de una forma intuitiva. Aunque en capítulos siguientes veremos con más detalle el tipo de datos con el que vamos a trabajar, así como la modelización propuesta para su imputación, cabe destacar que asumiremos un mecanismo de generación de valores ausentes *ignorable* y que la aparición de dichos valores ausentes es aleatoria (MAR). De esta forma podremos aprovechar la información proporcionada por los valores observados para la obtención de estimaciones de los valores ausentes.

2.4. Influencia de la imputación

La duda más importante que surge al imputar un banco de datos incompleto, bien sea mediante imputación simple, múltiple o bayesiana, es si se está alterando la influencia que dichos datos puedan tener sobre cualquier característica de estudio. Es decir, si los resultados de un estudio en el que intervienen estos datos incompletos imputados pueden considerarse los mismos que aquellos que se obtendrían si el estudio se realizara con los datos completos.

El problema es que no se puede dar una respuesta directa a esta cuestión porque no se dispone de los datos completamente observados. Sería por tanto necesario recurrir a un estudio con datos

simulados para acometer este propósito.

Si se enfoca este asunto desde el punto de vista bayesiano, se puede sacar ventaja del hecho de incluir el modelo de imputación como una capa más dentro del modelo jerárquico bayesiano planteado para la realización de dicho estudio. En consecuencia, la incertidumbre asociada a los valores ausentes es tenida en cuenta a la hora de obtener estimaciones de los parámetros de interés del estudio.

En Abellan (2005) se aborda este problema llegando a la conclusión de que, si se imputa convenientemente de acuerdo a la naturaleza y estructura de los datos, la imputación de valores ausentes no influye en los resultados obtenidos en futuros estudios en los que participe dicho banco de datos completado mediante imputación.

Para ello, se recurrió a un estudio por simulación en el que se generaban tres bancos de datos cuya influencia sobre una determinada variable respuesta (simulada también) fuera débil, normal y fuerte, respectivamente. Posteriormente se provocaban valores ausentes en los bancos de datos simulados y se imputaban para estudiar después si dicha influencia había variado. Los resultados ponían de manifiesto que el hecho de imputar los valores ausentes no modificaba la influencia original que dichos datos tenían sobre la variable respuesta.

No obstante, el estudio de la influencia de la imputación no es objetivo de esta tesis. Para este trabajo nos marcamos como meta el planteamiento, estudio y comparación de modelos con estructura espacio-temporal para la imputación de datos.

En el siguiente capítulo se resume brevemente los diferentes procesos en los que intervienen datos con correlación espacial y espacio-temporal. El objetivo de esta descripción no radica en la profundización de conceptos ni metodologías. Se pretende únicamente presentar el campo de investigación sobre el que se basan las ideas que han permitido el planeamiento de los modelos de imputación analizados en esta tesis.

Capítulo 3

Análisis espacial y espacio-temporal

Si la historia asociada a la existencia de datos ausentes es amplia, no lo es menos la asociada a la existencia de estructura espacial en la naturaleza de las observaciones a estudiar. El análisis de datos con correlación espacial surge por sí solo al trabajar con observaciones que de una manera u otra están ligadas a una localización geográfica. Un proceso de grabado de datos adecuado incluiría almacenar la información del lugar (e incluso el tiempo) en que dichos datos se han recopilado. Este simple hecho ha ocurrido y ocurre en muchas investigaciones, pese a que no siempre ha sido tenido en cuenta. En ocasiones debido sencillamente a que se ha ignorado que los datos están georeferenciados, y en ocasiones debido a la complejidad del tratamiento estadístico de datos con estructura espacial.

De hecho, hasta comienzos de 1990 los estudios estadísticos de datos con estructura espacial eran relativamente fáciles de plantear en la teoría pero casi imposibles de resolver en la práctica. No obstante, a lo largo de la historia ha habido casos importantes en los que se ha empleado la información espacial de los datos para analizar un problema, como por ejemplo el brote de cólera de 1854 en Londres donde John Snow (considerado desde entonces

padre de la epidemiología), con la ayuda de la georeferenciación de los casos, pudo localizar el foco del brote en una de las fuentes públicas de agua potable existentes en la época, figura 3.1. O los estudios de Danie Krige (considerado el padre de la geoestadística) que permitieron analizar los recursos mineros de diferentes regiones a principios de los 40. Es más, como la información espacial inicialmente aparece ligada a la geografía, en la literatura existen algunas referencias de principios de la segunda mitad del siglo XX (por ejemplo Garrison, 1959; Chorley, 1972) destacando Berry y Marble (1968), como el primer libro que realiza una revisión sobre el tema hasta la fecha y en el que se acuña el término *análisis espacial*.

Figura 3.1: Mapa de casos de cólera realizado en 1854 por John Snow

Sin embargo, no fue hasta principios de 1990, y nuevamente gracias a los avances en computación, cuando estos estudios pasaron de ser una bonita teoría o un laborioso trabajo de campo a ser una realidad práctica. Desde entonces investigadores de muy diferentes áreas llevan tiempo centrando sus estudios en datos con correlación espacial. Desde la arqueología hasta la zoología se ha ido profundizando en el análisis espacial de los datos y, en los últimos tiempos, en el análisis de datos con estructura espacio temporal; pues entender y analizar la dependencia entre observaciones en el espacio (y en el tiempo) es una parte crucial de la estadística.

Las primera referencias importantes que encontramos en la literatura contemporánea son Ripley (1981) y Cressie (1993). En estos libros se hace una completa y extensa revisión de los conceptos, las técnicas y las modelizaciones más utilizadas en el estudio de datos con estructura espacial.

En general, un proceso espacial se define mediante el conjunto de variables aleatorias

$$\{\mathbf{Z(s)} : \mathbf{s} \in D\}, \tag{3.1}$$

donde D es el espacio o región a estudio (generalmente de $\mathbf{R}^2$), $\mathbf{s}$ es el vector de localizaciones pertenecientes a la región D y $\mathbf{Z(s)}$ es el vector de observaciones. A la realización de dicho vector se le denota por $\mathbf{z(s)}$.

Dependiendo del carácter aleatorio de $\mathbf{s}$ o de $\mathbf{Z}$ se puede clasificar el proceso espacial en tres grandes grupos: geoestadística, procesos puntuales y redes de localizaciones.

El tratamiento y la modelización propuesta en esta tesis está enmarcada en el caso de redes de localizaciones (también conocido como datos en áreas o lattice). De este modo, en lo que resta de capítulo, primero se describe brevemente cada una de estas estructuras espaciales en la siguiente sección para, seguidamente, pasar a realizar una descripción más detallada de la cartografía de enfermedades, motor básico de los avances en modelización de datos en redes de localizaciones. Con esta descripción se pretende

presentar el marco teórico que inspira y sobre el que se basa la modelización propuesta para el tratamiento e imputación de los valores ausentes motivo de esta tesis.

Finalmente, previo repaso al proceso de evolución natural del análisis de datos espaciales al análisis de datos espacio-temporales en la sección 3.3, se presentan los datos objeto de análisis de esta tesis.

3.1. Tipos de datos espaciales

No es objetivo de esta tesis profundizar a nivel metodológico en la explicación de cada uno de los tres grandes grupos de datos espaciales, sino más bien describir brevemente sus características básicas con el fin de enmarcar posteriormente nuestra investigación.

Para una extensa información sobre el tema, Ripley (1981) y Cressie (1993) proporcionan una profunda revisión de cada uno de los tres casos no sólo desde el punto de vista metodológico sino también desde el punto de vista práctico ya que hacen uso de muchos y variados ejemplos para cada caso.

No obstante, también existen otras interesantes referencias sobre estadística espacial en las que se proporcionan explicaciones más detalladas sobre geoestadística, redes de localizaciones y procesos puntuales, por ejemplo Haining (2003), Schabenberger y Gotway (2005) o Banerjee *et al.* (2014).

3.1.1. Geoestadística

Se entiende por geoestadística el estudio de una superficie continua mediante la obtención de una muestra aleatoria de valores. Así, D es una región continua y fija mientras que $\mathbf{Z(s)}$ es un vector aleatorio de valores obtenidos en la muestra de localizaciones $\mathbf{s}$.

Entre las diferentes técnicas estadísticas que existen para analizar datos que por su naturaleza se enmarcan dentro de la geoestadística, destaca el Kriging (Diggle *et al.*, 1998). Esta técnica,

cuyo nombre viene por D. G. Krige, ingeniero de minas pionero en utilizarla para la estimación de reservas mineras (Krige, 1951), está basada en el variograma, piedra clave de la técnica. Para la obtención del variograma se calcula la variabilidad existente entre las observaciones de puntos distintos que se encuentran a una cierta distancia h, ($\mathbf{z(s)}$, $\mathbf{z(s+h)}$), es decir:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{s_i - s_j = h} (z(s_i) - z(s_j))^2, \tag{3.2}$$

donde $N(h)$ es el número de observaciones separadas una distancia h. Así, en función de h se observa la asociación espacial de las observaciones, esperando obtener poca variabilidad para valores pequeños de h y observar un incremento de dicha variabilidad a medida que h aumenta llegando un punto, conocido como rango, en el que la variabilidad es constante.

Modelizando el variograma se podrá llegar al objetivo del kriging que es proporcionar predicciones en aquellos puntos de la región de los cuales no se dispone de observaciones (pudiendo incluir covariables si fuera necesario). Incluso se podría obtener la estimación de dicha variable para todas las localizaciones de D, lo que equivaldría a obtener un mapa de superficie de la variable a estudio.

3.1.2. Procesos puntuales

Cuando las localizaciones donde un determinado evento a estudio son aleatorias estamos ante un proceso puntual. En este caso la variable aleatoria es $\mathbf{D}$, conociéndose como proceso de puntos. Las localizaciones $\mathbf{s}$ corresponden a observaciones de $\mathbf{D}$ y, en un proceso puntual espacial generalmente sólo se está interesado en dicha realización de las localizaciones, por tanto no suelen existir observaciones asociadas a $\mathbf{s}$, $\mathbf{Z(s)}$, o equivalentemente $\mathbf{Z(s)=1}$. En caso contrario se conoce como proceso espacial de puntos notable.

Procesos puntuales suelen encontrarse en estudios en los que

se analiza la aparición de casos de una determinada enfermedad en una población registrándose como dato la residencia del caso; o estudios de poblaciones de árboles en un bosque registrándose como datos la localización de la especie de interés.

Así, como se ha dicho, las localizaciones $\mathbf{s}$ son aleatorias y generalmente se está interesado únicamente en el lugar de ocurrencia del evento (salvo en un proceso puntual notable). El objetivo es llegar a clarificar si el proceso de aparición de puntos es puramente aleatorio o, si por el contrario, existe algún tipo de patrón subyacente que provoque la agrupación de puntos.

Generalmente, este objetivo se aborda asumiendo que la aparición de puntos en una región R sigue un proceso homogéneo Poisson con media $\lambda|R|$. Así el número esperado de eventos en la región viene determinado por λ, parámetro que recoge la intensidad del proceso (número de individuos por unidad de área), y $|R|$, área de la región R.

Posteriormente se procede a cuantificar la agrupación de eventos observada en dicha región destacando, entre las diferentes formas de llevarlo a cabo, la función de Ripley (Ripley, 1976):

$$K(h) = \tfrac{1}{\lambda}[\text{número eventos a una distancia } h \text{ de un punto arbitrario de } R]$$

Según el patrón puntual espacial en el que nos encontremos, existen valores teóricos de K(h) frente a los que comparar la estimación obtenida de la función. Por ejemplo, en caso de no existir correlación espacial se tendría que $K(h) = \pi h^2$, y si los datos se agruparan por encima de lo esperado significaría que $\hat{K}(h) > \pi h^2$, siendo el estimador más frecuente de la función de Ripley:

$$\hat{K}(h) = \frac{1}{\hat{\lambda}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} p_{ij}^{-1} I_h(h_{ij}), \tag{3.3}$$

donde n es el número de eventos en R, $\hat{\lambda} = \frac{n}{|R|}$, h_{ij} la distancia entre los puntos i y j, p_{ij} la proporción del círculo de centro i (y

que engloba a j) que cae dentro de R y $I_h(h_{ij})$ función indicador igual a 1 si $h_{ij} < h$ y 0 en otro caso.

La inclusión del término p_{ij} ayuda en la corrección del *efecto borde*, esto es, el posible sesgo en la estimación de la función en aquellos puntos situados a una distancia menor que h de la frontera del área de estudio.

Para una mayor profundización en el tratamiento de procesos puntuales espaciales así como en la metodología aplicada para la modelización y detección de clusters se recomienda Diggle (2003).

3.1.3. Redes de localizaciones

Más conocido como *lattice models* (Besag, 1974) es el caso en que se enmarca esta tesis. Mediante el análisis de datos en redes de localizaciones se estudia una variable aleatoria asociada a un conjunto de localizaciones fijo (por ejemplo el conjunto de municipios de una determinada región). En este caso D es un conjunto fijo contable de puntos o regiones (regulares o irregulares) en las cuales se recogen observaciones $\mathbf{z(s)}$, con $\mathbf{s} \in D$.

Al no entenderse continuidad en D, los modelos lattice resultan ser grafos cuyos nodos son las localizaciones a estudio en las que se observa una variable aleatoria de interés. Definidos de este modo, cuando se trabaja con datos en redes de localizaciones surge de manera natural el concepto de *vecindad*, es decir, cómo los nodos *vecinos* pueden influir en las observaciones registradas.

Así, el hecho de que los datos procedan de diferentes áreas o localizaciones geográficas induce a pensar en una posible correlación entre observaciones próximas en el espacio. Esto introduce la idea de que la variabilidad observada en la variable de interés está originada por dos componentes de distinta naturaleza: por un lado está la variabilidad asociada a la heterogeneidad propia de las localizaciones geográficas y que se conoce como variabilidad a gran escala y por otro está la variabilidad asociada a la relación existente entre observaciones cercanas en el espacio lo que se conoce como

variabilidad a pequeña escala.

Existen diferentes maneras de especificar una estructura de vecindad entre nodos o localizaciones, siendo las más extendidas aquellas que están basadas en la distancia y en la compartición de fronteras.

Respecto a la basada en la distancia entre localizaciones, aunque la más utilizada es la distancia euclídea existen otros tipos de medida de la distancia tales como la distancia al viajar de una localización a otra mediante algún medio de transporte, o el tiempo en llegar desde una localización a otra si se utiliza una combinación de éstos.

Respecto a la definición de vecindad basada en la compartición de fronteras, al estar los datos registrados por áreas o localizaciones (por ejemplo municipios de una comunidad autónoma o región) en general se establece que dos áreas se consideren vecinas si comparten alguna de sus fronteras.

Sea cual sea la definición de vecindad, modelizar esta correlación espacial de los datos no ha sido una tarea sencilla. Quizás la utilización de auto modelos sea lo más eficiente para este propósito, de ahí que sea la alternativa más extendida en la actualidad. Los auto modelos se basan en ajustar la correlación espacial entre observaciones vecinas mediante distribuciones condicionales auto regresivas (Besag y Kooperberg, 1995), siendo la más utilizada la distribución Normal.

Es decir, si $\mathbf{v} = \{v_i\}$ con $i = 1, ..., N$ y N el número de localizaciones, es la componente asignada para modelizar la estructura espacial de un conjunto de datos, el ajuste de dicha correlación espacial existente se consigue a través de la definición de la distribución conjunta de estas componentes. Esta se construiría a partir de las distribuciones condicionales,

$$v_i | v_{-i} \sim N(\sum \gamma_k v_k, \tau) \qquad i = 1...N,$$

siendo $v_{-i} = \{v_k : k \neq i\}$, τ la precisión y

$$\gamma_k = \begin{cases} \frac{1}{n_i} & \text{si la localización } k \text{ es vecina de } i \\ 0 & \text{en otro caso} \end{cases}$$

con n_i el número de vecinos de la localización i.

En otras palabras, la correlación espacial entre localizaciones vecinas es modelizada mediante la inclusión de una componente por cada localización cuya distribución se define Normal centrada en la media aritmética de las observaciones vecinas. Este tipo de distribuciones se suele denotar como CARNormal(τ), donde τ es la precisión.

Cabe destacar que por su definición este tipo de distribuciones son impropias, por lo que su utilización se restringe a la asignación de estas como distribuciones previas en un proceso bayesiano. De este modo, y bajo ciertas condiciones detalladas en Besag y Kooperberg (1995), entre las que se encuentra la simetría de la matriz de varianzas-covarianzas, se consigue que la distribución posterior sea propia.

3.2. Cartografía de enfermedades

La estructura espacial de los datos objeto de análisis en esta tesis es del tipo descrito como datos en redes de localizaciones, descrito en el apartado anterior 3.1.3. Además, la investigación asociada al estudio de cartografía de enfermedades ha sido, sin duda, crucial en el avance de la modelización de este tipo de datos. Estos hechos hacen necesario dedicar un espacio en este capítulo a la descripción de este tipo de modelizaciones familiarizándonos así con los conceptos y la notación involucrada en ellas.

Para el análisis de este tipo de datos, la hipótesis más extendida es asumir que las observaciones siguen, para cada localización, una distribución de Poisson (McCullagh y Nelder, 1989). La media de dicha distribución se descompone en el producto del valor esperado de eventos de una determinada causa de interés por un parámetro $\boldsymbol{\lambda}$, distinto para cada localización, conocido como riesgo relativo, es decir:

$$O_i \sim Poisson(\lambda_i E_i) \qquad i = 1 \dots N$$

siendo N el número de localizaciones geográficas y E_i y O_i el número de casos esperados y observados, respectivamente, en la localización i.

Es sobre este riesgo relativo, λ_i, asociado a cada localización geográfica i, sobre lo que se pretende realizar inferencia con el fin de determinar localizaciones cuyos valores de riesgos relativos sean sospechosamente altos o bajos, lo que equivaldría a obtener valores para $\boldsymbol{\lambda}$ superiores o inferiores a 1.

El principal problema de asumir este modelo es que en la mayoría de estudios se tiene que $Var(O_i) > \lambda_i E_i = E(O_i)$ lo que incumple la característica básica de la distribución de Poisson. Este fenómeno es conocido como sobredispersión o extravarianza (McCullagh y Nelder, 1989).

Dar solución a este inconveniente implica considerar los riesgos relativos, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_N)'$, como efectos aleatorios provenientes de una determinada distribución de riesgo.

Una de las opciones frecuentes es la utilización de distribuciones Gamma (Clayton y Bernardinelli, 1992; Clayton y Kaldor, 1987; Mollié, 1996), es decir:

$$\lambda_i \sim Gamma(\alpha, \beta) \qquad i = 1 \dots N$$

Los parámetros de esta distribución, α y β, pueden ser estimados mediante dos procesos diferentes (Bernardinelli y Montomoli, 1992; Carlin y Louis, 1997):

- Método de máxima verosimilitud, teniendo en cuenta que la distribución de los observados, $\mathbf{O}$, dados α y β, es una binomial negativa. Esto es lo que se conoce como proceso empírico bayes.

- Proceso completamente bayesiano, asignando una distribución previa para α y β, con el fin de obtener su distribución posterior.

El método empírico bayes de estimación de los riesgos relativos es utilizado con frecuencia debido a que la binomial negativa está implementada en casi todos los paquetes estadísticos y las expresiones que proporcionan la estimación máximo verosímil de los parámetros son conocidas y fáciles de implementar. De este modo, el proceso de estimación de los riesgos relativos resulta relativamente sencillo.

Otra opción muy extendida para modelizar $\boldsymbol{\lambda}$ es asumir un modelo lineal para su logaritmo. De este modo, se podría incluir en la modelización las componentes adecuadas según la naturaleza del problema que se está tratando, como el hecho de que los datos proceden de diferentes localizaciones geográficas. A través del predictor lineal se podría entonces incorporar términos que ajustaran la posible correlación entre observaciones próximas en el espacio, teniendo en cuenta así las dos fuentes de variabilidad descritas en el apartado 3.1.3.

Esta modelización, bajo la perspectiva bayesiana quedaría de la siguiente forma:

$$O_i \sim Poisson(\lambda_i E_i), \qquad i = 1...N$$

expresando el predictor lineal como

$$log\lambda_i = \mu + \theta_i + \phi_i \qquad i = 1...N \tag{3.4}$$

donde, μ es la interceptación, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)'$ es el efecto aleatorio encargado de reflejar la heterogeneidad de las localizaciones (este efecto es modelizado generalmente mediante una distribución Normal centrada en cero) y $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_N)'$ es el término que ajusta la correlación espacial entre observaciones próximas en el espacio, CARNormal, de la forma que se ha descrito en el apartado 3.1.3

De este modo, se construye un modelo jerárquico donde en sucesivas capas o niveles, se va introduciendo la información disponible en el estudio así como diversas componentes capaces de captar la estructura propia a la naturaleza de las observaciones.

Esta modelización es la más extendida para la realización de estudios de datos en redes de localizaciones. Encaja dentro de los modelos lineales generalizados mixtos y fue propuesta por Besag, York y Mollié (Besag *et al.*, 1991) pasando desde entonces a conocerse este tipo de modelos con el nombre de sus autores.

Además, dicha modelización puede generalizarse mediante la inclusión de covariables añadiendo al predictor lineal del modelo el término $\beta\mathbf{x}$ (o $\beta_1\mathbf{x_1} + \beta_2\mathbf{x_2} + \ldots + \beta_k\mathbf{x_k}$ en caso de tener k covariables). De este modo, se llega a lo que en la actualidad se conoce como modelo de regresión ecológica (Besag, York y Mollié con covariables).

Dada la complejidad de este tipo de modelos su implementación suele ser abordada desde la perspectiva bayesiana. Así, bastará con asignar distribuciones previas a aquellos parámetros desconocidos para completar la formulación del modelo y poder comenzar el análisis.

Puede consultarse Best *et al.* (2005) para obtener una mayor profundización en la modelización bayesiana utilizada con mayor frecuencia para la cartografía de enfermedades.

Cuando se trabaja con este tipo estructura espacial en los datos, la información va ligada puntualmente a cada una de las localizaciones, es decir, en general los datos con los que se trabaja son el resultado de agregar la información disponible por localización. Este hecho puede llevar a lo que se conoce como *falacia ecológica*. Aunque no siempre sucede (a pesar de que es difícil de analizar), cuando se estudia la posible influencia de un determinado factor sobre el conjunto de una población, realmente se tiene interés en estimar el efecto que dicho factor tiene de forma individual sobre cada uno de los miembros de la población. Sin embargo, el estudio se realiza utilizando datos agregados por localizaciones y se acaba analizando la asociación entre los valores agregados espacialmente de dicho factor y los valores de la variable respuesta de interés agregados también a nivel espacial. De este modo, podrían existir diferencias entre el efecto real a nivel individual y el efecto estimado

a nivel agregado.

3.3. Datos espacio-temporales

Como se ha comentado al comienzo del capítulo, llegado cierto momento en la evolución del análisis de datos con estructura espacial, se dejó de analizar dichas observaciones como datos planos. Se dejó de obviar la correlación espacial para, coincidiendo con el desarrollo computacional, empezar a tener en cuenta dicha naturaleza de los datos en los estudios y modelizaciones llevados a cabo desde entonces.

Cierto paralelismo ha ocurrido respecto a la dimensión temporal de las observaciones. De la misma forma que a partir de los 90 se empezó a tener en cuenta la correlación espacial incorporándose a las modelizaciones de datos georeferenciados, una evolución natural similar hacia la incorporación de la correlación temporal se dio más o menos a partir del 2000.

Se podría decir que empezó a considerarse que los datos con estructura espacial no suelen ser una instantánea de la realidad sino que las observaciones suelen ser recogidas a lo largo de un tiempo determinado y que, por tanto, la componente temporal también debía tenerse en cuenta.

En el pasado, generalmente se agregaba temporalmente dicha información para disponer de un único dato por cada localización pero, desde la primera década del 2000 la información temporal pasó a formar parte de muchas investigaciones, siendo el análisis espacio-temporal el tópico más frecuente hoy en día a la hora de analizar datos con estructura espacial observados durante un determinado periodo de tiempo.

Comprender y captar una estructura espacio-temporal resulta de vital importancia pues de este modo no sólo se tiene en cuenta la correlación existente entre observaciones próximas en el espacio sino que también se saca ventaja de la correlación existente entre observaciones vecinas en el tiempo de una misma localización,

llegando incluso a poder captar (si existiera) correlación entre observaciones de localizaciones vecinas en el espacio pero en tiempos anteriores y/o posteriores.

De este modo, el análisis espacio-temporal, a pesar de ser una evolución natural y lógica del análisis espacial, conlleva mayor complejidad de modelización pues, al hecho de incorporar una componente temporal a los modelos utilizados en el análisis espacial hay que sumarle el hecho de llegar a ser capaces de captar la interacción espacio-tiempo en la estructura de los datos, siendo en esta última parte donde radica la mayor dificultad.

De hecho, en los orígenes de la inclusión de la componente temporal, dicha interacción no fue tenida en cuenta. Por ejemplo, en Bernardinelli *et al.* (1995) se propone una asociación lineal entre la estructura espacial y el tiempo sin incluir interacción espacio-temporal, de este modo, se modifica el modelo visto en (3.4) para incluir el efecto temporal y plantear:

$$log\lambda_i = \mu + (\theta_i + \phi_i) \cdot j \qquad i = 1...N, j = 1...t$$

Sin embargo, en Waller *et al.* (1997), se propone anidar la estructura temporal dentro de la componente espacial reformulando la expresión (3.4) para dejarla en:

$$log\lambda_i = \mu + \theta_{i(j)} + \phi_{i(j)} \qquad i = 1...N, j = 1...t$$

Mientras que en Held y Besag (1998) se plantea extender el modelo (3.4) con una componente principal que refleje la evolución temporal, $\boldsymbol{\alpha}$. En este caso el modelo sería:

$$log\lambda_i = \mu + \theta_i + \phi_i + \alpha_j \qquad i = 1...N, j = 1...t$$

No fue hasta dos años más tarde cuando la interacción espacial y temporal empezó a recibir la atención necesaria. En Held (2000) se presenta por primera vez la modelización de datos con estructura espacio-temporal más completa al incluir la interacción espacio-temporal. La idea fue plantear una extensión natural del

modelo espacial de Besag, York y Mollié con el fin de tener en cuenta la evolución temporal. Así, además de la estructura espacial formada por las componentes de heterogeneidad y de correlación espacial, se incluye otra estructura similar de heterogeneidad y correlación temporal, $\boldsymbol{\gamma}$ y $\boldsymbol{\alpha}$ respectivamente, para captar el efecto del tiempo. Por último se añade una componente, $\boldsymbol{\delta}$, que pretendería modelizar la interacción espacio-temporal. De este modo el modelo quedaría:

$$log\lambda_i = \mu + \theta_i + \phi_i + \gamma_j + \alpha_j + \delta_{ij} \qquad i = 1...N, j = 1...t$$

pasando a considerarse desde entonces el modelo de referencia para todo análisis de datos con estructura espacio-temporal.

Todos estos ejemplos nacen bajo el amparo de la cartografía de enfermedades que, nuevamente, ha sido el motor de avance y desarrollo esta vez de la modelización espacio-temporal. Además de las citadas, en la literatura del tópico aparecen otras importantes referencias en las que se presentan modificaciones, ampliaciones e incluso modelizaciones alternativas a las citadas anteriormente. Sun *et al.* (2000), Richardson *et al.* (2006), Abellan *et al.* (2008), Martínez-Beneito *et al.* (2008) o Schrödle y Held (2011) son ejemplos de propuestas de modelos espacio-temporales que se están convirtiendo en el estándar de modelización, y no sólo en epidemiología sino que su uso llega a extenderse a otros ámbitos, véase por ejemplo Christensen y Yetkin (2005) o Gonzalez *et al.* (2012).

Sin embargo, y de manera análoga al caso de espacial, la variedad de estructuras espacio-temporales en las observaciones conlleva la distinción de diferentes procesos de análisis. Así, se podría hablar de procesos puntuales con dimensión temporal, de modelos dinámicos espacio-temporales, de modelización por bloques de nivel o de modelos de correlación espacio-temporal. No se pretende en este trabajo adentrarse en la definición y desarrollo de cada uno de estos procesos. Este hecho conllevaría una extensión no adecuada para el propósito de este capítulo cuya motivación radica

en la contextualización de las modelizaciones con las que vamos trabajar en lo sucesivo.

No obstante, en caso de necesitar una profundización en alguno de los procesos asociados a la existencia de estructura espacio temporal, Cressie y Wikle (2011) o Banerjee *et al.* (2014) contribuyen en la literatura con un revisión completa sobre el tema.

3.4. Datos medioambientales

La naturaleza de los datos con los que se trabaja en estudios medioambientales encaja con la ya comentada naturaleza de datos con estructura espacio-temporal. Las observaciones son tomadas en lugares geográficos diferentes y en instantes temporales distintos. Esta circunstancia implica la posible existencia de correlación entre observaciones próximas en el espacio y en el tiempo que debe ser tenida en cuenta. Consecuentemente, será necesario diseñar y aplicar modelos y procedimientos espacio-temporales que tengan en cuenta estas particularidades.

Si además de las características a tener en cuenta al trabajar con observaciones medioambientales, los datos están incompletos debido a la existencia de valores ausentes, la realización del estudio resulta aun más compleja a la vez que más atractiva estadísticamente.

En concreto, esta investigación está motivada por la aparición de valores ausentes en datos de calidad del agua potable en la Comunitat Valenciana. Estos datos recogen mediciones de concentración de diversas sustancias tales como magnesio, calcio y nitratos en el agua potable. El interés epidemiológico de la calidad del agua que bebemos reside en el análisis de la influencia que estos factores medioambientales tienen en la aparición de cánceres localizados en el aparato digestivo, Yang *et al.* (1998), o con enfermedades cardiovasculares, Ferrándiz *et al.* (2003).

Debido a esto, la Conselleria de Medio Ambiente de la Generalitat Valenciana mide anualmente, desde 1991, la

concentración de nitratos, calcio y magnesio, así como de otras sustancias, en el agua potable los 540 municipios de la Comunitat Valenciana.

Para el estudio planteado en esta tesis, de todas las componentes observadas en el banco de datos de calidad del agua únicamente se mostrará en este trabajo el proceso de imputación para los datos de concentración de nitratos y de magnesio. Estos conjuntos de datos resultan muy apropiados para este estudio pues ambos conjuntos de datos comparten a priori una misma estructura espacio-temporal pero, como más adelante se mostrará, las necesidades de modelización requeridas para cada uno de ellos son muy diferentes.

A nivel medioambiental, cabe destacar que el magnesio es un elemento indispensable para el crecimiento, de hecho el organismo humano ingiere gran cantidad de magnesio diariamente a través de los alimentos. Aunque no existe una evidencia científica de la toxicidad del magnesio, la Organización Mundial de la Salud (OMS) establece como concentración máxima deseable 50 mg/l. Sin embargo el magnesio, a elevadas concentraciones, sí resulta ser un contaminante del agua pues contribuye notablemente a caracterizar su dureza y por lo tanto su calidad para el consumo humano. No obstante, el contenido en magnesio de un agua depende casi exclusivamente de los terrenos que atraviesa, pudiendo variar desde muy pocos mg/l a varios cientos de mg/l, por lo que su origen no es vinculado con actividades del ser humano.

Respecto a los nitratos, representan un problema creciente de contaminación para los acuíferos. Al contrario que el magnesio, estos sí suponen un riesgo para la salud, sobretodo a nivel infantil (enfermedad de los bebés azules) y su excesiva presencia viene causada fundamentalmente por el uso masivo de fertilizantes nitrogenados y por explotaciones ganaderas sin una adecuada gestión de los purines. En varias regiones de España de gran tradición agrícola y ganadera, entre las que se encuentra la Comunitat Valenciana, las aguas contienen cada vez mayores

niveles de nitratos. En nuestra Comunitat existen regiones donde se alcanzan concentraciones de nitratos muy por encima de los 50 mg/l permitidos por ley. Este límite viene, de nuevo, por la recomendación de la OMS como umbral máximo para el agua de consumo humano. Incluso existen administraciones más restrictivas que sitúan este límite por debajo de los 50mg/l, como la Agencia para la Protección del Medio Ambiente Norteamérica (EPA) que lo fija en 10 mg/l.

A nivel estadístico, el interés por estos datos radica en el hecho de están incompletos debido a la existencia de valores ausentes. Esto impide que estas observaciones puedan ser incluidas en estudios epidemiológicos para estudiar su influencia en salud pública. Por tanto se hace necesaria la imputación de los valores ausentes para completar los datos y poder así proceder a su posterior explotación epidemiológica.

A modo de resumen se presentan a continuación el cuadro 3.1 y los histogramas de ambos conjuntos de datos, figura 3.2 .

	Nitratos	Magnesio
Mínimo	0.00	0.00
1er Cuartil	5.00	17.00
Mediana	12.00	28.00
Media	27.50	30.45
3er Cuartil	33.00	39.00
Máximo	343.00	143.00
N	5400	5400
Ausentes	1336	1523
% valores $\leq$ OMS	83.83	87.88

Tabla 3.1: Resumen de la concentración de nitratos y magnesio

Se observa que los valores máximos, 343mg/l para nitratos y 143mg/l para magnesio, están muy por encima de las recomendaciones de la OMS. De hecho, aproximadamente el 15 % de los valores de nitratos y el 12 % de magnesio están por encima

del umbral máximo recomendado de 50mg/l. Esto no representaría un problema de salud pública en el caso de las concentraciones de magnesio. No obstante, para las concentraciones de nitratos superar los umbrales máximos sí podría llegar a suponer un problema de salud pública, además de poner de manifiesto que existen actividades ganaderas o agrícolas que no están cumpliendo con la reglamentación vigente.

A nivel estadístico, a simple vista se aprecia una marcada asimetría en ambos casos, sobre todo en la concentración de nitratos. Los valores medios son de 27.5mg/l para las concentraciones de nitratos y de 30.45mg/l para las de magnesio, mientras que las medianas son de 12mg/l y 28mg/l, respectivamente. Este es un comportamiento típico en las observaciones de factores medioambientales, el cual debe tenerse en cuenta para su adecuada modelización.

Por lo que respecta al principal interés que esta tesis tiene por estos datos, se observa un alto porcentaje de valores ausentes en ambos conjuntos, un 24.74 % para las observaciones de concentración de nitratos y un 28.20 % para las de concentración de magnesio. Ante tal proporción de información faltante se haría necesaria la imputación de los valores ausentes en caso de inclusión de estas variables medioambientales en cualquier estudio de salud pública.

Si nos centramos en la información disponible para cada municipio, en la tabla 3.2 se presenta la distribución del número de municipios según el número de valores ausentes. Destaca el hecho que la gran mayoría de municipios posee al menos un valor ausente de concentración de magnesio, mientras que casi un tercio de los municipios disponen de todas las observaciones de concentración de nitratos.

Teniendo en cuenta que se han observado 10 años, se aprecia que hay municipios con muy poca información disponible, llegando al extremo de dos municipios que no disponen de ningún valor observado ni de concentración de nitratos ni de concentración

de magnesio. Para estos casos la correlación espacial adquiere un papel principal, pues la información que puedan aportar los municipios vecinos es la única disponible a la hora de proporcionar imputaciones de los valores ausentes.

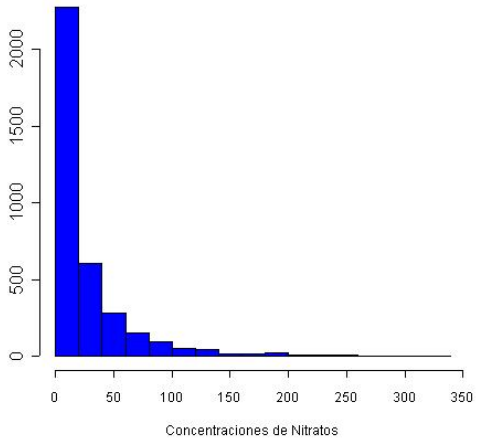
	Nº de municipios	
Nº observaciones ausentes	Nitratos	Magnesio
0	104	11
1	79	106
2	109	142
3	97	118
4	72	91
5	42	41
6	19	17
7	12	9
8	4	2
9	0	1
10	2	2

Tabla 3.2: Número de municipios según número de observaciones ausentes

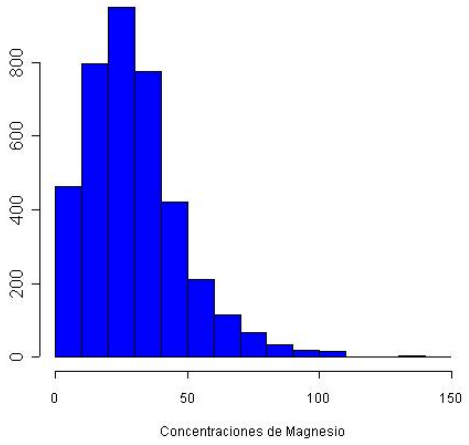
En las figuras 3.3, 3.4 y 3.5 se muestra la distribución espacial (a través de mapas de la Comunitat Valenciana) únicamente de las concentraciones de nitratos en agua potable en la Comunitat Valenciana para los diez años disponibles (1991-2000). Como se detallará más adelante, los datos de concentración de nitratos resultan ser muy atractivos desde el punto de vista estadístico, de ahí que la visualización de su distribución espacio-temporal sea más interesante que la de los datos de concentración de magnesio.

Los mapas mostrados ayudan, de una forma visual y rápida, a hacerse una idea de la posible correlación espacial y temporal entre las observaciones de concentración de nitratos así como de la gran cantidad de información ausente, destacando en este sentido el año 1992.

A pesar de que el proceso de recogida de datos sea espacio-temporal, no siempre conlleva la necesidad de implementar modelos con dicha estructura para su tratamiento. Debido a esto, en el siguiente capítulo planteamos un estudio previo mediante el cual decidiremos acerca de la naturaleza de las componentes que formarán parte de las modelizaciones para la imputación de los valores ausentes.



(a) Concentración de nitratos



(b) Concentración de magnesio

Figura 3.2: Histogramas de la concentración de nitratos y magnesio en el agua potable

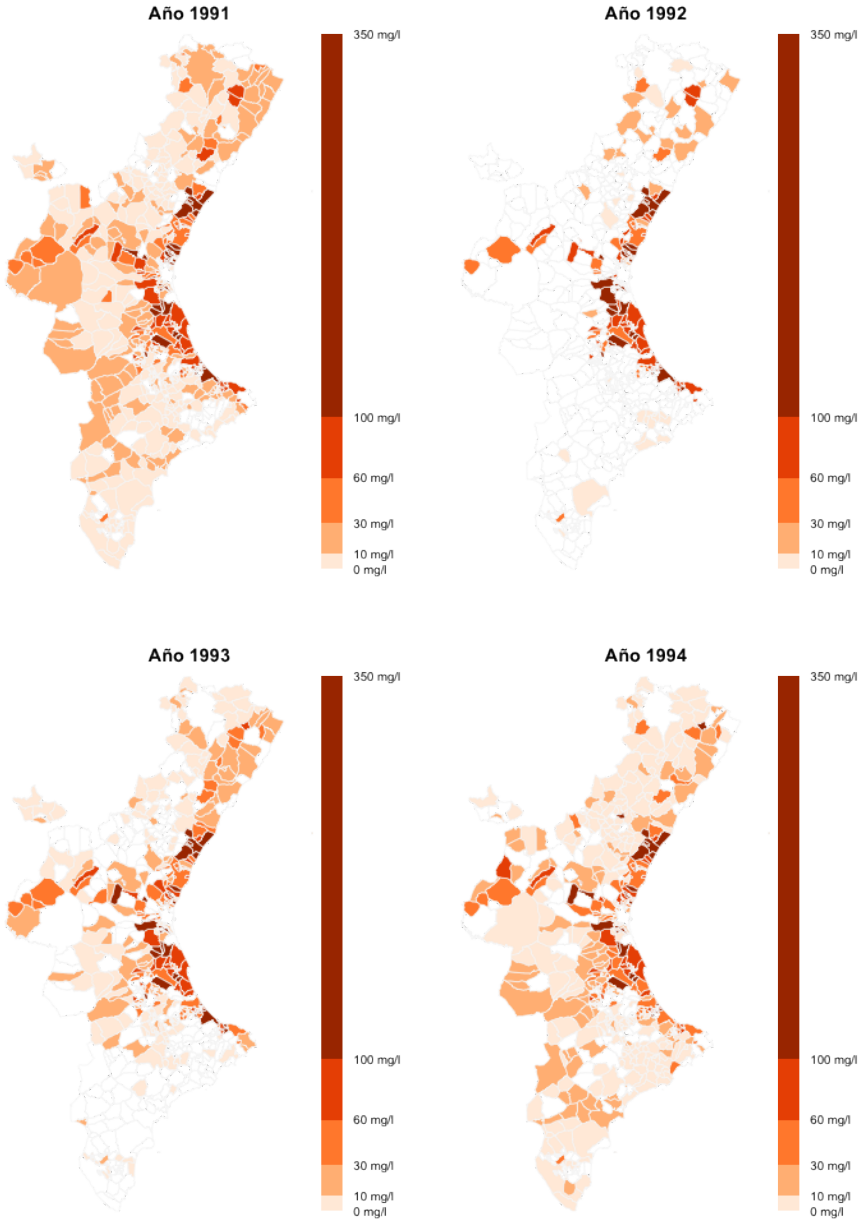


Figura 3.3: Distribución espacio-temporal de las concentraciones de nitratos en la Comunitat Valenciana. 1991-1994

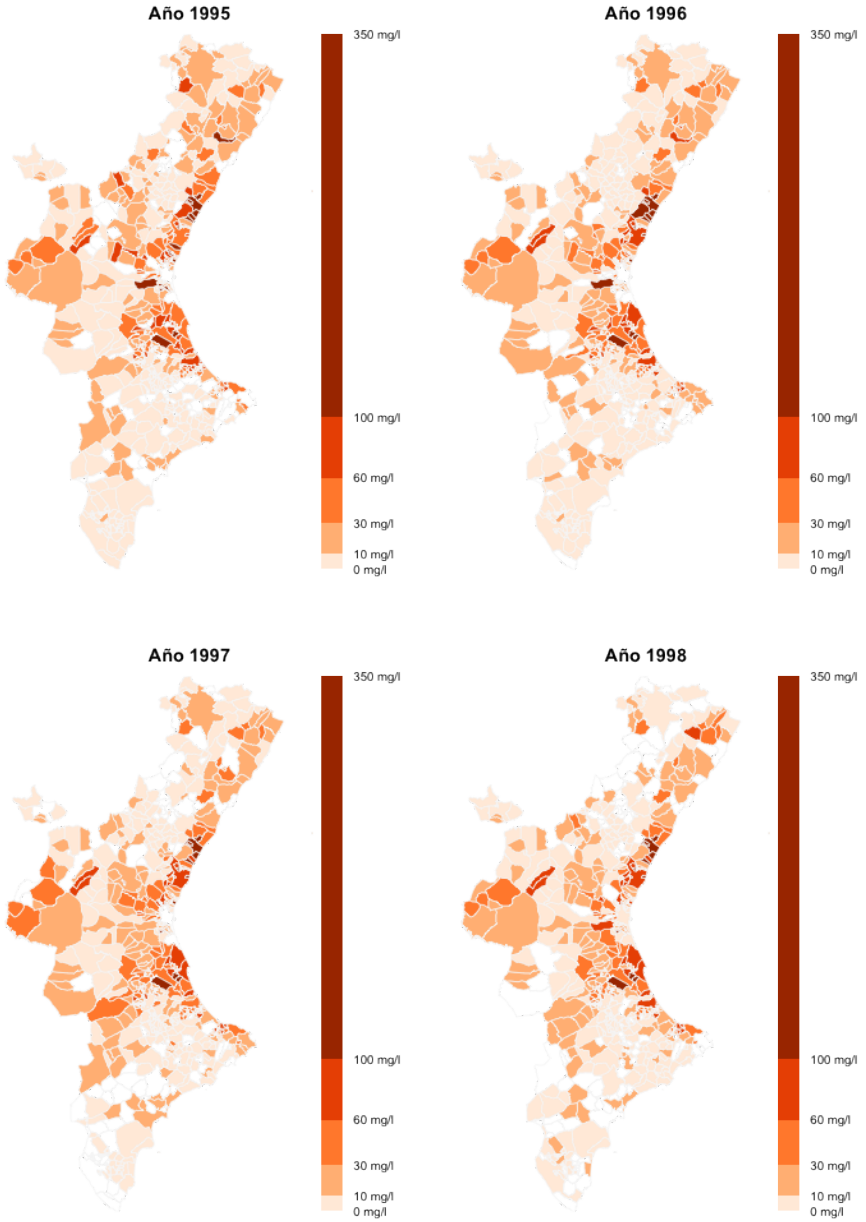


Figura 3.4: Distribución espacio-temporal de las concentraciones de nitratos en la Comunitat Valenciana. 1995-1998

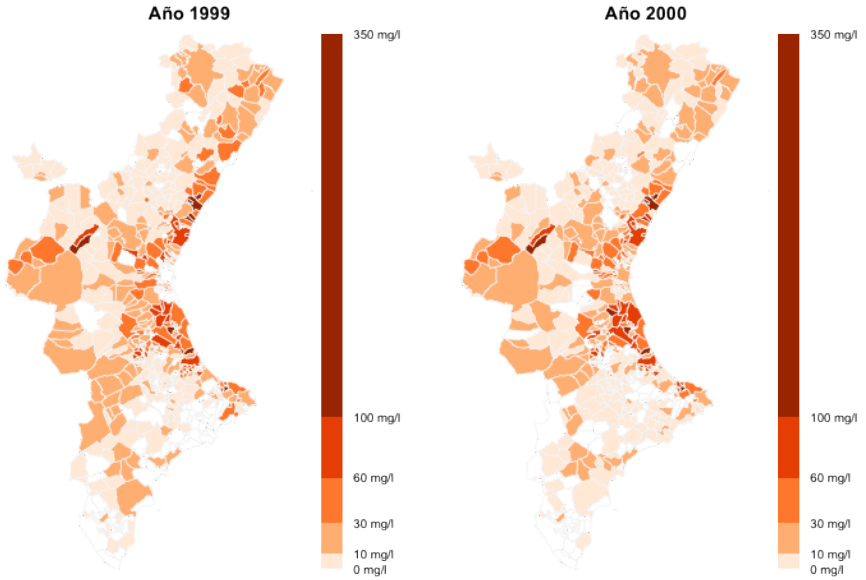


Figura 3.5: Distribución espacio-temporal de las concentraciones de nitratos en la Comunitat Valenciana. 1999-2000

Capítulo 4

Estudio de la estructura del modelo de imputación

Generalmente, y tal y como se ha descrito en el capítulo anterior, un banco de datos compuesto de observaciones recogidas en un conjunto de localizaciones y a lo largo de un cierto periodo de tiempo, requeriría para su análisis de un modelo espacio-temporal en coherencia con la estructura de dichos datos. Además, en la realización de análisis en los que intervienen datos medioambientales, como es el caso de esta tesis, la correlación existente entre observaciones próximas en el espacio y en el tiempo podría ayudar a realizar imputaciones adecuadas puesto que permitiría incorporar la información que proporcionan las observaciones vecinas tanto en el espacio como en el tiempo.

No obstante, este hecho no siempre es cierto, pues existen situaciones en las que a pesar de que los datos hayan sido recogidos de forma espacio-temporal es posible que dicha estructura no aporte información útil ya sea para realizar una imputación o para su análisis mediante el planteamiento de un modelo espacio-temporal.

Por ello, y centrándonos ya en nuestros datos, en el siguiente apartado se estudia la necesidad de proporcionar un modelo espacio-temporal de imputación que utilice toda la información que ofrecen los datos observados para resolver el problema de la

existencia de datos ausentes en la base de datos de calidad del agua potable de la Comunitat Valenciana.

4.1. Modelos básicos de imputación

Partimos de la definición de la verosimilitud de nuestros datos. Es decir, si llamamos Y_{it} a los datos que pretendemos completar mediante imputación, siendo $i = 1, ..., 540$ municipios de la Comunitat Valenciana y $t = 1, ..., 10$ los años representando el periodo 1991-2000. Se parte de

$$Y_{it} \sim f(\mu_{it}, \tau_Y) \tag{4.1}$$

donde μ_{it} es la media de las observaciones del municipio i en el año t y τ_Y es la precisión, que se considera constante para todos los municipios y años.

Atendiendo a lo comentado anteriormente respecto a la asimetría que suele presentarse en las observaciones de factores medioambientales, y con el fin de corregir este comportamiento, en nuestro caso se emplea $f(\mu_{it}, \tau_Y) = LogNormal(\mu_{it}, \tau_Y)$. De este modo se asume normalidad para los logaritmos de las concentraciones de nitratos y magnesio en el agua potable.

Como se ha visto en el resumen de los datos, tabla 3.1, existen municipios con un valor de 0mg/l ya sea en concentración de magnesio, 7 municipios, o de nitratos, 66 municipios. Dada la dificultad de que realmente existan 0mg/l de magnesio o nitratos en cualquier muestra de agua potable, y con el fin de poder aplicar la transformación logarítmica a los datos, sustituimos dichas concentraciones de 0mg/l por el valor 0.5mg/l. Este cambio no implica variaciones sensibles a nivel numérico pues en cualquier caso, 0mg/l o 0.5mg/l, se entiende que el nivel de magnesio o nitratos es muy bajo. Sin embargo este cambio sí nos permite

calcular el logaritmo de todos los valores presentes en el banco de datos.

Así, con el objetivo de estudiar la estructura de modelización necesaria para la adecuada imputación de los datos de nitratos y magnesio, se proponen cinco estructuras distintas que, partiendo de un modelo sencillo (modelo de heterogeneidad), se irá ampliando con la incorporación sucesiva de componentes más complejas que reflejen mejor la realidad espacio-temporal de los datos. Para ello se hará uso de modelos jerárquicos que permitirán introducir en sucesivas capas toda la información relevante sobre los datos.

Considerando que nuestro primer nivel es la distribución de los datos, expresada en (4.1), pasaremos al segundo nivel modelizando las medias μ_{it}.

Concretamente, las modelizaciones propuestas son:

1. Modelo de heterogeneidad (**Modelo H**). Consideramos que μ_{it} es siempre la misma en cada municipio y que en cada uno de ellos se expresa como la suma de un valor constante c y una componente aleatoria específica del municipio, a la que llamaremos componente de heterogeneidad. Se pretende con esta componente aleatoria ajustar las diferencias de valores entre los diferentes municipios debidas al azar o a cualquier factor oculto que pueda provocar dichas diferencias.

2. Modelo temporal (**Modelo T**). Puesto que las observaciones fueron realizadas anualmente durante el periodo 1991-2000, es lógico pensar en la posible existencia de una relación temporal entre las observaciones. Debido a esto, se incorpora al modelo **H** un término que recoja la evolución de las observaciones con el tiempo. En esta modelización suponemos que esta componente temporal es la misma en todos los municipios.

3. Modelo espacio temporal (**Modelo ET**). Según lo comentado acerca de la posible correlación existente entre observaciones próximas en el espacio, cabe esperar valores similares de

concentración de una determinada componente (nitratos o magnesio) en municipios colindantes. Así pues, incorporamos al modelo **T** una componente espacial que permita introducir esas relaciones de vecindad en el problema. Esta estructura espacial tiene el valor añadido de que hará posible la imputación en municipios con un número bajo de observaciones basándose en las observaciones de los municipios vecinos.

4. Modelo temporal completo (**Modelo TC**). Es una modificación del modelo **T**, pero incorporando una componente temporal específica para cada municipio. De esta forma, la componente temporal ya no es común y cada municipio tiene su propia evolución temporal. Esto pretende reflejar diferentes relaciones temporales para cada municipio, las cuales pueden ser debidas, por ejemplo, a diferentes intervenciones de aumento de la contaminación o de mayor tratamiento del agua potable realizadas durante el periodo estudiado.

5. Modelo espacio temporal completo (**Modelo ETC**). La inclusión de la componente espacial descrita para el modelo **ET** y las componentes temporales específicas del modelo **TC** proporciona al modelo la estructura espacio temporal más completa de ajuste de las concentraciones de nitratos y magnesio en el agua potable.

En el cuadro 4.1 se muestra con mayor detalle los cinco modelos, empleando la siguiente notación para los distintos términos:

μ: constante
θ_i: componente de heterogeneidad
α_j: componente temporal común
α_{it}: componente temporal específica
ϕ_i: componente espacial

El término de heterogeneidad $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_{540})'$, componente aleatoria presente en las cinco modelizaciones propuestas, se

Modelo H	$\mu_{it} = \mu + \theta_i$
Modelo T	$\mu_{it} = \mu + \theta_i + \alpha_j$
Modelo ET	$\mu_{it} = \mu + \theta_i + \alpha_j + \phi_i$
Modelo TC	$\mu_{it} = \mu + \theta_i + \alpha_{it}$
Modelo ETC	$\mu_{it} = \mu + \theta_i + \alpha_{it} + \phi_i$

Tabla 4.1: Modelización propuesta

considera que sigue una distribución normal centrada en cero y precisión τ_θ.

$$\theta_i \sim N(0, \tau_\theta) \qquad i = 1, \ldots, 540$$

Para la componente temporal común $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{10})'$, que aparece en los Modelos **T** y **ET**, se establece una distribución condicional autoregresiva CARNormal, ya comentada en el apartado 3.2, cuya densidad se construye a través de las distribuciones condicionadas de la siguiente forma:

$$\alpha_t | \alpha_{-t} \sim N(\textstyle\sum_k \beta_{tk}\alpha_k, \tau_\alpha) \qquad t = 1, \ldots, 10$$

siendo $\alpha_{-t} = \{\alpha_k : k \neq t\}$ el conjunto que incluye todas las componentes de $\boldsymbol{\alpha}$ excepto la que aparece en el subíndice, τ_α la precisión y los β_{tk} contienen la estructura de vecindad temporal que se considere oportuna definir. En nuestro caso, para un mismo municipio, se considera que una observación en un instante de tiempo concreto será vecina de la observación temporal anterior y posterior de dicha variable. De esa forma, β_{tk} será distinto de cero únicamente cuando $k = t - 1$ ó $k = t + 1$, valdrá $\beta_{tk} = 1/2$ cuando t pertenezca al conjunto $\{2, 3, \ldots, 9\}$ y será $\beta_{t2} = 1$ para $t = 1$ y $\beta_{t9} = 1$ cuando $t = 10$.

Para la componente espacial $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{540})'$ presente en los Modelos **ET** y **ETC**, se propone también una distribución CARNormal definida de la siguiente forma:

$$\phi_i|\phi_{-i} \sim N(\textstyle\sum_k \gamma_{ik}\phi_k, \tau_\phi) \qquad i = 1, \ldots, 540$$

siendo $\phi_{-i} = \{\phi_k : k \neq i\}$ y τ_ϕ la precisión. En este caso γ_{ik} será distinto de cero si el municipio k es vecino del municipio i, entendiendo por vecindad la contigüidad de los términos municipales. Para todos los vecinos k del municipio i se tendrá que $\gamma_{ik} = 1/n_i$, donde n_i es el número de vecinos que tiene el municipio i.

Para las 540 componentes temporales específicas de cada municipio definidas en los Modelos **TC** y **ETC**, se utilizan 540 distribuciones CARNormal, una para cada municipio:

$$\alpha_{it}|\alpha_{i-t} \sim N(\textstyle\sum_k \beta_{tk}\alpha_{ik}, \tau_\alpha^i) \qquad t = 1, \ldots, 10;\ i = 1, \ldots, 540$$

siendo $\alpha_{i-t} = \{t_{ik}, k \neq j\}$, $i = 1, \ldots, 540$ y τ_α^i la precisión, que depende del municipio considerado. En este caso, $\beta_{tk} = 1/2$ si t pertenece al conjunto $\{2, 3, \ldots, 9\}$ y $k = t-1$ ó $k = t+1$, mientras que $\beta_{t2} = 1$ para $t = 1$ y $\beta_{t9} = 1$ cuando $t = 10$.

Para la constante μ que aparece en todos los modelos se le asigna una distribución normal centrada en 0 y de amplia varianza.

Por último, bastará definir las distribuciones previas de los hiperparámetros desconocidos para obtener la completa modelización del problema.

Existen en la literatura diferentes formas de asignar distribuciones previas a las precisiones. En muchos estudios se asigna la distribución previa a la desviación típica, y no a la precisión, siendo la distribución Normal truncada y la distribución Uniforme las más habituales. Sin embargo, en otros estudios se asignan distribuciones previas de la familia Gamma directamente a las precisiones, Mollié (1996). La idea es definir los parámetros de estas distribuciones Gamma cercanos a cero de forma que, independientemente donde quede centrada la distribución, a su vez esté dotada de una amplia varianza. El motivo de esta asignación radica en la intención de definir distribuciones propias pero poco informativas de forma que el mayor peso del proceso caiga sobre la verosimilitud de los datos.

En este caso, empleamos distribuciones de la familia Gamma introduciendo unas ligeras modificaciones en pro de la estabilidad y convergencia de los modelos en el proceso de simulación. Concretamente centramos las distribuciones Gamma en valores a nuestro criterio más lógicos. Así, pese a dejar de ser distribuciones poco informativas, al asignársele una gran varianza el mayor peso del proceso sigue recayendo sobre la verosimilitud de los datos.

El cuadro 4.2 muestra el resumen de los parámetros que componen los modelos de imputación y en el cuadro 4.3 se detallan las distribuciones previas asignadas.

Constante	μ
Precisión de las observaciones	τ_Y
Precisión de la heterogeneidad	τ_θ
Precisión de la componente temporal común	τ_α
Precisión de la componente espacial	τ_ϕ
Precisión de las componentes temporales específicas	$\tau_\alpha^i \ \forall i$

Tabla 4.2: Resumen de los parámetros desconocidos en los modelos de imputación

Parámetro	Distribución previa
μ	N(0,0.001)
τ_Y	Ga(2,0.01)
τ_θ	Ga(5,0.01)
τ_α	Ga(2,0.01)
τ_ϕ	Ga(2,0.01)
$\tau_\alpha^i \ \forall i$	Ga(2,0.01)

Tabla 4.3: Distribuciones previas para los parámetros e hiperparámetros de los modelos de imputación

4.2. Selección de la estructura del modelo de imputación

Una vez planteados completamente los modelos de imputación, es necesario elegir aquel que se ajuste mejor a la naturaleza de los datos atendiendo a su estructura para proporcionar así las mejores imputaciones de los valores ausentes.

Para la selección del mejor modelo, es habitual el uso de criterios basados en la complejidad (haciendo uso del número de parámetros) y en la adecuación del propio modelo para explicar la variable respuesta (calidad del ajuste). Desde el punto de vista bayesiano, el criterio propuesto por Spiegelhalter *et al.* (2002), conocido por DIC (Deviance Information Criterion) es el más utilizado para comparar diferentes modelizaciones teniendo en cuenta tanto el ajuste del modelo como su complejidad. No obstante, este criterio no puede ser aplicado de la manera usual cuando se trabaja con valores ausentes.

Aunque existen estudios de adaptaciones del DIC al caso de la existencia de valores ausentes, veáse Mason *et al.* (2012a), nosotros entendemos que una buena forma de comparar los modelos de imputación propuestos es comparando las imputaciones que proporcionan cada uno de ellos. Para ello, una evidencia adecuada para comparar dichas imputaciones es la medida del error cometido al imputar dichos valores ausentes.

Por otro lado, es imposible determinar cómo de buena resulta ser una imputación si no se dispone de los valores que se están estimando para poder calcular así una medida del error cometido.

Así, para poder comparar las imputaciones que proporciona cada modelo se decide crear más valores ausentes de los que ya existen. De esta forma, podremos realizar también la imputación de éstos y así estaremos en condiciones de calcular una medida del error del imputación. En concreto se amplia el conjunto de datos ausentes en 500 valores, tanto para las concentraciones de nitratos como para las de magnesio. De este modo, de los 1336 valores ausentes inicialmente existentes en el conjunto de

las concentraciones de nitratos se pasa a 1836, y de los 1523 valores ausentes inicialmente existentes en el conjunto de las concentraciones de magnesio se pasa a 2023. Sobre estos 500 valores ausentes extra calcularemos la medida del error de imputación cometido por cada modelo.

Esta medida vendrá determinada por la expresión:

$$\sum_{Observados}(ValorImputado - ValorObservado)^2$$

Es decir, se calcula el cuadrado de la diferencia entre el valor imputado y el observado. Este error cuadrático lo sumamos para los 500 valores observados que se han eliminado para este propósito, proporcionando así una medida del error cometido por la imputación.

Como valor imputado probaremos con los dos posibles candidatos más lógicos: la media y la mediana de la muestra de la distribución posterior obtenida para los valores ausentes.

Ajustamos los modelos haciendo uso del software WinBUGS, ver código en el Anexo 3. Lanzamos un total de 40000 iteraciones en dos cadenas descartando las 35000 primeras y quedándonos con una iteración de cada 5. Posteriormente a la comprobación de la convergencia de los modelos, con la muestra total de 2000 observaciones de la distribución posterior de los valores ausentes guardada, calculamos los errores de imputación.

Los resultados obtenidos se muestran en los cuadros 4.4 y 4.5.

Modelo	Media	Mediana
H	2.55	2.60
T	2.40	2.46
ET	2.39	2.44
TC	1.32	1.21
ETC	1.24	1.17

Tabla 4.4: Error de imputación (multiplicado por 10^{-6}) para los nitratos

Modelo	Media	Mediana
H	6.46	6.16
T	6.49	6.14
ET	6.69	6.39
TC	6.56	6.10
ETC	6.74	6.29

Tabla 4.5: Error de imputación (multiplicado por 10^{-4}) para el magnesio

Si observamos los resultados obtenidos de las dos imputaciones, media y mediana, para los datos de magnesio, se aprecia que la imputación con la mediana es, en todos los modelos, la que comete siempre el menor error.

Lo mismo sucede en el caso de los nitratos en los modelos **TC** y **ETC**.

Además, viendo el cuadro 4.4, se observa que se produce un descenso considerable en el error cometido para los modelos que incorporan la componente temporal específica para cada municipio. En este caso, es el Modelo **ETC**, el más complejo de los cinco, y a través de la mediana, el que consigue el error más pequeño en la imputación. Este modelo es el que se aproxima mejor a la realidad de nuestro problema: incluye elementos específicos de cada municipio, de su evolución temporal y de su posible correlación con los municipios vecinos.

No ocurre lo mismo en el caso de las concentraciones de magnesio. En contra de lo esperado (en principio la estructura de las observaciones de magnesio no difiere de la de nitratos) la componente espacial no mejora las imputaciones de magnesio. Sí lo hace la inclusión del término temporal específico por municipio, aunque la diferencia de error de imputación con el modelo de heterogeneidad (el más sencillo) es pequeña.

Por tanto, en vista de los resultados obtenidos, resulta necesario plantear modelos con estructura espacio-temporal para la

imputación de las concentraciones de nitratos ausentes en el banco de datos de calidad del agua potable, pues es la estructura que mejor refleja la realidad espacio-temporal de las observaciones cometiendo un error de imputación inferior a los otros modelos más sencillos.

No obstante, para realizar la imputación de las concentraciones de magnesio ausentes, optaríamos por utilizar el modelo de heterogeneidad ya que, siendo el modelo más sencillo, proporciona imputaciones similares a las del mejor modelo, el temporal específico.

En el siguiente capítulo se profundiza en el estudio del modelo espacio-temporal más adecuado de imputación para las concentraciones de nitratos basándonos en las modelizaciones espacio-temporales más recurrentes en la literatura.

Capítulo 5

Modelos de imputación

Vistos los resultados obtenidos en el apartado 4.2, se llega a la conclusión de que el modelo adecuado para la imputación de los valores ausentes de concentraciones de nitratos en la Comunidad Valenciana debe tener una estructura espacio-temporal.

Como se ha visto en el apartado 3.2, el uso de modelos con estructura espacio-temporal es muy común en el campo de la salud pública y medio ambiente. Existen múltiples modelizaciones con dicha estructura que pretenden analizar la influencia de uno o varios factores medioambientales sobre la salud pública (regresión ecológica) o estudiar la distribución espacial y su evolución temporal de determinadas variables de interés epidemiológico (cartografía de enfermedades).

En nuestro caso, y ante la escasa existencia en la literatura de este tipo de modelizaciones para el tratamiento de valores ausentes, adaptamos estos modelos tan conocidos en otros ámbitos para realizar la imputación de las observaciones ausentes de concentración de nitratos.

Del amplio abanico de modelizaciones espacio-temporales, elegimos dos de las más importantes en la literatura con el fin de estudiar y comparar el comportamiento de estos modelos en el campo de la imputación de valores ausentes. Las referencias a estos modelos son muchas dentro de la cartografía de enfermedades,

pero prácticamente inexistentes en cuanto al uso de este tipo de modelizaciones con el fin de proporcionar imputaciones en un banco de datos con valores ausentes.

El hecho de seleccionar estas modelizaciones espacio-temporales para realizar la imputación radica no sólo en su recurrencia en la literatura sino también en que son modelizaciones diferenciadas en cuanto a su estructura conceptual de la modelización del espacio-tiempo.

Cabe destacar que no sólo no se han encontrado referencias de la utilización de modelos espacio-temporales para la imputación de valores ausentes, sino que no existen suficientes referencias en la literatura (únicamente en Martínez-Beneito *et al.*, 2008) en las que se planteen una comparativa entre estas dos modelizaciones, ya sea para la cartografía de enfermedades o para imputación de valores ausentes. Por tanto, el análisis que se plantea en esta tesis tiene una doble componente de interés: por un lado la adaptación de este tipo de modelos al ámbito de la imputación de datos y, por otro, el planteamiento de una nueva comparativa entre los tipos de modelización más frecuentes en estudios espacio-temporales.

Así, lo que resta de capítulo lo dedicaremos a la exposición y desarrollo de cada una de las tres modelizaciones que proponemos para abordar la existencia de valores ausentes en la concentración de nitratos de la Comunitat Valenciana.

5.1. Interacción espacio-temporal, IET

Planteamos aquí el primer modelo espacio-temporal para la imputación de los valores ausentes. Proponemos una adaptación de la estructura espacio-temporal más común en la literatura y cuyos mimbres fueron propuestos inicialmente en (Besag *et al.*, 1991) en al ámbito de la cartografía de enfermedades, véase apartado 3.2. Desde entonces, como se ha comentado en el apartado 3.3, dicha estructura espacio-temporal ha sido utilizada en innumerables estudios en los que se han analizado diferentes variaciones, adaptaciones y

evoluciones. No obstante, la estructura base de la modelización ha permanecido estable a lo largo de todo este tiempo. Leonhard Held (2000), con revisión y actualización en Schrödle y Held (2011), resume a modo de compendio dicha modelización espacio-temporal. Para ello parte de la estructura base del modelo propuesto en Besag *et al.* (1991) y clasifica, atendiendo a cómo interactua el espacio y el tiempo, las diferentes interacciones en cuatro grandes grupos.

Así, el modelo, adaptado a nuestro estudio, parte nuevamente de la verosimilitud de nuestros datos, vista en el apartado 4.1,

$$Log(Y_{it}) \sim Normal(\mu_{it}, \tau_Y) \qquad i = 1, \ldots, N \;\; t = 1, \ldots, T$$

donde τ_Y es la precisión de las concentraciones de nitratos y donde μ_{it} se descompone de la siguiente forma,

$$\mu_{it} = \mu + \gamma_t + \alpha_t + \theta_i + \phi_i + \delta_{it}$$

siendo μ la media global de la concentración de nitratos, a escala logarítmica, γ_t y α_t forman el bloque temporal en el que γ_t es el término que ajusta la variabilidad temporal a gran escala, sin estructura de vecindad o término de heterogeneidad, y α_t modeliza la variabilidad a pequeña escala, dotándolo de estructura de vecindad temporal. θ_i y ϕ_i formarían el bloque espacial donde θ_i es la heterogeneidad y ϕ_i el término con estructura de vecindad espacial. Por último, δ_{it} representa la interacción espacio-temporal que pretende ajustar la variabilidad a pequeña escala.

El bloque espacial y el bloque temporal forman la comentada estructura base del modelo espacio-temporal. Son los efectos principales que pretenden ajustar la correlación espacial y la correlación temporal.

Por tanto, seguidamente planteamos las distribuciones previas a las diferentes componentes de dichos efectos principales.

Para el bloque temporal,

$$\gamma_t \sim Normal(0, \tau_\gamma) \qquad t = 1, \ldots, T$$

donde τ_γ es la precisión, y

$$\boldsymbol{\alpha} \sim CARNormal(\tau_\alpha)$$

Nuevamente hacemos uso de la distribución CARNormal, vista en el apartado 3.2 y utilizada también en el aparado 4.1, para la modelización de la correlación temporal. En este caso,

$$\alpha_t|\alpha_{-t} \sim N(\sum_k \beta_{tk}\alpha_k, \tau_\alpha) \qquad t = 1, \ldots, T$$

con $\alpha_{-t} = \{\alpha_k : k \neq t\}$, τ_α la precisión y β_{tk} será distinto de cero únicamente cuando $k = t - 1$ ó $k = t + 1$, valdrá $\beta_{tk} = 1/2$ cuando t pertenezca al conjunto $\{2, 3, \ldots, 9\}$ y será igual a 1 para $t = 1$ y para $t = 10$. Mediante esta definición de estructura temporal asumimos que cada valor de nitratos en un municipio determinado está correlado con las observaciones del año anterior y posterior.

Para el bloque espacial,

$$\theta_i \sim Normal(0, \tau_\theta) \qquad i = 1, \ldots, N$$

siendo τ_θ la precisión, y

$$\boldsymbol{\phi} \sim CARNormal(\tau_\phi)$$

donde,

$$\phi_i|\phi_{-i} \sim N(\sum_k \gamma_{ik}\phi_k, \tau_\phi) \qquad i = 1, \ldots, N$$

con $\phi_{-i} = \{\phi_k : k \neq i\}$ y τ_ϕ la precisión. γ_{ik} será distinto de cero si el municipio k es vecino del municipio i, entendiendo por vecindad espacial la contigüidad de los términos municipales. Para todos los vecinos k del municipio i se tendrá que $\gamma_{ik} = \frac{1}{n_i}$, donde n_i es el número de vecinos que tiene el municipio i.

Para las precisiones presentes en el modelo, τ_Y, τ_γ, τ_α, τ_θ y τ_ϕ, así como para τ_δ que veremos a continuación, asignamos

distribuciones poco informativas Gamma siguiendo el razonamiento visto en el apartado 4.1,

$$\tau_Y, \tau_\gamma, \tau_\alpha, \tau_\theta, \tau_\phi, \tau_\delta \sim Gamma(a, b).$$

Por último asignamos

$$\mu \sim N(0, 0{,}001).$$

Partiendo de esta estructura de modelización base, la componente de interacción, δ_{it}, marcará la diferencia entre modelizaciones pues es la componente que ajustará las diferentes relaciones entre el espacio y el tiempo presentes en la naturaleza de los datos. Así, en Schrödle y Held (2011) se clasifica la interacción espacio-temporal en cuatro tipos diferentes atendiendo a dicha naturaleza:

I No existe interacción espacio-temporal, por tanto

$$\delta_{it} \sim Normal(0, \tau_\delta) \qquad i = 1, \ldots, N \;\; t = 1, \ldots, T$$

siendo τ_δ la precisión.

II Cada municipio tiene una evolución temporal diferente e independiente de sus vecinos, así para cada municipio i

$$\boldsymbol{\delta_i} \sim CARNormal(\tau_\delta) \qquad i = 1, \ldots, N$$

es decir,

$$\delta_{it}|\delta_{i-t} \sim N(\sum_k \beta_{tk}\delta_{ik}, \tau_\delta) \qquad t = 1, \ldots, T$$

siendo $\delta_{i-t} = \{\delta_{ik} : k \neq t\}$ y τ_δ la precisión. β_{tk} se define de la misma forma que para la componente $\boldsymbol{\alpha}$ del efecto principal temporal.

III En cada unidad de tiempo la correlación espacial es diferente e independiente del resto de unidades temporales, así para cada año t

$\boldsymbol{\delta_t} \sim CARNormal(\tau_\delta) \qquad t = 1, \ldots, T$

es decir,

$$\delta_{it}|\delta_{-it} \sim N(\sum_k \gamma_{ik}\delta_{kt}, \tau_\delta) \qquad i = 1, \ldots, N$$

donde $\delta_{-it} = \{\delta_{kt} : k \neq i\}$ y τ_δ la precisión. γ_{ik} queda definido de la misma forma que en la componente ϕ_i del efecto principal espacial.

IV Existe correlación espacial y temporal para todos los municipios y para cada año, en este caso, llamando $h = i \cdot t$ con $i = 1, \ldots, N$ y $t = 1, \ldots, T$

$\boldsymbol{\delta} \sim CARNormal(\tau_\delta)$

siendo,

$$\delta_h|\delta_{-h} \sim N(\sum_k \gamma_{hk}\delta_k, \tau_\delta) \qquad h = 1, \ldots, N \cdot T$$

con $\delta_{-h} = \{\delta_k : k \neq h\}$ y τ_δ la precisión. En este caso, en lugar de hablar de municipios vecinos y vecindad temporal, hablamos de observaciones vecinas. Es decir, γ_{hk} será distinto de cero si la observación k, correspondiente a un municipio i_k en un año t_k, es vecina de la observación h, que corresponde al municipio i en el año t. Definimos entonces, como observaciones vecinas de la observación h las observaciones del mismo municipio i el año anterior y posterior, $t-1$, $t+1$, las observaciones de los municipios contiguos para el mismo año t y las observaciones de los municipios contiguos en los años anterior y posterior. De esta forma asumimos que no sólo existe correlación entre las observaciones anterior y posterior

en el tiempo de un municipio dado y entre las observaciones de municipios contiguos, sino que también existe correlación entre las observaciones de dicho municipio y la de los años anterior y posterior de sus municipios contiguos.

Por último, para todas las observaciones vecinas k de la observación h se tendrá que $\gamma_{hk} = \frac{1}{n_h}$, donde n_h es el número de observaciones vecinas que tiene la observación h.

Así, atendiendo a esta clasificación descrita, el modelo **ETC** presentado en el apartado 4.1 encajaría en la modelización de la interacción presentada aquí como tipo **II**. No obstante, la interacción tipo **II**, aún conservando la idea base del modelo **ETC** de asumir evoluciones temporales independientes entre municipios, supone una ampliación de este último a través de la inclusión del bloque principal temporal (término de heterogenenidad y término con estructura de vecindad temporal) a la vez que asume la misma precisión para todas las evoluciones temporales.

Cabe recordar que dicho modelo **ETC**, resultó ser el que mejores imputaciones proporcionaba de los valores ausentes de concentración de nitratos en el estudio preliminar realizado en el capítulo 4. Y que este estudio preliminar nos ayudó a entender que las concentraciones de nitratos requerían de una modelización adecuada a su estructura espacio-temporal.

Por este motivo, seleccionamos de entre los cuatro tipos de interacción espacio-temporal descritos para nuestro propósito de comparar su comportamiento en la imputación de los valores ausentes, el tipo **II**, que ya hemos visto que resulta adecuado para este caso, y el tipo **IV**, para explorar la definición alternativa de vecindad entre observaciones. A estos modelos añadiremos una tercera opción que expondremos en la siguiente sección.

Así, las dos modelizaciones elegidas en esta sección aunque comparten la misma estructura espacio-temporal ajustan la interacción de forma diferente. En el caso de la interacción tipo **II**, se asume que únicamente las observaciones anterior y posterior en el tiempo aportan información para la imputación de los

valores ausentes de un determinado municipio, siendo independiente esta información del resto de municipios; es decir, evoluciones temporales diferentes e independientes entre los municipios. En el caso de la interacción tipo **IV**, se asume una correlación más fuerte al considerar que, además de las observaciones anterior y posterior en el tiempo de un municipio dado, también aportan información las observaciones de los municipios colindantes no sólo en ese mismo año sino también en el año anterior y posterior.

En resumen, las dos primeras modelizaciones propuestas para la imputación de los valores ausentes quedan de la siguiente forma:

Modelo IET1

$$Log(Y_{it}) \sim Normal(\mu_{it}, \tau_Y) \qquad i = 1, \ldots, N \;\; t = 1, \ldots, T$$

$$\mu_{it} = \mu + \gamma_t + \alpha_t + \theta_i + \phi_i + \delta_{it}$$

$$\begin{aligned}
&\gamma_t \sim Normal(0, \tau_\gamma) \\
&\boldsymbol{\alpha} \sim CARNormal(\tau_\alpha) \\
&\theta_i \sim Normal(0, \tau_\theta) \\
&\boldsymbol{\phi} \sim CARNormal(\tau_\phi) \\
&\boldsymbol{\delta_i} \sim CARNormal(\tau_\delta) \\
&\mu \sim N(0, 0{,}001) \\
&\tau_Y, \tau_\gamma, \tau_\alpha, \tau_\theta, \tau_\phi, \tau_\delta \sim Gamma(0{,}5, 0{,}0005)
\end{aligned}$$

Modelo IET2

$$Log(Y_{it}) \sim Normal(\mu_{it}, \tau_Y) \qquad i = 1, \ldots, N \;\; t = 1, \ldots, T$$

$$\mu_{it} = \mu + \gamma_t + \alpha_t + \theta_i + \phi_i + \delta_{h=i \cdot t}$$

$$\begin{aligned}
&\gamma_t \sim Normal(0, \tau_\gamma) \\
&\boldsymbol{\alpha} \sim CARNormal(\tau_\alpha) \\
&\theta_i \sim Normal(0, \tau_\theta) \\
&\boldsymbol{\phi} \sim CARNormal(\tau_\phi) \\
&\boldsymbol{\delta} \sim CARNormal(\tau_\delta) \\
&\mu \sim N(0, 0{,}001) \\
&\tau_Y, \tau_\gamma, \tau_\alpha, \tau_\theta, \tau_\phi, \tau_\delta \sim Gamma(1, 0{,}02)
\end{aligned}$$

5.2. Autoregresivo espacio-temporal, ARET

Proponemos ahora una tercera modelización basada en el modelo descrito en Martínez-Beneito *et al.* (2008). Esta modelización es novedosa en comparación a la propuesta en el apartado anterior, es un modelo planteado también para la cartografía de enfermedades y en esta tesis lo adaptamos a la imputación de valores ausentes, comparando además su comportamiento respecto a los modelos ya descritos.

El interés por adaptar la idea que esta modelización plantea a la imputación de datos radica en que el ajuste de la correlación espacio-temporal es, de inicio, conceptualmente diferente al planteado en la sección 5.1.

El modelo parte de nuestra verosimilitud para las concentraciones de nitratos,

$$Log(Y_{it}) \sim Normal(\mu_{it}, \tau_Y) \qquad i = 1, \ldots, N \;\; t = 1, \ldots, T$$

donde τ_Y es la precisión y se modeliza μ_{it} como

$$\begin{aligned} &\text{si t} = 1 \quad \mu_{i1} = \mu + \alpha_1 + \tfrac{1}{\sqrt{1-\rho^2}}(\theta_{i1} + \phi_{i1}) \\ &\forall \text{ t} \neq 1 \quad \mu_{it} = \mu + \alpha_t + \rho(\mu_{i(t-1)} - \mu - \alpha_{t-1}) + \theta_{it} + \phi_{it} \end{aligned}$$

donde μ es la media global de la concentración de nitratos a escala logarítmica, α_t es la componente temporal, común para todos los municipios, θ_{it} es la heterogeneidad espacial y temporal y ϕ_{it} es la componente espacial dotada de estructura de vecindad.

Así, se modeliza la media de concentración de nitratos de cada municipio y año mediante un modelo que incluye un término temporal como efecto principal α_t, un bloque espacial, compuesto por un término de heterogeneidad θ_{it} y por un término con estructura vecinal ϕ_{it}, y la concentración de nitratos del año anterior pre-multiplicado por un coeficiente de correlación ρ. Este último

bloque modeliza la conexión en el tiempo de manera similar a una serie temporal auto regresiva de orden 1. Éste es el motivo por el que aparece el término $\frac{1}{\sqrt{1-\rho^2}}$, pues así nos aseguramos la estabilidad del modelo, véase Martínez-Beneito *et al.* (2008).

Esta modelización plantea que la estimación de la concentración de nitratos en una determinada localización depende de las estimaciones de las concentraciones de nitratos en localizaciones vecinas y de las propias estimaciones en años anteriores. De esta manera, haciendo depender la concentración de nitratos de la estimación obtenida en el año anterior, la evolución temporal también tiene estructura espacial, es decir, localizaciones vecinas tienden a tener una evolución temporal de concentración de nitratos parecida.

Se puede apreciar entonces, que este modelo ajusta la correlación espacio-temporal de las observaciones de una forma conceptualmente diferente a la planteada en el modelo anterior. La idea que subyace en esta modelización podría ser equivalente a la planteada en el modelo IET2 pero tiene al añadido de ser capaz de contar con la información de todos los años anteriores para proporcionar imputaciones en un año dado. Esta ventaja se hace más valiosa a medida que se avanza en el tiempo en contraste con el modelo IET2 que sólo tiene en cuenta el año anterior y posterior.

Si profundizamos en el modelo desarrollándolo obtenemos, llamando $f(\rho) = \frac{1}{\sqrt{1-\rho^2}}$

$\forall$ **i,t=1** $\mu_{i1} = \mu + \alpha_1 + f(\rho)(\theta_{i1} + \phi_{i1})$

$\forall$ **i,t=2** $\mu_{i2} = \mu + \alpha_2 + \rho(\mu_{i1} - \mu - \alpha_1) + \theta_{i2} + \phi_{i2} =$
$= \mu + \alpha_2 + \rho(\mu + \alpha_1 + f(\rho)(\theta_{i1} + \phi_{i1}) - \mu - \alpha_1) + \theta_{i2} + \phi_{i2} =$
$= \mu + \alpha_2 + \rho f(\rho)(\theta_{i1} + \phi_{i1}) + \theta_{i2} + \phi_{i2}$

$\forall$ **i,t=3** $\mu_{i3} = \mu + \alpha_3 + \rho(\mu_{i2} - \mu - \alpha_2) + \theta_{i3} + \phi_{i3} =$
$= \mu + \alpha_3 + \rho(\mu + \alpha_2 + \rho f(\rho)(\theta_{i1} + \phi_{i1}) + \theta_{i2} + \phi_{i2} - \mu - \alpha_2) +$
$\theta_{i3} + \phi_{i3} = \mu + \alpha_3 + \rho(\rho f(\rho)(\theta_{i1} + \phi_{i1}) + \theta_{i2} + \phi_{i2}) + \theta_{i3} + \phi_{i3} =$
$= \mu + \alpha_3 + \rho^2 f(\rho)(\theta_{i1} + \phi_{i1}) + \rho(\theta_{i2} + \phi_{i2}) + \theta_{i3} + \phi_{i3}$

$\forall$ **i,t=4** $\mu_{i4} = \mu + \alpha_4 + \rho(\mu_{i3} - \mu - \alpha_3) + \theta_{i4} + \phi_{i4} =$
$= \mu + \alpha_4 + \rho(\mu + \alpha_3 + \rho^2 f(\rho)(\theta_{i1} + \phi_{i1}) + \rho(\theta_{i2} + \phi_{i2}) + \theta_{i3} +$
$\phi_{i3} - \mu - \alpha_3) + \theta_{i4} + \phi_{i4} =$
$= \mu + \alpha_4 + \rho^3 f(\rho)(\theta_{i1} + \phi_{i1}) + \rho^2(\theta_{i2} + \phi_{i2}) + \rho(\theta_{i3} + \phi_{i3}) + \theta_{i4} + \phi_{i4}$

$\forall$ **i,t** $\mu_{it} = \mu + \alpha_t + \rho^{t-1} f(\rho)(\theta_{i1} + \phi_{i1}) + \sum_{k=2}^{t} \rho^{t-k}(\theta_{ik} + \phi_{ik})$

Se puede observar que la media de concentración de nitratos en un año dado depende de la media global μ, del término temporal común de ese año, del bloque espacial formado por la componente de heterogeneidad y la componente con estructura de vecindad de ese año y de dichos bloques de todos los años anteriores (esta última dependencia está controlada por el parámetro ρ). Por tanto, a medida que se avanza en el tiempo, se dispone de más información para el ajuste de la concentración de nitratos.

Si de la expresión general de μ_{it} separamos, por un lado los términos espaciales con estructura de vecindad y, por otro, los términos de heterogeneidad se obtiene:

$$\mu_{it} = \mu + \alpha_t + \rho^{t-1} f(\rho)\theta_{i1} + \sum_{k=2}^{t} \rho^{t-k}\theta_{ik} + \sum_{k=2}^{t} \rho^{t-k}\phi_{ik}$$

Por tanto, la media de concentración de nitratos depende de tres bloques claramente distinguibles, la formada por la media global de concentración de nitratos y la componente temporal común para todos los municipios, el bloque formado por los términos de heterogeneidad de todos los años anteriores y gobernado por el parámetro ρ y el bloque formado por los términos con estructura espacial de todos los años anteriores que también está gobernado por el parámetro ρ. Se podría considerar que el bloque formado por las heterogeneidades y el formado por los términos con estructura espacial estuvieran gobernados por diferentes parámetros.

Para esta modelización asignamos las siguientes distribuciones previas, para la componente temporal

$$\boldsymbol{\alpha} \sim CARNormal(\tau_\alpha)$$

siendo,

$$\alpha_t|\alpha_{-t} \sim N(\sum_k \beta_{tk}\alpha_k, \tau_\alpha) \qquad t = 1, \ldots, T$$

con $\alpha_{-t} = \{\alpha_k : k \neq t\}$, τ_α la precisión y β_{tk} definido de la misma forma que para el modelo X, es decir, será distinto de cero únicamente cuando $k = t - 1$ ó $k = t + 1$, valdrá $\beta_{tk} = 1/2$ cuando k pertenezca al conjunto $\{2, 3, \ldots, 9\}$ y será $\beta_{t2} = 1$ para $t = 1$ y $\beta_{t9} = 1$ cuando $t = 10$.

Para el bloque espacial,

$$\theta_{it} \sim Normal(0, \tau_\theta) \qquad i = 1, \ldots, N \;\; t = 1, \ldots, T$$

siendo τ_θ la precisión, y para cada t$\in \{1, 2, \ldots, T\}$

$$\boldsymbol{\phi_t} \sim CARNormal(\tau_\phi)$$

es decir, para cada t$\in \{1, 2, \ldots, T\}$

$$\phi_{it}|\phi_{-it} \sim N(\sum_k \gamma_{itk}\phi_t k, \tau_\phi) \qquad i = 1, \ldots, N$$

con $\phi_{-it} = \{\phi_k t : k \neq i\}$ y τ_ϕ la precisión. γ_{itk} será distinto de cero si el municipio k es vecino del municipio i, entendiendo nuevamente por vecindad espacial la contigüidad de los términos municipales. De igual modo, para todos los vecinos k del municipio i se tendrá que $\gamma_{itk} = \frac{1}{n_i}$, donde n_i es el número de vecinos que tiene el municipio i.

Para el parámetro que ajusta la correlación temporal asignamos

$$\rho \sim Uniforme(-1, 1).$$

Para la media global

$$\mu \sim N(0, 0{,}01).$$

Finalmente asignamos distribuciones Gamma como hiperprevias de las precisiones del modelo de la misma forma que en el modelo IET,

$$\tau_Y, \tau_\alpha, \tau_\theta, \tau_\phi \sim Gamma(a, b).$$

Resumiendo, el tercer modelo propuesto para la imputación de los valores ausentes queda:

Modelo ARET

$$Log(Y_{it}) \sim Normal(\mu_{it}, \tau_Y) \qquad i = 1, \ldots, N \;\; t = 1, \ldots, T$$

$$\begin{aligned} \mu_{i1} &= \mu + \alpha_1 + \tfrac{1}{\sqrt{1-\rho^2}}(\theta_{i1} + \phi_{i1}) && t = 1 \\ \mu_{it} &= \mu + \alpha_t + \rho(\mu_{i(t-1)} - \mu - \alpha_{t-1}) + \theta_{it} + \phi_{it} && t > 1 \end{aligned}$$

$$\begin{aligned} &\boldsymbol{\alpha} \sim CARNormal(\tau_\alpha) \\ &\theta_{it} \sim Normal(0, \tau_\theta) \\ &\boldsymbol{\phi_t} \sim CARNormal(\tau_\phi) \\ &\rho \sim Uniforme(-1, 1) \\ &\mu \sim N(0, 0{,}01) \\ &\tau_Y, \tau_\alpha, \tau_\theta, \tau_\phi \sim Gamma(0{,}5, 0{,}005) \end{aligned}$$

En el siguiente capítulo se muestran los resultados obtenidos al ajustar las tres modelizaciones propuestas para la imputación de los valores ausentes de concentración de nitratos. Asimismo se expondrá la comparativa entre las modelizaciones descritas incluyéndose el análisis de cómo ajusta cada uno de los modelos propuestos la realidad espacio-temporal de las observaciones.

Capítulo 6

Resultados

En este capítulo se muestran los diferentes resultados obtenidos a partir de la implementación de los modelos propuestos en el capítulo anterior. En primer lugar, se presenta el banco de datos de concentración de nitratos con los valores ausentes imputados por los diferentes modelos. Posteriormente, se procede a analizar el comportamiento de los tres modelos de imputación propuestos. Para ello, valoraremos primero el error cometido por cada uno de los modelos al imputar los valores ausentes mediante una imputación simple. Después mostraremos una medida de cómo ajusta cada modelo los datos de concentración de nitratos. Completaremos el análisis de la imputación de los valores ausentes con los resultados obtenidos al imputar estos mediante imputación múltiple. Finalmente, se ilustra mediante algunos ejemplos las particularidades del conjunto de datos con el que se está trabajando, a la vez que se analiza cómo se comporta cada modelo en estos casos.

6.1. Imputación de los valores ausentes

No debemos olvidar que, al enfrentarse con un banco de datos incompleto el objetivo final de estudiar y analizar, con la profundidad que se pretende en esta tesis, el problema de la

existencia de valores ausentes es proporcionar una estimación de estos para su uso posterior.

Como se ha comentado en el apartado 2.3, dependiendo de la utilidad del banco de datos se puede optar por realizar una imputación simple, múltiple o bayesiana. Nos decantaríamos por una imputación simple si el propósito fuera utilizar los datos para diversos estudios con propósitos diferentes. Así, imputando los datos ausentes con un único valor se completa el banco de datos y se asume totalmente observado para su utilización en posteriores análisis. Este tipo de imputación sería incluso necesaria si el proveedor de los datos y el usuario final de estos fuertan entidades diferentes. Si, por el contrario, es posible el tratamiento en profundidad de la ausencia de valores en el banco de datos con el que se va a trabajar, se puede realizar un imputación múltiple o bayesiana. Desde el punto de vista frecuentista se puede realizar una imputación múltiple asignando a cada dato ausente una muestra de m imputaciones, véase apartado 2.3.2. Desde el punto de vista bayesiano se puede incluir el modelo de imputación dentro del modelo principal y obtener resultados habiendo tenido en cuenta así la incertidumbre asociada al desconocimiento de los valores ausentes.

En nuestro caso, el principal interés radica en la comparación de los tres modelos espacio-temporales propuestos en la imputación de los valores ausentes. Continuamos con nuestra propuesta de que una adecuada forma de llevar a cabo esta comparación es a través del error de imputación que estos cometen al estimar los datos faltantes. Dicho error de imputación lo calcularemos considerando tanto una imputación simple como una imputación múltiple. Mediante la imputación simple mediremos el error que comenten los modelos cuando proporcionan una única estimación para cada valor ausente, mientras que con la imputación múltiple calcularemos el error cometido cuando proporcionan una muestra de m estimaciones para cada dato ausente. Asimismo, como los modelos de imputación proporcionan valores no sólo para los datos

ausentes sino también para los datos observados, aprovecharemos esta circunstancia para calcular una medida del ajuste de cada modelo a los datos de concentración de nitratos.

Este proceso de comparación de las modelizaciones espacio-temporales de imputación lo realizamos bajo la perspectiva bayesiana.

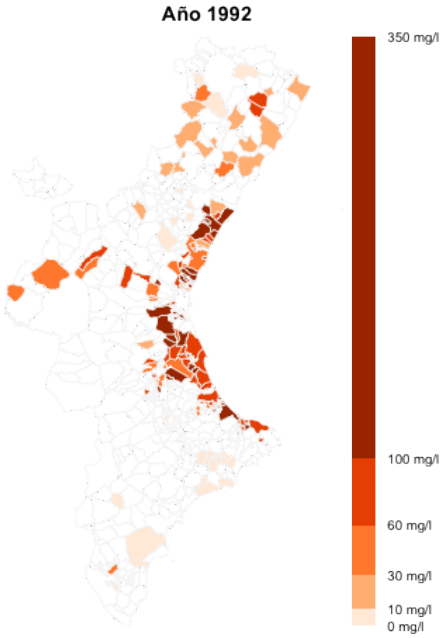
Así pues, se procede al ajuste de los modelos mediante el uso del software libre WinBUGS, ver código en el Anexo 3. Para ello lanzamos el modelo IET1 un total de 500000 iteraciones, el modelo IET2 un total de 800000 iteraciones y el modelo ARET un total de 200000 iteraciones. En todos los casos se lanzan dos cadenas, tomando una iteración de cada 5 y quedándonos con las 1000 últimas iteraciones de cada cadena. Se comprueba convergencia de los modelos mediante gráficos de las cadenas y mediante el estadístico *Rhat* (Plummer *et al.*, 2006).

De este modo, obtenemos una muestra de tamaño 400 de la distribución posterior de los diferentes parámetros desconocidos en cada modelo. En lo que respecta a los hiperparámetros, en el Anexo 1 se muestra el resumen de dichas distribuciones posteriores obtenidas tras el proceso de simulación. Para cada modelo propuesto se presenta una tabla incluyendo la descriptiva básica así como el valor del estadístico *Rhat* obtenido.

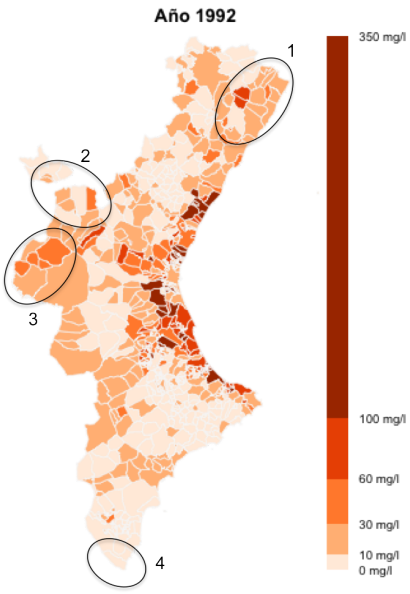
Asimismo, y como paso previo al posterior estudio detallado de la imputación, gracias a esta distribución posterior obtenida en el ajuste de los modelos, estamos en condiciones de proporcionar una imputación simple de los valores ausentes de las concentraciones de nitratos. Para realizar dicha imputación simple recurrimos a la media y a la mediana de la muestra obtenida de su distribución posterior.

Por tanto, se procede a completar el banco de datos de nitratos imputando los valores ausentes con las imputaciones obtenidas por cada uno de los modelos y, a continuación, se presenta los mapas resultantes después de la imputación de los datos faltantes. Por economía de espacio, para cada uno de los modelos y para cada

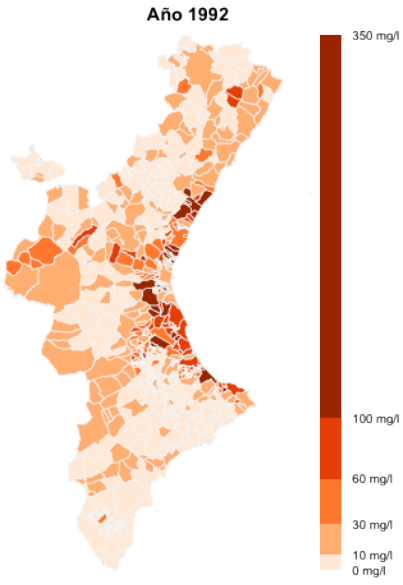
una de las imputaciones, media y mediana, se muestran a modo de ejemplo únicamente los mapas del año 1992, por ser en el que más información ausente existe. En primer lugar se muestra el mapa original en el que se observa un elevado número de municipios sin datos de concentraciones de nitratos.



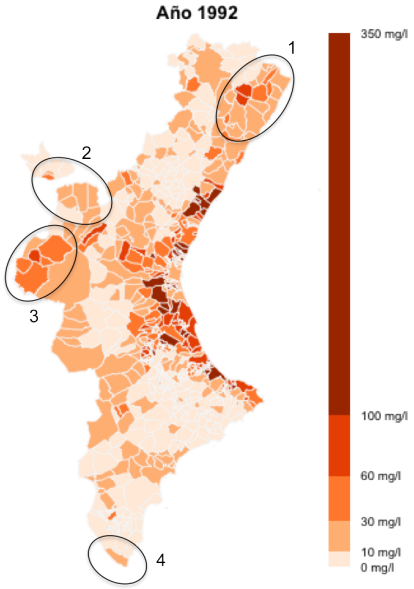
(a) Datos de concentración de nitratos



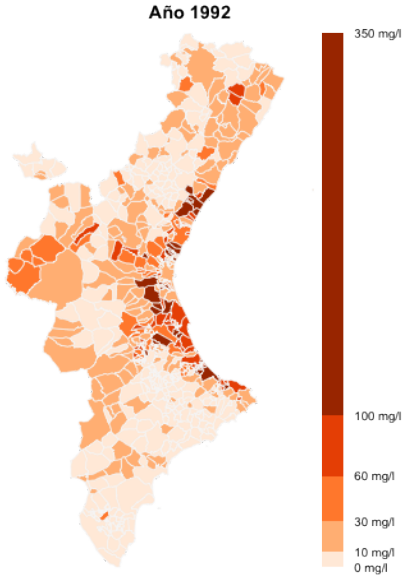
(b) Modelo IET1. Media



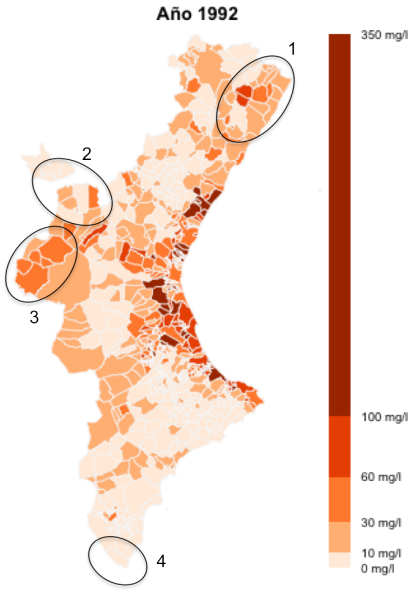
(c) Modelo IET1. Mediana



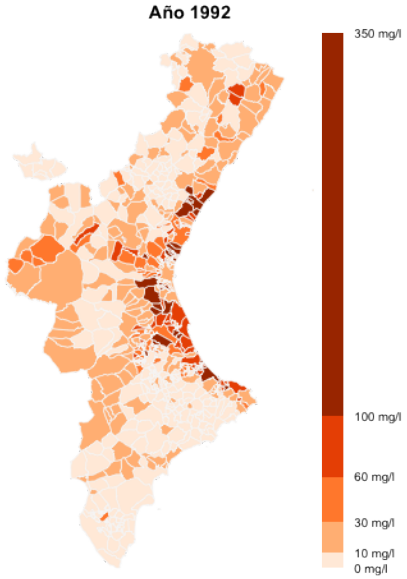
(d) Modelo IET2. Media



(e) Modelo IET2. Mediana



(f) Modelo ARET. Media



(g) Modelo ARET. Mediana

A simple vista no parecen existir grandes diferencias entre las imputaciones de concentración de nitratos proporcionadas por los diferentes modelos. No obstante, si se observa con detalle pueden apreciarse algunos aspectos de cierta relevancia.

En primer lugar, se observa que en la distribución espacial de los datos en este año mostrado 1992, en comparación con el año 1991, gráfico 3.3, la totalidad de los valores de concentración de nitratos más elevados, por encima de 100mg/l, forman parte del conjunto de valores observados. Esto podría ser debido a que en el año 1992 se decidiera observar sólo los municipios que presentaron mayores concentraciones de nitratos en 1991. Si así fuera, la suposición de que el mecanismo de aparición de valores ausentes es aleatorio, apartado 2.2, podría ser cuestionada. No obstante, al ser sólo una hipótesis, pues se desconoce el verdadero motivo de esta circunstancia, y al observar que este hecho ocurre únicamente en el año 1992, se considera que asumir aleatoriedad en la aparición de los valores ausentes es adecuado. En cualquier caso, si este hecho hubiera sido observado en el resto de años, o existiera alguna otra fuente de duda acerca de la suposición de MAR, se podría plantear un modelo de generación de valores ausentes, pues la hipótesis de ignorabilidad quedaría en entredicho. Dicho modelo se incluiría en la modelización propuesta para la imputación siguiendo el esquema visto en la figura 2.2.

En segundo lugar, si se centra la atención entre imputación con la media o con la mediana, se aprecia una diferencia clara y lógica: la mediana proporciona estimaciones de los valores ausentes de concentración de nitratos inferiores a la media. Llegando incluso a cambiar a una categoría inferior si el valor obtenido con la media está cerca del límite inferior.

Profundizando en la observación de los mapas de imputación por la media, se pueden distinguir pequeñas diferencias en las imputaciones obtenidas por cada uno de los modelos. Con este fin, y a modo de ejemplo, se destacan 4 zonas en las que se observa con claridad el comportamiento de cada uno ellos a la hora de

proporcionar imputaciones de concentración de nitratos.

En la zona 1, dos municipios que obtienen imputaciones dentro de la misma categoría por parte de los modelos IET1 y ARET, obtienen valores de una categoría superior por parte del modelo IET2. En el caso del municipio situado más al norte, San Jorge, el modelo IET1 imputa con un valor de 19.77mg/l, el modelo ARET con 21.72mg/l, ambos en categoría 10-30mg/l, mientras que el modelo IET2 imputa con un valor de 32.75mg/l que entraría en la categoría 30-60mg/l. En el municipio más al sur, Cuevas de Vinroma, los valores son 4.67mg/l, 6.86mg/l y 12.04mg/l respectivamente, es decir, se pasa de categoría 0-10mg/l a 10-30mg/l. Este hecho es, como se verá más adelante, característico del modelo IET2. Es un modelo más flexible y puede llegar a simular valores más elevados que los modelos IET1 y ARET.

En esta misma zona, se observa un municipio, situado entre los dos anteriores, Cervera del Maestre, para el que los modelo IET2 y ARET proporcionan valores de la misma categoría, 31.79mg/l y 33.67mg/l mientras que el modelo IET1 imputa un valor de categoría inferior, 29.95mg/l. En este caso las imputaciones son similares y la diferencia en el mapa de debe a que éstas están situadas en la frontera entre las categorías 10-30mg/l y 30-60mg/l.

En la zona 2, se observa que, en el municipio Alpuente, en el centro, el modelo IET2 proporciona imputaciones superiores a los modelos IET1 y ARET. En concreto se pasa de 8.32mg/l del modelo IET1 y 9.23mg/l del modelo ARET a 10.10mg/l del modelo IET2. También se observa una subida progresiva del valor de la imputación en el municipio Casas Bajas, en el sur del Rincón de Ademuz. El modelo ARET imputa con un valor casi nulo, $4{\cdot}10^{-6}$, en categoría 0-10mg/l, el modelo IET1 proporciona un valor de 26.49mg/l, de categoría 10-30mg/l y el modelo IET2 imputa con 30.68mg/l subiendo a categoría 30-60mg/l. Cabe destacar en esta zona el caso del municipio La Yesa, a la derecha de Alpuente, en el que los modelos IET1 y ARET proporcionan imputaciones superiores al modelo IET2, pasando de 32.19mg/l del modelo IET1y 33.08mg/l

del modelo ARET, ambos en categoría 30-60mg/l, a 23.81mg/l del modelo IET2, en la categoría inferior 10-30mg/l.

Respecto a la zona 3, nuevamente se observa que el modelo IET2 imputa valores de concentraciones de nitratos superiores a los modelos IET1 y ARET. Así, en el municipio Fuenterrobles, en el centro de la zona, el modelo IET1 imputa un valor de 57.69mg/l y el modelo ARET de 58.84mg/l, ambos en categoría 30-60mg/l mientras que el modelo IET2 imputa 64,85mg/l, de categoría 60-100mg/l. Para el municipio Venta del Moro, en el sur de la zona, el modelo IET1 proporciona una imputación de 29.15mg/l, categoría 10-30mg/l, frente a 31.47mg/l del modelo ARET y 38.76mg/l del IET2, ambos de categoría 30-60mg/l.

En la zona 4, municipio de Pilar de la Horadada, el modelo IET1 imputa un valor de 7.72mg/l, el modelo ARET de 8.56mg/l, ambos en categoría 0-10mg/l y el modelo IET2 vuelve a imputar por encima proporcionando un valor de concentración de nitratos de 10.29mg/l, categoría 10-30mg/l.

Así, mediante estos ejemplos se entiende que el modelo IET2 suele simular valores de concentración de nitratos superiores a los modelos IET1 y ARET. Esto es debido a la capacidad del modelo de simular valores más extremos. Más adelante, en el apartado 6.5, se profundiza en el comportamiento de cada modelo.

6.2. Cálculo del error de imputación

Como se ha comentado en el apartado 4.2, la forma usual de comparar diferentes modelizaciones es mediante el uso de criterios que tienen en cuenta la complejidad de cada modelo así como la calidad de ajuste que proporcionan. Desde el punto de vista bayesiano, el DIC es un claro ejemplo de este tipo de criterios. Sin embargo, también se ha apuntado que este criterio no es el más adecuado cuando se trabaja con valores ausentes.

Así, de nuevo, consideramos que la mejor estrategia para poder comparar las modelizaciones propuestas pasa por cuantificar el

error de imputación que comete cada una de ellas al proporcionar estimaciones de los valores ausentes.

Lamentablemente, es imposible valorar el error cometido al estimar un dato faltante pues no se tiene la observación real que se está estimando. Debido a esto, se procede a intervenir sobre los datos con el fin de posibilitar la obtención de una medida del error de imputación cometido con los modelos propuestos.

Nuevamente, se decide ampliar el número de valores ausentes. Sin embargo, a diferencia del apartado 4.2, en este caso diseñamos una estrategia de creación de valores ausentes más exhaustiva. Originalmente existen 4064 datos observados y 1336 datos ausentes. El subconjunto de valores observados lo dividimos en tres partes, dos de 1355 datos y una tercera de 1354. El proceso que se sigue es, de los 4064 valores observados se toma una muestra aleatoria de tamaño 1355, estos valores observados se convierten en ausentes, formando así un conjunto de datos con 2661 valores ausentes (1336 originales más 1355 creados por nosotros y de los cuales tenemos el verdadero valor) y 2709 observados. A este nuevo conjunto lo llamamos partición 1.

Seguidamente procedemos a extraer otra muestra aleatoria de tamaño 1355 de los 4064 datos observados originales, con la condición de que los seleccionados anteriormente para la partición 1 no pueden volver a ser seleccionados. Convertimos de nuevo en ausentes los valores observados seleccionados, creando un nuevo conjunto, en lo sucesivo partición 2, con 2661 datos faltantes (1336 originales más 1355 nuevos creados por nosotros y de los cuales tenemos el verdadero valor) y 2709 observados.

Por último, los 1354 valores originalmente observados que no han sido seleccionados ni en la partición 1 ni en la partición 2 se convierten en ausentes, formando la partición 3 que tendrá 2690 valores ausentes (los 1336 originales más 1354 creados y de los cuales tenemos el verdadero valor) y 2710 observados.

Así, se crean 3 particiones que comparten los valores ausentes originales pero que contienen un subconjunto de datos ausentes,

diferente en cada una, y de los cuales conocemos su verdadero valor. De esta forma es posible calcular una medida del error cometido por cada uno de los modelos propuestos al imputar los valores ausentes de cada una de las particiones de los que sí conocemos su verdadero valor.

Previamente al cálculo del error de imputación en cada una de las particiones creadas y para cada uno de los modelos propuestos, es conveniente plantear qué tipo de cálculo de medida de error sería el más conveniente. En el apartado 4.2 hemos utilizado únicamente el error cuadrático, pues el objetivo básico del análisis planteado en el capítulo 4 era discernir la estructura base del modelo de imputación. No obstante, habiendo asumido la necesidad de plantear un modelo espacio-temporal para la imputación de las concentraciones de nitratos ausentes, es posible que, dado que la estructura de las modelizaciones es similar, calcular el error de una forma u otra pueda derivar en diferentes conclusiones.

Así pues, en este caso, con el objetivo de evitar cualquier tipo de confusión asociada a la idoneidad o no de un método de cálculo de error sobre nuestros datos y poder así garantizar la mayor objetividad posible, planteamos tres métodos diferentes para la obtención de la medida del error de imputación.

El planteamiento de cada uno de ellos obedece a las diferentes particularidades de nuestros datos así como de las de los modelos de imputación.

En primer lugar se considera el error cuadrático planteado en el apartado 4.2,

$$\sum_{Observados}(ValorImputado - ValorObservado)^2$$

no obstante, dicho cálculo penalizaría en exceso las imputaciones alejadas de los valores altos de nuestros datos. Es decir, como sabemos que existen concentraciones de nitratos elevadas, más de un 4 % superiores a 100, si algún modelo de imputación proporciona estimaciones alejadas de dichas concentraciones, el error cometido

se vería agrandado por la magnitud que se está estimando y el hecho de elevar al cuadrado la diferencia.

Esta circunstancia no sería indicativa de que el modelo no imputa adecuadamente, pues podría suceder que en este pequeño porcentaje de valores elevados la influencia de las observaciones vecinas tanto en el tiempo como en el espacio indicaran que la concentración de nitratos no debería ser tan alta, mientras que la realidad observada no obedeciera a dicho patrón. Para el resto de concentraciones el modelo podría esta funcionando correctamente y sin embargo su error de imputación sería elevado por culpa de esta circunstancia.

Este hecho motiva que también se plantee calcular el error cuadrático relativo,

$$\sum_{Observados} \frac{(ValorImputado - ValorObservado)^2}{ValorObservado}$$

que al estar dividida la diferencia cuadrática por el valor que se está imputando, el error cometido por los diferentes modelos ya no se ve influenciado por la magnitud del valor que estiman.

Además, de la misma forma que la asimetría de las concentraciones de nitratos nos lleva a plantear una verosimilitud LogNormal, este hecho nos lleva a plantear un tercer método para el cálculo del error de imputación,

$$\sum_{Observados} (log(ValorImputado) - log(ValorObservado))^2$$

así, al calcular la diferencia cuadrática entre el logarítmo del valor observado y el logarítmo del valor estimado, en lugar de hacerlo para los valores sin transformar, la asimetría de las concentraciones de nitratos no influirá en el cálculo del error de imputación, pues al tomar logaritmos la magnitud de los valores se reduce.

6.3. Resultados obtenidos

Se ajusta cada uno de los tres modelos espacio-temporales propuestos de la misma forma que en al apartado 6.1. Esta vez, el

ajuste se lleva a cabo con cada una de las tres particiones descritas en la sección 6.2. Para cada una de ellas nos quedamos con la media y la mediana de la muestra obtenida de la distribución final de los valores ausentes. Finalmente calculamos, para cada partición y para cada modelo, el error de imputación cometido haciendo uso de los tres métodos de cálculo de error presentados en el apartado 6.2.

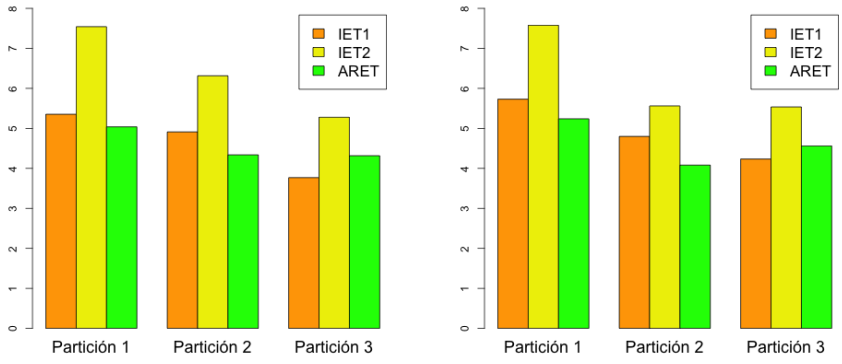
En las tablas 6.1, 6.2 y 6.3, acompañadas de sus respectivas figuras, se muestran los valores obtenidos. Estos resultados muestran que en lineas generales el modelo ARET es el que mejores imputaciones proporciona. Se observa que para la Partición 3 el modelo IET1 comete un error de imputación inferior al cometido por el modelo ARET, aunque si el cálculo del error está basado en el método logarítmico no se observa este hecho.

El modelo IET2 comete los mayores errores de imputación con diferencias respecto a los modelos IET1 y ARET elevadas en algunos casos.

Así pues, a la vista de los resultados obtenidos, el modelo ARET proporciona las mejores imputaciones de los valores ausentes de las concentraciones de nitratos. El modelo IET1 realiza imputaciones ligeramente peores pues no existen grandes diferencias entre los errores obtenidos por este modelo y los obtenidos por el modelo ARET. El modelo IET2 es, ante la evidencia de los errores de imputación obtenidos, el modelo que peor imputaría los datos faltantes de la concentración de nitratos.

	Modelo	Media	Mediana
Partición 1	IET1	5.35	5.73
	IET2	7.54	7.58
	ARET	5.04	5.24
Partición 2	IET1	4.91	4.80
	IET2	6.32	5.56
	ARET	4.34	4.08
Partición 3	IET1	3.77	4.23
	IET2	5.28	5.54
	ARET	4.32	4.56

Tabla 6.1: Error cuadrático cometido en la imputación (multiplicado por 10^{-5}) por los diferentes modelos.

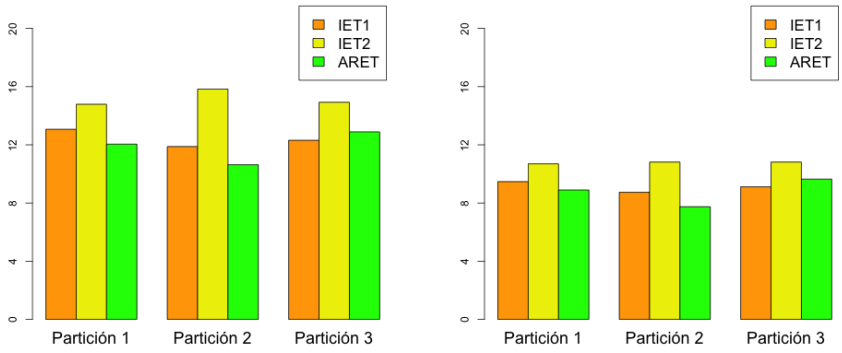


(h) Imputación con la media (i) Imputación con la mediana

Figura 6.1: Error cuadrático de imputación cometido por cada uno de los modelos propuestos y multiplicado por 10^{-5}.

	Modelo	Media	Mediana
Partición 1	IET1	13.07	9.47
	IET2	14.79	10.69
	ARET	12.05	8.90
Partición 2	IET1	11.88	8.74
	IET2	15.83	10.82
	ARET	10.63	7.74
Partición 3	IET1	12.31	9.11
	IET2	14.93	10.82
	ARET	12.89	9.64

Tabla 6.2: Error relativo cometido en la imputación (multiplicado por 10^{-3}) por los diferentes modelos.

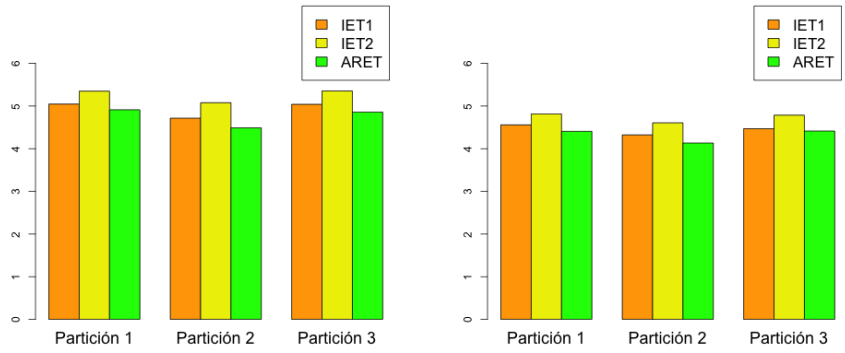


(a) Imputación con la media (b) Imputación con la mediana

Figura 6.2: Error relativo de imputación cometido por cada uno de los modelos propuestos y multiplicado por 10^{-3}.

	Modelo	Media	Mediana
Partición 1	IET1	5.05	4.56
	IET2	5.35	4.82
	ARET	4.91	4.41
Partición 2	IET1	4.72	4.32
	IET2	5.08	4.61
	ARET	4.49	4.13
Partición 3	IET1	5.04	4.47
	IET2	5.35	4.79
	ARET	4.86	4.41

Tabla 6.3: Error logarítmico cometido en la imputación (multiplicado por 10^{-2}) por los diferentes modelos.



(a) Imputación con la media (b) Imputación con la mediana

Figura 6.3: Error logarítmico de imputación cometido por cada uno de los modelos propuestos y multiplicado por 10^{-2}.

6.4. Ajuste de los modelos propuestos

Además del evidente interés por comparar los modelos entre sí a partir del error de imputación cometido, sería interesante también profundizar en la comparación viendo cómo ajusta cada uno de ellos la realidad espacio-temporal existente en los datos de concentración de nitratos.

Con este fin se calcula, para cada uno de los 2709 valores observados en las particiones 1 y 2 y para los 2710 de la partición 3, la diferencia entre la estimación proporcionada por los modelos y el valor real. Esta diferencia se calcula mediante los tres tipos de errores propuestos en el apartado 6.2.

En las tablas 6.4, 6.5 y 6.6, acompañadas de sus respectivas figuras, se presentan los resultados obtenidos. En ellos se observa que el modelo que mejor capta la naturaleza espacio-temporal presente en las concentraciones de nitratos es el modelo IET2. Las diferencias en esta medida de ajuste entre este modelo y los modelos IET1 y ARET son elevadas. Esto, unido al hecho de que es el modelo que peor imputaciones proporciona, induce a pensar que el modelo IET2 es lo suficientemente flexible como para estimar valores de concentración de nitratos extremos. Sin embargo, esta misma flexibilidad provocaría que en aquellos casos en los que se presentan situaciones atípicas, sobretodo en cuanto a la evolución temporal de las concentraciones de nitratos en un mismo municipio, el modelo comete errores elevados.

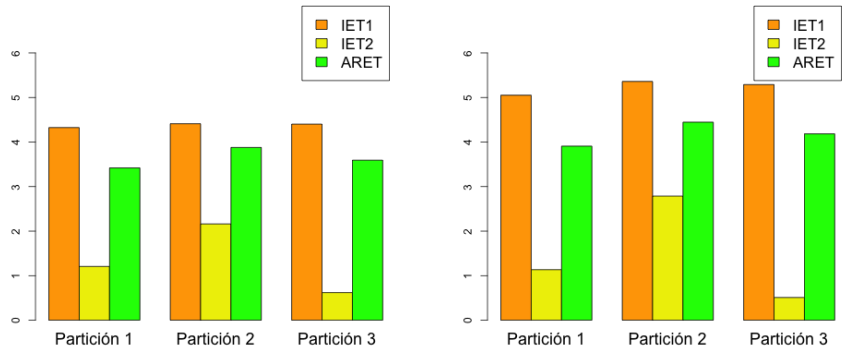
Más adelante mostraremos con algunos ejemplos la situación que se apunta aquí a la vista de los resultados obtenidos.

El modelo IET1 es el que peor ajuste de los datos realiza, mientras que el modelo ARET finalmente resulta el modelo que mejor se adecuaría a la imputación de las concetraciones de nitratos. Es menos flexible que el modelo IET2 pero ajusta mejor que el modelo IET1, además comete menor error de imputación que ambos, modelo IET1 y IET2. Así, el modelo ARET es el que mejor conjuga el compromiso entre un adecuado ajuste y una imputación precisa de la información faltante, siendo, este último aspecto, el

de mayor interés y el objetivo final de este trabajo.

	Modelo	Media	Mediana
Partición 1	IET1	4.32	5.05
	IET2	1.21	1.13
	ARET	3.42	3.91
Partición 2	IET1	4.41	5.36
	IET2	2.16	2.79
	ARET	3.88	4.45
Partición 3	IET1	4.40	5.29
	IET2	0.62	0.51
	ARET	3.59	4.18

Tabla 6.4: Medida del ajuste de los diferentes modelos (multiplicado por 10^{-5}) basada en el error cuadrático.

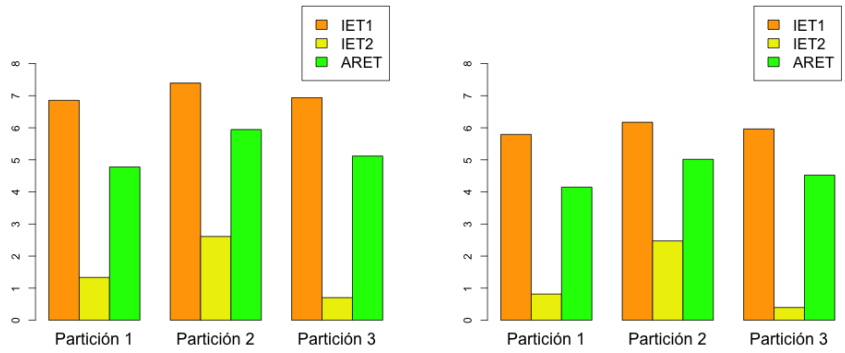


(a) Imputación con la media (b) Imputación con la mediana

Figura 6.4: Medida del ajuste proporcionado por cada uno de los modelos propuestos, calculado mediante el error cuadrático y multiplicado por 10^{-5}.

	Modelo	Media	Mediana
Partición 1	IET1	6.86	5.79
	IET2	1.33	0.82
	ARET	4.78	4.15
Partición 2	IET1	7.40	6.17
	IET2	2.61	2.47
	ARET	5.94	5.02
Partición 3	IET1	6.94	5.96
	IET2	0.70	0.40
	ARET	5.12	4.52

Tabla 6.5: Medida del ajuste de los diferentes modelos (multiplicado por 10^{-3}) basada en el error relativo.

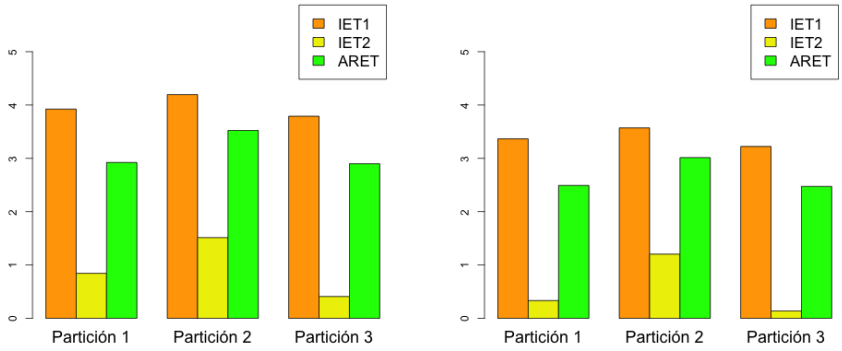


(a) Imputación con la media (b) Imputación con la mediana

Figura 6.5: Medida del ajuste proporcionado por cada uno de los modelos propuestos, calculado mediante el error relativo y multiplicado por 10^{-3}.

	Modelo	Media	Mediana
Partición 1	IET1	3.92	3.37
	IET2	0.84	0.33
	ARET	2.92	2.49
Partición 2	IET1	4.20	3.57
	IET2	1.51	1.20
	ARET	3.52	3.01
Partición 3	IET1	3.79	3.22
	IET2	0.41	0.14
	ARET	2.90	2.47

Tabla 6.6: Medida del ajuste de los diferentes modelos (multiplicado por 10^{-2}) basada en el error logarítmico.



(a) Imputación con la media (b) Imputación con la mediana

Figura 6.6: Medida del ajuste proporcionado por cada uno de los modelos propuestos, calculado mediante el error logarítmico y multiplicado por 10^{-2}.

6.5. Comportamiento de los modelos

A la vista de los resultados vistos en cuanto a error cometido en la imputación, sección 6.3, y ajuste de los datos por parte de los modelos, sección 6.4, parece lógico preguntarse cómo el modelo IET2, que proporciona medidas de ajuste mucho mejores que los modelos IET1 y ARET, realiza las peores imputaciones.

Las figuras 6.7, 6.8 y 6.9 nos ayudan a entender mejor el comportamiento de cada modelo. En ellas se muestra, a modo de ejemplo, la evolución temporal de las concentraciones de nitratos de 6 de los 540 municipios. Sarratella y Estivella se muestran en la figura 6.7, Puçol y Albuixech en la figura 6.8 y Villargordo del Cabriel y Beniparrell en la figura 6.9. Estos municipios han sido seleccionados debido a que, o bien presentan una gran variabilidad en su evolución temporal, o bien su evolución temporal es atípica, o bien presenta ambas circunstancias.

En general, la variabilidad temporal intramunicipal es elevada. Como se observa en la tabla 6.7, casi un 25 % de los municipios presenta una desviación típica superior a 10. Cerca de un 4 % supera una desviación típica de 50, llegando a los valores extremos 87.06, 158.16 (máximo) y 107.75 que presentan respectivamente tres de los municipios seleccionados, Abuixech, Villargordo del Cabriel y Beniparrell.

Si nos fijamos en evoluciones temporales atípicas, con importantes oscilaciones de un año a otro o con incrementos/descensos exageradamente elevados en el periodo a estudio, podríamos decir que todos los municipios seleccionados presentan alguna de éstas características. En los municipios Sarratella, Estivella y Beniparrell, se aprecian oscilaciones en el tiempo imprevisibles para cualquier modelo, mientras que en los municipios Puçol, Albuixech, Villargordo del Cabriel e incluso Beniparrell se observa un descenso muy elevado en su evolución temporal, de más de 200mg/l a menos de 10mg/l en menos de 10 años.

En cada gráfico se puede ver, en negro los valores observados

Mínimo	0.00
1er cuartil	1.58
Mediana	3.76
Media	9.60
3er cuartil	9.58
Máximo	158.20
Ausentes	2

Tabla 6.7: Descriptiva de las desviaciones típicas en la evolución temporal de las concetraciones de nitratos de los 540 municipios.

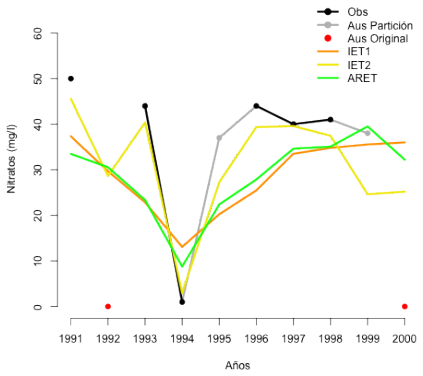
de concentración de nitratos. Con un punto rojo se indica los valores originalmente ausentes. En gris se muestran aquellos valores que, siendo datos observados inicialmente, han sido convertidos en valores ausentes en la partición correspondiente. En colores las imputaciones de todos los valores del municipio proporcionadas por cada modelo, naranja el modelo IET1, amarillo el modelo IET2 y verde el modelo ARET. Por economía de espacio se muestra únicamente la imputación haciendo uso de la media.

Así, se observa que el modelo IET2 muestra un ajuste casi perfecto. La imputación que dicho modelo proporciona, en amarillo, va casi ligada a las observaciones, en negro, en todos los municipios, a pesar de las grandes variaciones que se presentan en cada uno de ellos. No obstante, si la concentración de nitratos está ausente, en gris, el modelo IET2 siempre comete errores de imputación elevados en comparación con los modelos IET1 y ARET en todos los casos. Esto se puede observar en la figura 6.7, en el año 1999 en Sarratella, gráfico (a) y en el año 1991 en Estivella, gráfico (b). También en la figura 6.8, en 1992 y 1993 en Puçol, gráfico (a) y en los años 1995 y 1996 en Albuixech , gráfico (d). En la figura 6.9, se aprecia en Villargordo del Cabriel, gráfico (a), de 1991 a 1993 y en los años 1991 y 2000 en Beniparrell, gráfico (b).

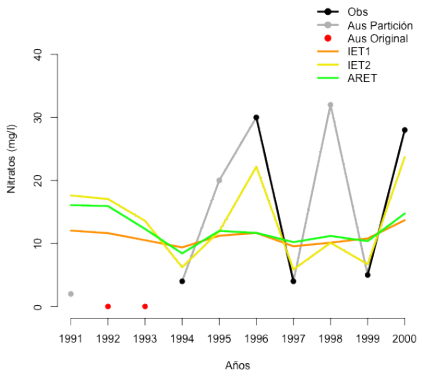
Este hecho puede interpretarse a partir de la estructura del modelo IET2. La definición del modelo implica que, ante un valor

ausente en un municipio y año determinado, los valores de los municipios contiguos en el espacio para el año en curso, el anterior y el posterior, y los valores propios del municipio el año anterior y posterior, adquieren vital importancia a la hora de estimar dicha concentración de nitratos faltante. Por tanto, la correlación espacial es fundamental para el modelo IET2 a la hora de proporcionar una imputación. Este hecho puede ser perjudicial si la realidad apunta a que el peso de la correlación temporal es mayor que el de la correlación espacial. Sin embargo, para el modelo ARET, a la hora de imputar, la correlación temporal resulta ser más importante, tanto mayor cuanto más se avanza en el tiempo, que la espacial. De ahí que sus errores de imputación sean menores que los del modelo IET2 en casos similares a los mostrados.

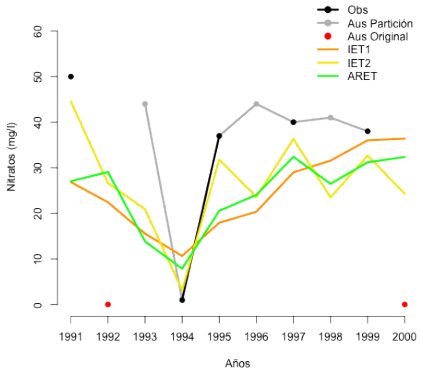
Asimismo, se aprecia en los ejemplos la flexibilidad del modelo IET2. Mientras que las evoluciones en las imputaciones proporcionadas por los modelo IET1 y ARET son suaves en el sentido de carecer de cambios bruscos en la estimación de los valores de un año a otro, se observa que el modelo IET2 sí proporciona variaciones bruscas en sus imputaciones. Así, el modelo IET1 y sobretodo el modelo ARET, realizan mejores imputaciones en los casos de existir un valor ausente entre dos años con valores observados, mientras que el modelo IET2 comete un mayor error. Esto se puede apreciar en la figura 6.7 en 1994 del municipio Sarratella, gráfico(e), y en el mismo año en la figura 6.8 en Albuixech, gráfico (f).



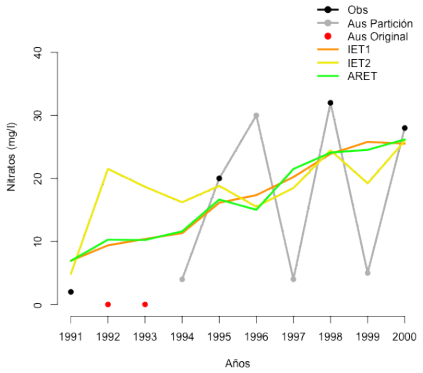
(a) Partición 1. Sarratella



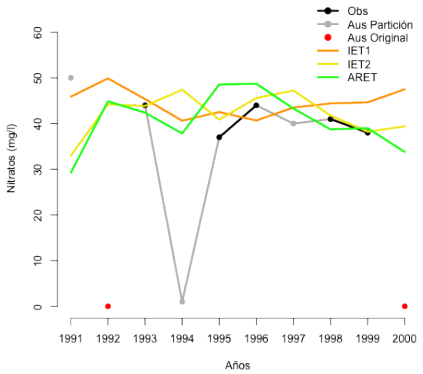
(b) Partición 1. Estivella



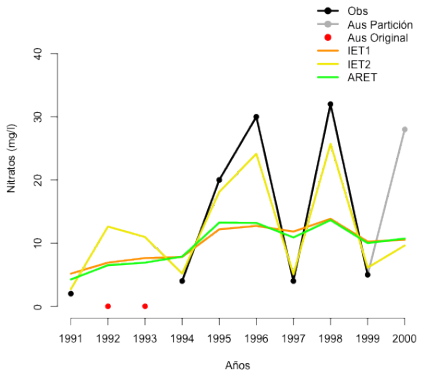
(c) Partición 2. Sarratella



(d) Partición 2. Estivella

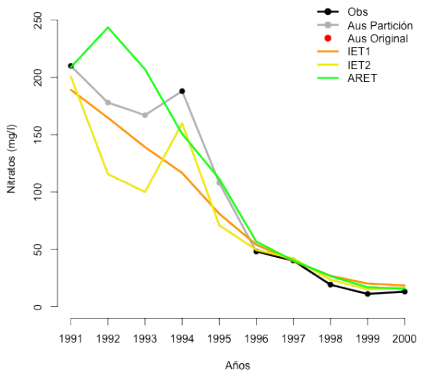


(e) Partición 3. Sarratella

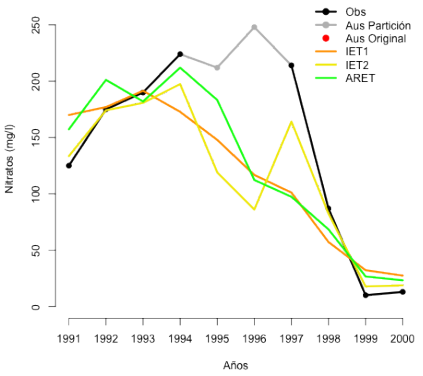


(f) Partición 3. Estivella

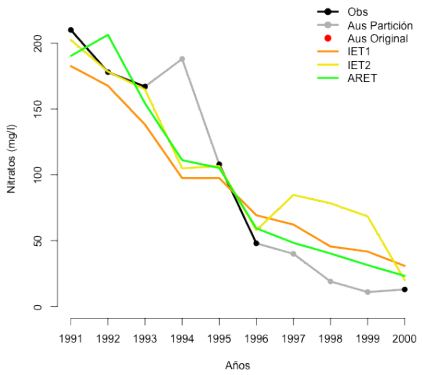
Figura 6.7: Evolución temporal de las concentraciones de nitratos en Sarratella y Estivella



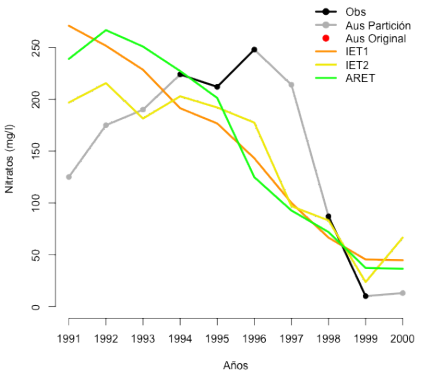
(a) Partición 1. Puçol



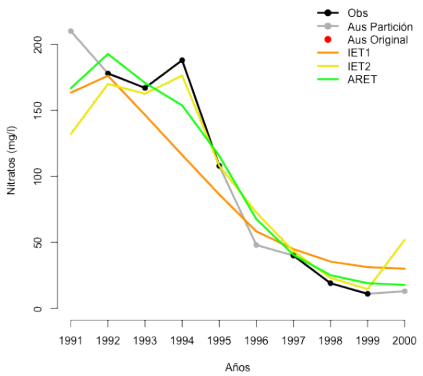
(b) Partición 1. Albuixech



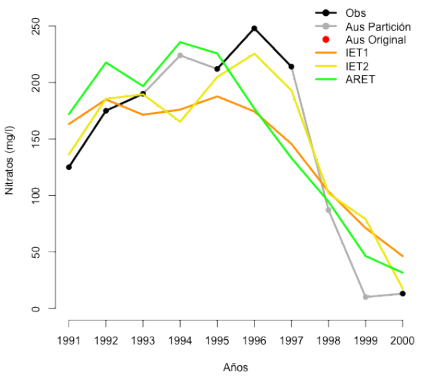
(c) Partición 2. Puçol



(d) Partición 2. Albuixech

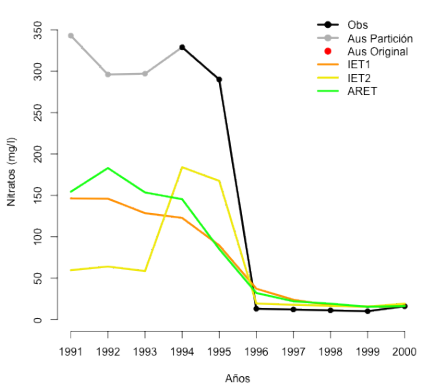


(e) Partición 3. Puçol

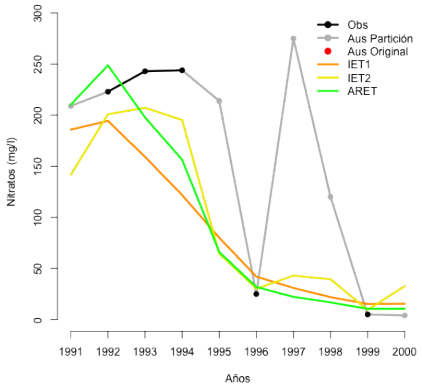


(f) Partición 3. Albuixech

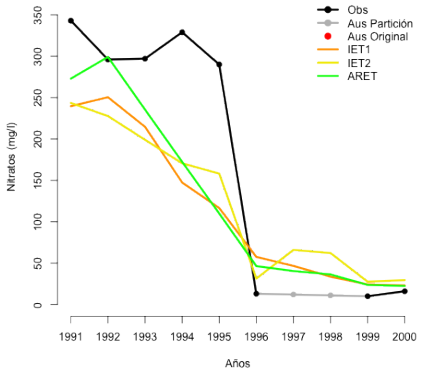
Figura 6.8: Evolución temporal de las concentraciones de nitratos en Puçol y Albuixech



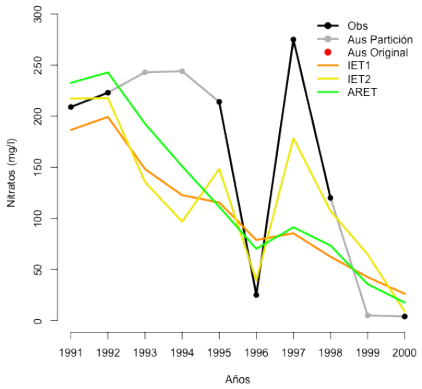
(a) Partición 1. Villargordo del Cabriel



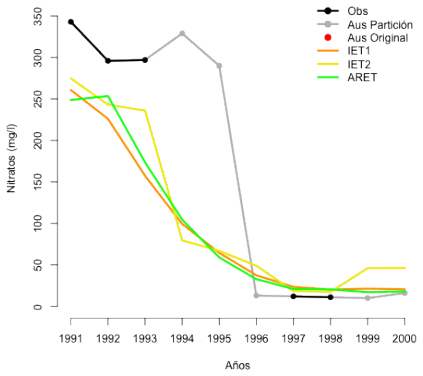
(b) Partición 1. Beniparrell



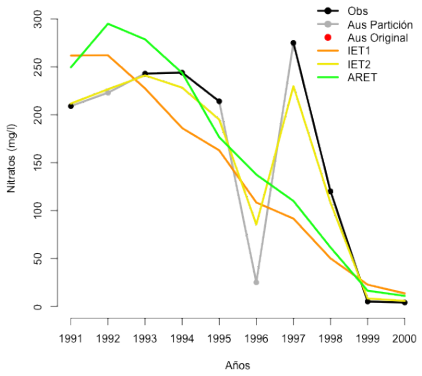
(c) Partición 2. Villargordo del Cabriel



(d) Partición 2. Beniparrell



(e) Partición 3. Villargordo del Cabriel



(f) Partición 3. Beniparrell

Figura 6.9: Evolución temporal de las concentraciones de nitratos en Villargordo del Cabriel y Beniparrell

Cabría decir que somos conscientes que este ejemplo del comportamiento de los modelos es una simplificación de su funcionamiento. Se muestra únicamente la evolución temporal de cada municipio y sus imputaciones, sin tener en cuenta la correlación entre observaciones vecinas en el espacio y en el tiempo ni cómo gestiona dichas correlaciones cada uno de los modelos.

Aun así, estos ejemplos mostrados a modo de simplificación del funcionamiento de los modelos de imputación, ayudan a entender la realidad con la que estamos trabajando. Además, aportan luz en la comprensión de los valores obtenidos en cuanto a error de imputación y ajuste de los datos para cada una de las modelizaciones propuestas.

6.6. Distribución posterior del error de imputación

Complementariamente al estudio del comportamiento de cada modelo, se procede a calcular en cada una de las iteraciones de la simulación el error de imputación cometido. Para ello, se introduce en el código WinBUGS cada una de las tres expresiones de cálculo del error vistas en el apartado 6.2. Así, cada vez que el modelo simula un valor nuevo de concentración de nitratos, inmediatamente se tiene, en esa misma iteración, la diferencia entre dicho valor simulado y el verdadero valor. Por tanto, para cada una de las tres particiones, al finalizar la simulación de cada modelo se dispone de una muestra de la distribución posterior del error de imputación cometido por este.

Mediante este procedimiento, por un lado, se profundiza en cómo estima cada uno de los modelos los valores ausentes de concentración de nitratos. El interés radica, no tanto en los valores medios de los errores cometidos, sino en la variabilidad de estos. Una variabilidad elevada sería indicativa de que el modelo recorre un amplio rango de valores a la hora de imputar lo que, a su vez,

provocaría un amplio rango de errores cometidos. Y un amplio rango de valores simulados por un modelo sería indicativo de que el modelo no está siendo todo lo preciso que nos gustaría a la hora de imputar los valores ausentes. Por el contrario, una menor variabilidad estaría indicándonos que el modelo está simulando valores más cercanos al verdadero valor, tendría así menos dudas a la hora de imputar y por tanto su imputación sería más precisa.

Por otro lado, este procedimiento ayuda también a valorar futuras inferencias si se planteara un proceso de imputación múltiple. Es decir, si el objetivo final fuera obtener una muestra de tamaño m para cada uno de los valores ausentes. Sabríamos que, aquel modelo o modelos con mayor variabilidad en el error de imputación, inducirían mayor variabilidad en la inferencia de los parámetros de interés. Al existir menos precisión en la imputación de los valores ausentes, el rango de las estimaciones de futuros parámetros de interés sería mayor.

En la misma línea, este procedimiento resulta también útil si se piensa en un proceso de imputación completamente bayesiana. Es decir, incluir el modelo de imputación dentro del modelo global mediante el cual se ajusta un determinado parámetro de interés. Esto implicaría añadir a la incertidumbre asociada a la propia existencia de los valores ausentes, el exceso de variabilidad en las imputaciones que proporciona un modelo frente a otro. Así, al tener dicha variabilidad extra cuantificada, se optaría por el modelo cuya distribución posterior de errores de imputación fuera menos variable.

En las tablas 6.8, 6.9 y 6.10 con sus respectivos gráficos pueden verse los resultados obtenidos.

Puede apreciarse que el modelo IET2 es claramente el que peor comportamiento muestra a la hora de proporcionar imputaciones. Estas resultan ser muy poco precisas pues existe una elevada variabilidad en los errores de imputación. La variabilidad de más que muestra el modelo IET2 respecto al modelo IET1 y al modelo ARET, sería claramente perjudicial en un proceso de imputación

múltiple o completamente bayesiano, pues habría un exceso de incertidumbre añadida.

Este comportamiento del modelo IET2 va en la línea de lo ya observado en los apartados anteriores. Su mayor flexibilidad, aunque le permite ajustar correctamente los valores observados, le lleva a cometer errores de imputación elevados. Lo mismo sucede con su mayor dependencia de la correlación espacial. Aunque en teoría pueda ser una idea adecuada en el tratamiento de datos espacio-temporales, también le induce a proporcionar peores imputaciones en aquellos casos en los que la correlación temporal es más fuerte que la espacial.

Por tanto, a la vista de los resultados puede decirse que el modelo IET2 no resulta ser un método adecuado para la imputación de las concentraciones de nitratos de la Comunidad Valenciana.

Asimismo, se puede afirmar que el modelo ARET sería el más adecuado para este fin. En general es el modelo que menor error de imputación comete. Como los otros dos modelos, aúna en su definición la correlación espacial y la correlación temporal pero resulta ser el más estable ajustando adecuadamente la tendencia temporal de cada municipio. Además resulta ser más preciso en el proceso de imputación, lo que implica que proporciona un rango de imputaciones cercano al valor real que se quiere estimar. Este hecho nos permite asegurar también que el modelo ARET sería el modelo más adecuado para una imputación múltiple o una imputación completamente bayesiana.

	Modelo	Media	Mediana	2.5 %	97.5 %
Partición 1	IET1	1.74	1.66	1.24	2.74
	IET2	2.21	2.02	1.50	3.88
	ARET	1.68	1.60	1.18	2.67
Partición 2	IET1	1.57	1.49	1.11	2.49
	IET2	2.35	1.98	1.35	5.45
	ARET	1.67	1.51	1.07	3.22
Partición 3	IET1	1.54	1.47	1.10	2.41
	IET2	2.24	1.87	1.32	5.20
	ARET	1.58	1.49	1.09	2.67

Tabla 6.8: Error cuadrático de imputación (multiplicado por 10^{-6}) cometido por los modelos propuestos.

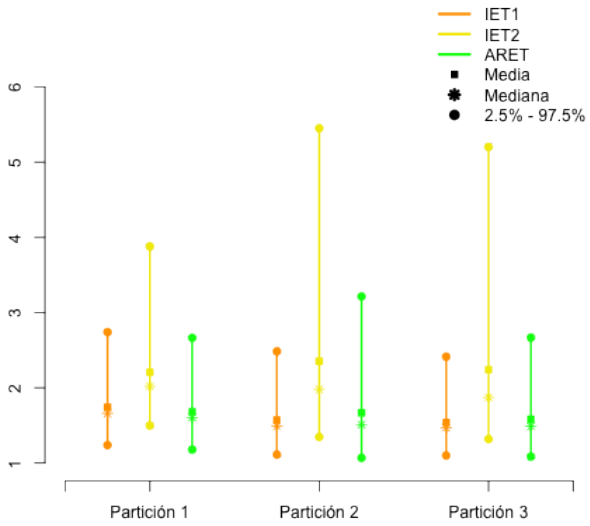


Figura 6.10: Error cuadrático de imputación (multiplicado por 10^{-6}) cometido por los modelos propuestos.

	Modelo	Media	Mediana	2.5 %	97.5 %
Partición 1	IET1	3.64	3.40	2.56	6.20
	IET2	4.12	3.92	2.96	6.39
	ARET	3.37	3.21	2.41	5.30
Partición 2	IET1	3.42	3.31	2.55	4.86
	IET2	4.73	4.33	3.09	8.56
	ARET	3.26	3.12	2.36	4.95
Partición 3	IET1	3.50	3.36	2.55	5.29
	IET2	4.36	4.01	2.91	7.65
	ARET	3.37	3.23	2.40	5.12

Tabla 6.9: Error relativo de imputación (multiplicado por 10^{-4}) cometido por los modelos propuestos.

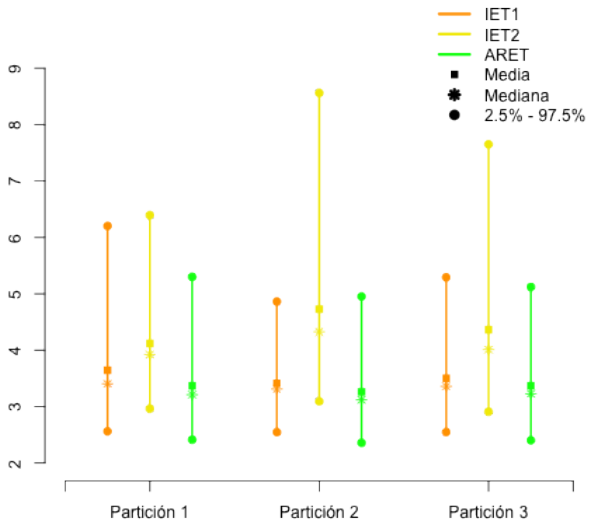


Figura 6.11: Error relativo de imputación (multiplicado por 10^{-4}) cometido por los modelos propuestos.

	Modelo	Media	Mediana	2.5 %	97.5 %
Partición 1	IET1	9.44	9.43	8.72	10.19
	IET2	9.94	9.93	9.19	10.72
	ARET	9.06	9.05	8.34	9.81
Partición 2	IET1	9.15	9.15	8.46	9.87
	IET2	9.73	9.72	9.00	10.48
	ARET	8.77	8.77	8.10	9.47
Partición 3	IET1	9.36	9.35	8.63	10.15
	IET2	9.92	9.91	9.16	10.75
	ARET	9.00	9.00	8.30	9.76

Tabla 6.10: Error logarítmico de imputación (multiplicado por 10^{-2}) cometido por los modelos propuestos.

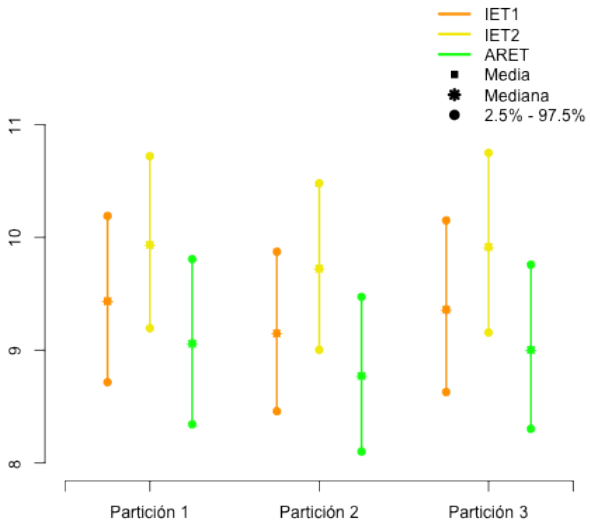


Figura 6.12: Error logarítmico de imputación (multiplicado por 10^{-2}) cometido por los modelos propuestos.

Cabe recordar que los resultados mostrados han sido obtenidos al aplicar sobre los datos de concetraciones de nitratos las modelizaciones propuestas. No obstante, sería conveniente exponer estos modelos espacio-temporales de imputación a estudio sobre otros bancos de datos con el fin de observar si mantienen el comportamiento observado. De este modo se afianzarían las conclusiones a las que se ha llegado a la vista de los resultados presentados en este capítulo.

Este es el objetivo del siguiente capítulo. Comparar de nuevo el comportamiento de los modelos IET1, IET2 y ARET al imputar bancos de datos espacio-temporales distintos creados mediante un proceso de simulación.

Capítulo 7

Estudio simulado

Los resultados mostrados en el capítulo anterior han sido obtenidos gracias a la creación de las tres particiones a partir de los datos originales de concentración de nitratos. Sin este proceso no hubiera sido posible estudiar el ajuste e imputación de cada uno de los modelos propuestos. Tampoco hubiera sido posible comparar entre sí las modelizaciones y sus comportamientos a la hora de imputar los valores ausentes.

No obstante, la creación de las particiones supone, por un lado el aumento del conjunto de valores ausentes, y, por otro, la pérdida de información al prescindir de valores observados. Ambas circunstancias implican una mayor incertidumbre a la hora de proporcionar imputaciones de los valores ausentes de concentración de nitratos.

Además, una completa comparación de modelos espacio-temporales de imputación implicaría estudiar su comportamiento frente a diferentes y diversos bancos de datos.

Por todo esto, en este capítulo se plantea la simulación de un conjunto de datos espacio-temporal de estructura similar a la del conjunto de concentraciones de nitratos con el que se ha trabajado hasta ahora. Serán matrices de 540 filas y 10 columnas simulando la estructura de municipios y años de las concentraciones de nitratos. De este modo, tendremos conjuntos de datos con

correlación espacio-temporal y completamente observados.

Concretamente se crean tres conjuntos de datos espacio-temporales, uno por cada modelo analizado en este trabajo. Una vez creados los bancos de datos, se crean valores ausentes eliminando de forma aleatoria un subconjunto de observaciones, respetando la proporción existente en el banco de datos de concentraciones de nitratos.

Este estudio simulado no pretende ser un exhaustivo proceso de simulación que nos permita decantarnos por alguna de las tres modelizaciones comparadas como modelización de referencia. De hecho, consideramos que no existe ningún modelo capaz de abordar futuros problemas de información faltante en bancos de datos con estructura espacio-temporal de una forma incuestionablemente mejor que cualquier otro. Más bien al contrario, el espíritu de esta tesis lleva implícito que plantearse una comparativa de modelizaciones previa con el ánimo de esclarecer qué estructura de modelización es la más adecuada para los datos con los que se está trabajando, es lo más saludable por el bien del tratamiento de la existencia de valores ausentes en este contexto.

Por este motivo, este proceso de simulación nos resulta muy útil para dar soporte al estudio comparativo visto en el anterior capítulo. Reporta evidentes ventajas, por un lado, al no tener que prescindir de un subconjunto extra de valores observados como en el caso de las particiones, se posee mayor información para un mejor ajuste de los modelos. Se podría esperar así una mejor calidad en las imputaciones.

Por otro, los errores de imputación cometidos por los modelos se pueden calcular directamente con los verdaderos valores, pues el conjunto de datos inicialmente estaba completo. Además, este cálculo estará basado en un ajuste del modelo con mayor número de observaciones, ya que se tendrá 4064 valores observados frente a los 2709 de las particiones. Esta mayor información se entiende que jugará en favor de la robustez del cálculo del error.

Finalmente, el disponer de tres conjuntos de datos con

estructura espacio-temporal diferentes y estudiar nuevamente cómo se comportan los tres modelos de imputación analizados, induce a pensar que las conclusiones que de este estudio se deduzcan no dependerán de los datos utilizados. Así, atendiendo a esta diversidad de conjuntos de datos espacio-temporales, las evidencias que se obtengan sobre la adecuación de los modelos de imputación se pueden considerar concluyentes.

En lo sucesivo se procede, en primer lugar, a explicar con detalle el proceso de generación de los bancos de datos espacio-temporales para, posteriormente, mostrar los resultados obtenidos al imputar los datos simulados con los diferentes modelos propuestos.

7.1. Simulación

Generamos tres bancos de datos, uno con cada uno de los modelos que se están estudiando: modelo IET1, modelo IET2 y modelo ARET.

Hemos visto que cada modelo presenta comportamientos distintos a la hora de ajustar e imputar los datos de concentración de nitratos. Dicho comportamiento obedece a la propia naturaleza del modelo, su estructura y definición. Así, parece lógico pensar que los tres bancos de datos simulados, pese a tener el denominador común de la correlación espacio-temporal, presentarán diferencias que permitan considerar que son tres bancos de datos espacio-temporales distintos. Por lo tanto, como se ha apuntado anteriormente, este proceso de simulación presenta la ventaja añadida de que se tiene la oportunidad de analizar nuevamente los modelos de imputación esta vez al realizar la imputación de diferentes bancos de datos con estructura espacio-temporal. Además, el hecho de realizar la imputación de los tres bancos de datos simulados con cada uno de los tres modelos de imputación, entendemos que aumenta la fiabilidad de los resultados, pues éstos no dependerán del proceso de generación de datos simulados.

El proceso para crear los tres bancos de datos consiste en, con los datos originales de concentraciones de nitratos, lanzar cada modelo y, una vez conseguida la convergencia, guardarnos los valores de nitratos simulados por ese modelo. Concretamente, como valores de nitratos nos quedamos con la mediana de una muestra de tamaño 400 de la distribución posterior de cada uno de los datos de nitratos (distribución posterior de Y_{ij}, $\forall ij$, en cada uno de los modelos).

Posteriormente, se provocan los valores ausentes en cada uno de los bancos de datos simulados. Para ello, se extrae una única muestra aleatoria de tamaño 1336 (mismo número de valores ausentes que en los datos originales) del conjunto de índices de los datos, el cual va de 1 (municipio 1, año 1) a 5400 (municipio 540, año 10). Los valores asociados a los índices muestreados son convertidos en ausentes en los tres bancos de datos simulados. De esta manera, sabemos que los errores de imputación obtenidos por los tres modelos son comparables ya que se han calculado sobre los mismos valores ausentes.

Seguidamente, se realiza la imputación de los valores ausentes de estos tres bancos de datos espacio-temporales simulados. Cada banco de datos es completado mediante imputación con cada uno de los tres modelos de imputación, modelo IET1, modelo IET2 y modelo ARET.

7.2. Resultados obtenidos

El banco de datos simulado por el modelo IET1 es completado mediante imputación por el propio modelo IET1 pero también por el modelo IET2 y el modelo ARET. De la misma forma, el banco de datos generado mediante el modelo IET2 es imputado por el modelo IET1, el propio modelo IET2 y el modelo ARET. Finalmente, el banco de datos simulado por el modelo ARET es a su vez completado por el modelo de imputación IET1, el modelo IET2 y el propio modelo ARET.

Este proceso se lleva a cabo lanzando los modelos con los

mismos parámetros MCMC que se han descrito en al apartado 6.1. Nuevamente nos quedamos con la media y la mediana de la distribución posterior obtenida para los valores ausentes. Estos valores serán entonces las imputaciones con las que calcularemos el error de imputación cometido por cada modelo y para cada banco de datos.

Así, en estas nueve imputaciones se calcula el error de imputación haciendo uso de los tres expresiones vista en el apartado 6.2: error cuadrático, relativo y logarítmico.

En la tabla 7.1 y figura 7.1 puede verse el error cuadrático de imputación cometido por cada modelo al imputar cada uno de los tres bancos de datos simulados. En la tabla 7.2 y figura 7.2 se muestra el error de imputación calculado por el método relativo. Por último, en la tabla 7.3 y figura 7.3 se presentan los errores logarítmicos. En todas las tablas y figuras se muestra el error cometido al imputar con la media y con la mediana de la distribución posterior de los valores ausentes.

En ellas puede apreciarse como el modelo ARET sigue proporcionando las mejores imputaciones en términos generales. Si nos fijamos en el error calculado por el método cuadrático, se observa que el modelo que mejor imputa los datos simulados con el modelo IET1 es el propio modelo IET1. Lo mismo ocurre para los datos simulados por el modelo ARET, el propio modelo es el que menor error cuadrático comete al completarlos. No obstante, para los datos generados por el modelo IET2 el modelo que mejores imputaciones proporciona es el modelo IET1. El modelo IET2 comete los mayores errores en todos los casos, incluso con diferencias muy elevadas como en el caso de los datos simulados por el modelo IET1.

Este comportamiento se observa tanto para la media como para la mediana. Viéndose como, en este caso de error cuadrático, los valores obtenidos al imputar mediante la mediana son superiores a los obtenidos al imputar con la media. Algo que no ocurre para los otros dos métodos de cálculo de error.

Si nos fijamos en ellos, error relativo y logarítmico, puede apreciarse como el modelo ARET es el que mejores imputaciones proporciona en todos los casos. Los valores de los errores de imputación para este modelo son siempre inferiores, sea cual sea el banco de datos imputado y tanto si la imputación se realiza con la media o con la mediana.

Para el error relativo se observa que el error cometido por el modelo IET2 al imputar los datos generados por el modelo ARET es inferior al cometido por el modelo IET1, algo que sólo ocurre en este caso.

En todos los casos puede apreciarse como la magnitud del error es descendente, es decir, al imputar los datos generados por el modelo IET1 se cometen errores más elevados que los cometidos al imputar los valores simulados por el modelo IET2. Y a su vez estos son superiores a los errores obtenidos al imputar los datos generados por el modelo ARET.

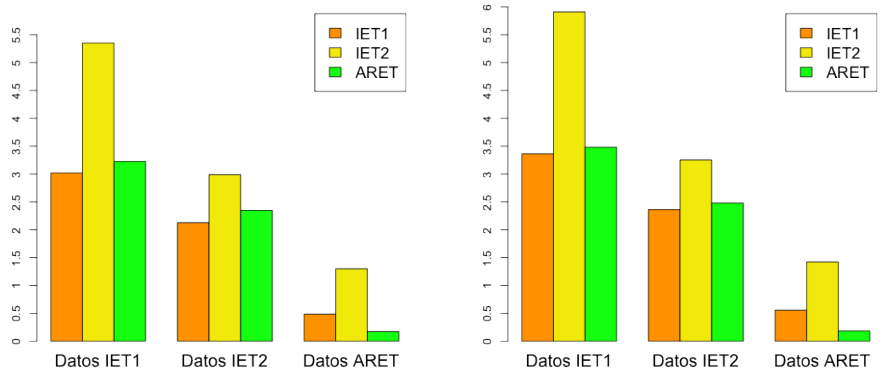
		Datos simulados		
		IET1	IET2	ARET
Modelo imputación	IET1	3.02	2.13	0.49
	IET2	5.35	2.99	1.30
	ARET	3.23	2.35	0.18

(a) Imputación con la media

		Datos simulados		
		IET1	IET2	ARET
Modelo imputación	IET1	3.36	2.36	0.56
	IET2	5.91	3.25	1.42
	ARET	3.48	2.48	0.19

(b) Imputación con la mediana

Tabla 7.1: Error cuadrático de imputación (multiplicado por 10^{-5}) cometido al imputar cada banco de datos simulado (columnas) por cada uno de los modelos de imputación propuestos (filas).



(a) Imputación con la media (b) Imputación con la mediana

Figura 7.1: Error cuadrático de imputación cometido por cada uno de los modelos propuestos multiplicado por 10^{-5}.

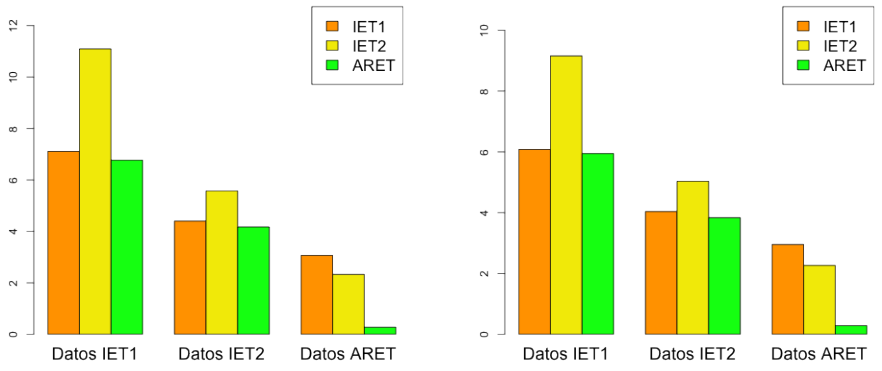
		Datos simulados		
		IET1	IET2	ARET
Modelo imputación	IET1	7.11	4.40	3.07
	IET2	11.10	5.58	2.33
	ARET	6.77	4.17	0.27

(a) Imputación con la media

		Datos simulados		
		IET1	IET2	ARET
Modelo imputación	IET1	6.08	4.04	2.95
	IET2	9.16	5.04	2.27
	ARET	5.95	3.84	0.29

(b) Imputación con la mediana

Tabla 7.2: Error relativo de imputación (multiplicado por 10^{-3}) cometido al imputar cada banco de datos simulado (columnas) por cada uno de los modelos de imputación propuestos (filas).



(a) Imputación con la media (b) Imputación con la mediana

Figura 7.2: Error relativo de imputación cometido por cada uno de los modelos propuestos multiplicado por 10^{-3}.

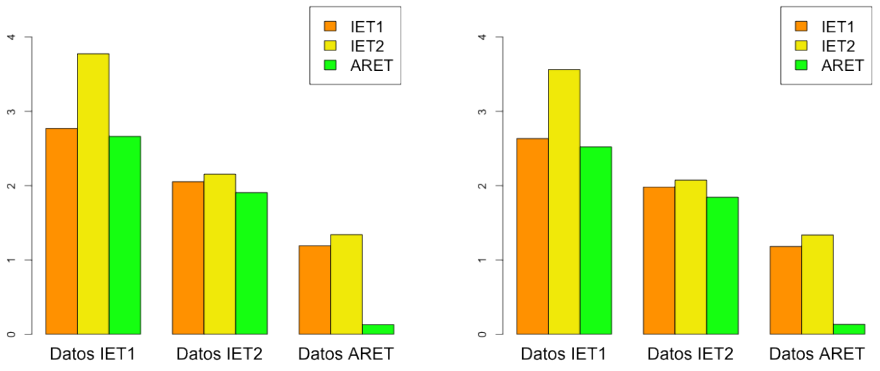
		Datos simulados		
		IET1	IET2	ARET
Modelo imputación	IET1	2.77	2.05	1.19
	IET2	3.77	2.15	1.34
	ARET	2.66	1.91	0.13

(a) Imputación con la media

		Datos simulados		
		IET1	IET2	ARET
Modelo imputación	IET1	2.63	1.98	1.18
	IET2	3.56	2.07	1.34
	ARET	2.52	1.84	0.13

(b) Imputación con la mediana

Tabla 7.3: Error logarítmico de imputación (multiplicado por 10^{-2}) cometido al imputar cada banco de datos simulado (columnas) por cada uno de los modelos de imputación propuestos (filas).



(a) Imputación con la media (b) Imputación con la mediana

Figura 7.3: Error logarítmico de imputación cometido por cada uno de los modelos propuestos multiplicado por 10^{-2}.

Los valores obtenidos para cada uno de los tipos de error y tanto para la imputación con la media como con la mediana, son inferiores a los obtenidos en el apartado 6.3. Esto es debido a que los datos simulados no presentan tanta variabilidad como las concentraciones de nitratos originales. Los modelos suavizan la realidad espacio-temporal observada en los datos originales para proporcionar valores sin cambios tan bruscos en las evoluciones temporales como los vistos a modo de ejemplo en el apartado 6.5.

Por último, a la vista de los resultados obtenidos, puede afirmarse que el modelo ARET es un serio candidato para completar mediante imputación un banco de datos con correlación espacio-temporal. Se ha observado que proporcionaba mejores imputaciones que el modelo IET1 y el modelo IET2 en el caso del banco de datos de concentraciones de nitratos. También comete errores inferiores, en termino general, al imputar los bancos de datos simulados con cada uno de los modelos. La consistencia de estos últimos resultados es evidente, pues los datos simulados están generados por modelos diferentes. Por tanto, el hecho de sea el modelo ARET el que mejor imputaciones reporta en todos los casos, le proporciona un valor añadido. Por todo esto, puede considerarse el modelo ARET como una de las mejores opciones de imputación de valores ausentes en presencia de correlación espacio-temporal.

Capítulo 8

Conclusiones y líneas futuras

En esta tesis se ha abordado en profundidad, desde el punto de vista bayesiano, el tratamiento de la información faltante en bancos de datos con correlación espacio-temporal. El problema de la existencia de valores ausentes es muy común en el análisis de datos y no lo es menos cuando estos están dotados de una estructura espacio-temporal. No obstante, no existen referencias específicas de cómo enfrentarse a este problema en la literatura especializada. Con este trabajo, se ha pretendido iniciar una línea de investigación que se ocupe de proponer posibles soluciones.

Para ello, como primer y segundo objetivo nos hemos planteado un estudio que parte de la base de datos de calidad del agua potable en la Comunitat Valenciana. Esta información, al estar recogida a nivel municipal y entre los años 1991-2000, se considera susceptible de ser tratada considerando una posible correlación espacio-temporal. La tesis se ha centrado sobre las concentraciones de magnesio y de nitratos existentes en el agua potable. Se ha visto que ambos conjuntos de datos presentan un importante porcentaje de valores ausentes, un 28 % las concentraciones de magnesio y un 24 % en el caso de las concentraciones de nitratos. En el apartado 3.4 se han mostrado los mapas de concentraciones de nitratos, banco

de datos sobre el que se ha extendido el estudio motivo de esta tesis. En ellos puede apreciarse la existencia de valores ausentes, destacando el año 1992 sobre el resto debido a que el porcentaje de valores ausentes es especialmente elevado.

Para estos datos de concentraciones de magnesio y de concentraciones de nitratos, se ha planteado en primer lugar un estudio preliminar en el que se analiza si existe la necesidad de tratarlos con modelos que incluyan componentes espacio-temporales. Este estudio previo ha consistido en cuantificar el error cometido al imputar los valores ausentes con modelos que, partiendo de una estructura sencilla de heterogeneidad, van aumentado su complejidad hasta llegar a plantear un modelo espacio-temporal. El cálculo de dicho error se ha planteado como la suma de la diferencia cuadrática entre el valor imputado y el verdadero valor.

Para poder calcular este error de imputación, aumentar el número de valores ausentes ha sido clave pues de lo contrario no se disponía de los verdaderos valores con los que enfrentar los valores imputados.

De este modo, los errores obtenidos en este pre-análisis han apuntado a la necesidad de plantear modelos con estructura espacio-temporal para la imputación de las concentraciones de nitratos. Sin embargo, no ha ocurrido lo mismo en el caso de las concentraciones de magnesio, para las que un modelo sencillo de heterogeneidad resulta proporcionar mejores imputaciones.

Es a partir de estos resultados cuando se ha centrado la atención sobre los datos de concentraciones de nitratos con el fin de estudiar la adecuación de modelos espacio-temporales de imputación.

Así, nuestro objetivo número 1 ha sido plantear una comparativa de modelos espacio-temporales. Se ha expuesto que el uso de modelos con estructura espacio-temporal es muy común en el campo de la salud pública y medio ambiente. Existen múltiples modelizaciones con estructura espacio-temporal que pretenden analizar la influencia de uno o varios factores medioambientales

sobre la salud pública (regresión ecológica) o estudiar la distribución espacial y su evolución temporal de determinadas variables de interés epidemiológico (cartografía de enfermedades). No obstante observamos que no ocurre lo mismo con el tratamiento de la información faltante en bancos de datos con esta estructura. Así, ante la inexistencia en la literatura de este tipo de modelizaciones para realizar imputación de valores ausentes, adaptamos estos modelos tan conocidos en otros ámbitos para realizar la imputación de los valores ausentes de concentración de nitratos. Del amplio abanico de modelizaciones espacio-temporales, hemos elegido las dos tipologías de modelización espacio-temporal más importantes en la literatura con el fin de estudiar y comparar su comportamiento en el campo de la imputación de valores ausentes.

El hecho de haber elegido estas dos modelizaciones radica no sólo en su recurrencia en la literatura sino también en que son modelizaciones diferenciadas conceptualmente respecto a cómo tratan de ajustar la correlación espacio-temporal.

Por un lado, hemos planteado la modelización que ajusta dicha correlación a través de la interacción entre la componente espacial y la componente temporal. Con esta estructura de modelización, a la que hemos denominado modelización IET, hemos planteado dos variantes: el modelo IET1 que asume evoluciones temporales diferenciadas e independientes para cada municipio, y el modelo IET2 que supone una estructura espacio-temporal más fuerte al modelizar que existe correlación entre observaciones vecinas en el tiempo, año anterior y posterior, y en el espacio, municipios contiguos, tanto para el año en curso como para el anterior y posterior.

Por otro lado, hemos planteado la modelización, a la que hemos denominado ARET, que basa el ajuste de la correlación espacio-temporal de forma conceptualmente similar a una serie temporal auto regresiva de orden 1. Así observaciones vecinas en el espacio, municipios contiguos, tienden a tener una evolución temporal similar.

De la misma forma que no existen referencias de la utilización de modelos espacio-temporales para la imputación de valores ausentes, la existencia en la literatura de algún tipo de comparativa entre estos dos tipos de modelizaciones, ya sea para la cartografía de enfermedades o para imputación de valores ausentes es escasa. Por tanto, el análisis que se ha planteado en esta tesis tiene una doble componente de interés: la adaptación de este tipo de modelos al ámbito de la imputación de datos y la comparación entre los dos tipos de modelizaciones más frecuentes en estudios espacio-temporales.

Dicha comparativa se ha planteado en términos del cálculo del error que comete cada una de las modelizaciones a la hora de imputar los valores ausentes de las concentraciones de nitratos de la Comunidad Valenciana. Este planteamiento coincide con el llevado a cabo en el estudio previo de la necesidad de utilizar modelos con estructura espacio-temporal. No obstante, para esta comparativa se han planteado tres métodos diferentes para el cálculo del error de imputación. El objetivo ha sido asegurarnos que los resultados obtenidos eran consistentes y no dependían del método con el que se cuantificaba el error.

Asimismo, para llevar a cabo este cálculo, si para el estudio previo ha sido clave la ampliación del número de valores ausentes, más aún lo ha sido para esta comparativa la creación de las tres particiones. De este modo, junto con los diferentes cálculos del error de imputación, se ha pretendido comprobar la coherencia de los resultados a la vez que afianzar la fiabilidad estos. Pues procediendo de este modo se ha tratado de evitar posibles dependencias entre los resultados obtenidos y los valores ausentes que se imputaban y entre estos y el método de calculo elegido.

También se ha empleado el cálculo de estos errores para medir cómo se ajusta cada modelo a la realidad espacio-temporal de los datos de concentraciones de nitratos. Y esto se ha llevado a cabo comparando los valores proporcionados por las modelos con los verdaderos valores observados disponibles.

Respecto al cálculo de dicho ajuste, se ha podido comprobar que el modelo IET2 es el que mejor se ajusta a la naturaleza espacio-temporal de las observaciones.

No obstante, respecto a la imputación de los valores ausentes, los errores obtenidos han mostrado que el modelo IET2 es claramente el que peor comportamiento tiene a la hora de proporcionar estimaciones de estos. Las imputaciones que proporciona el modelo IET2 resultan ser muy poco precisas pues existe una elevada variabilidad en los errores de imputación. La variabilidad de más que muestra el modelo IET2 respecto al modelo IET1 y al modelo ARET, además sería claramente perjudicial en un proceso de imputación múltiple o completamente bayesiano, pues habría un exceso de incertidumbre añadida.

Asimismo, los resultados han apuntado a que el modelo ARET es el que menor error de imputación comete. Al igual que los otros dos modelos, en su definición se conjuga la correlación espacial y la correlación temporal pero este modelo resulta ser el más estable ajustando adecuadamente la tendencia temporal de cada municipio así como su correlación espacial con los municipios vecinos. Además resulta ser más preciso en el proceso de imputación, lo que implica que proporciona un rango de imputaciones cercano al valor real que se quiere estimar. Este hecho nos permite asegurar también que el modelo ARET sería el modelo más adecuado para una imputación múltiple o una imputación completamente bayesiana.

Por último, como tercer objetivo de esta tesis, se ha planteado un proceso de simulación de datos espacio-temporales con el fin de comparar de nuevo las tres modelizaciones propuestas. El procedimiento de comparación entre los modelos ha sido el mismo que para el caso de las concetraciones de nitratos, nos hemos basado de nuevo en el error de imputación y hemos utilizado los mismos tres métodos de cálculo del error de imputación.

Es importante destacar que este estudio simulado ha resultado útil para corroborar aquello que los resultados de la comparación planteada en el objetivo 1 apuntaban. Sin pretensión de ser

un proceso de simulación capaz de alumbrar un modelo de referencia para futuros casos de información faltante en datos con estructura espacio-temporal, este estudio, a la vista de los resultados obtenidos, ha mostrado de nuevo que el modelo ARET es un serio candidato para completar mediante imputación un banco de datos con correlación espacio-temporal. Se ha visto que comete errores inferiores, en términos generales, al imputar los bancos de datos, tanto los simulados por él mismo como los simulados con cada uno de los otros dos modelos.

Este hecho, unido a que el modelo ARET también proporciona mejores imputaciones que el modelo IET1 y el modelo IET2 en el caso del banco de datos de concentraciones de nitratos, permite concluir que modelo ARET es un serio candidato para la imputación de valores ausentes en datos con estructura espacio-temporal. Siendo muy conveniente, por tanto, tenerlo en cuenta ante un nuevo problema de datos espacio-temporales incompletos debido a la existencia de valores ausentes.

No obstante, honestamente consideramos que no existe un único modelo de referencia que nos permita abordar con confianza cualquier problema futuro de existencia de valores ausentes en datos con correlación espacio-temporal. En cambio, sí somos firmes defensores de que ante dicho problema, el tratamiento adecuado de la información faltante es importante. El global de esta tesis pretende aportar luz a este hecho, haciendo hincapié en que lo más aconsejable para hacerle frente es plantearse más de una opción de modelización de los valores ausentes e implementar una comparativa previa antes de proceder a su imputación.

Líneas futuras de investigación

Toda investigación es susceptible de mejoras. Incluso cuando estas se implementan, ya sea debido a sugerencias de colegas o revisiones bibliográficas, siempre existe algún punto sobre el que se podría avanzar todavía. Este es uno de los motivos que hacen que el trabajo de un investigador sea tan motivador como exigente. Y que nunca tenga otro final que no sea el propio límite temporal que uno se fije para su finalización.

Partiendo de esta base de revisión constante, entendemos que este trabajo de investigación no es ajeno a este proceso y proponemos una serie de líneas de investigación futuras que sin duda ayudarían en la mejoría y avance de este trabajo.

Hemos supuesto que el mecanismo de generación de valores ausentes es ignorable. No obstante, aunque las situaciones en las que esta hipótesis no puede ser asumida son poco frecuentes, podría analizarse cómo se comportan las modelizaciones espacio-temporales presentadas en esta tesis a la hora de imputar valores ausentes bajo no ignorabilidad e incluso bajo MNAR. Este hecho proporcionaría un completo estudio de sensibilidad de la imputación de los datos no observados pues se estudiaría esta bajo todas las suposiciones posibles de aparición de valores ausentes.

A la vista de los resultados, el modelo IET2 ha resultado tener un comportamiento sorprendente. Por un lado resulta ser el que mejor se ajusta a la realidad espacio-temporal de los datos pero, por otro lado, es el modelo que peores imputaciones proporciona. Debido a esto, sería conveniente profundizar en la definición de vecindad del modelo IET2 con el fin de estudiar una posible mejoraría en su comportamiento a la hora de imputar los valores ausentes. Esta labor no sería sencilla, pues la casuística de posibles definiciones de vecindad a nivel espacial y temporal es muy amplia.

Asimismo se podría ampliar las modelizaciones espacio-temporales a comparar. Para ello cabría la posibilidad de incluir el ajuste de la correlación espacio-tiempo clasificado como tipo III en el apartado 5.1. Entendemos que para los

datos utilizados en esta tesis, dicha definición de correlación espacio-temporal no encajaría, pero sí podría probarse con otros datos más adecuados o podría plantearse como una extensión del estudio con datos simulados.

Al respecto de este estudio simulado, sería también interesante profundizar en él con la intención de convertirlo en un proceso de simulación capaz de proporcionar conclusiones por sí mismo sobre la comparativa de modelizaciones. Esto pasaría posiblemente por aumentar el número de banco de datos simulados así como repetir el muestreo de los valores observados varias veces con el fin de controlar que la imputación no depende de dichos valores convertidos a ausentes.

También sería muy conveniente adaptar las modelizaciones espacio-temporales propuestas a INLA. De este modo el tiempo de computación se reduciría drásticamente con la consiguiente ventaja de poder realizar mayor número de pruebas a los modelos, ampliar el abanico de modelizaciones a estudio así como abordar la mencionada profundización del estudio simulado.

Por último, en un intento de cerrar el círculo, resultaría también muy interesante plantear de nuevo un análisis de la influencia de la imputación. Se elegiría, tal y como hemos realizado, el modelo de imputación que mejor se adapte a los datos y mejores imputaciones de los valores ausentes proporcione. Posteriormente, se estudiaría como dicha imputación influye en la asociación existente entre los datos imputados y una variable respuesta. Esta influencia se estudiaría simulando diferentes variables respuestas, cada una con una asociación diferente con los datos imputados. Posteriormente se analizaría cómo varía dicha asociación al ajustarse con el banco de datos completo real y con el banco de datos completado mediante imputación.

Referencias

Abellan, C. (2005). Imputación espacio temporal de datos medioambientales en estudios de regresión ecológica. Tesis de máster, Universitat de València.

Abellan, J. J., Richardson, S., y Best, N. G. (2008). Use of space-time models to investigate the stability of patterns of disease. *Environmental Health Perspectives*, 116(8):1111–1119.

Afifi, A. y Elashoff, R. (1966). Missing observations in multivariate statistics I: Review of the literature. *Journal of the American Statistical Association*, 61:595–604.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52(278):200–203.

Banerjee, S., Gelfand, A. E., y Carlin, B. P. (2014). *Hierarchical modeling and analysis for spatial data. Second edition.* Chapman and Hall/CRC.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., y Songini, M. (1995). Bayesian analysis of space—time variation in disease risk. *Statistics in Medicine*, 14(21-22):2433–2443.

Bernardinelli, L. y Montomoli, C. (1992). Empirical bayes versus

fully bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, 11:983–1007.

Berry, B. y Marble, D. (1968). *Spatial Analysis: A Reader in Statistical Geography*.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236.

Besag, J. y Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746.

Besag, J., York, J., y Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20.

Best, N. G., Richardson, S., y Thompson, A. (2005). A comparison of bayesian spatial models for disease mapping. *Statistics in Medicine*, 14(1):35–39.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 22:302–306.

Carlin, B. P. y Louis, T. A. (1997). Bayes and empirical bayes methods for data analysis. *Statistics and Computing*, 7(2):153–154.

Carpenter, J. y Kenward, M. (2013). *Multiple Imputation and its Application*. Wiley.

Chorley, R. (1972). *Spatial analysis in geomorphology*. London: Methuen.

Christensen, W. F. y Yetkin, F. Z. (2005). Spatio-temporal analysis of auditory cortex activation as detected with silent event related fMRI. *Statistics in Medicine*, 24:2539–2556.

Clayton, D. y Bernardinelli, L. (1992). *Bayesian methods for mapping disease risk. En: Geographical and Environmental Epidemiology: Methods for Small Area Studies*, capítulo 18, pp. 205–220. Oxford University Press.

Clayton, D. y Kaldor, J. (1987). Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43(3):671–681.

Cressie, N. (1993). *Statistics for Spatial Data, revised edition*. Wiley, original edition was published in 1991 edición.

Cressie, N. y Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Wiley.

Daniels, M. J. y Hogan, J. W. (2008). *Missing Data In Longitudinal Studies Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman and Hall.

Dempster, A. P., Laird, N. M., y Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 39:1–38.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. Oxford University Press.

Diggle, P. J., Tawn, J. A., y Moyeed, R. A. (1998). Model-Based geostatistics. *Applied Statistics*, 47(3):299–350.

Ferrándiz, J., López-Quílez, A., Gómez-Rubio, V., Sanmartín, P., Martínez-Beneito, M. A., Melchor, I., Vanaclocha, H., Zurriaga, O., Ballester, F., Gil, J., Pérez-Hoyos, S., y Abellan, J. J. (2003). Statistical relationship between hardness of drinking water and cerebrovascular mortality in valencia: a comparison of spatiotemporal models. *Environmetrics*, 14(5):491–510.

Garrison, W. (1959). Spatial structure of the economy, II. *Annals of the Association of American Geographers*, 49:471–482.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., y Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC.

Gonzalez, J. R., Abellan, C., y Abellan, J. J. (2012). Bayesian model to detect phenotype-specific genes for copy number data. *BMC Bioinformatics*, 13(1):130.

Haining, R. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge university press.

Hartley, H. y Hocking, R. (1971). The analysis of incomplete data. *Biometrics*, 27:7783–808.

Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19:2555–2567.

Held, L. y Besag, J. (1998). Modelling risk from a disease in time and space. *Statistics in Medicine*, 17(18):2045–2060.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., y Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52(6):119–139.

Lee, K. J. y Carlin, J. B. (2010). Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation. *American Journal of Epidemiology*, 171(5):624–632.

Little, R. J. A. y Rubin, D. B. (2002). *Statistical analysis with missing data (second edition)*. Chichester: Wiley.

Martínez-Beneito, M., López-Quílez, A., y Botella-Rocamora, P. (2008). An autoregressive approach to spatio-temporal disease mapping. *Statistics in Medicine*, 27:2874–2889.

Mason, A., Richardson, S., y Best, N. (2012a). Two-pronged strategy for using dic to compare selection models with non-ignorable missing responses. *Bayesian Analysis*, 7:109–146.

Mason, A., Richardson, S., Plewis, I., y Best, N. (2012b). Strategy for modelling nonrandom missing data mechanisms in observational studies using bayesian methods. *Journal of Official Statistics*, 28(2):279–302.

McCullagh, P. y Nelder, J. A. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC.

Molenberghs, G. y Kenward, M. G. (2007). *Missing Data in Clinical Studies*. John Wiley and Sons.

Mollié, A. (1996). Bayesian mapping of disease. *Markov Chain Monte Carlo in Practice*, pp. 359–379.

Orchard, T. y Woodbury, M. (1972). A missing information principle: theory and applications. En Cam, L. M. L., Neyman, J., y Scott, E. L., editores, *Proceedings of the Sixth Berkely Symposium on Mathematics, Statistics and Probability, Volume 1.*, pp. 697–715. Berkeley: University of California Press.

Plummer, M., Best, N., Cowles, K., y Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11.

Resseguier, N., Giorgi, R., y Paoletti, X. (2011). Sensitivity Analysis When Data Are Missing Not-at-random. *Epidemiology*, 22(2):282.

Richardson, S., Abellan, J. J., y Best, N. G. (2006). Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in yorkshire (UK). *Statistical Methods in Medical Research*, 15:385–407.

Ripley, B. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13:255–266.

Ripley, B. D. (1981). *Spatial Statistics*. John Wiley and Sons.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Rubin, D. B. (1978). Multiple imputations in sample surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 20–34.

Rubin, D. B. (1979). *Illustrating the use of multiple imputations to hand- le nonresponse in sample surveys*, volumen 2. Proceedings of the 42nd Session of the International Statistical Institute.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18 years. *Journal of the American Statistical Association*, 91:473–490.

Rue, H., Martino, S., y Chopin, N. (2009). Approximate bayesian inference for latent gaussian models using integrated nested laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, (71):319–392.

Schabenberger, O. y Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman and Hall/CRC.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, (8):3–15.

Schafer, J. L. y Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7:147–177.

Schrödle, B. y Held, L. (2011). Spatio-temporal disease mapping using inla. *Environmetrics*, (22):725–734.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., y Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Sun, D., Tsutakawa, R. K., Kim, H., y Zhuoquiong, H. (2000). Spatio-temporal interaction with disease mapping. *Statistics in Medicine*, 19(15):2015–2035.

VanBuuren, S. (2012). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC.

Waller, L. A., Carlin, B. P., Xia, H., y Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92:607–617.

Wood, A. M., White, I. R., y Thompson, S. G. (2004). Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1(4):368–376.

Yang, C. Y., Cheng, M. F., Tsai, S. S., y Hsieh, Y. L. (1998). Calcium, magnesium, and nitrate in drinking water and gastric cancer mortality. *Cancer Science*, 89(2):124–130.

Anexo 1

Resumen de la distribución posterior de los hiperparámetros

A continuación se presentan los resultados obtenidos tras la simulación de los modelos IET1, IET2 y ARET. Dicha simulación se ha llevado a cabo mediante el software WinBUGS.

Se han lanzado dos cadenas para cada modelo. Para el modelo IET1 se han lanzado un total de 500000 iteraciones por cadena, para el modelo IET2 un total de 800000 iteraciones por cadena y para el modelo ARET un total de 200000 iteraciones por cadena. En todos los casos nos hemos quedado con una iteración de cada 5 y hemos tomado como muestra posterior las 1000 últimas iteraciones de cada cadena.

Para cada modelo se muestra una tabla que incluye la descriptiva básica de cada hiperparámetro que forma parte de su definición. También se incluye el estadístico *Rhat* obtenido en la simulación y que nos ayuda a asumir la convergencia de los modelos.

Modelo IET1

	Media	DT	2.5 %	97.5 %	Rhat
τ_Y	4.904	0.173	4.578	5.234	1.013
τ_γ	1375.010	1405.650	120.600	5277.050	1.001
τ_α	497.300	448.000	72.150	1801.001	1.002
τ_θ	8.490	4.151	4.470	20.400	1.002
τ_ϕ	0.499	0.063	0.399	0.637	1.001
τ_δ	15.400	1.331	13.040	18.690	1.004

Tabla 1: Hiperparámetros modelo IET1

Modelo IET2

	Media	DT	2.5 %	97.5 %	Rhat
τ_Y	6.327	0.559	5.452	7.758	1.055
τ_γ	84.290	58.950	13.970	221.002	1.002
τ_α	148.500	75.660	37.270	330.700	1.002
τ_θ	6.572	2.012	3.807	11.550	1.046
τ_ϕ	2.171	0.274	1.707	2.784	1.017
τ_δ	0.505	0.054	0.407	0.604	1.001

Tabla 2: Hiperparámetros modelo IET2

Modelo ARET

	Media	DT	2.5 %	97.5 %	Rhat
τ_Y	5.711	0.212	5.329	6.131	1.008
τ_α	242.703	136.093	64.39	598.915	1.009
τ_θ	30.6	4.128	23.859	40.043	0.999
τ_ϕ	7.261	0.689	5.87	8.651	1.015
ρ	0.953	0.003	0.946	0.96	1.045

Tabla 3: Hiperparámetros modelo ARET

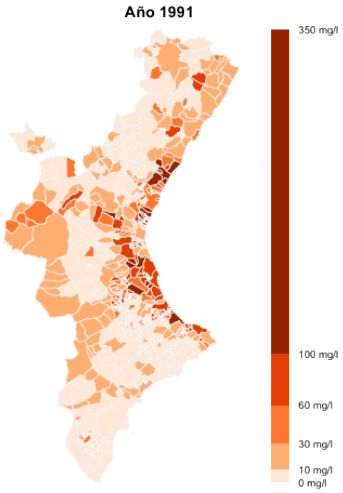
Anexo 2

Mapas de las concentraciones de nitratos completadas mediante imputación

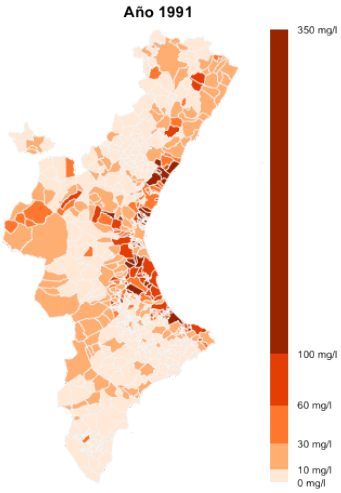
Se muestra en esta sección la distribución espacio-temporal de las concentraciones de nitratos. Estos datos se presentan ya sin la existencia de valores ausentes pues estos han sido estimados con las estimaciones obtenidas tras el proceso de imputación. Para cada año se incluyen los mapas correspondientes a las imputaciones obtenidas con cada uno de los modelos: modelo IET1, modelo IET2 y modelo ARET. Para cada uno de los modelos se presenta el mapa de concentraciones de nitratos completado por imputación mediante la media y la mediana de la distribución posterior de los valores ausentes.

Dicha distribución posterior ha sido obtenida mediante el uso del software WinBUGS. Para ello, para el modelo IET1 se han simulado un total de 500000 iteraciones, para el modelo IET2 un total de 800000 iteraciones y para el modelo ARET un total de 200000 iteraciones. En todos los casos se han lanzando dos cadenas, tomando una iteración de cada 5 y quedándonos con las 1000 últimas iteraciones de cada cadena.

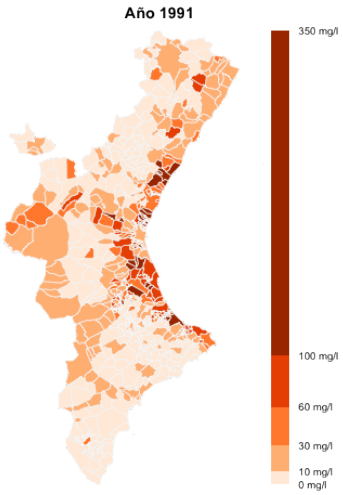
Año 1991



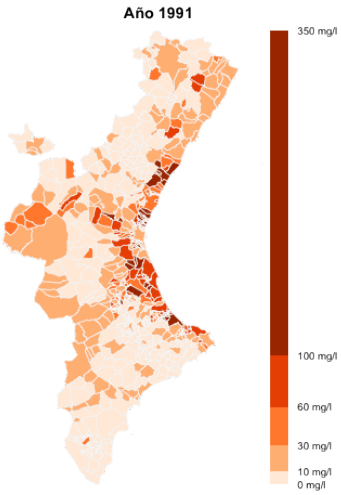
(a) Modelo IET1. Media



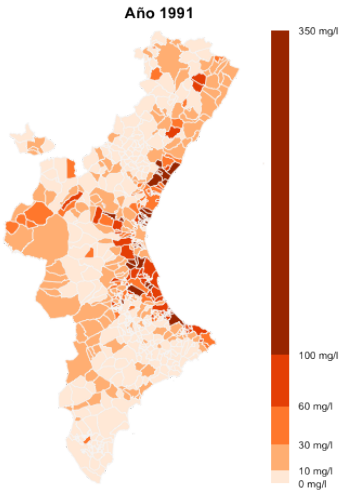
(b) Modelo IET1. Mediana



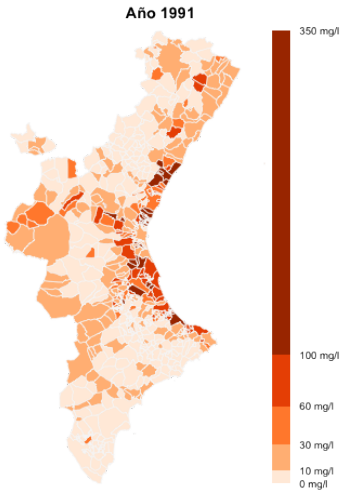
(c) Modelo IET2. Media



(d) Modelo IET2. Mediana

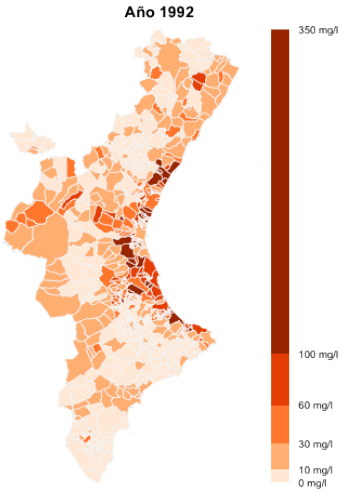


(e) Modelo ARET. Media

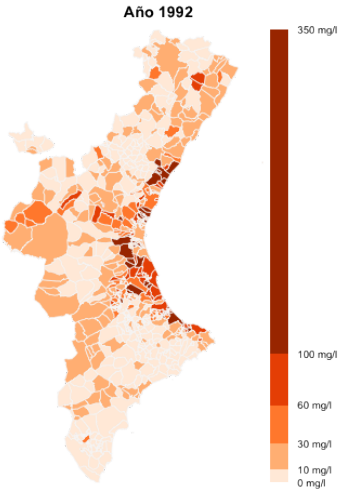


(f) Modelo ARET. Mediana

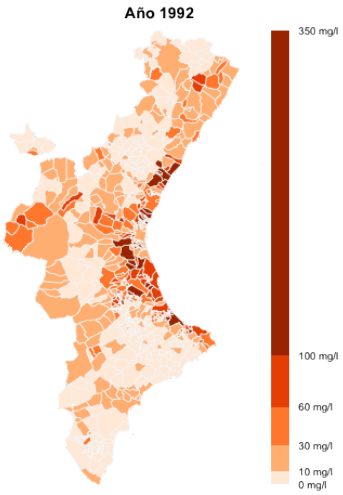
Año 1992



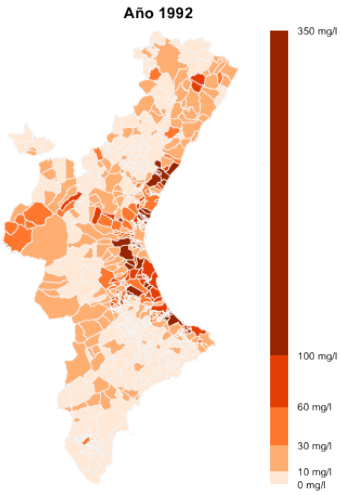
(a) Modelo IET1. Media



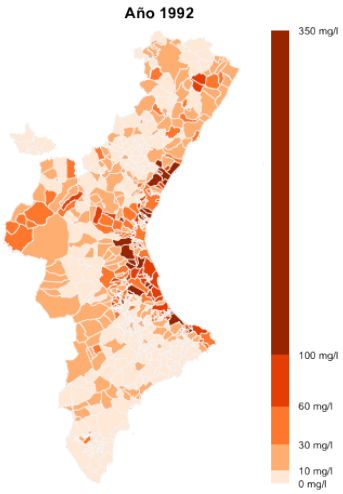
(b) Modelo IET1. Mediana



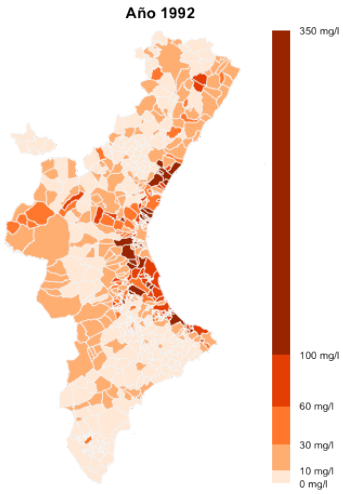
(c) Modelo IET2. Media



(d) Modelo IET2. Mediana

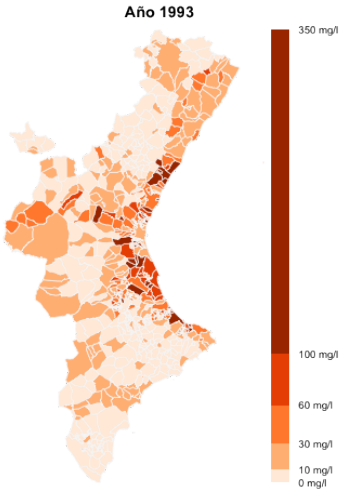


(e) Modelo ARET. Media

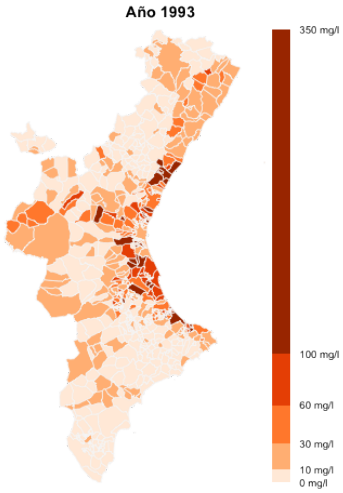


(f) Modelo ARET. Mediana

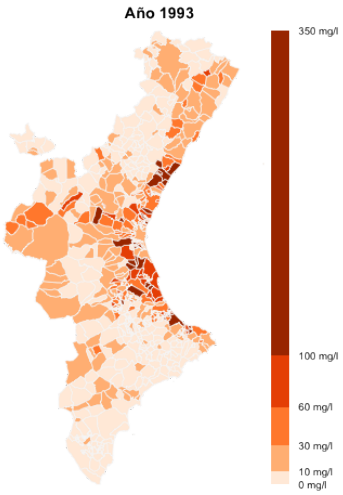
Año 1993



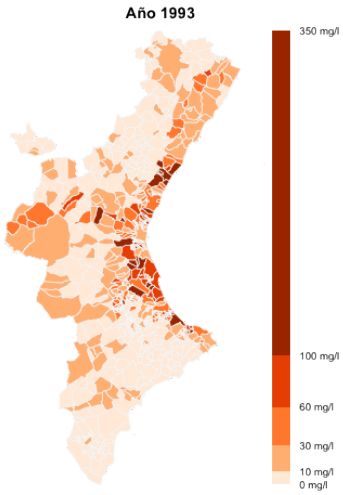
(a) Modelo IET1. Media



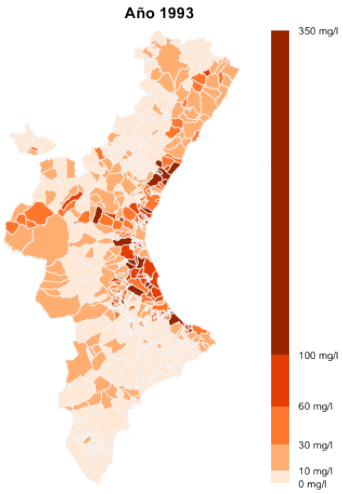
(b) Modelo IET1. Mediana



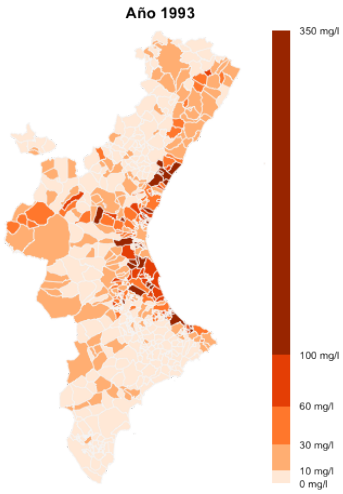
(c) Modelo IET2. Media



(d) Modelo IET2. Mediana

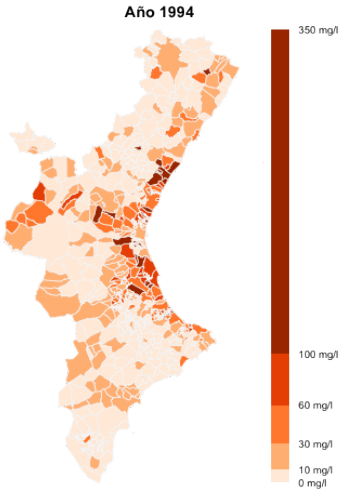


(e) Modelo ARET. Media

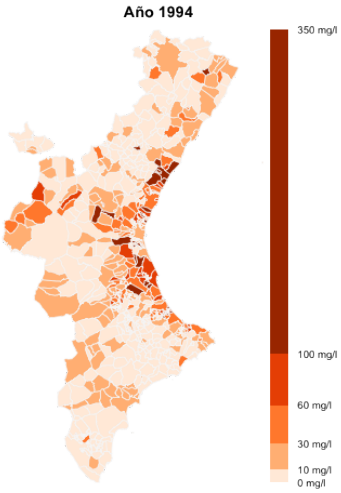


(f) Modelo ARET. Mediana

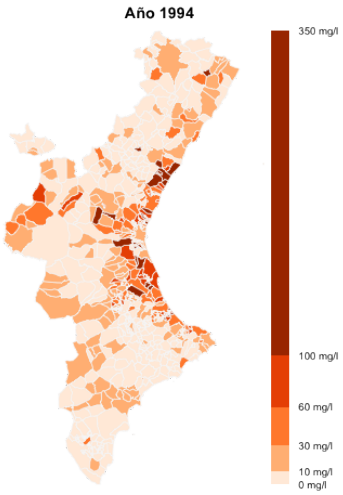
Año 1994



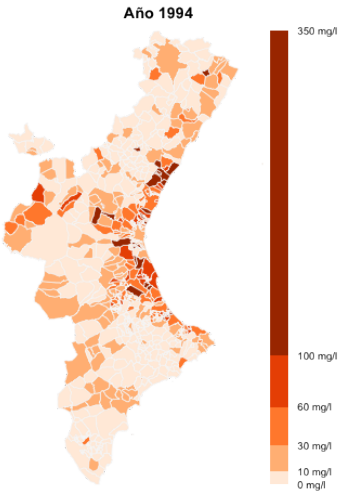
(a) Modelo IET1. Media



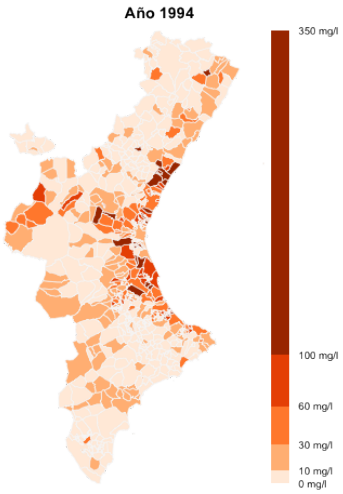
(b) Modelo IET1. Mediana



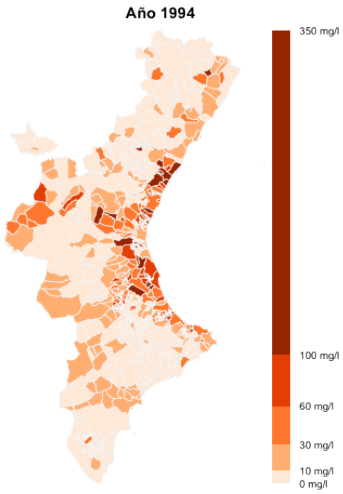
(c) Modelo IET2. Media



(d) Modelo IET2. Mediana

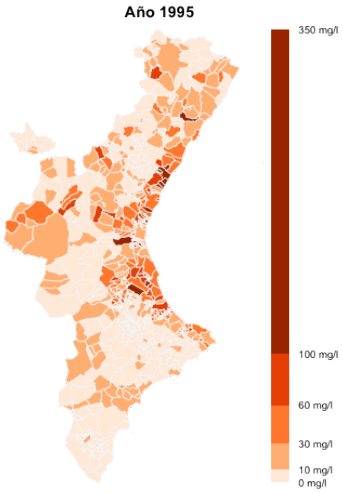


(e) Modelo ARET. Media

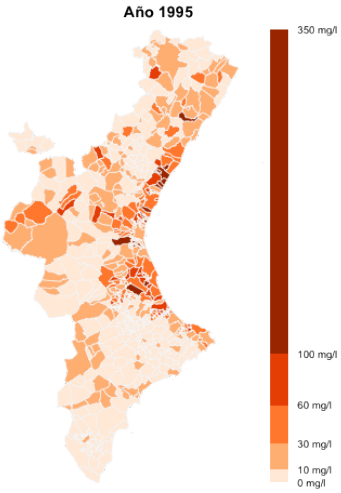


(f) Modelo ARET. Mediana

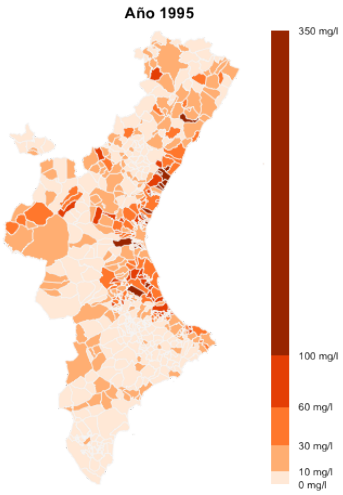
Año 1995



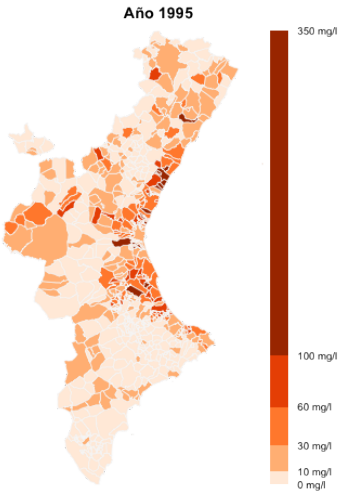
(a) Modelo IET1. Media



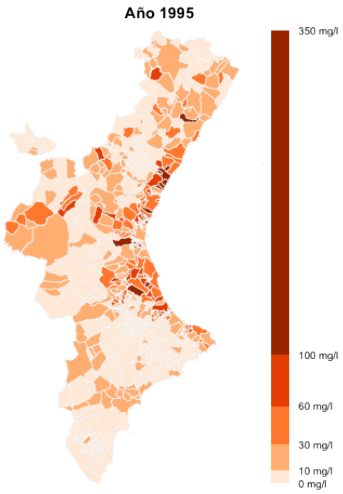
(b) Modelo IET1. Mediana



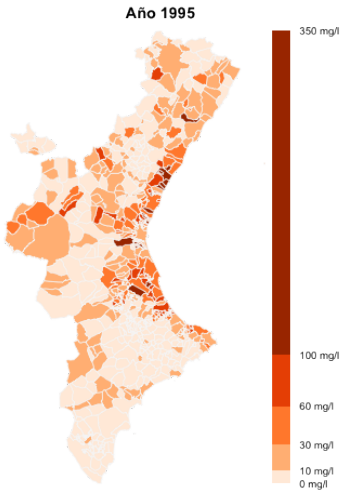
(c) Modelo IET2. Media



(d) Modelo IET2. Mediana

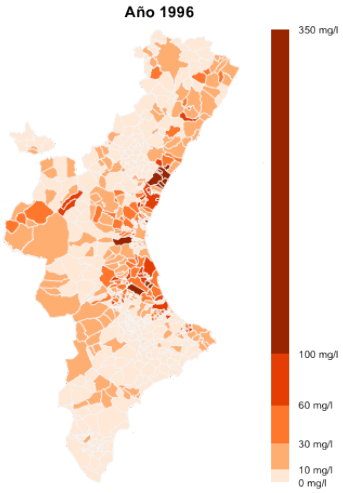


(e) Modelo ARET. Media

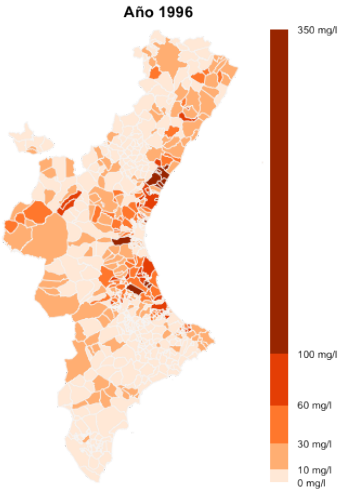


(f) Modelo ARET. Mediana

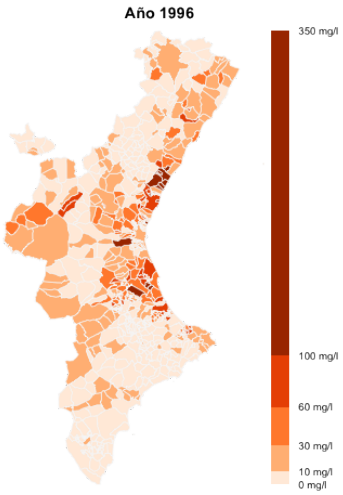
Año 1996



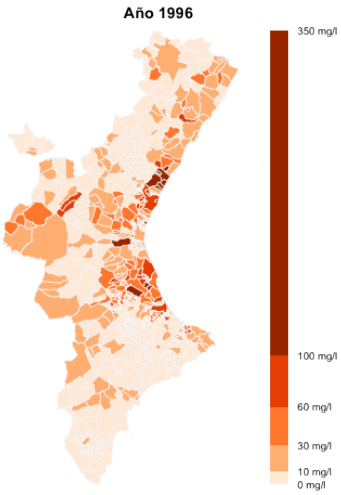
(a) Modelo IET1. Media



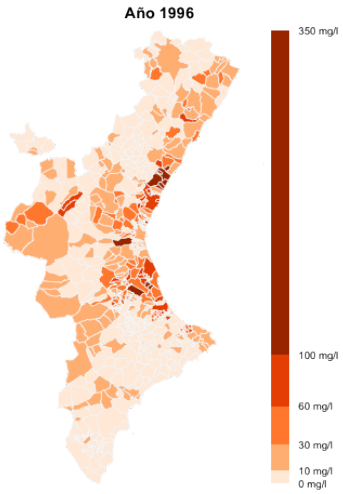
(b) Modelo IET1. Mediana



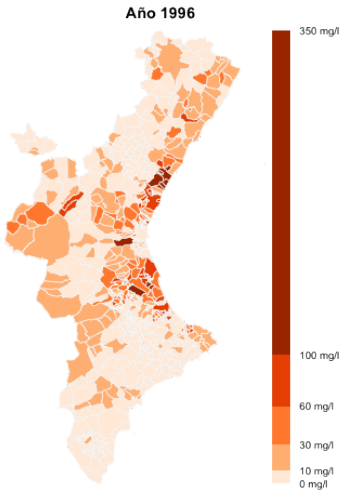
(c) Modelo IET2. Media



(d) Modelo IET2. Mediana

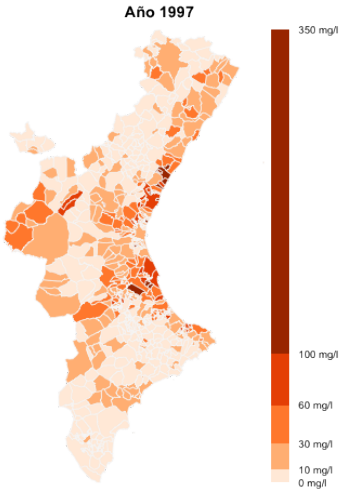


(e) Modelo ARET. Media

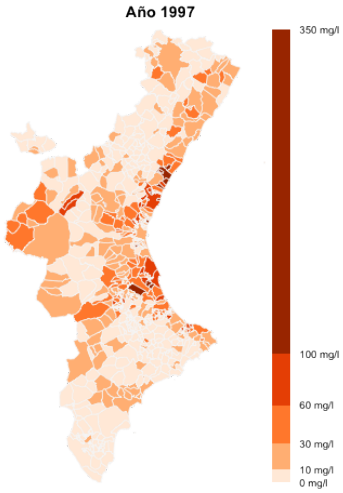


(f) Modelo ARET. Mediana

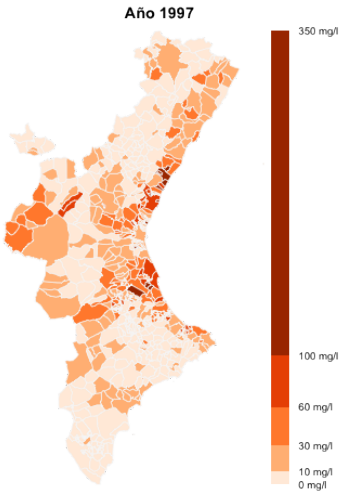
Año 1997



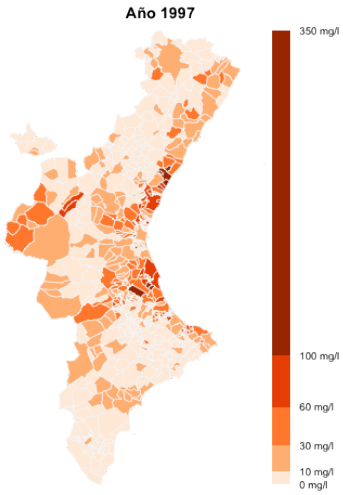
(a) Modelo IET1. Media



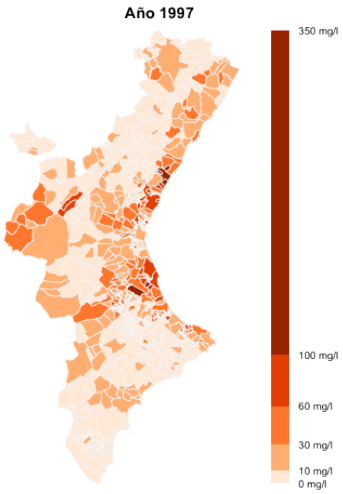
(b) Modelo IET1. Mediana



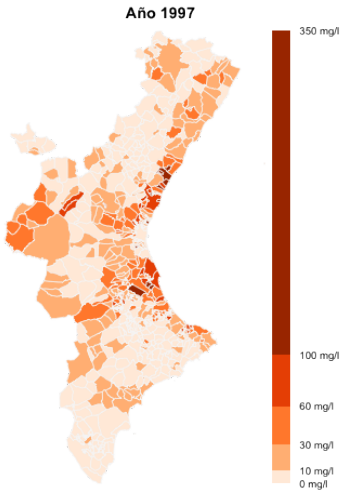
(c) Modelo IET2. Media



(d) Modelo IET2. Mediana

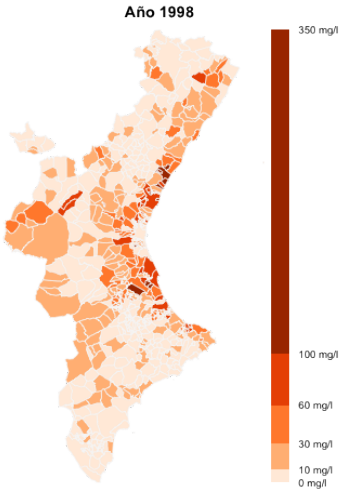


(e) Modelo ARET. Media

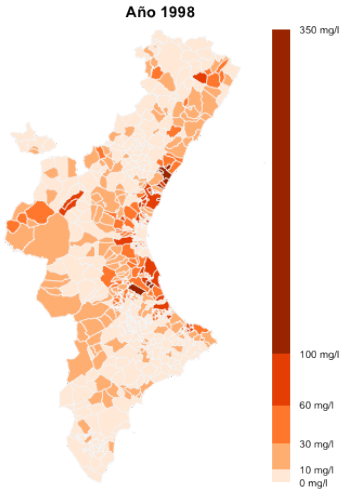


(f) Modelo ARET. Mediana

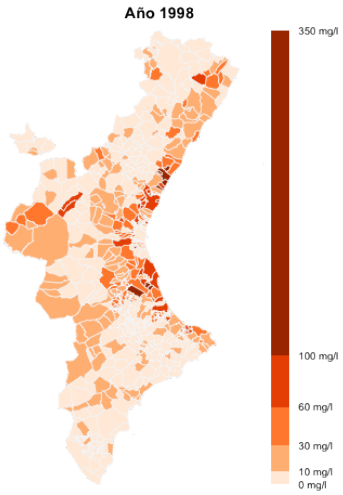
Año 1998



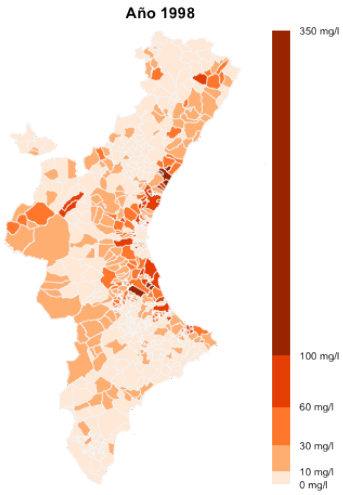
(a) Modelo IET1. Media



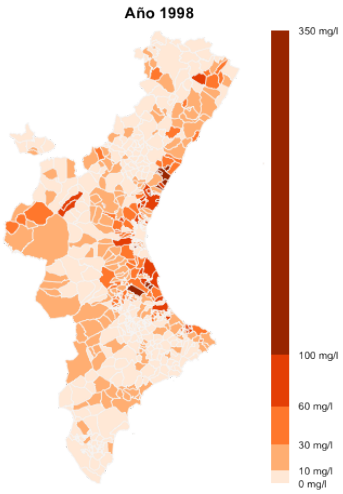
(b) Modelo IET1. Mediana



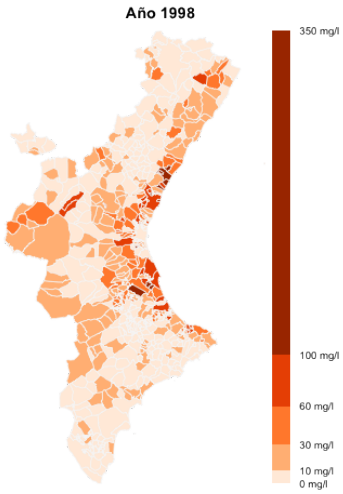
(c) Modelo IET2. Media



(d) Modelo IET2. Mediana

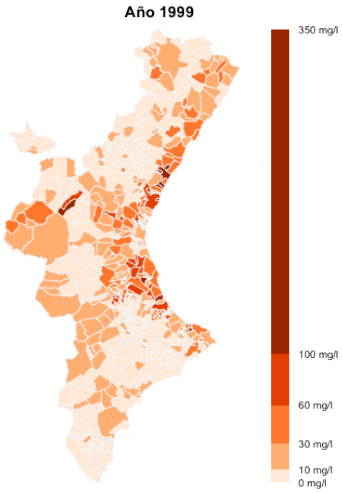


(e) Modelo ARET. Media

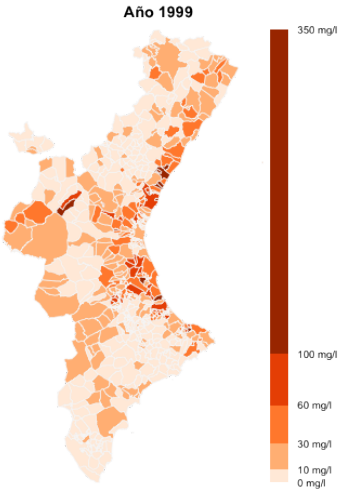


(f) Modelo ARET. Mediana

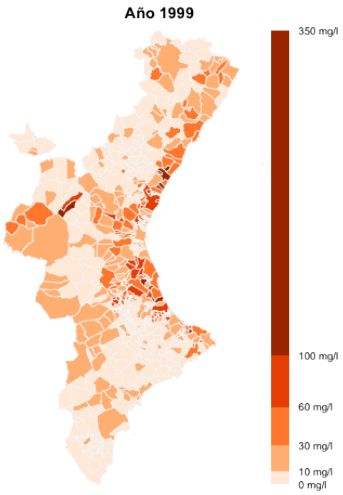
Año 1999



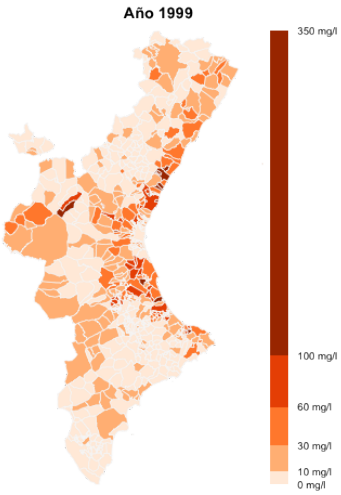
(a) Modelo IET1. Media



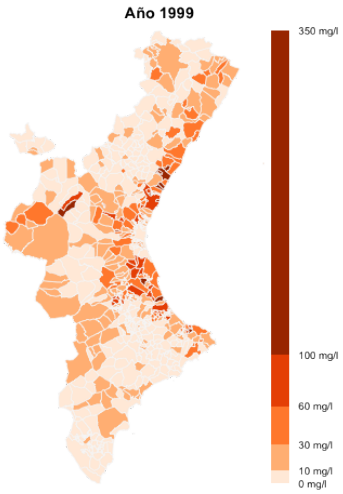
(b) Modelo IET1. Mediana



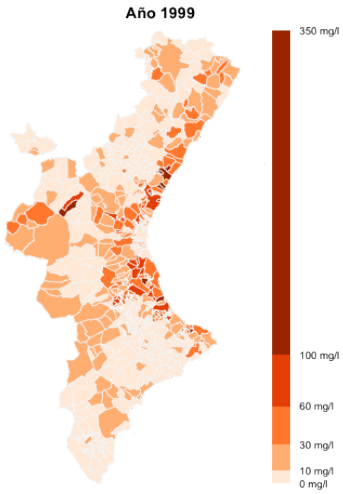
(c) Modelo IET2. Media



(d) Modelo IET2. Mediana

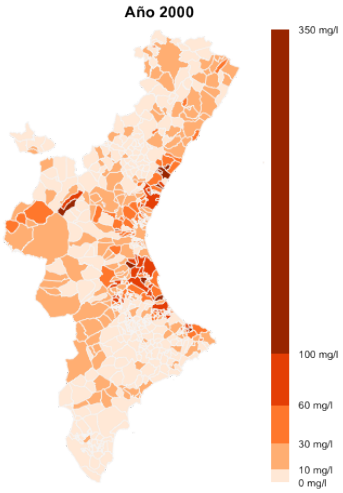


(e) Modelo ARET. Media

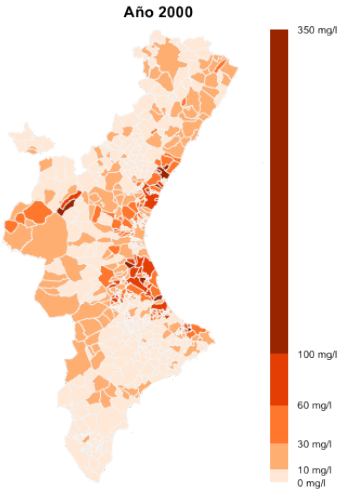


(f) Modelo ARET. Mediana

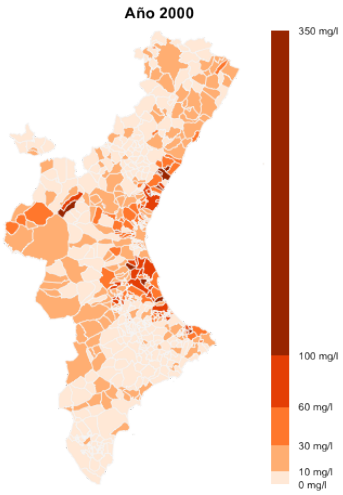
Año 2000



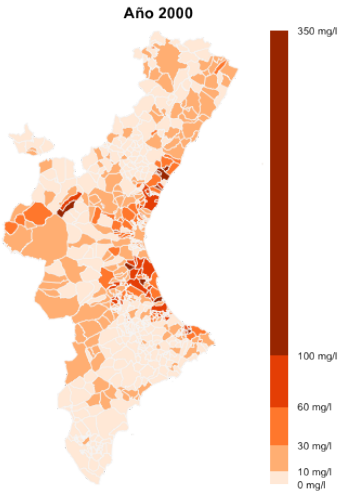
(a) Modelo IET1. Media



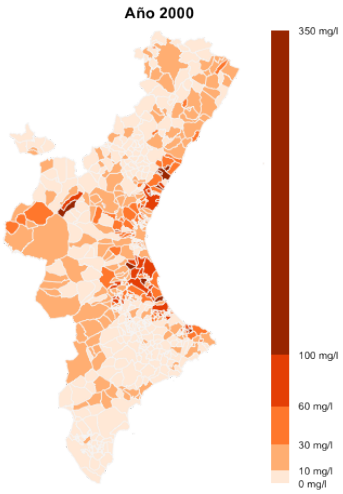
(b) Modelo IET1. Mediana



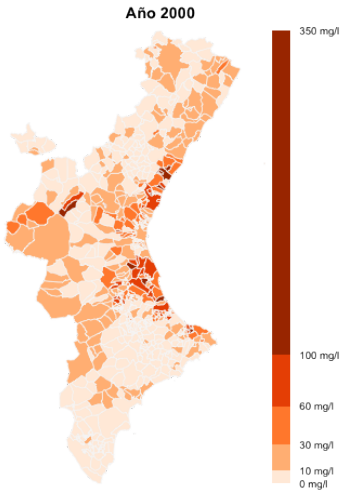
(c) Modelo IET2. Media



(d) Modelo IET2. Mediana



(e) Modelo ARET. Media



(f) Modelo ARET. Mediana

Anexo 3

Modelos de imputación

En este anexo se presenta el código WinBUGS utilizado para la implementación y ajuste de todos los modelos que aparecen en la tesis, tanto los utilizados para el estudio previo de la estructura del modelo de imputación como los modelos comparados motivo de este trabajo.

Modelo H

```
model {
#Modelización de la imputación
 for (i in 1:nregions){
    for (j in 1:ntemp){
       y[i,j] ~ dlnorm(mu[i,j],xi)
       mu[i,j] <- c + h[i]
    }
    #Heterogeneidad
    h[i] ~ dnorm(0,zeta)
 }
#Interceptación
 c ~ dflat()
#Precisión de los datos
 xi ~ dgamma(2,0.01)
#Precisión de la heterogeneidad
 zeta ~ dgamma(5,0.01)
}
```

Modelo T

```
model {
#Modelización temporal de la imputación
 for (i in 1:nregions){
    for (j in 1:ntemp){
       y[i,j] ~ dlnorm(mu[i,j],xi)
       mu[i,j] <- c + h[i] + t[j]
    }
    #Heterogeneidad
    h[i] ~ dnorm(0,zeta)
 }
#Interceptación
 c ~ dflat()
#Componente temporal común
 t[1:ntemp] ~ car.normal(tiempo[], p[], nvectemp[], eta)
#Precisión del magnesio
 xi ~ dgamma(2,0.01)
#Precisión de la heterogeneidad
 zeta ~ dgamma(5,0.01)
#Precisión de la componente temporal
 eta ~ dgamma(2,0.01)
}
```

Modelo ET

```
model {
#Modelización espacio temporal de la imputación
 for (i in 1:nregions){
    for (j in 1:ntemp){
       y[i,j] ~ dlnorm(mu[i,j],xi)
       mu[i,j] <- c + h[i] + t[j] + e[i]
    }
    #Heterogeneidad
    h[i] ~ dnorm(0,zeta)
 }
#Interceptación
 c ~ dflat()
#Componentes espaciales
 e[1:nregions] ~ car.normal(adj[], pesos[], num[], varsigma)
#Componente temporal común para todos los municipios
 t[1:ntemp] ~ car.normal(tiempo[], p[], nvectemp[], eta)
#Precisión de los datos
 xi ~ dgamma(2,0.01)
#Precisión de las componentes espaciales
 varsigma ~ dgamma(2,0.01)
#Precisión de la heterogeneidad
 zeta ~ dgamma(5,0.01)
#Precisión de la componente temporal
 eta ~ dgamma(2,0.01)
}
```

Modelo TC

```
model {
#Modelización temporal de la imputación
 for (i in 1:nregions){
    for (j in 1:ntemp){
       y[i,j] ~ dlnorm(mu[i,j],xi)
       mu[i,j] <- c + h[i] + t[i,j]
    }
    #Heterogeneidad
    h[i] ~ dnorm(0,zeta)
 }
#Interceptación
 c ~ dflat()
#Componentes temporales específicas por municipio
 for (i in 1:nregions){
    t[i,1:ntemp] ~ car.normal(tiempo[], p[], nvectemp[], eta[i])
 }
#Precisión de los datos
 xi ~ dgamma(2,0.01)
#Precisión de la heterogeneidad
 zeta ~ dgamma(5,0.01)
#Precisiones de las componentes específicas por municipio
 for (i in 1 : nregions) {
    eta[i] ~ dgamma(2,0.01)
 }
}
```

Modelo ETC

```
model {
#Modelización espacio temporal de la imputación
 for (i in 1:nregions){
    for (j in 1:ntemp){
       y[i,j] ~ dlnorm(mu[i,j],xi)
       mu[i,j] <- c + h[i] + t[i,j] + e[i]
    }
    #Heterogeneidad
    h[i] ~ dnorm(0,zeta)
 }
#Interceptación
 c ~ dflat()
#Componentes espaciales
 e[1:nregions] ~ car.normal(adj[], pesos[], num[], varsigma)
#Componentes temporales específicas por municipio
 for (i in 1:nregions){
    t[i,1:ntemp] ~ car.normal(tiempo[], p[], nvectemp[], eta[i])
 }
#Precisión de los datos
 xi ~ dgamma(2,0.01)
#Precisión de las componentes espaciales
 varsigma ~ dgamma(2,0.01)
#Precisión de la heterogeneidad
 zeta ~ dgamma(5,0.01)
#Precisión de las componentes temporales específicas por
municipio
 for (i in 1 : nregions) {
    eta[i] ~ dgamma(2,0.01)
 }
}
```

Modelo IET1

```
model {

#La verosimilitud
  for (i in 1:N) {
    for (j in 1:T) {
       nit[i,j] ~ dlnorm(mu[i,j], tau.nit)
#Modelización de la media para cada municipio y año
       mu[i,j] <- inter+s[i]+t[j]+st[i,j]
    }
  }

#Distribuciones previas:
  #Interceptación:
    inter ~ dnorm(0, 0.001)
#Bloque espacial:
  for (i in 1:N) {
    s[i] ~ dnorm(mu.s[i], tau.s)
  }
  mu.s[1:N] ~ car.normal(adj[],weights[],num[],tau.mu.s)
#Bloque temporal:
  for(j in 1:T) {
    t[j] ~ dnorm(mu.t[j], tau.t)
  }
  mu.t[1:T] ~ car.normal(adj.t[],weights.t[],num.t[],tau.mu.t)
#Interacción:
  for(i in 1:N) {
    st[i,1:T] ~ car.normal(adj.t[],weights.t[],num.t[],tau.st)
  }

#Distribuciones previas para los hiperparámetros:
  tau.nit ~ dgamma(0.5, 0.0005)
  tau.s ~ dgamma(0.5, 0.0005)
  tau.t ~ dgamma(0.5, 0.0005)
```

$$
\begin{aligned}
&\quad \text{tau.st} \sim \text{dgamma}(0.5,\ 0.0005)\\
&\quad \text{tau.mu.s} \sim \text{dgamma}(0.5,\ 0.0005)\\
&\quad \text{tau.mu.t} \sim \text{dgamma}(0.5,\ 0.0005)\\
&\}
\end{aligned}
$$

Modelo IET2

```
model {

  for (i in 1:N) {
    for (j in 1:T) {
       nit[i, j] ~ dlnorm(mu[i, j], tau.nit)
#Modelización de la media para cada municipio y año
mu[i, j]<- inter + s[i] + t[j] + st[i,j]
    }
  }

# Distribuciones previas:
  inter ~ dnorm(0, 0.001)
# Bloque espacial:
  for (i in 1:N){
    s[i] ~ dnorm(mu.s[i], tau.s)
  }
  mu.s[1:N] ~ car.normal(adj[],w[],num[],tau.mu.s)
# Bloque temporal:
  for (j in 1:T){
    t[j] ~ dnorm(mu.t[j], tau.t)
  }
  mu.t[1:T] ~ car.normal(adj.t[],w.t[],num.t[],tau.mu.t)
# Interacción espacial y temporal:
  for (i in 1:N) {
    for (j in 1:T) {
       st[i,j]<- v.d[((i-1)*10+j)]
    }
 }
  v.d[1:(N*T)] ~ car.normal(adj.v[],w.v[],num.v[],tau.st)

# Distribuciones previas para los hiperparámetros:
  tau.nit ~ dgamma(1, 0.02)
  tau.s ~ dgamma(1, 0.02)
```

```
    tau.t ~ dgamma(1, 0.02)
    tau.st ~ dgamma(1, 0.02)
    tau.mu.s ~ dgamma(1, 0.02)
    tau.mu.t ~ dgamma(1, 0.02)
  }
```

Modelo ARET

```
model {
  for(i in 1:nmuni){
    for(j in 1:nperiods){
      nit[j,i]~dlnorm(mu[j,i],prec.nit)
#Modelización de la media para cada municipio y año
      mu[j,i]<- mediainter + inter[j] + theta.ST[j,i]
    }
  }
#Distribuciones previas
#Efecto espacio-temporal para el primer año:
  theta.S[1,1:nmuni]~car.normal(map[],w[],nvec[],prec.spat)
  for(i in 1:nmuni){
    BYM[1,i]~dnorm(theta.S[1,i],prec.het)
  }
  for(i in 1:nmuni){theta.ST[1,i]<-pow(1-ro*ro,-0.5)*BYM[1,i]}
#Efecto espacio-temporal para los años siguientes:
  for(j in 2:nperiods){
    for(i in 1:nmuni){
      theta.ST[j,i]<-ro*theta.ST[j-1,i]+BYM[j,i]
      BYM[j,i]~dnorm(theta.S[j,i],prec.het)
    }
    theta.S[j,1:nmuni]~car.normal(map[],w[],nvec[],prec.spat)
  }
#Distribución previa para la media de todos los municipios y
años
  mediainter~dnorm(0,0.01)
#Distribución previa para la tendencia temporal global:
  inter[1:nperiods]~car.normal(mapT[],wT[],nvecT[],prec.inter)
#Distribución previa para los hiperparámetros:
  prec.nit~dgamma(0.5,0.005)
  prec.inter~dgamma(0.5,0.005)
  prec.het~dgamma(0.5,0.005)
  prec.spat~dgamma(0.5,0.005)
```

#Distribución previa para el parámetro de dependencia temporal:

ro∼dunif(-1,1)

}