Trabajo Fin de Máster - Curso 2021/2022

Protocolo para evaluar la modelización de los índices de biomasa relativos

Autora: **Alba Fuster Alonso**

Tutora: Maria Grazia Pennino
Tutor académico : David Conesa Guillen

Máster en Bioestadística

Agradecimientos

Mis más sinceros agradecimientos a todas las personas que han estado y siguen estando implicadas en el desarrollo de este trabajo.

Primero, agradecer a todo el profesorado que he tenido a lo largo de mi carrera académica por cada minuto de clase impartida. A fin de cuentas, son ellos los que me han proporcionado las herramientas necesarias para ir logrando las metas que me he propuesto.

También, he de agradecer al Centro Oceanográfico de Vigo (IEO-CSIC) por brindarme la oportunidad de realizar prácticas en sus instalaciones, sobretodo, agradecer la acogida de todo el personal investigador y el ambiente agradable de trabajo. He de hacer mención especial al equipo MERVEX, ya que, siempre han estado ahí, con paciencia infinita, resolviendo las dudas y problemas que iban surgiendo o simplemente acompañándome durante el camino, gracias Fran Izquierdo, Marta Cousido, Santiago Cerviño, Amina Tifoura y Maria Grazia Pennino, sin vosotros mi estancia en el centro y este trabajo no habrían sido lo mismo. También, agradecer a Iosu Paradinas por haberme dedicado su tiempo y haber estado disponible siempre que he necesitado su ayuda.

Además, he de destacar el esfuerzo y la dedicación que mis dos tutores, Maria Grazia Pennino y David Conesa, han puesto en este trabajo y en mi formación. Desde el primer momento, ellos han sido mi motor, no solo en el TFM, sino en mis primeros pasos en el mundo de la investigación. Como decía Isaac Newton "si he visto más lejos, ha sido al pararme sobre los hombros de gigantes que me han guiado".

Por último, gracias al apoyo de mi padre, madre, hermano, tía Concha y Jorge, han sido pilares fundamentales durante el desarrollo de este TFM.

Este trabajo se lo dedico a mis abuelos y abuelas, a los que les debo mucho y estaré eternamente agradecida. En especial, quiero dedicar este trabajo a mi tía María, nadie me ha dedicado más tiempo que tú, por ello, no encuentro las palabras adecuadas con las que agradecerte, simplemente te dedicaré cada logro, porque sé que será gracias a ti.

Resumen

El medio marino ha sido objeto de explotación para satisfacer las necesidades humanas desde los inicios de la humanidad. En consecuencia, las actividades pesqueras han ido generando un impacto sobre el ecosistema, de forma que multitud de hábitats marinos se han ido deteriorando, y cada vez más especies de interés pesquero se encuentran en un estado crítico de conservación. De ahí nace la necesidad de **evaluar** el estado de las poblaciones de especies marinas de interés pesquero, con el fin de implantar medidas sobre la actividad pesquera y evitar así el colapso de los sistemas marinos.

La evaluación de los *stocks* [1] pesqueros recae sobre los científicos marinos, o especialistas en ciencias marinas. Sin embargo, la figura responsable de tomar decisiones sobre el *stock* es el gestor. Normalmente, los gestores siguen el consejo proporcionado por los expertos, y en base a este toman decisiones sobre la conservación del *stock* y las actividades pesqueras. Pero, puede ocurrir que el consejo científico sea erróneo o mejorable y, por consiguiente, las medidas para gestionar el *stock* no estén siendo las adecuadas, pudiendo provocar que la especie acabe fuera de los límites sostenibles de explotación.

En vista a la importancia del asesoramiento científico, surge el proyecto **IMPRESS** *Improving scientific advice to fishery management for resources of interest for Spain in Atlantic waters* en el que se enmarca este trabajo fin de máster (TFM). IMPRESS es un proyecto nacional español, en el que se pretenden afrontar los problemas presentes en la evaluación del estado de las poblaciones de especies marinas de interés pesquero. Por lo tanto, el objetivo general de IMPRESS es mejorar la calidad del asesoramiento científico, con el fin de evitar una mala gestión de los *stocks* pesqueros.

En la actualidad, la evaluación de las pesquerías, se basa en **modelizar** características biológicas gestionables del *stock* a lo largo del tiempo, p.ej. la biomasa de una población de peces. Pero, ¿por qué recurrir a la modelización? Un buen modelo provee al investigador de una visión profunda de los procesos biológicos ligados a especies susceptibles de explotación pesquera, lo que puede ser muy útil a la hora de tomar decisiones sobre su conservación.

[1] Un *stock* se define como una población de peces que está siendo explotada y gestionada.

Existen diferentes tipos de modelos de evaluación del *stock*, uno de los más empleados son los SPMs, del inglés *Surplus Production Models*, cuyo objetivo es estimar la serie temporal de biomasa en un periodo de tiempo determinado. El fin de los modelos SPMs es evaluar cuánta biomasa queda 'disponible' para la pesca, de manera que se mantengan unos límites de explotación sostenibles, y así, evitar que la actividad pesquera provoque el colapso del sistema.

Pero, ¿podemos mejorar la estimación de la biomasa de estos modelos? ¿cómo? Existen infinidad de vías para mejorar un modelo de evaluación del *stock*. Una de las vías para mejorar las estimaciones que proporcionan los modelos SPMs es estudiando la calidad de sus *inputs*, y es aquí donde se centra nuestro trabajo. En sí, estos modelos necesitan alimentarse de dos *inputs*: (1) una serie de capturas y (2) una serie temporal de índices de biomasa relativa o de CPUE (capturas por unidad de esfuerzo), que sea representativa de la biomasa.

Los índices de biomasa relativa o de CPUE provienen, respectivamente, de datos de campañas oceanográficas y de pesquerías. Así pues, estos índices pueden depender de factores ambientales, procesos espacio-temporales subyacentes, etc. Por lo tanto, para conseguir que estos índices sean representativos de la biomasa real del *stock* necesitamos la ayuda de la modelización estadística. Sin embargo, en la literatura nos encontramos con infinidad de modelos y procesos de inferencia y predicción dedicados a conseguir esa serie temporal de índices representativa de la biomasa. Ahora bien, ¿qué modelización de los índices captura mejor el comportamiento de la biomasa? ¿Inferimos y predecimos en frecuentista o en bayesiano?

En base a lo mencionado, nuestro trabajo propone la simulación de un escenario de biomasa (espacio-tiempo), de forma que a partir de la biomasa simulada, podamos reproducir las principales fuentes de información empleadas en la evaluación de pesquerías (índices de biomasa relativa y de CPUE). Es entonces, cuando utilizando dichas fuentes de información, se proponen distintos modelos para los índices de biomasa relativa y de CPUE, que son analizados tanto desde la perspectiva frecuentista como de la bayesiana, permitiendo evaluar qué modelo ha conseguido capturar mejor el comportamiento de la biomasa real del *stock*.

El trabajo se divide en 6 capítulos. El primero está dedicado a contextualizar las pesquerías, matizando en los problemas que derivan de esta actividad, y cómo la estadística puede ser una herramienta muy útil para la gestión de los recursos pesqueros. El segundo capítulo es un marco teórico donde se explica la estadística espacial y temporal. Además, el segundo capítulo presenta la modelización, la inferencia y la predicción en el contexto de la estadística bayesiana y, detalla algunas de las herramientas con las que podemos inferir y predecir en bayesiano. El tercer capítulo expone el protocolo diseñado para poder evaluar la calidad de los *inputs* (índices de biomasa relativa y de CPUE) que alimentan los modelos SPMs. El cuarto capítulo, recoge los resultados más relevantes que se han obtenido en el trabajo. Por último, los capítulos quinto y sexto, se centran en discutir los resultados obtenidos y extraer las principales conclusiones del trabajo, así como, remarcar algunas de las limitaciones con las que nos hemos ido encontrando y las lineas futuras de investigación.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Pesquerías y estadística: problemas y soluciones

En este capítulo se manifiesta la problemática ligada a la sobreexplotación de los recursos pesqueros, a la vez que narramos como surge y evoluciona la evaluación de los *stocks*, y como la estadística puede servir de gran ayuda para entender en profundidad características susceptibles de gestión.

1.1. Sobreexplotación de recursos pesqueros

Según Guerra Sierra y Sánchez Lizaso (1998) la pesca puede definirse como una actividad humana enfocada en la extracción de organismos del medio acuático utilizando distintos instrumentos (redes, cañas, arpones, poteras, etc.). En referencia a los productos derivados de la pesca, desde los comienzos de esta actividad costera, se utilizan fundamentalmente para el consumo humano (Guerra Sierra y Sánchez Lizaso, 1998). Es más, la pesca ha sido y es un recurso fundamental en la historia de la humanidad, proporcionando alimento y empleo a millones de personas (Pauly y Zeller, 2003). Ya en el Paleolítico encontramos evidencias de que los organismos marinos eran importantes en la dieta de los habitantes de zonas costeras (Guerra Sierra y Sánchez Lizaso, 1998).

Sin embargo, a pesar del gran impacto socio-económico de los recursos pesqueros, estos están siendo degradados y sobreexplotados desde hace siglos (Iversen, 1996). De hecho, algunas pesquerías han provocado el colapso de *stocks* pesqueros, llegando a situaciones de no retorno en el sistema (Pauly, 1996). Ya en el siglo XV comenzaban a aparecer signos de declive en algunas especies explotadas en las costas españolas (p.ej. la ballena vasca o franca en el golfo de Vizcaya

o el atún y la sardina). Este escenario forzó a los españoles a buscar nuevos caladeros, llegando a tierras lejanas como Canadá (Guerra Sierra y Sánchez Lizaso, 1998). Otras pesquerías que colapsaron en épocas más recientes son, por ejemplo, la anchoveta de Perú o el arenque del Mar del Norte a principios de los años 80 y, la pesquería del bacalao de Terranova cuya situación todavía no se ha recuperado (Hutchings, 1996; Walters y Maguire, 1996; Shelton y Lilly, 2000). Es más, según la FAO en 1990 un 90 % de los *stocks* pesqueros se encontraban dentro de los niveles sostenibles biológicos de explotación, mientras que, en 2017 únicamente un 65.8 % se encontraba dentro de estos límites sostenibles (FAO, 2020).

Desde el punto de vista socio-económico y biológico, la sobrepesca es un inconveniente (Clark, 2006). En cambio, es este enfoque económico el que incita a malas prácticas pesqueras, y deriva en que la sobreexplotación de los recursos marinos ocurra con cierta frecuencia (Clark, 2006). Por ello, en el siglo XVIII en vista a la proyección económica de la pesca, los políticos ilustrados de la época comenzaron a solicitar informes sobre el estado de las especies marinas explotadas. Es en ese momento cuando nace la Biología Pesquera, encargada de la evaluación de los *stocks* (Guerra Sierra y Sánchez Lizaso, 1998). Así pues, en el siglo XVIII se toman las primeras medidas de regulación, como la prohibición y autorización de ciertos artes de pesca [1] (Guerra Sierra y Sánchez Lizaso, 1998).

A pesar de que en el siglo XVIII ya se comenzaban a evaluar y gestionar los recursos pesqueros y, que la preocupación por el bienestar de los ecosistemas marinos ha traído consigo mejoras en la gestión de los recursos pesqueros, sobretodo en los últimos años (FAO, 2020). Hoy en día, dichas mejoras no han conseguido revertir la tendencia a la sobreexplotación de especies marinas de interés pesquero (FAO, 2020). Incluso, algunos científicos creen que nos enfrentamos a una crisis mundial en la pesca y, que la escasez de estos recursos podría tener consecuencias devastadores para el ser humano (Pauly *et al.*, 2002; Myers y Worm, 2003).

De entre todas las razones que suelen llevar al fracaso la conservación de una especie de interés pesquero, hay dos que destacan y que suelen ser las más frecuentes: (1) la dificultad para detectar la sobreexplotación del *stock* a tiempo y (2) el inconveniente de reducir la presión sobre el *stock* debido a intereses económicos, aunque ya se tenga consciencia de que la situación de la especie es insostenible (Hilborn y Walters, 2013; Cerviño, 2004).

En definitiva, el problema fundamental de la gestión pesquera, es que puede ser un arma de doble filo, puesto que se ha de preservar la salud del ecosistema, pero también la salud de la economía asociada a los recursos (Björndal *et al.*, 2004). Más aún, Larkin (1989) recalca que el problema de la sobrepesca ha de resolverse desde un punto de vista antropocéntrico, de modo que se establezca un equilibrio entra la conservación del ecosistema y el sustento económico que los recursos proporcionan, sustento del que depende la economía de muchos países.

[1] Los artes de pesca son los distintos métodos con los que se realiza la extracción de los organismos marinos, p.ej. palangre, arrastre, nasas, etc.

En la actualidad, la gestión de los recursos pesqueros ha ido evolucionando, de manera que, normalmente, son los científicos marinos, o especialistas en ciencias marinas, los encargados de evaluar el estado del *stock*. A pesar de ello, los científicos no son los responsables de tomar las decisiones sobre el *stock*. Es la figura del gestor, la encargada de tomar las decisiones sobre el recurso pesquero. Es más, los gestores pueden seguir el consejo de los científicos o no. De hecho, ocurre con reiteración que el gestor ignora los informes y consejos proporcionados por los expertos, aunque, también puede ocurrir que el consejo científico sea erróneo o mejorable Cerviño (2004), y es aquí donde se centra el interés de nuestro trabajo.

1.2. La evaluación de *stocks*

La evaluación de los recursos pesqueros ha ido ganando fuerza con los años por una cuestión de necesidad, al ver como los *stocks* de especies marinas están siendo mermados por las malas prácticas relacionadas con la pesca y la gestión (King, 2013). Conceptualmente, la evaluación de pesquerías consiste en adquirir la información biológica sobre el estado presente, pasado y futuro del *stock*, a fin de aconsejar a los gestores sobre la regulación de las distintas pesquerías, y procurar que las capturas de una especie de interés pesquero se mantengan dentro de los límites de explotación sostenibles (Cerviño, 2004). Pero, ¿cómo podemos evaluar el estado de un *stock*? En si mismo, para poder evaluar un *stock* necesitamos conocer en profundidad aquellas características biológicas que puedan ser gestionables, p.ej. la biomasa, la estructura de edades, la estructura de tallas, la abundancia, etc. Y como no, una estrategia para adquirir dicha información biológica es a través de la modelización. Conceptualmente, la modelización estadística, se basa en intentar reproducir la realidad de un sistema, de forma que entendamos en detalle el comportamiento del mismo.

Uno de los primeros estudios en los que utilizaron técnicas estadísticas para conocer características de una especie de interés pesquero fue desarrollado por Hjort *et al.* (1933) en la pesquería de ballenas. En concreto, utilizaron un método capaz de calcular el tamaño de la población equilibrando el número de nacimientos y el de muertes por causas naturales. A partir de aquí, se fueron desarrollando modelos y ecuaciones matemáticas, cada vez más complejos, que nos permiten valorar el estado de los *stocks* pesqueros.

En general, en el campo de la gestión pesquera, la modelización estadística ha sido y es una herramienta fundamental para los científicos en la evaluación de un *stock* pesquero (Peterman, 1990). Actualmente, los modelos que se aplican en la evaluación dependen, sobretodo, de la información de la que disponemos sobre la especie. En sí, los modelos de evaluación del *stock* pueden agruparse en cuatro categorías:

1. *Models for data-poor stocks*. Este tipo de modelos utiliza información sobre la estructura

de tallas (longitudes de distintos especímenes) y algunos parámetros biológicos de la especie (*life-traits*).

2. *Surplus production biomass models* (SPMs). Estos modelos se centran en la estimación de la biomasa real del *stock*. Así pues, requieren de dos *inputs*, la serie de capturas de la especie de interés y una serie temporal de índices de biomasa relativa o de CPUE representativa de la biomasa.

3. *Age-structured population models*. Este enfoque consiste en analizar cada una de la cohortes [2] que componen el *stock*.

4. *Integrated models*. Este último tipo de modelo es más flexible en lo que a *inputs* se refiere, puede diferenciar por tallas, edad, añadir una estructura espacial, etc. Pero, también necesita disponer de más información en comparación con el resto de modelos.

En definitiva, dependiendo de la información de la que disponemos podremos aplicar distintos modelos de evaluación del *stock*. Pero, ¿de dónde se consigue la información para obtener estos *inputs*? Por lo común, podríamos diferenciar los datos sobre especies de interés pesquero como aquellos que provienen de la pesca y aquellos que provienen de campañas oceanográficas. A continuación, enumeramos las fuentes de datos más características con las que podemos alimentar los modelos de evaluación del *stock*, a la vez que remarcamos la información espacial asociada a cada tipo de fuente:

1. Datos de desembarques. En principio, los capitanes de todos los buques pesqueros que lleguen a puerto con productos que sean susceptibles a la venta en lonja han de declararlos en el momento del desembarque. Las declaraciones en puertos proporcionan información sobre el peso y la especie, sin embargo, no se tiene información espacial sobre la pesca, únicamente el puerto de desembarque.

2. Datos del cuaderno de bitácora. La actividad pesquera de un buque queda recogida en los cuadernos de bitácora o *logbook*. Dichos cuadernos contienen mucha información, como puede ser el histórico de capturas (en algunos cuadernos se incluyen los descartes), el esfuerzo de pesca en cada operación, información sobre el propio buque (nombre, nacionalidad, número de registro, etc.), el tipo de arte de pesca con algunas especificaciones (p.ej. tamaño de la red), la fecha en la que se realiza el lance de pesca, el puerto de desembarque y en algunos casos se añade el lugar de captura. Por tanto, es posible conseguir información espacial, que puede venir dada como una zona de pesca (división de la región en rectángulos) o, si el pescador lo desea, mediante la georreferenciación del lance (localización exacta del punto en el que se ha pescado).

[2] Una cohorte, se corresponde a todos los peces nacidos en el mismo período, normalmente dentro de un mismo año.

3. *Vessel Monitoring System* (VMS). Los datos de capturas obtenidos por el cuaderno de bitácora pueden ser asociados al sistema VMS, el cual reporta la posición del buque. Normalmente, este tipo de datos tiene una mayor precisión geográfica que los anteriores mencionados.

4. Observadores a bordo. Los observadores a bordo son personal científico-técnico que se encarga de controlar y documentar las capturas que se realizan en los distintos pesqueros. Este tipo de datos son de gran calidad y riqueza, incluso pueden contener datos biológicos de muestras de peces (determinaciones de edad y sexo, mediciones de longitud y peso, etc.). De hecho, los datos de observadores a bordo proporcionan información georreferenciada sobre las capturas y la biología de las especies de interés pesquero.

5. Campañas oceanográficas. Hasta ahora hemos hablado de datos derivados de las pesquerías. Sin embargo, cabe destacar que durante todo el año parte del personal investigador pasa periodos en el mar para recopilar datos sobre especies de interés pesquero (longitud, peso, talla, sexo, georreferenciación de los lances, etc.).

A partir de las fuentes de información mencionadas, vamos a profundizar en cómo la evaluación de un *stock* pesquero utiliza dicha información derivada de pesquerías y de campañas oceanográficas como *input* en los distintos modelos de evaluación, concretamente en SPMs.

1.3. *Inputs* ligados a SPMs: índices de biomasa relativa o de CPUE

Tal y como hemos comentando, este trabajo se centra en valorar la calidad de los *inputs* que alimentan los modelos SPMs. El motivo para su elección, es que, los *Surplus production biomass models* se suelen considerar el método más completo de evaluación del *stock* cuando la información sobre el mismo es limitada, puesto que son el único método que proporciona una evaluación completa del *stock* [3] (Cousido-Rocha *et al.*, 2022).

Así pues, remarcar que, los *inputs* que alimentan los SPMs se basan en: (1) una serie de capturas, se trata de una recopilación de capturas de todas las flotas dedicadas a la pesca de una especie en concreto y (2) una serie temporal de índices de biomasa relativa o de capturas por unidad de esfuerzo (CPUE). De hecho, este último *input* se basa en seleccionar aquellas flotas de las que podamos tener más información sobre la especie o, una campaña oceanográfica,

[3] En una evaluación completa del *stock* se obtienen los puntos de referencia, es decir, los límites para una explotación sostenible del recurso.

y modelizar las capturas de manera que ajustemos una serie temporal de biomasa relativa o de CPUE representativa de la biomasa real en un periodo de tiempo (Cousido-Rocha *et al.*, 2022).

De los dos *inputs* necesarios para llevar a cabo la evaluación del *stock* con un SPMs, nos vamos a enfocar en cómo obtener la serie temporal de índices de biomasa relativa derivados de campañas oceanográficas y de los índices de CPUE derivados de las pesquerías. En sí, la estimación de una serie temporal de biomasa relativa o de CPUE, que sea representativa de la biomasa real del *stock* trae consigo una modelización que se ha de examinar y valorar, ya que ajustar estos índices puede ser todo un reto.

Por lo que respecta a los índices de biomasa relativa, que provienen de campañas oceanográficas bien diseñadas, suelen ser proporcionales a la biomasa real, y por ende, precisos (Maunder *et al.*, 2020). Igualmente, estos índices pueden verse afectados por distintos factores ambientales (batimetría, clorofila, temperatura, etc.) o presentar un proceso espacio-temporal subyacente, siendo conveniente su modelización previa a la inferencia y predicción. Por otra parte, no siempre disponemos de información proveniente de campañas oceanográficas. Por lo tanto, muchas evaluaciones de *stocks* utilizan índices de CPUE derivados de la pesca para obtener la serie temporal. A diferencia de los índices de biomasa relativa derivados de campañas oceanográficas, los índices de CPUE son dependientes de la pesca, de modo que pueden estar afectados por diversidad de factores y por el propio muestreo, lo que complica la modelización.

Normalmente la modelización de ambos índices, tanto los índices de biomasa relativa provenientes de campañas como los índices de CPUE, utilizan *Generalized Linear Models* GLMs o *Generalized Additive models* GAMs para intentar paliar los efectos de la dependencia de la pesca u otros factores, y conseguir una estimación representativa de la serie temporal de biomasa para alimentar a los modelos de evaluación. No obstante, Maunder *et al.* (2020) advierten de la necesidad de emplear modelos más complejos, como son los modelos espacio-temporales geoestadísticos, para poder estimar de un modo preciso las series temporales derivadas de estos índices, sobretodo, los índices de CPUE.

Zhou *et al.* (2019) apuntaban que los modelos espacio-temporales resueltos mediante aproximaciones basadas en geostadística, en comparación con modelos GLMs o GAMs, mejoraban sustancialmente la estimación de la serie temporal de biomasa. Así mismo, Stock *et al.* (2020) comparaban entre distintos tipos de modelos para predecir los descartes en pesquerías. Los resultados de esta investigación concluyeron que es preferible utilizar modelos espacio-temporales (geostadística) y *Random Forest* RF, en lugar de modelos GLM o GAM, de los cuales el GLM fue el que obtuvo peores resultados.

En resumen, en los últimos años la introducción de modelos más complejos para la estimación de las series temporales de los índices de biomasa relativa y de CPUE ha ido haciéndose notar. Uno de los primeros trabajos que utilizó modelos espacio-temporales geoestadísticos fue el de Thorson *et al.* (2015), donde se ajustaba la serie temporal de índices de biomasa relativa

con un Delta-GLMM para un total de 28 especies demersales de la costa oeste de Estados Unidos. Del mismo modo, Cavieres y Nicolis (2018) plantean un modelo espacio-temporal geoestadístico para la langosta amarilla (*Cervimunida johni*). Más tarde, Grüss y Thorson (2019) emplearon un modelo Delta-GLMM, parecido al propuesto por Thorson *et al.* (2015), para el Pargo rojo (*Lutjanus campechanus*) del Golfo de México. En pocas palabras, un buen número de trabajos relacionados con la evaluación de *stocks* han comenzado a manejar modelos complejos en los que se trabaja con efectos espacio-temporales de modelos geoestadísticos (Thorson y Barnett, 2017; Tremblay-Boyer *et al.*, 2017, 2018; Kai, 2019; Xu *et al.*, 2019).

1.4. Motivación y objetivos

En vista a todo lo mencionado, la modelización de la serie temporal de índices de biomasa relativa y de CPUE es todo un desafío, y existen infinidad de propuestas, desde modelos más sencillos, como puede ser un GLM, hasta modelos complejos, como puede ser un modelo espacio-temporal geoestadístico. Por ello, nuestro trabajo va a enfocarse en valorar, a través de distintas modelizaciones de los índices de biomasa relativa y de CPUE, qué modelo captura mejor el comportamiento de la biomasa real del *stock*. En otras palabras, ¿son realmente necesarios modelos más complejos que incluyan efectos espacio-temporales? ¿Se llega a las mismas conclusiones de modelado con los índices derivados de campañas oceanográficas que con los índices provenientes de pesquerías? ¿Qué modelo consigue representar mejor la biomasa? ¿Inferimos y predecimos sobre los parámetros del modelo en frecuentista o en bayesiano? Todas estas preguntas y más fijan una serie de objetivos que marcan la motivación de este trabajo.

El objetivo general o principal del trabajo consiste en:

- La elaboración de un protocolo que nos permita determinar qué modelización de los índices de biomasa relativa e índices de CPUE consigue la predicción que mejor representa el comportamiento de la biomasa simulada del *stock*.

Atendiendo al objetivo principal, se pueden desglosar una serie de objetivos específicos que se abordarán en el presente trabajo:

- Abordar la simulación de una escenario de biomasa y la reproducción de las principales fuentes de información es pesquerías (índices de biomasa relativa y de CPUE).
- Proponer una serie de modelos, donde los índices de biomasa relativa y de CPUE son la variable de interés, que serán ajustados en el contexto frecuentista y bayesiano para poder comparar los resultados.

- Modelizar los índices de CPUE derivados de la actividad pesquera como un patrón puntual marcado preferencial en el espacio y el tiempo, para luego inferir y predecir sobre sus parámetros.

Para concluir con la motivación, abordar los objetivos propuestos requerirá de elaborar un protocolo que se basará en los siguientes pasos:

1. Para poder valorar qué *input* ha conseguido una mejor estimación de la biomasa, hay que conocer la biomasa real en un espacio y tiempo determinado, por consiguiente, se propone la simulación de un modelo de biomasa.

2. A continuación, a partir de la biomasa simulada recreamos dos escenarios que reproduzcan datos ligados a campañas oceanográficas, donde a partir de un muestreo aleatorio (independiente de la pesca) obtenemos los índices de biomasa relativa y, datos ligados a las pesquerías, donde a partir de un muestreo preferencial (dependiente de la pesca) obtenemos los índices de CPUE.

3. Una vez tenemos los bancos de datos ya podemos modelizar los índices de biomasa relativa o de CPUE para, inmediatamente después, inferir y predecir sobre los parámetros del modelo.

4. Por último, las predicciones de los índices de biomasa relativa y de CPUE se compararán con la serie de biomasa simulada, con el fin de comprobar si han conseguido capturar el comportamiento de la biomasa del *stock*.

Capítulo 2

Marco teórico

En el siguiente capítulo, se desarrolla el marco teórico que hay detrás del protocolo a seguir en este trabajo para alcanzar nuestros objetivos. Como ya se ha citado, nuestro trabajo se centra en cómo modelizamos los índices de biomasa relativa y de CPUE, ya que en la literatura la modelización de estos índices puede variar desde modelos más sencillos como es un GLM hasta modelos más complejos como son los geoestadísticos.

En primer lugar, se explicará qué es un **modelo estadístico**, abordando algunos de los modelos a los que uno puede enfrentarse en el contexto de las pesquerías. Una vez, entendemos qué es un modelo y contemplamos el abanico de posibilidades que nos abre la modelización de una variable, hay que **inferir y predecir** sobre los parámetros del modelo. En general, existen dos perspectivas principales con las que llevar a cabo un proceso inferencial y predictivo: (1) la inferencia frecuentista o (2) la **inferencia bayesiana**. En este trabajo, se abordarán las diferencias entre ambas, profundizando en la estadística bayesiana. Igualmente, la inferencia bayesiana puede emplear diferentes técnicas de aproximación, p.ej. *Markov chain Monte Carlo* MCMC o INLA (*Integrated Nested Laplace Approximation*). Matizamos en qué es INLA y el por qué hemos utilizado esta herramienta para resolver la inferencia y predicción en el contexto bayesiano.

Por otro lado, Maunder *et al.* (2020) exponen la necesidad de incluir en los modelos la variabilidad espacial asociada a los índices de biomasa relativa y, sobretodo, a los índices de CPUE derivados de pesquerías. El motivo por el que se propone contemplar la dependencia espacial, es que, en pesquerías los fenómenos a modelizar no están controlados en un laboratorio, es decir, los procesos suelen cambiar a lo largo de una región. Este hecho, seguramente conlleve una dependencia espacial, que al menos, habrá que plantear cuando queramos explicar mediante un modelo la variable de interés en cuestión. Como resultado, dedicamos una sección al contexto teórico de **la estadística espacial**.

Del mismo modo, los bancos de datos de los que dispone el investigador en pesquerías suelen contener observaciones a lo largo de un periodo de tiempo determinado. Es más, cuando hablamos de la evaluación de un *stock* siempre se trabaja con series de tiempo. En consecuencia, se dedica una sección a contextualizar las **series temporales**, más concretamente, al ser objeto de nuestro estudio los comportamientos autorregresivos en el tiempo.

2.1. Modelización

Un modelo consiste en una representación a pequeña escala de la realidad, de manera que se pretende describir el comportamiento de una variable aleatoria a través de una ecuación matemática. En concreto, los modelos estadísticos son modelos que nos permiten incorporar la variabilidad presente en la vida real utilizando el azar (Figura 2.1). Si el modelo consigue capturar la naturaleza de la variable puede darnos una visión profunda de la misma, siendo muy útil para resolver problemas o cuestiones. Sin embargo, no es trivial encontrar un modelo que describa a la perfección la vida real.

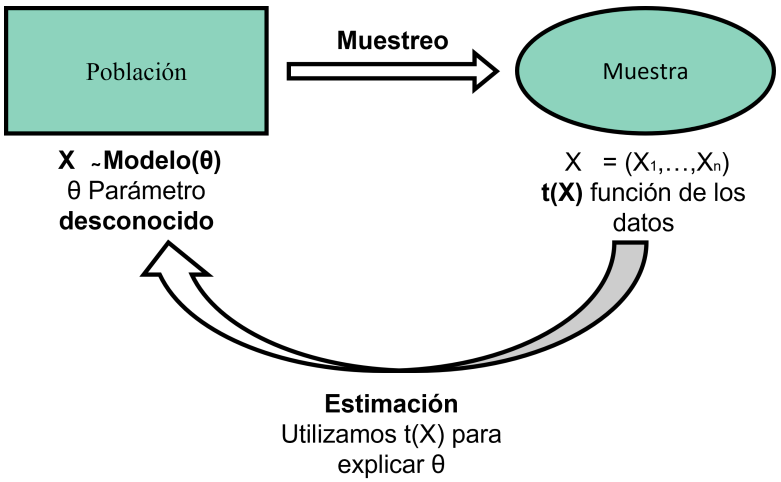

Figura 2.1: Proceso estadístico.

Box dijo "Esencialmente, todos los modelos son incorrectos, pero algunos pueden ser útiles", incluso Einstein dijo "La formulación del problema es más esencial que la solución en sí misma, la cual puede ser simplemente una cuestión de habilidades matemáticas o experimentales". Esto ya nos da una idea de lo importante que es partir de un buen modelo, porque sin eso, no estaremos capturando el comportamiento del fenómeno que queremos modelar y llegaremos a conclusiones erróneas. Así lo deja por escrito Anabel Forte en su blog BAYESANA, "El problema es que muchas veces nos centramos tanto en la búsqueda de la respuesta que nos

olvidamos de formular la pregunta correcta. No somos conscientes que un cuestionamiento erróneo nos lleva a respuestas sin utilidad real o incluso completamente equivocadas".

Así pues, la modelización estadística abarca desde modelos más simples como puede ser un modelo lineal, donde asumimos que la variable tiene una distribución de probabilidad Normal y los efectos asociados tienen un comportamiento lineal, hasta modelos más complejos como pueden ser los modelos espacio-temporales, donde se incluye la variabilidad espacial y temporal del proceso. ¿Cómo escogemos que modelo explica mejor la variable de interés? Esto es una cuestión de entender el proceso que queremos modelar. En esta sección nos centramos en la teoría de tres tipos de modelos que se abordaran más adelante: modelos lineales generalizados o modelos lineales generalizados mixtos (GLMs y GLMMs del inglés *Generalized Linear Models* y *Generalized Linear Mixed Models*, respectivamente), modelos aditivos generalizados o modelos aditivos generalizados mixtos (GAMs y GAMMs del inglés *Generalized Additive Models* y *Generalized Additive Mixed Models*, respectivamente) y modelos espacio-temporales.

Modelos lineales generalizados (GLMs o GLMMs)

Los modelos lineales asumen que la variable respuesta tiene un comportamiento Normal, esta asunción conlleva que la variable de interés es cuantitativa, se mueve en un intervalo continuo de $(-\infty, +\infty)$ y que los datos presentan una curva simetría. No obstante, a priori podemos pensar en una infinidad de variables a modelar que no cumplen estas condiciones, p.ej. conteos de enfermedades, estudios de sí o no, variables positivas y con curvas asimétricas, etc. En definitiva, un modelo lineal no es siempre una buena opción.

De esta problemática nacen los GLMs. Este tipo de modelos nos permiten asociar a la variable respuesta una distribución de probabilidad que no sea Normal, p.ej. una distribución Bernoulli para estudios de sí o no, una Poisson para conteos, una Gamma para variables positivas y asimétricas, etc. La clave de un GLM reside en la función de enlace, que relaciona el predictor lineal con la esperanza de la distribución. Además, podemos complicar el modelo incluyendo efectos aleatorios en el predictor lineal (GLMMs), estos son factores que pueden suponer una fuente de variabilidad y han de ser incluidos para controlarla. Por ejemplo, en estudios donde la variable tiene asociada una distribución de probabilidad con un solo parámetro es muy útil para evitar la sobredispersión del parámetro.

Supongamos una variable aleatoria X, con media $\mu_i = (\mu_1, ..., \mu_n)$. Cada μ_i puede ser enlazada al predictor lineal con una función de enlace $g(\cdot)$, tal que, $g(\mu_i) = \eta_i$:

$$\begin{aligned} \eta_i &= \beta_0 + \sum_m^M \beta_m Y_{m_i} + a_j, \\ a_j &\sim \text{Normal}(0, \sigma_{a_j}), \end{aligned} \tag{2.1}$$

donde, η_i es la función de enlace para los parámetros de la distribución de la variable respuesta, β_0 se corresponde con el intercepto, $\beta = \{\beta_1, ..., \beta_M\}$ hace referencia a los coeficientes de las covariables $Y = (Y_1, ..., Y_M)$ y a_j se corresponde con una efecto aleatorio distribuido como una Normal de media cero y desviación típica σ_{a_j}.

Modelos aditivos generalizados (GAMs o GAMMs)

Los GLMs o GLMMs únicamente permiten relaciones paramétricas entra la variable respuesta y las variables explicativas. Por el contrario, en los GAMs o GAMMs pueden asumirse relaciones no paramétricas a través de funciones suaves [1]. Algunos métodos de suavizado son la regresión polinómica local, el suavizado con *kernels* y el más utilizado los *splines*. Este último es más flexible y suave que otras técnicas y se comporta mejor en relación a la extrapolación (Yee, 2015). En el caso anterior, podríamos suponer relaciones no paramétricas en el predictor utilizando suavizado con *splines*. Así mismo, es posible combinar funciones suaves (suavizado bivariante) entre dos variables. Esto es muy útil para intentar incluir parte de la variabilidad que podemos tener en el espacio:

$$\eta_{ij} = \beta_0 + \sum_{m}^{M} \beta_m Y_{mij} + \sum_{l=1}^{L} f_l(Z_{lij}) + a_{ij}, \tag{2.2}$$
$$a_{ij} \sim \text{Normal}(0, \sigma_{a_{ij}}),$$

donde, η_i es la función de enlace para los parámetros de la distribución de la variable respuesta, β_0 se corresponde con el intercepto, $\beta = \{\beta_1, ..., \beta_M\}$ hace referencia a los coeficientes de las covariables $Y = (Y_1, ..., Y_M)$, $f = \{f_1(\cdot), ..., f_L(\cdot)$ hace referencia a las funciones suaves aplicadas a las covariables $Z = Z_1, ..., Z_L$ y a_j se corresponde con un efecto aleatorio distribuido como una Normal de media cero y desviación típica σ_{a_j}.

Modelos espacio-temporales

Para finalizar, muchas veces no es suficiente con los GLMMs o incluso con los GAMMs para capturar el comportamiento de un fenómeno si este tiene un proceso espacio-temporal subyacente. Por ello, necesitamos recurrir a técnicas más complejas relacionadas con la estadística espacial.

La estadística espacial es la rama de la estadística que aborda procesos en los que existe una **dependencia** a lo largo de una región (Ripley, 2005). Esta dependencia en el espacio se ve reflejada en los datos que recogemos cuando hacemos un muestreo. En función de cómo

[1] Se dice que una función es suave, o de clase C^∞, si todas sus derivadas, de cualquier orden, existen.

se presente la información se puede diferenciar entre distintos tipos de datos espaciales, de forma que esta diferenciación será clave para entender el fenómeno y modelar, inferir y predecir correctamente sobre los parámetros del modelo.

En vista a la importancia de la estadística espacial, dedicaremos una sección a diferenciar entre los tipos de datos espaciales con los que podemos encontrarnos en la naturaleza y sus características más relevantes.

2.2. Estadística espacial

En estadística espacial tenemos tres tipos de datos espaciales, de forma que la naturaleza del fenómeno junto al muestreo determinan el tipo de dato espacial: (1) datos geoestadísticos o de localización continua, (2) datos *lattice/areal* o red de localizaciones fijas, y (3) patrones puntuales. Cada tipo de datos se analiza con técnicas diferentes, por ello, hay que saber diferenciarlos (Banerjee, 2016).

Geoestadística

La geoestadística tiene su origen en la segunda mitad del siglo pasado. Su desarrollo se debe a su aplicación en ingeniería de minas para predecir las reservas de mineral a partir de observaciones espacialmente distribuidas en una región. La característica común de cualquier modelo geoestadístico es que los datos pueden verse como una realización de un proceso estocástico, o parcialmente estocástico, sobre una región **continua** (Pawlowsky-Glahn y Olea, 2004).

En la actualidad, existen una gran variedad de problemas que pueden resolverse con métodos geoestadísticos, siendo normalmente el objetivo predecir en toda la región. La clave en un modelo geoestadístico es el variograma, que será objeto de modelización y estimación para describir adecuadamente el fenómeno observado (García, 2004). El variograma es una herramienta que permite analizar el comportamiento espacial de una variable en una región continua, de manera que representa la influencia de un punto sobre otro en función de la distancia entre ellos (García, 2004).

La principal característica de interés para el estudio en una región continua D de un proceso estocástico, o parcialmente estocástico, $Z(s) \in D$ es la **función de covarianza**, que determina, para cada par de puntos, la covarianza entre las variables aleatorias correspondientes:

$$Cov(Z(s_1), Z(s_2)). \tag{2.3}$$

Para poder estimar la función de covarianza, y en consecuencia, que la predicción sea posible, el proceso tiene que tener un comportamiento estable en toda la región de estudio D, es decir, ha de ser estacionario e isotrópico. Por un lado, un proceso estacionario implica que la distribución de probabilidad del proceso, en una posición, sea constante en el resto de posiciones en el plano de coordenadas cartesianas. Por otro lado, un proceso isotrópico implica el comportamiento anterior en todas las direcciones (estacionariedad en una tercera dimensión del plano). En nuestro caso, como los datos son simulados sobre una región continua ambas condiciones se cumplen y podemos proponer la modelización geoestadística de un proceso y predecir en toda la región. Además, en pesquerías mayoritariamente se suelen cumplir ambas condiciones.

Red de localizaciones fijas

Puede darse el caso en el que las observaciones provienen de un conjunto fijo de localizaciones, p.ej. número de enfermos en cada municipio. En consecuencia, la predicción en otros puntos del espacio no tiene sentido, puesto que, el fenómeno observado únicamente ocurre en las localizaciones de una red o, cuando es observado de forma agregada. Podríamos definir una red de localizaciones o retículo como una colección finita de localizaciones espaciales, distribuidas en el espacio regular o irregularmente (provincias, municipios, ciudades, regiones de pesca establecidas por ICES (*International Council for the Exploration of the Sea*), etc.) (Banerjee, 2016).

La característica principal de este tipo de datos es la **relación de vecindad** en las localizaciones. Una relación de vecindad consiste en establecer que localizaciones de la red son dependientes entre ellas. Esto puede depender de infinidad de factores, bien sea la distancia, comunicación, rasgos comunes, etc. Por tanto, si dos localizaciones de la red se establecen como vecinas estas tendrán una dependencia entre ellas, y por ende, lo que ocurra en una influencia a la otra (Bivand *et al.*, 2008).

Patrones puntuales

Este último tipo de datos espaciales, aborda el estudio de fenómenos que ocurren aleatoriamente en diferentes puntos de una región. Así pues, un patrón puntual es una colección de puntos que nos indican dónde está ocurriendo el fenómeno (Baddeley *et al.*, 2007). La distribución de los puntos puede dar lugar a distintos tipos de patrón puntual: (1) patrón puntual aleatorio, en este caso los puntos aparecen completamente al azar alrededor de una región, p.ej. campaña oceanográfica de arrastre, (2) patrón puntual regular, los puntos presentan una separación regular, p.ej. campañas oceanográficas acústicas, y (3) patrón puntual agrupado, los puntos están agrupados en zonas determinas, p.ej. las pesquerías.

En vista a lo comentado, el interés de un patrón puntual reside en la cuantificación del número medio de sucesos por unidad de área en su entorno, en otras palabras, la intensidad con la que ocurre el suceso a lo largo de un espacio (Møller y Waagepetersen, 2007; Krainski *et al.*, 2018). Por consiguiente, el objetivo suele ser estimar y predecir la función de intensidad $\lambda(s)$ asociada al patrón de puntos:

$$\lambda_A = \int_A \lambda(s)\mathrm{d}s, \tag{2.4}$$

donde A forma parte de la región de estudio y λ es la intensidad del proceso en dicha región.

En el presente trabajo, nos centramos en datos de tipo geoestadístico y patrón puntual, de manera que modelizamos las variables de interés teniendo en cuenta que pueden tener un proceso espacial subyacente.

2.3. Series temporales

Las series temporales se analizan para entender el pasado y predecir el futuro. Tener una serie temporal se traduce en una medida secuencial de la variable en el tiempo a intervalos equiespaciados (Chatfield, 2003). El intervalo en el que se mida la variable puede variar desde una serie anual, hasta una serie diaria. Obviamente cuanto más corto sea el intervalo más compleja será la estadística que necesitemos para la estimación y predicción de sus componentes.

Para ilustrar algunos de los comportamientos que pueden darse sobre un proceso, ligado a la distribución de un *stock* pesquero, en un periodo de tiempo, mostramos la clasificación propuesta por Paradinas *et al.* (2017). En la Figura 2.2 podemos observar un comportamiento oportunista, es decir, no existe una dependencia en la serie temporal, el proceso es puramente estocástico. A continuación, en la Figura 2.3 la distribución del proceso es persistente en el tiempo, en consecuencia el comportamiento es determinista (no cambia con el tiempo). Por último, en la Figura 2.4 vemos una distribución donde el proceso cambia progresivamente en el tiempo (autorregresivo), este comportamiento es una combinación de un proceso estocástico y determinista. Este último comportamiento progresivo, será en el que nos centraremos en este trabajo, ya que en pesquerías es la situación que se da por excelencia (Paradinas *et al.*, 2017; Izquierdo *et al.*, 2021; Pennino *et al.*, 2022).

Así pues, se van a utilizar modelos que nos permitan estimar y predecir comportamientos parcialmente estocásticos. Los modelos por excelencia que se utilizan en este tipo de series temporales son los ARIMA *AutoRegresive Integrated Moving Average*, ya que han mostrado ser uno de los métodos de ajuste de series temporales más valiosos desde que fueron formalizados en 1976 en el libro *Time series analysis, forecasting and control* (Box George *et al.*, 1976).

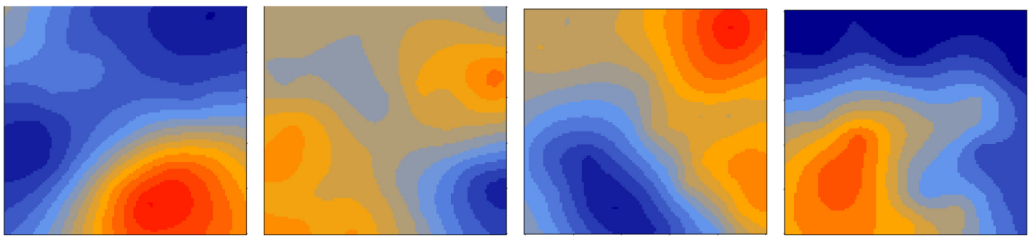

Figura 2.2: Comportamiento oportunista.

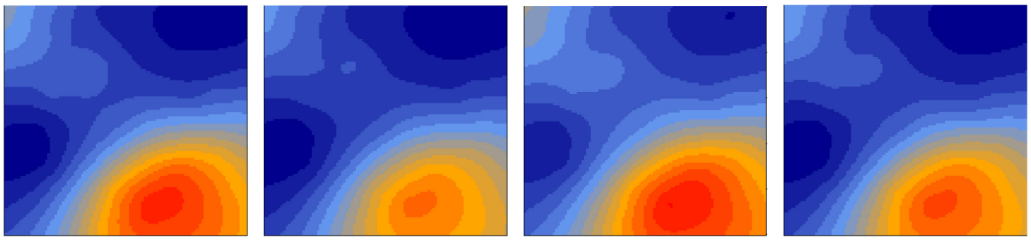

Figura 2.3: Comportamiento persistente.

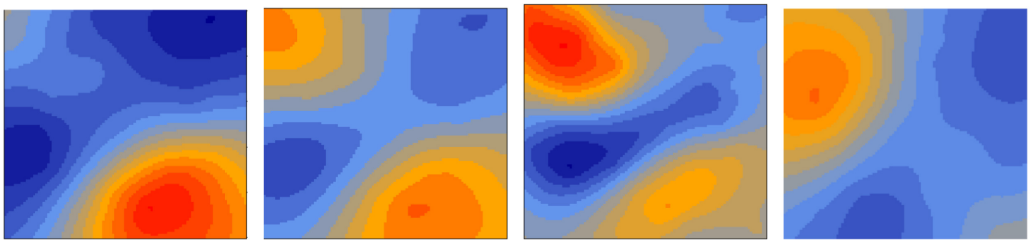

Figura 2.4: Comportamiento autorregresivo.

Es más, nosotros únicamente nos centraremos en un proceso autorregresivo *AR(p)*:

$$\mathrm{x}(t) = c + \rho_1 \mathrm{x}(t-1) + \rho_2 \mathrm{x}(t-2) + ... + \rho_p \mathrm{x}(t-p) + \epsilon(t), \tag{2.5}$$

donde, c es una constante, ρ_p es el parámetro a estimar que indica lo correlacionadas que están las observaciones en el tiempo y $\epsilon(t)$ es el error distribuido como una Normal de media cero y varianza la unidad.

Después de detallar algunos de los modelos con los que podríamos enfrentarnos en el contexto de las pesquerías y la importancia de la estadística espacial y temporal, se han de buscar cuáles son las herramientas estadísticas que nos permiten inferir y predecir sobre sus parámetros. Por lo común, la inferencia y predicción pueden resolverse desde dos aproximaciones: con la estadística clásica/frecuentista o con la **estadística bayesiana**.

2.4. Inferencia y predicción bayesiana

En general, cuando realizamos inferencia y predicción en estadística existen dos aproximaciones: la frecuentista y la bayesiana. La principal diferencia entre ambas es cómo interpretan la probabilidad. Cuando se estima un parámetro/s, p.ej. θ, los métodos frecuentista y bayesiano lidian con la probabilidad de la siguiente manera:

1. La aproximación frecuentista se centra en la probabilidad de los datos ($x = x_1, ..., x_2$) dado el parámetro/s θ, p(x|θ), consiguiendo una estimación fija de cada uno de los parámetros del modelo θ con un intervalo de confianza al 95 %.

2. Por el contrario, la aproximación bayesiana considera cada componente desconocida como una variable aleatoria, y trata de obtener la distribución de probabilidad asociada a esa variable. Por tanto, la inferencia bayesiana se centra en p($\theta|x$), es decir, la probabilidad de θ dado un muestra $x = (x_1, ..., x_n)$ representativa de la población. La inferencia bayesiana se apoya en el teorema de Bayes para estimar la probabilidad de los parámetros del modelo dado los datos observados:

$$\mathrm{p}(\theta|x) = \frac{\mathrm{p}(x,\theta)}{\mathrm{p}(x)} = \frac{\mathrm{p}(\theta)\mathrm{p}(x|\theta)}{\mathrm{p}(x)} = \frac{\mathrm{p}(\theta)\mathrm{p}(x|\theta)}{\int \mathrm{p}(\theta)\mathrm{p}(x|\theta)d\theta}, \tag{2.6}$$

   Como p(x) no depende de β:

$$\mathrm{p}(\theta|x) \propto \mathrm{p}(\theta) \times \mathrm{p}(x|\theta) \tag{2.7}$$

   - p($\theta|x$) es la distribución a posteriori del parámetro.
   - p(θ) es la distribución a priori del parámetro.
   - p($x|\theta$) es la función de verosimilitud del modelo.

Tradicionalmente, la aproximación frecuentista ha sido la más popular. Este hecho, se debía a la limitada capacidad de resolver modelos complejos desde la perspectiva bayesiana, es decir, la inferencia bayesiana suele resultar en complejas expresiones prácticamente imposibles de resolver analíticamente, y por ende, no podían obtenerse las distribuciones a posteriori p($\theta|x$) de los parámetros.

No obstante, en las últimas décadas gracias a los avances en computación se han desarrollado aproximaciones computacionales, como los métodos Markov chain Monte Carlo (MCMC) o como el método *Integrated Nested Laplace Approximation* (INLA), que son capaces de obtener una aproximación precisa de las distribuciones a posteriori de los parámetros. Con ayuda de este tipo de técnicas la estadística bayesiana ha ido ganando más resolutividad. Es más, algunas de las ventajas del método bayesiano son:

1. La distribución a posteriori $p(\theta|x)$ proporciona toda la información sobre los parámetros, p.ej., media, mediana, cuantiles, probabilidad de no ser cero, etc. y además, la interpretación es muy simple.

2. Permite incorporar conocimiento a priori $p(\theta)$ de una manera muy sencilla. Si el conocimiento previo es pobre, se asumen distribuciones a priori vagas o no informativas, de manera que todo el peso de la posteriori recae sobre la verosimilitud (los datos).

3. En comparación con algunas técnicas clásicas, la inferencia bayesiana tiene la capacidad de ajustar un gran número de modelos complejos de forma eficiente y rápida.

Por todo lo comentado, el presente trabajo, utilizará ambas perspectivas, la estadística bayesiana y frecuentista, para inferir y predecir sobre los parámetros de los modelos, aunque se hará más énfasis en la estadística bayesiana. En consecuencia, vamos a desmenuzar los términos del teorema de Bayes (2.7), con el objetivo de entender como funciona la estadística bayesiana.

Función de verosimilitud $\mathbf{p(x|\theta)}$

A pesar de que la inferencia bayesiana es más conocida por incorporar la información previa, no debemos olvidar la importancia de la otra fuente de información: los **datos**. La información que contienen los datos se expresa a través de la función de verosimilitud $p(x|\theta)$. Esta función de los datos o verosimilitud, nos proporciona una medida de cómo de probable es cada valor del parámetro. Así pues, si tenemos una muestra aleatoria $x = (x_1, ..., x_n)$ de una variable aleatoria $X \sim F(x|\theta)$ siendo θ desconocido, la función de verosimilitud se define como: [2]

$$l(\theta) = f_x(x|\theta) = f_{x_1,...,x_n}(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f_x(x|\theta). \tag{2.8}$$

Distribución a priori $\mathbf{p(\theta)}$

Los métodos bayesianos requieren establecer una distribución a priori $p(\theta)$ sobre los parámetros desconocidos del modelo. En principio, no existen limitaciones a la hora de escoger una distribución a priori, sin embargo, es de vital importancia no condicionar con la previa el valor de los parámetros, ya que una previa mal escogida puede llevarnos a conclusiones erróneas. Por ello, existen distintas posibilidades para fijar una distribución previa:

[2] Si $f_x(x|\beta)$ es la función de densidad de una variable continua.

- Información a priori **objetiva** o no informativa: si a priori no tenemos información sobre el parámetro de interés, así que, necesitaremos una distribución de probabilidad para el parámetro que indique desconocimiento. Las distribuciones a priori no informativas pueden ser de gran utilidad en aquellas situaciones en las que la información previa es inexistente o no estamos seguros del comportamiento del parámetro. Suelen ser distribuciones de probabilidad constantes en todos los valores del espacio paramétrico y pueden ser impropias [3].

  Un ejemplo de distribución a priori no informativa son las distribuciones a priori de Jeffreys, invariantes frente a transformaciones, en otras palabras, su forma no se ve alterada por una reparametrización. Tienen la forma:

$$\mathrm{p}(\theta) \propto [\mathrm{I}(\theta)]^{\frac{1}{2}}, \tag{2.9}$$

  donde $\mathrm{I}(\theta) = -\mathrm{E}^{(x|\theta)}[\frac{\partial^2 \ln \mathrm{p}(x|\theta)}{\partial \theta^2}]$, representa la información de Fisher de θ y $\mathrm{p}(x|\theta)$ la función de verosimilitud.

- Información a priori **subjetiva** o informativa: a partir del consejo de expertos en el área o de experimentos previos que se hayan podido realizar sobre el estudio podemos especificar una previa más informativa. Por ejemplo, sabemos por otros estudios que la clorofila tiene un efecto positivo sobre la biomasa de especies pesqueras, así que, podríamos seleccionar una previa con una distribución de probabilidad que únicamente tenga valores positivos.

Otro punto a tener en cuenta al escoger una previa, es que la familia para la función de probabilidad escogida puede ser conjugada [4]. Algunas de las razones por las que conviene utilizar una distribución previa conjugada son: (1) simplifican los cálculos en la obtención de la posteriori, (2) facilita la descripción de los resultados y (3) son de gran utilidad en la construcción de modelos más complicados.

Es cierto que la elección de una previa conjugada facilitaba la inferencia sobre los parámetros del modelos. Sin embargo, (Simpson *et al.*, 2017) proponen el uso de PC-priors del inglés *Penalized complexity priors*, un enfoque novedoso que ha ido ganando fuerza con los años. Las PC priors se describieron para modelos con relaciones aditivas definidos en distintos niveles o capas, de forma que proponen distribuciones a priori que penalizan la complejidad del modelo. Además, un atractivo de las PC priors, es que, utilizan la probabilidad para establecer que valores a priori pueden tomar los parámetros de un modelo.

[3] Una distribución impropia: integra infinito en el espacio paramétrico: $\int_{-\infty}^{+\infty} f_x(x)dx = +\infty$.

[4] Definición: una clase P de distribuciones a priori es una familia conjugada para F (la clase de todas las funciones de densidad $\mathrm{p}(x|\theta)$ del parámetro θ), si la distribución a posteriori $\mathrm{p}(\theta|x)$ está en la clase P para todo x del espacio paramétrico y para toda distribución a priori de P.

Distribución a posteriori $\mathrm{p}(\theta|x)$

La aplicación del teorema de Bayes (2.7) da como resultado la distribución a posteriori $\mathrm{p}(x|\theta)$ de cada uno de los parámetros de interés. La distribución a posteriori de un parámetro puede variar en función de la distribución a priori y del tamaño muestral. Así pues, si disponemos de una muestra pequeña la distribución a priori tendrá más fuerza que la verosimilitud. Por el contrario, si la muestra es lo suficientemente grande dominará la verosimilitud sobre la previa.

Las distribuciones a posteriori no siempre tienen una forma sencilla, de manera que no es posible resolver de forma analítica el teorema de Bayes. A pesar de ello, es posible aproximar la distribución utilizando, por ejemplo, métodos MCMC o INLA

Distribución predictiva

Una de las mayores utilidades de un modelo es poder utilizarlo para predecir, por ejemplo, en un modelo geoestadístico la predicción nos permite observar que está ocurriendo en toda el área de estudio.

Recordamos que en estadística bayesiana se considera todo lo desconocido como un variable aleatoria, por ello, a la hora de predecir simplemente buscamos la distribución de probabilidad de una nueva realización de la variable de interés condicionada al conocimiento que tengamos sobre los parámetros del modelo. Dicha distribución expone el valor más y menos probable, y se llama distribución predictiva. Podemos tener dos tipos de distribuciones predictivas en función de si se ha realizado o no el experimento:

- **Predictiva a priori**. La distribución predictiva a priori es la distribución de una nueva realización de la variable de interés antes de realizar el experimento utilizando únicamente la información previa sobre los parámetros:

$$m(X_{pred}) = \int \mathrm{p}(X|\theta)\mathrm{p}(\theta)d\theta, \tag{2.10}$$

  donde m es la dimensión del vector paramétrico y X_{pred} una nueva realización. La distribución predictiva a priori de un modelo cuando la a priori es objetiva no tiene mucho sentido ya que no tenemos ningún tipo de información sobre los parámetros.

- **Predictiva a posteriori**. La distribución predictiva a posteriori es la distribución de una nueva realización de la variable de interés después de realizar el experimento, utilizando pues la información aportada por la verosimilitud. Por tanto, si $x = (x_1, ..., x_n)$ es la realización de una muestra aleatoria $X = (X_1, ..., X_n)$ de una variable aleatoria de interés $X \sim \mathrm{F}(x|\theta)$ siendo θ desconocido y, $\mathrm{p}(\theta|x)$ es la distribución a posteriori de cada uno de

los parámetros que rigen la variable, entonces la distribución predictiva a posteriori de un nuevo X es:

$$m(X_{pred}|x) = \int f(X|\theta)\mathrm{p}(\theta|x)d\theta. \tag{2.11}$$

Observar que todo el proceso que hemos ido desarrollando, desde escoger una distribución previa para el parámetro hasta que estimamos las distribuciones predictivas a posteriori, se trata de un proceso de aprendizaje. Nuestro trabajo, empleará todo este proceso de aprendizaje bayesiano en cada uno de los modelos que se planteen para así obtener las distribuciones a posteriori y las distribuciones predictivas a posteriori.

Para finalizar, hasta ahora hemos abordado la estadística bayesiana desde una perspectiva univariante. Sin embargo, en muchas áreas como la epidemiología, climatología, marketing y ecología, los datos pueden ser multivariantes (diversidad de efectos y más de una variable respuesta), lo que complica la modelización y su posterior inferencia y predicción. Por ejemplo, un modelo geoestadístico con covariables presentará una serie de efectos que se pueden diferenciar en distintos niveles o capas, una capa podrían ser los efectos fijos y otra los hiperparámetros del modelo.

Así pues, con el fin de abordar modelos más complejos, diferenciados en capas o niveles, surgen los Modelos Jerárquicos. El análisis de los modelos jerárquicos, puede resolverse desde las perspectiva clásica o frecuentista, sin embargo, el paradigma bayesiano puede resultar muy conveniente gracias a la evaluación de la incertidumbre en cada capa o nivel. Por lo tanto, en este trabajo se llevará a cabo el análisis de modelos jerárquicos desde el enfoque bayesiano. Un Modelo Jerárquico Bayesiano suele basarse en 3 etapas fundamentales:

1. Verosimilitud de los datos.
2. Modelización del parámetro/s de interés.
3. Modelización de los hiperparámetros.

2.5. INLA vs métodos MCMC

El siguiente apartado explica como funcionan los métodos MCMC e INLA, ambas herramientas son ampliamente usadas en estadística para aproximar las distribuciones a posteriori de los Modelos Jerárquicos Bayesianos. Por ello, procedemos a exponer las ventajas e inconvenientes que presentan dichas aproximaciones. Del mismo modo, acabaremos centrándonos en INLA, puesto que, ha sido la escogida para resolver la mayoría de los modelos que se planteen en este trabajo.

Métodos MCMC

De entre las distintas aproximaciones propuestas para estimar distribuciones a posteriori destacan las basadas en métodos MCMC, implementados en softwares como `WinBUGS` (Lunn *et al.*, 2000) o más recientemente `JAGS` (*Just Another Gibb sampling*) (Depaoli *et al.*, 2016). La base de estas aproximaciones es la simulación, es decir, los métodos MCMC combinan la simulación Monte Carlo con las cadenas de Markov. Digamos que la unión de ambas da como resultado una cadena simulada de valores (simulación por Monte Carlo), en la que cada valor depende, única y exclusivamente, del anterior (cadena Markov). En consecuencia, cualquier método basado en MCMC lo que hace es, a través de la simulación y con ayuda de la propiedad Markoviana, muestrear de la distribución a posteriori $p(\theta|x)$. Los dos algoritmos más conocidos de simulación Monte Carlo que utilizan Cadenas de Markov son el muestreo *Gibbs* y *Metropilis-Hastings*. A pesar, de que estos métodos suelen traer consigo buenos resultados, presentan algunas desventajas:

1. Los métodos MCMC simulan para todos los parámetros del modelo, es decir, lidian con una distribución multivariante. En consecuencia, se obtiene la distribución a posteriori conjunta de los parámetros del modelo. Esto puede suponer una desventaja por dos razones, una, no siempre estamos interesados en todos los parámetros del modelo, y otra, puede conllevar un alto coste computacional.

2. Del mismo modo, estos métodos pueden requerir de un elevado número de simulaciones para que la inferencia sea válida, lo que supone a su vez un alto coste computacional.

3. Además, se ha de verificar que el periodo de *burn-in* ha terminado, es decir, que hemos logrado una muestra representativa de la distribución a posteriori.

En contraposición, la principal ventaja de algunas aproximaciones basadas en métodos MCMC, sobretodo, la implementada en el software `WinBUGS`, es que, la única limitación es tu capacidad de plasmar correctamente el modelo y el tiempo del que dispongas. Esto supone una gran ventaja, puesto que, hay modelos muy complejos prácticamente imposibles de resolver con otros softwares basados en métodos MCMC o con la aproximación INLA.

INLA

Rue *et al.* (2009) proponen una nueva vía alternativa a los métodos MCMC para aproximar las distribuciones a posteriori. En concreto, proporcionan una herramienta que permite aproximar **las distribuciones a posteriori marginales** de cada uno de los parámetros del modelo en el contexto bayesiano.

INLA es una alternativa computacionalmente más rápida a los métodos MCMC. En gran parte, esta velocidad computacional se debe a que, no necesitamos simular de la posteriori, simplemente aproximarla numéricamente. La única condición que presenta INLA es que el modelo estadístico ha de ser un *Latent Gaussian Model* (LGM). De hecho, un LGM es un tipo de Modelo Jerárquico Bayesiano, de forma que la mayoría de los modelos estadísticos habituales pueden ser formulados para que cumplan esta condición, p.ej. modelos espacio-temporales, modelos espaciales, GLM, GAM, modelos de *random walk* (de primer y segundo orden), modelos LGCP, modelos geoestadísticos y geoaditivos, etc.

En vista a que INLA es computacionalmente más rápido frente a los métodos MCMC y permite ajustar los modelos a proponer en este trabajo de una forma precisa, se ha escogido como herramienta principal para inferir y predecir sobre los modelos. Por consiguiente, se van a detallar los elementos necesarios para entender como funciona INLA.

2.5.1. Como funciona la aproximación INLA

En primer lugar, hacer mención al libro *Bayesian inference with INLA*, en el que se desarrollan las bases de INLA y `R-INLA`, en base a este libro se ha resumido en el siguiente apartado dicha aproximación (Gómez-Rubio, 2020). Para comprender el funcionamiento de INLA, necesitamos familiarizarnos con tres conceptos clave para esta aproximación:

1. *Latent Gaussian Models* (LGMs)
2. *Gaussian Markov Random Fields* (GMRFs)
3. Aproximación de Laplace

Latent Gaussian Models

INLA necesita que el modelo sobre el que se quiera inferior o predecir sea un LGM, de lo contrario, no podrá aproximar las distribuciones a posteriori de los parámetros. Esto implica que los parámetros asociados a la verosimilitud del modelo deben seguir una distribución Normal. Para ilustrar un LGM, supongamos una variable aleatoria X a la que se le asocia una distribución de probabilidad, la verosimilitud sería:

$$\mathrm{p}(x|\theta, \psi_1) = \sum_{i=1}^{n} \mathrm{p}(x_i|\eta_i(\theta), \psi_1), \tag{2.12}$$

donde, $x_i = (x_1, ..., x_n)$ es una realización de la muestra, $\theta = (\theta_1, ..., \theta_n)$ es un *latent field* [5], ψ_1 es el hiperparámetro de la distribución asociada y $\eta_i(\theta)$ es la función de enlace del predictor lineal que conecta los datos al *latent field*.

Para la verosimilitud del modelo en INLA hay que transforma el *latent field* en un *latent Gaussian field*. Para ello, simplemente suponemos una distribución a priori que sea Normal de media 0 y matriz de precisión $Q^{-1}(\psi_i)$ [6] para cada uno de los parámetros del *latent field* θ:

$$\theta|\psi_i \sim \mathrm{N}(0, Q^{-1}(\psi_i)) \quad i = 1 \text{ y } 2. \tag{2.13}$$

Además, si podemos asumir que θ es condicionalmente independiente [7], entonces dicho *latent Gaussian field* se considera un *Gaussian Markov Random Field (GMRF)* (2.18).

Finalmente, se ha de asumir una distribución a priori para los hiperparámetros, tal que, $\psi_i \sim \mathrm{p}(\psi_i)$. Las distribuciones a priori de los hiperparámetros no han de seguir una distribución normal.

Gaussian Markov Random Fields (GMRFs)

En un *Gaussian Markov Random Field* asumimos que los parámetros (θ) son una normal multivariante, y además suponemos propiedades Markovianas para θ. Asumir un comportamiento Markoviano implica una importante mejora a nivel computacional, puesto que, se simplifican los cálculos numéricos al suponer que únicamente las relaciones vecinas tienen un valor en la matriz de covarianza distinto de cero.

Rue *et al.* (2009) demostraron como la propiedad de independencia condicionada puede ser codificada en la matriz de precisión, de manera que se simplifica considerablemente el cálculo en comparación con una matriz de covarianza original:

$$\begin{aligned} &i \neq j,\ \theta_i \perp \theta_j|\theta_{ij}, \\ &\theta_i \perp \theta_j|\theta_{ij} \leftrightarrow Q_{ij} = 0. \end{aligned} \tag{2.14}$$

Por tanto, asumir un comportamiento Markoviano en el GF resulta en una matriz de precisión con muchos ceros y es lo que hace de INLA una herramienta extremadamente rápida en comparación con los métodos MCMC.

[5] θ representa el conjunto de parámetros que contiene el predictor lineal y queremos estimar y predecir, p.ej., β_0, β_1, u_i, etc. A dicho conjunto se le denomina *latent field*.

[6] Q^{-1}: En INLA se trabaja en términos de precisión (τ) no de desviación típica.

[7] Dos sucesos A y B son condicionalmente independientes de un suceso Y si $\mathrm{p}(\mathrm{A} \cap \mathrm{B}|\mathrm{Y}) = \mathrm{p}(\mathrm{A}|\mathrm{Y}) \times \mathrm{p}(\mathrm{B}|\mathrm{Y})$.

Aproximación de Laplace

La aproximación de Laplace puede ser utilizada para estimar cualquier distribución $\mathrm{p}(\theta)$ con una distribución de probabilidad Normal. Dicha aproximación utiliza los tres primeros términos de la expansión de Taylor en torno a la moda de una función para aproximar su logaritmo. Por tanto, mediante Laplace, $\mathrm{p}(\theta)$ puede aproximarse usando una distribución Gaussiana con media y moda θ^* y varianza con información de Fisher, $\frac{-1}{\frac{d^2 \log(p(\theta^*))}{d\theta^2}}$.

$$\mathrm{p}(\theta) \approx \mathrm{N}\left(\theta^*, \frac{-1}{\frac{d^2 \log(p(\theta^*))}{d\theta^2}}\right). \tag{2.15}$$

Distribuciones a posteriori marginales en INLA

Una vez entendidos los conceptos básicos, remarcar que INLA tiene como objetivo obtener las distribuciones a posteriori marginales del *latent field* θ y de los hiperparámetros:

$$\begin{aligned} \mathrm{p}(\theta|x) &= \int \mathrm{p}(\theta|\psi, x) \cdot \mathrm{p}(\psi|x) d\psi \,, \\ \mathrm{p}(\psi|x) &= \int \mathrm{p}(\psi|x) d\psi \,, \end{aligned} \tag{2.16}$$

donde, $\mathrm{p}(\theta|x)$ se corresponde con las distribuciones a posteriori marginales del *latent field* y $\mathrm{p}(\psi|x)$ hace referencia a las marginales a posteriori de los hiperparámetros. Como resultado, debemos aproximar numéricamente las siguientes expresiones:

1. Para poder calcular las distribuciones a posteriori marginales de los hiperparámetros $\mathrm{p}(\psi|x)$ y del *latent field* $\mathrm{p}(\theta|x)$ necesitamos aproximar la **distribución a posteriori conjunta de los hiperparámetros** $\mathrm{p}(\psi|x)$.

2. Del mismo modo, para calcular las distribuciones a posteriori marginales del *latent field* $\mathrm{p}(\theta|x)$ necesitamos aproximar las **marginales de la distribución condicional completa** de θ, es decir, $\mathrm{p}(\theta|\psi, x)$.

Por un lado $\mathrm{p}(\psi|x)$. Para poder calcular la distribución a posteriori conjunta de los hiperparámetros se plantea la siguiente ecuación:

$$\tilde{p}(\psi|x) := \frac{\mathrm{p}(\theta, \psi|x)}{\mathrm{p}_G(\theta|\psi, x)}|_{\theta=\theta^*(\psi)}, \tag{2.17}$$

donde, $\mathrm{p}_G(\theta|\psi, x)$ se corresponde con una aproximación Gaussiana de $\mathrm{p}(\theta|\psi, x)$ dado el método Laplace. Así mismo, $\theta^*(\psi)$ es la moda de $\mathrm{p}(\theta|\psi, x)$ para un ψ dado. Esta aproximación es exacta si $\mathrm{p}(\theta|x, \psi)$ tiene una verosimilitud Gaussiana.

Por otro lado $\mathrm{p}(\theta|\psi, x)$. El cálculo de las marginales de la distribución condicional completa puede llevarse a cabo a través de tres estrategias distintas:

1. **Aproximación Gaussiana**. Las distribuciones a posteriori condicionadas $\mathrm{p}(\theta|\psi, x)$ se aproximan directamente como las marginales de $\mathrm{p}_G(\theta|\psi, x)$. Es la vía más rápida computacionalmente pero con posibles errores en la localización de la media a posteriori.

2. **Laplace approximation**. El vector θ se reescribe como $\theta = (\theta_i, \theta_{-i})$, de forma que la aproximación Laplace se utiliza para cada elemento del *latent field*:

$$\tilde{p}(\psi|x) := \frac{\mathrm{p}(\theta, \psi|x)}{\mathrm{p}_{LG}(\theta_{-i}|\theta_i, \psi, x)}|_{\theta_{-i}=\theta^*_{-i}(\theta_i,\psi)}, \tag{2.18}$$

   donde, $\mathrm{p}_{LG}(\theta_{-i}|\theta_i, \psi, x)$ es la aproximación de Laplace Gaussiana de $\mathrm{p}(\theta_{-i}|\theta_i, \psi, x)$ y θ_{-i} es su moda. Esta estrategia se considera la más precisa, pero, también la que más tiempo consume.

3. **Aproximación de Laplace simplificada**. Esta última vía de cálculo está basada en la serie de expansión de Taylor de tercer orden. Es un alternativa rápida y lo suficientemente precisa a la estrategia anteriormente mencionada.

Por último, el algoritmo implementado en INLA utiliza el método de Newton para explorar la distribución a posteriori conjunta de los hiperparámetros $\tilde{\mathrm{p}}(\psi, x)$, con el fin de encontrar aquellos puntos que son adecuados para la integración numérica. En consecuencia, una vez aproximadas $\tilde{\mathrm{p}}(\psi|x)$ y $\tilde{\mathrm{p}}(\theta|\psi, x)$, las distribuciones a posteriori marginales para el *latent field* $\tilde{\mathrm{p}}(\theta|x)$ son calculadas vía integración numérica:

$$\tilde{p}(\theta|x) = \int \tilde{p}(\theta|\psi, x)\tilde{p}(\psi|x)d\psi \approx \sum_{k=1}^{K} \tilde{p}(\theta|\psi^{(k)}, x)\tilde{p}(\psi^{(k)}|x)\Delta_k. \tag{2.19}$$

Las distribuciones marginales a posteriori para los hiperparámetros ψ_j se aproximan usando los puntos de integración previamente construidos.

Toda esta metodología desarrollada por Rue *et al.* (2009) se encuentra implementada en el software `R` (Team, 2013) a través del paquete `R-INLA`, para más información sobre la instalación, uso y aplicaciones del paquete clicar el siguiente enlace https://www.r-inla.org/.

2.5.2. Estadística espacial con INLA

Por el momento, se ha marcado el contexto teórico del trabajo y se ha explicado la herramienta principal con la que se realiza la inferencia y predicción sobre la mayoría de los modelos a proponer: INLA. De modo que, en vista a que se contempla abordar modelos complejos como son los modelos geoestadísticos o los patrones puntuales correlados en el tiempo, vamos a adentrarnos en cómo INLA aborda este tipo de modelos.

Datos geoestadísticos

En primer lugar, supongamos que tenemos una serie de localizaciones n con coordenadas $s_1, ..., s_n$ y en cada una de ellas hay un valor de la variable de interés $X(s)$. Esto se traduce, en una muestra $x(s_1), ..., x(s_n)$ de un proceso estocástico $X(s)$ definido en un dominio continuo D. Si asumimos que $x(s_1), ..., x(s_n)$ está distribuido como una Normal multivariante, entonces nos encontramos ante un *Gaussian field* (GF) con media cero y matriz de covarianza Σ. Por ejemplo, supongamos un modelo, tal que:

$$\begin{aligned} X(s) &\sim \text{Normal}(\mu(s), \sigma^2)\,, \\ \mu(s) &= \beta_0 + U(s)\,, \\ U(s) &\sim \text{GMRF}(0, \Sigma). \end{aligned} \tag{2.20}$$

En la ecuación anterior, lo que nos interesa es la presencia del término $U(s)$, ya que utilizamos este término para incorporar dependencia espacial al modelo. Hemos asumido que el término espacial $U(s)$ sigue una distribución Normal multivariante de media cero y **matriz de covarianza** Σ. Ahora, nos enfrentamos al problema de cómo estimar Σ en el *Gaussian Field* (GF). Zuur *et al.* (2017) resumían los pasos para calcular un efecto aleatorio espacial en un dominio continuo:

1. Primero, disponemos de n localizaciones de muestreo $s = s_1, ..., s_n$.
2. En cada una de las localizaciones tenemos un efecto aleatorio $U(s)$.
3. Asumimos que $U(s_1), ..., U(s_n)$ están distribuidos como una normal multivariante de media 0 y matriz de covarianza Σ [8], es decir, tenemos un *Gaussian Field* (GF).
4. A continuación, para simplificar el cálculo de la matriz de covarianza asumimos un comportamiento Markoviano. Un comportamiento Markoviano implica que únicamente las

[8] Vamos a hablar de matriz de covarianza por facilitar la comprensión, pero, recordamos que en INLA se trabaja con precisión Q.

localizaciones vecinas están correlacionadas (explicábamos en la sección anterior que utilizamos la independencia condicionada del *latent field* para conseguir esta propiedad). Esto repara en que $U(s)$ ya no es solo un GF, sino que, tenemos un *Gaussian Markovian Random Field* (GMRF), cuya matriz de covarianza presenta una gran cantidad de ceros.

5. Ya con nuestro GMRF, para cuantificar aquellas localizaciones cuyo valor sea distinto de cero en la matriz de covarianza Σ utilizamos la matriz de correlación. Dicha matriz tiene un parámetro k (kappa) relacionado con el rango [9] que se ha de estimar:

$$\text{cor}_{Matern}(U(s_i), U(s_j)) = k \times ||s_i - s_j|| \times K_1(k \times ||s_i - s_j||)\,, \tag{2.21}$$

$$\Sigma = \text{cov}_{Matern}(U(s_i), U(s_j)) = \sigma_U^2 \times \text{cor}_{Matern}(U(s_i), U(s_j))\,,$$

donde $\text{cor}_{Matern}(U(s_i), U(s_j))$ es la matriz de correlación entre dos localizaciones, con un parámetro desconocido k multiplicado por la distancia entre dos localizaciones s_i y s_j multiplicadas a su vez por una función matemática que depende del mismo parámetros desconocido k y la distancia. El parámetro desconocido k se denomina kappa y está relacionado con el rango ($r = \frac{\sqrt{(8\times \text{v})}}{k}$, donde v suele valer 1).

6. En consecuencia, INLA solamente debe estimar kappa k y la varianza del término espacial σ_U para obtener la matriz de covarianza. Pero, tenemos un problema, INLA no es capaz de ajustar GMRFs continuos. La solución es utilizar la aproximación SPDE *Stochastic partial differencial equation* (Jentzen y Kloeden, 2009), ya que los parámetros de esta aproximación están relaciones con los que nosotros necesitamos:

$$(k^2 - \Delta)^{\frac{\alpha}{2}} \tau U(s) = W(s)\,, \tag{2.22}$$

donde Δ es el operador Laplace y $W(s)$ es un proceso de ruido blanco espacial Gaussiano (cuando hablamos de ruido queremos decir que no existe una correlación). Si resolvemos la ecuación conseguimos estimar los valores de kappa k y la varianza σ_{U}^2.

7. Con la combinación de las aproximaciones SPDE y *Finite element approach* (Lindgren, 2001), que nos permite proyectar en un grid irregular, conseguimos estimar los parámetros que necesitamos:

$$U(s) = \sum_{k=1}^{G} a_k(s_i) \times w_k, \tag{2.23}$$

donde $U(s)$ es el efecto espacial que andamos buscando, a_k se denomina matriz de proyección y nos permite proyectar las localizaciones de muestreo en un grid irregular al que llamamos *mesh* sobre el que se realizaran las estimaciones pertinentes y w_k es campo espacial que necesitamos para resolver la aproximación SPDE.

[9] Denominamos rango a la distancia que hay entre dos localizaciones que están correladas.

Obviamente nosotros no tenemos que resolver cada uno de estos pasos, `R-INLA` tiene todo este proceso implementado a través de una serie de funciones. Así pues, los pasos a seguir en `R-INLA` para inferir y predecir sobre un modelo geoestadístico son los siguientes:

1. Elaborar un *mesh*. Un *mesh* no es más que una división irregular del área en un número finito de triángulos. Se ha de tener en cuenta que cuanto menor sea el tamaño del triángulo mayor será el coste computacional. No obstante, si el triángulo es demasiado grande parte de la variabilidad espacial quedará enmascarada en la variabilidad de la propia variable de interés. El modelo se evalúa en los nodos que intereseccionan.

```
# Mesh
loc <- cbind(datos$xcoord, datos$ycoord) # Localizaciones

bound <- inla.nonconvex.hull(loc) # Límite

mesh <- inla.mesh.2d(
  boundary = bound,
  max.edge = c(0.63, 2.5), # Parámetros del mesh
  cutoff = 0.05
)
```

2. Definir matriz de proyección a_{ik}. Necesitamos construir una matriz de proyección para así proyectar las observaciones que se han recogido sobre los triángulos que componen el *mesh*.

```
# Matriz de proyección
A_est <- inla.spde.make.A(
  mesh = mesh,
  loc = cbind(datos$xcoord, datos$ycoord)
  )
)
```

3. Definir la aproximación SPDE. Con esta aproximación conseguimos estimar k y σ_U^2. Para facilitar los cálculos en `R-INLA` podemos utilizar PC-priors *Penalized complexity priors* (Simpson *et al.*, 2017).

```
# Aproximación SPDE
spde <- inla.spde2.pcmatern(
  mesh = mesh,
  prior.range = c(0.5, 0.01), # P(range <0.5) = 0.01
  prior.sigma = c(1.5, 0.01)  # P(sigma >1.5) = 0.01
)
```

4. Definir el campo espacial w_k. Generamos un índice o una lista para el modelo SPDE, simplemente especificamos el nombre del efecto espacial $U(s)$ y el número de vértices en el modelo SPDE.

```r
# Matriz de pesos (campo espacial)
iset <- inla.spde.make.index("i",
  n.spde = spde$n.spde
)
```

5. Elaborar un *stack*. Un *stack* nos permite indicarle a `R-INLA` donde están los puntos observados de los que tenemos información sobre la variable respuesta y el valor de las variables explicativas que forman el predictor. Así mismo, además de elaborar un stack para estimar los parámetros del modelo, si queremos predecir hemos de elaborar otro stack en el que le indiquemos en qué puntos del *mesh* queremos predecir y cual es el valor de los componentes del predictor en esos nodos. La variable respuesta en el stack de predicción serán NA.

```r
# Stack
# Estimación
stack_est <- inla.stack(
  data = list(y = datos$variable_respuesta, link = 1),
  A = list(A_est, 1, 1), # Efectos
  effects = list(iset,
    covariable = data$covariable,
    intercept = rep(1, length(data$variable_respuesta))
  ),
  tag = "est"
)

# Predicción
stack_pred <- inla.stack(
  data = list(y = datos_pred$variable_respuesta, link = 1), #
  A = list(A_pred, 1, 1), # Efectos
  effects = list(iset,
    covariable = datos_pred$covariable,
    intercept = rep(1, length(data_pred$variable_respuesta))
  ),
  tag = "pred"
)

# Juntar ambos stacks
stack <- inla.stack(stack_est, stack_pred)
```

6. Especificar la formula. Simplemente le decimos a `R-INLA` cual es nuestra variable respuesta y el predictor lineal asociado a la media de dicha variable.

```r
# Fórmula
formula <- y ~ -1 + intercept + f(covariable, model = "rw2") + f(i,
  model = spde, group = i.group)
```

7. Correr el modelo en `R-INLA`.

```
modelo <- inla(formula,
  data = inla.stack.data(stack),
  family = "gamma", # Distribución de probabilidad
  control.predictor = list(
    compute = TRUE,
    A = inla.stack.A(stack), link = 1
  ), # Prediccón
  verbose = TRUE, # Salida interna R-INLA
  control.compute = list(waic = TRUE, cpo = TRUE, dic = TRUE),
  num.threads = 2 # Número de cadenas
)
```

8. Inspeccionar los resultados.

Con todo lo descrito en el apartado ya podríamos plantear un modelo geoestadísticos y resolverlo utilizando `R-INLA`.

Patrones puntuales

A continuación, vamos a explicar el segundo tipo de datos que vamos a abordar, un patrón de puntos. Recordamos que, el interés de un patrón puntual reside en la cuantificación del número medio de sucesos por unidad de área en su entorno, es decir, la intensidad con la que ocurre el suceso a lo largo de un espacio (Baddeley *et al.*, 2007; Møller y Waagepetersen, 2007; Krainski *et al.*, 2018). Para ello, hacemos uso de la modelización, con el fin de estimar y predecir la función de intensidad $\lambda(s)$ asociada al patrón de puntos. Esta función puede llegar a modelizarse con un predictor que incluya covariables y otros efectos.

En la actualidad, uno de los modelos más extendidos para la estimación y predicción de la función de intensidad es el modelo LGCP (log-Gaussian Cox process). Un proceso de Cox no es más que un proceso de Poisson donde la intensidad de los sucesos $\lambda(s)$ varía en el espacio (Baddeley *et al.*, 2007; Møller y Waagepetersen, 2007; Krainski *et al.*, 2018). En otros términos, dado un área A, la probabilidad de observar un número concreto de sucesos en el área sigue una distribución Poisson con intensidad variable. Además, se considera un proceso log-Gaussiano porque se modela el logaritmo de la intensidad $\log(\lambda(s))$ como un campo Gaussiano:

$$\begin{aligned} \log(\lambda(s)) &= \beta_0 + S(s), \\ S(s) &\sim \text{GMRF}(0, \Sigma), \end{aligned} \tag{2.24}$$

donde, β_0 corresponde con el intercepto y $S(s)$ es un proceso espacial Gaussiano con matriz de covarianza y media cero.

Una de las primeras aproximaciones implementadas en INLA y otros softwares estadísticos para estimar y predecir un modelo LGCP consistía en dividir la región de estudio en celdas regulares y contar el número de puntos en cada una de ellas (Krainski *et al.*, 2018). Dichos conteos, podían modelarse con una distribución Poisson condicionada a un predictor lineal Gaussiano. Sin embargo, Simpson *et al.* (2016) proponen una nueva aproximación que considera los modelos SPDE, explicados en la sección anterior, para aproximar la matriz de covarianza, y por ende, el efecto espacial $S(s)$.

Los pasos a seguir para ajustar un patrón puntual con `R-INLA` son muy parecidos a los comentados en el apartado anterior. Con esta aproximación los sucesos observados del patrón puntual se modelan considerando su localización exacta en lugar de agruparlos en celdas y conseguimos una predicción de la intensidad $\lambda(s)$ en toda la zona de estudio, pudiendo concluir en que zonas en más probable que ocurra el suceso.

2.5.3. `inlabru`

El paquete de `inlabru`, implementado en `R` (Team, 2013), se desarrolla como parte del proyecto *Modelling spatial distribution and change from wildlife survey data* (Bachl *et al.*, 2019). La base del paquete `inlabru` sigue siendo la aproximación INLA, pero, con algunas modificaciones que permiten abordar distintos problemas que `R-INLA` no puede o necesita de más esfuerzo computacional para resolverlo. Así pues, `inlabru` está enfocado a datos espaciales de tipo patrón puntual, no obstante, permite ajustar una gran variedad de modelos, llegando a ser más rápido computacionalmente que `R-INLA` si se utiliza adecuadamente (Bachl *et al.*, 2019).

En lo referente a las ventajas del paquete `inlabru`, si pensamos en `R-INLA`, para poder utilizarlo el usuario necesita tener cierto conocimiento sobre qué aproximaciones hay detrás. En contraposición, `inlabru` pretende evolucionar a una versión de `R-INLA` más accesible y amigable para el usuario (Bachl *et al.*, 2019). Del mismo modo, cuando abordamos un modelo con un proceso de patrón puntual subyacente, en `R-INLA` se asume que la probabilidad de detección del punto es conocida y constante (Bachl *et al.*, 2019). Este hecho, no siempre ocurre, ya que en temas biológicos la detección de un punto puede depender de infinidad de factores. Por ello, `inlabru` permite modelizar dicha probabilidad de detección como una probabilidad desconocida, para luego estimarla.

Gracias a las ventajas mencionadas y la habilidad de `inlabru` para ajustar modelos espacio temporales combinando distintos tipos de datos espaciales, p.ej. patrones puntuales y datos geoestadísticos, se va a utilizar este paquete para inferir y predecir sobre algunos de los modelos que se propondrán más adelante.

2.6. Modelos de producción excedentaria

Para concluir el marco teórico, el objetivo de nuestro trabajo consiste en evaluar qué modelización de los índices de biomasa relativa y de CPUE captura mejor el comportamiento de la biomasa real. Pero, ¿por qué queremos que estos índices sean representativos de la biomasa? Los índices de biomasa relativa y de CPUE suelen emplearse como *input* en modelos de producción excedentaria, en inglés *Surplus Production Models* (SPMs). Así pues, examinar la calidad de los *inputs* puede ayudarnos a conseguir mejoras en los modelos de evaluación del *stock* pesquero, en concreto, del tipo SPMs.

Los SPMs se centran en modelar la evolución en el tiempo de la biomasa agregada [10], combinando el efecto del crecimiento, el reclutamiento [11] y la mortalidad en una única función de producción. En sí, los SPMs estiman los cambios en la biomasa a lo largo de un periodo, como una función de la biomasa en el tiempo anterior, la producción excedentaria de biomasa (la biomasa 'disponible' para la pesca) y las capturas. Por todo lo mencionado, los SPMs se conocen también como *Biomass Dynamic Models*.

La formulación de los SPMs tiene su base en la ecuación de Russell (1931), donde se establece que los cambios en la biomasa de un *stock* a lo largo de un periodo dependen del reclutamiento, crecimiento, mortalidad natural y capturas que se dan en un *stock* :

$$B_{t+1} = B_t + f(B_t) - C_t, \tag{2.25}$$

donde, B_{t+1} es la biomasa del *stock* al final del año t o al principio del año $t+1$, B_t es la biomasa del *stock* al comenzar el año t, $f(B_t)$ es una función de producción de biomasa (recoge los reclutamientos y el crecimiento menos la mortalidad natural del *stock*) y C_t son las capturas durante el año t.

En vista a la ecuación (2.25), para estimar la serie de biomasa de un modelo de SPMs necesitamos relacionar B_t con los datos de los que disponemos, es decir, los índices de biomasa relativa o de CPUE. Esto se realiza a través del coeficiente de capturabilidad q:

$$\hat{I}_t = C_t/E_t = qB_t, \tag{2.26}$$

donde $\hat{I}_t$ es un índice estimado de biomasa relativa o de CPUE para el año t. En teoría, la serie temporal de índices predicha con los distintos modelos se considera proporcional a la biomasa del *stock* B_t, de manera que $\hat{I}_t$ y B_t se relacionan por el coeficiente de capturabilidad q, el cual se supone constante en el espacio y tiempo. C_t y E_t son las capturas y el esfuerzo de pesca respectivamente.

[10] Hablamos de biomasa agregada porque no hay diferenciación por características biológicas (tallas, sexo, edades, etc.).

[11] Llamamos reclutas a los individuos que se incorporan por primera vez en un stock.

A continuación, tenemos que estimar la función de producción $f(B_t)$ (2.25). Existen diferentes formas para estimar $f(B_t)$, de entre ellas destaca la formulación de Pella y Tomlinson (1969), ya que proponen una estimación en la que se ve implicada la tasa de crecimiento r del *stock*, la capacidad de carga del ecosistema K y un parámetro de asimetría de la curva p (que permite curvas de producción asimétricas generalizando la curva de Schaefer (1954)):

$$f(B_t) = \frac{r}{p} B_t \left(1 - \left(\frac{B_t}{K}\right)^2\right). \tag{2.27}$$

En resumen, haciendo uso de las ecuaciones propuestas (2.26 y 2.27), a través de la inferencia bien sea desde la perspectiva frecuentista o bayesiana, se estiman los diferentes parámetros q, K, r y p junto a la serie de biomasa.

Una vez explicado el contexto general de los modelos SPMs, se ha de mencionar que existen diferentes métodos de estimación de los parámetros mencionados. La principal diferencia entre las distintas propuestas suele residir en dónde asumen el error residual, es decir, si el error se encuentra en los datos, en el modelo o en ambos:

- **Error de proceso**, asumir un error de proceso significa que las observaciones se han recogido con un error, y que el error se encuentra en la ecuación de la dinámica del *stock* (2.25).
- **Error de estimación**, se asume que todos los errores se encuentran en las observaciones (capturas e índices de biomasa), mientras que, la ecuación de la dinámica del *stock* es determinista y sin error.

De entre los SPMs más relevantes destacan: (1) ASPIC (*A Surplus-Production Model Incorporating Covariates*) (Prager, 1992, 1994), (2) SPiCT (*Surplus-Production model in Continuous Time*) (Pedersen y Berg, 2017) y (3) JABBA (*Just Another Bayesian Biomass Assessment*) (Winker *et al.*, 2018). Cada uno de estos SPMs presenta una serie de ventajas y desventajas. En nuestro caso, hemos escogido aplicar un modelo SPiCT, sobre el que entraremos en detalle a continuación.

2.6.1. SPiCT

De entre todos los SPMs disponibles ¿por qué hemos decidido utilizar SPiCT? Normalmente, este modelo suele ser el más completo y obtiene buenos resultados en la estimación de la serie de biomasa. En principio, podríamos enumerar las siguiente ventajas de SPiCT frente a otros SPMs:

1. Es un modelo en tiempo **continuo**, por lo general los bancos de datos de los que disponemos contienen series de capturas durante todo el año, así que, a diferencia de otros SPMs, SPiCT no necesita utilizar aproximaciones inexactas para modelar la relación entre la biomasa y las capturas (utiliza integración numérica).

2. SPiCT considera error de observación. El error de observación lo incluye en las capturas y, también supone error de observación en la serie de índices de biomasa relativa o de CPUE:

$$\log(C_t) = \log\left(\int_t^{t+\Delta_t} F_s B_s ds\right) + \epsilon_t, \tag{2.28}$$

   donde C_t son las capturas en un intervalo de tiempo Δ_t y con un error de observación $\epsilon \sim \mathrm{N}(0, \sigma_C^2)$. Además B_t es la biomasa explotable del *stock* y F_t la tasa de mortalidad instantánea.

$$\log(I_{t,i}) = \log(q_i B_t) + e_{t,i}, \tag{2.29}$$

   donde $e_{t,i} \sim \mathrm{N}(0, \sigma_{I,i}^2)$ es un error distribuido como una Normal, $I_{t,i}$ son los distintos índices i en el tiempo t, q_i el coeficiente de capturabilidad y B_t la biomasa en el tiempo t.

3. También se considera error de proceso. El error de proceso se incluye en la formulación del modelo que está basada en la ecuación de Pella y Tomlinson (1969), pero con una reparametrización más estable para la estimación de los parámetros:

$$\frac{dB_t}{dt} = \left(\gamma m \frac{B_t}{K} - \gamma m \left(\frac{B_t}{K}\right)^n - F_t B_t\right) dt + \sigma_B B_t dW_t, \tag{2.30}$$

   donde σ_B es la desviación estándar del proceso residual y W_t es un movimiento browniano.

4. Permite modelar patrones estacionales en la mortalidad por pesca F.

Capítulo 3

Protocolo para evaluar modelos en pesquerías.

En este capítulo se ilustra el protocolo que hemos desarrollado para comparar entre distintos modelos utilizados en gestión de pesquerías, con el fin de estimar y predecir una serie temporal de índices de biomasa relativa o de capturas por unidad de esfuerzo (CPUE) [1] representativa de la biomasa del *stock*. Posteriormente, dichas series de índices relativos alimentaran modelos de evaluación del *stock* del tipo SPMs.

Cabe destacar que, para poder señalar qué modelización de los índices obtiene un mejor *input* para el modelo de evaluación, se ha de valorar la similitud entre la serie de índices y la biomasa real. Para ello, se requiere conocer la biomasa absoluta de la especie a lo largo del espacio y el tiempo. No obstante, a nivel práctico muestrear todo un *stock* durante años es inverosímil. Como consecuencia, hemos elaborado una **simulación de un modelo** de biomasa, en la que pueden asumirse distintos escenarios en el espacio y durante un número de años a determinar por el usuario.

A partir de la biomasa simulada hemos recreado distintos escenarios de **muestreo** con los que poder reproducir los bancos de datos de los que suele disponer el investigador para llevar a cabo sus análisis. Es necesario mencionar que, los bancos de datos pesqueros no contienen la variable de biomasa absoluta de un *stock*, sino, una serie de capturas de la especie a la que denominan biomasa relativa (cuando el muestreo es independiente de la pesca, p.ej. campañas oceanográficas) o capturas por unidad de esfuerzo (cuando el muestreo es dependiente de la pesca, p.ej. pesquerías). Por tanto, para poder reproducir con exactitud los bancos de datos se han de multiplicar los valores de biomasa simulada obtenidos en el muestreo por una constante

[1] En el contexto de las pesquerías de arrastre, el esfuerzo en pesquerías se refiere al tiempo que permanece un arte de pesca activo. Matizar que, la medida del esfuerzo no tiene porque ser tiempo, también puede ser número de redes, número de anzuelos, etc.

de proporcionalidad a la que llamaremos constante de capturabilidad q.

Una vez disponemos de los bancos de datos ya podemos proceder a **modelizar** las distintas opciones con las que podríamos encontrarnos en la literatura (GLMMs, GAMMs, modelos espacio-temporales, etc.). Para a continuación, **inferir** y **predecir** sobre los parámetros de los modelos propuestos, donde los índices de biomasa relativa o de CPUE actúan como variable respuesta.

En lo referente a la inferencia y predicción de los índices, se ha resuelto mediante ambas perspectivas: la inferencia la **frecuentista** y la inferencia **bayesiana**. Así mismo, se han utilizado como herramientas para inferir y predecir en el contexto bayesiano los paquetes `R-INLA`, `inlabru` y `R2BayesX` y, para la inferencia y predicción clásica el paquete `mgcv`. Por último, las serie de índices de biomasa relativa y CPUE predichas con cada uno de los modelos se compararán a través de medidas de error como RMSE (error cuadrático medio) y MAPE (error porcentual absoluto medio) con la biomasa simulada. Para así, utilizar aquella predicción con menor RMSE y MAPE como *input* en un modelo SPiCT y, mostrar los distintos *outputs* que ofrecen este tipo de modelos para la gestión de un *stock* pesquero.

3.1. Simular de un modelo

Para poder simular la biomasa o cualquier variable aleatoria es necesario entender y controlar que procesos están detrás de su comportamiento. Una forma de integrar y controlar los procesos que afectan a una variable aleatoria es a través de la modelización estadística. Por consiguiente, lo primero que necesitamos es simular de un modelo de biomasa.

3.1.1. Modelización de la biomasa

En primer lugar, cuando uno quiere modelizar una variable ha de plantearse que procesos están condicionando el comportamiento de la variable que se quiere modelar. En este caso, según la literatura y el criterio de expertos, la biomasa de un *stock* pesquero suele estar condicionada por una fuerte componente espacio-temporal y por diferentes variables ambientales. Entre ellas, la batimetría es la variable que más suele afectar a los *stocks* (Hinton y Maunder, 2004; Stock *et al.*, 2019; Izquierdo *et al.*, 2021). Por lo tanto, se establece un modelo geoestadístico con un efecto espacial correlado a través de un proceso autorregresivo de orden 1 y una tendencia temporal para la biomasa, es decir, se recogen tres efectos clave: (1) el efecto de la batimetría,

(2) el efecto espacio-temporal y (3) la tendencia temporal de la biomasa.

$$\begin{gathered} \text{Biomasa(s,t)} \sim \text{Gamma}(\mu(s,t), \phi) \\ \log(\mu(s,t)) = \beta_0 + f(\text{Batimetría}) + f(\text{tiempo}) + v(s,t)\,, \\ v(s,t) = \rho \times v(s,t-1) + U(s,t)\,, \\ U(s,t) \sim \text{GMRF}(0, \Sigma) \end{gathered} \tag{3.1}$$

donde, Biomasa(s,t) es la variable respuesta biomasa en un espacio s y tiempo t y, $\mu(s,t)$ y ϕ son la media y la dispersión de la distribución, respectivamente. De igual manera, $\mu(s,t)$ viene enlazada a un predictor lineal por la función logaritmo, donde β_0 es el intercepto, f(Batimetría) es una función suave para la covariable batimetría, f(tiempo) es una función suave que recoge la tendencia temporal y $v(s,t)$ es el efecto aleatorio espacio-temporal, donde el efecto del tiempo tiene un comportamiento autorregresivo AR de orden uno y el efecto espacial $U(s,t)$ está distribuido como un *Gaussian Markovian Random Field* GMRF de media cero y matriz de covarianza Σ.

Con el modelo de biomasa planteado ya podemos simular los distintos escenarios de la variable biomasa en el espacio-tiempo que nos permitirán llevar a cabo la comparación con la biomasa estimada.

3.1.2. Simulación

La simulación se basa en reproducir de manera artificial el comportamiento de una variable aleatoria bajo unas condiciones controladas. En consecuencia, al simular una variable conocemos su valor en el total de la población, lo que nos permite jugar con los modelos, pudiendo valorar cómo de bueno es el ajuste (Figura 3.1). Diversos estudios han simulado con anterioridad el comportamiento espacial y espacio-temporal de la variable biomasa (Paradinas *et al.*, 2017; Pennino *et al.*, 2019). Paradinas *et al.* (2017) simulaban la biomasa en distintos escenarios para testear la unión de verosimilitudes en modelos geoestadísticos. Del mismo modo, Pennino *et al.* (2019) simulaban la biomasa para mostrar la efectividad del ajuste de modelos preferenciales.

Pueden existir una infinidad de razones por las que es necesario simular una variable, en nuestro caso como el objetivo del trabajo requiere conocer la biomasa a lo largo del tiempo y, conseguir este dato en la vida real es básicamente imposible, hemos desarrollado un método que nos permite **simular de un modelo** de biomasa (3.1) y partiendo de dicha biomasa simulada reproducir distintos escenarios de muestreo con los que obtener las fuentes de información típicas en pesquerías para poder modelar, inferir y predecir.

Desglosando como se ha simulado la biomasa, se ha elaborado un *script* en el que se han de fijar una serie de parámetros, a fin de ir simulando cada una de las componentes del predictor lineal (3.1):

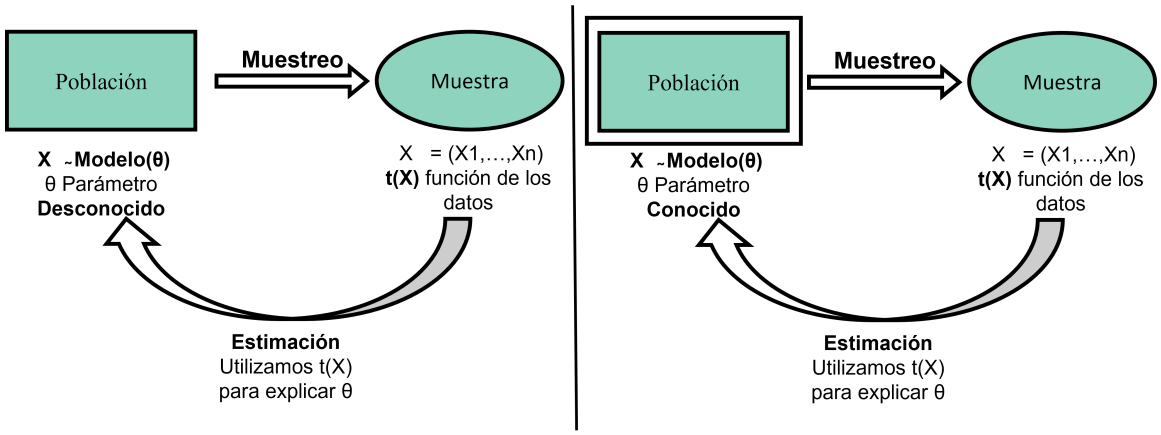

Figura 3.1: Población desconocida frente a población conocida.

1. **Efecto espacial correlacionado** $v(s,t)$. Krainski *et al.* (2018) utilizan un ejemplo de una simulación espacial correlacionada en el tiempo, en el que ilustran la función *book.rspde* disponible en el archivo *sdpe-book-function.R*. Con ayuda de esta función hemos lleva a cabo la simulación del término espacial como un *Gaussian Markovian Random Field* GMRF de matriz de covarianza Σ y media 0. Además, el efecto espacial simulado se correlaciona en el tiempo como un AR de primer orden (3.1) (Figura 3.2).

2. **Batimetría** $f(\textbf{Batimetría}(s,t))$. El rango de batimetrías en el que se pesca está condicionado por la biología de la especie. Por consiguiente, al no especificar una especie, se ha escogido un rango de batimetría de 0 a 800 metros y uno de los efectos más comunes sobre la biomasa, siendo que las localizaciones de máxima biomasa se encuentran a batimetría intermedias (Figura 3.3).

3. **Tendencia temporal** $f(\textbf{tiempo})$. Normalmente la biomasa cambia a lo largo del tiempo, por tanto, se ha añadido un término que recoge esa tendencia a lo largo del periodo de estudio. Este término es un vector de valores fijado, de manera que se le suma un valor distinto a cada año.

Una vez simulados todos los términos del predictor lineal, se ha de construir la variable biomasa a partir de las componentes que conforman la media de la distribución $\mu(s,t)$ (3.1). Para ello, primero han de sumarse cada una de las partes simuladas del predictor (intercepto, efecto espacial, batimetría y tendencia temporal).

En el siguiente código se muestra como $\mu(s,t)$ está formada por un intercepto (β_0), un polinomio de grado dos para la batimetría (β_1 y β_2) (se añade un polinomio para que la relación no sea lineal), el término espacial correlacionado en el tiempo $v(s,t)$ y una tendencia temporal *vector tiempo*:

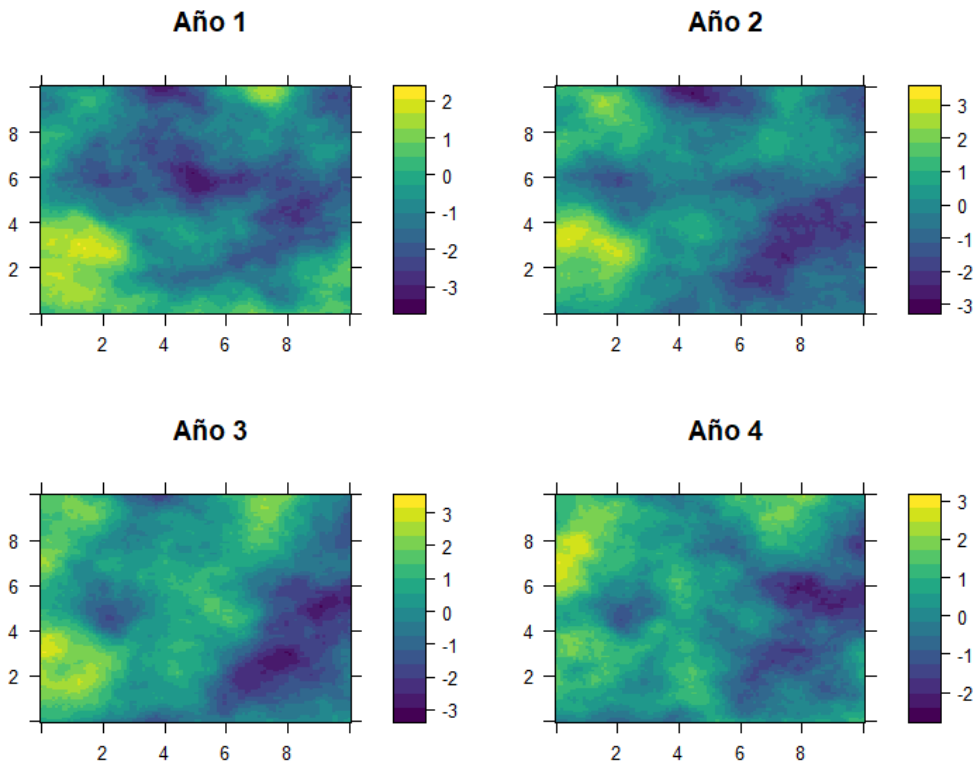

Figura 3.2: Simulación del efecto espacio-temporal.

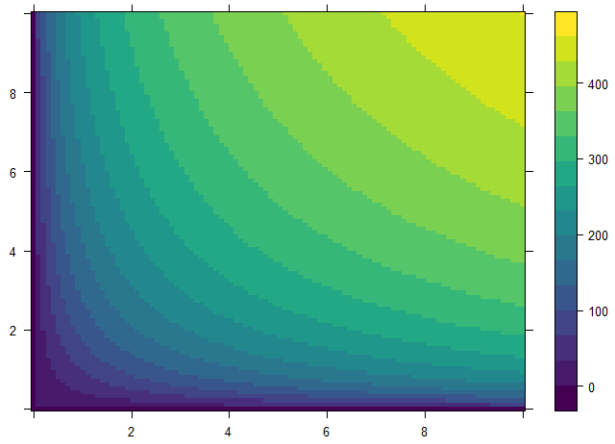

Figura 3.3: Simulación de la batimetría.

```r
# Media
    # predictor lineal
    lin_pred_mu <- list()
    for (i in 1:k) {
        lin_pred_mu[[i]] <-
            exp(beta_0 + beta_1 * bat + beta_2 * bat^2 + v + vector_tiempo)
    }
```

Cuando ya tenemos la media de la distribución $\mu(s,t)$, con ayuda de la función *rgamma* [2] simulamos la variable biomasa. En la Figura 3.4 ilustramos un ejemplo de biomasa simulada, donde es posible observar como a batimetrías intermedias aparecen parches de alta biomasa. Si nos fijamos en este ejemplo, los parches de máxima biomasa se van desplazando en el espacio-tiempo, mostrando una comportamiento autorregresivo (Figura 3.4).

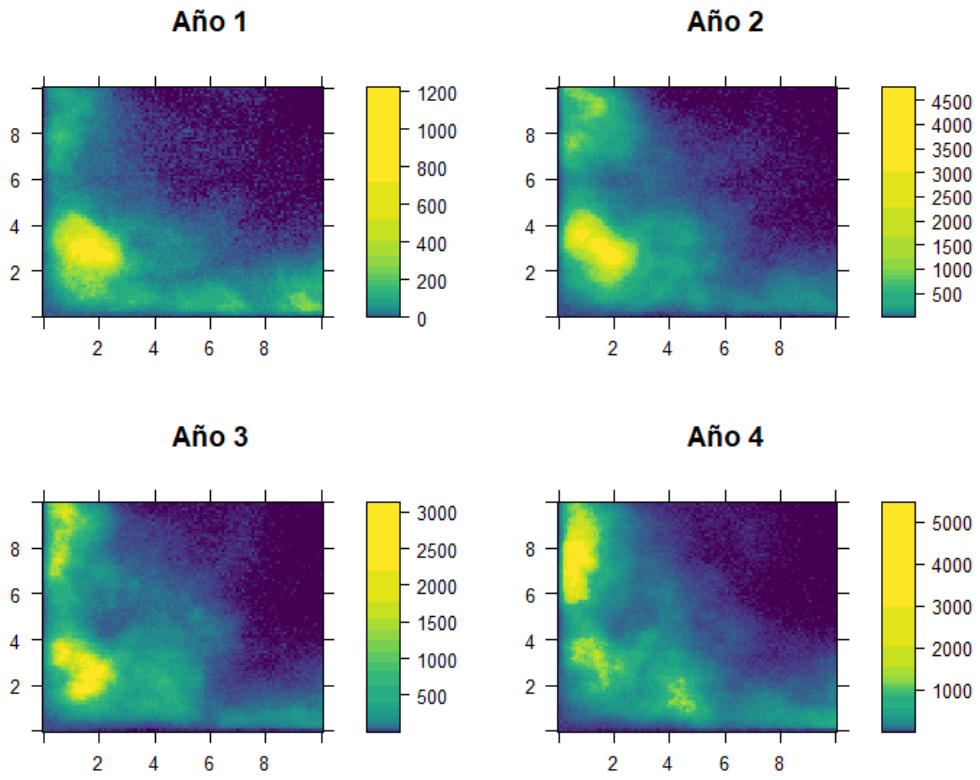

Figura 3.4: Simulación de la biomasa.

Ya simulada la biomasa absoluta de un *stock* pesquero, se pretenden recrear los bancos de datos de los que dispone el personal investigador para llevar a cabo la evaluación del *stock*.

[2] En R la distribución Gamma viene dada en términos de α y β, por tanto, se ha reparametrizando la distribución para poder introducir la media $\mu(s,t)$ de la distribución, y así simular la biomasa.

3.2. Reproducir bancos de datos pesqueros

Hasta este punto se ha ejemplificado la simulación de la biomasa en un espacio-tiempo. No obstante, el investigador únicamente dispone de una fotografía de unas cuantas localizaciones con un valor proporcional a la biomasa que denominamos biomasa relativa o capturas por unidad de esfuerzo (CPUE). En la vida real, estos bancos de datos pueden conseguirse a través de fuentes de información como son las campañas oceanográficas o las pesquerías. De manera que para recrear el conjunto de datos, primero debemos imitar el comportamiento en el muestreo de una campaña oceanográfica y de las pesquerías.

3.2.1. Muestrear de la simulación

En estadística, para conseguir un banco de datos que nos permita describir, modelar y realizar inferencia y predicción sobre una variable aleatoria, p.ej. la biomasa, es necesario tener una muestra de la población. Por consiguiente, en este trabajo se van de reproducir diferentes tipos de muestreos [3] recurrentes en pesquerías, con el fin de recrear las principales fuentes de datos con las que habitualmente trabajan los científicos en gestión pesquera:

- **Muestreo aleatorio**: consiste en recrear una campaña oceanográfica (datos independientes de la pesca). Son los propios investigadores los que se embarcan durante periodos para pescar, de forma que diseñan todo un experimento que les permite conseguir una muestra representativa de la biomasa de la población. Dicho experimento consiste en realizar una serie de lances de pesca aleatorios y georreferenciados con un esfuerzo constante. Al ser el muestreo aleatorio y el esfuerzo constante, los datos que conseguimos se pueden considerar una medida relativa de la biomasa (índices de biomasa relativa).
- **Muestreo preferencial**: consiste en recrear la labor de los observadores en pesquerías (datos dependientes de la pesca), se trata de un muestreo preferencial georreferenciado. Cuando hablamos de un muestreo preferencial, esto quiere referirse al tipo de patrón de puntos que ocasiona el comportamiento de los pescadores. En otras palabras, el capitán de un buque pesquero, debido a su experiencia, siempre recurre a los mismos caladeros de pesca, ya que conocen donde están los parches de máxima biomasa. Este comportamiento condiciona las zonas de muestreo, generando una muestreo agrupado, donde únicamente se han muestreado las regiones con una biomasa elevada. Además, el esfuerzo de pesca no es constante, por tanto, la información viene dada como índices de captura por unidad de esfuerzo (CPUE).

[3] Los muestreos están reproducidos pensando en un arte de pesca concreto: **arrastre**.

En lo referente al muestreo de la simulación, para el muestreo aleatorio escogemos 100 localizaciones al azar (Figura 3.5) con la función *sample*. Por el contrario, para el muestreo preferencial se seleccionan 100 localizaciones de máxima biomasa (Figura 3.6) añadiendo un vector de probabilidades a la función *sample*.

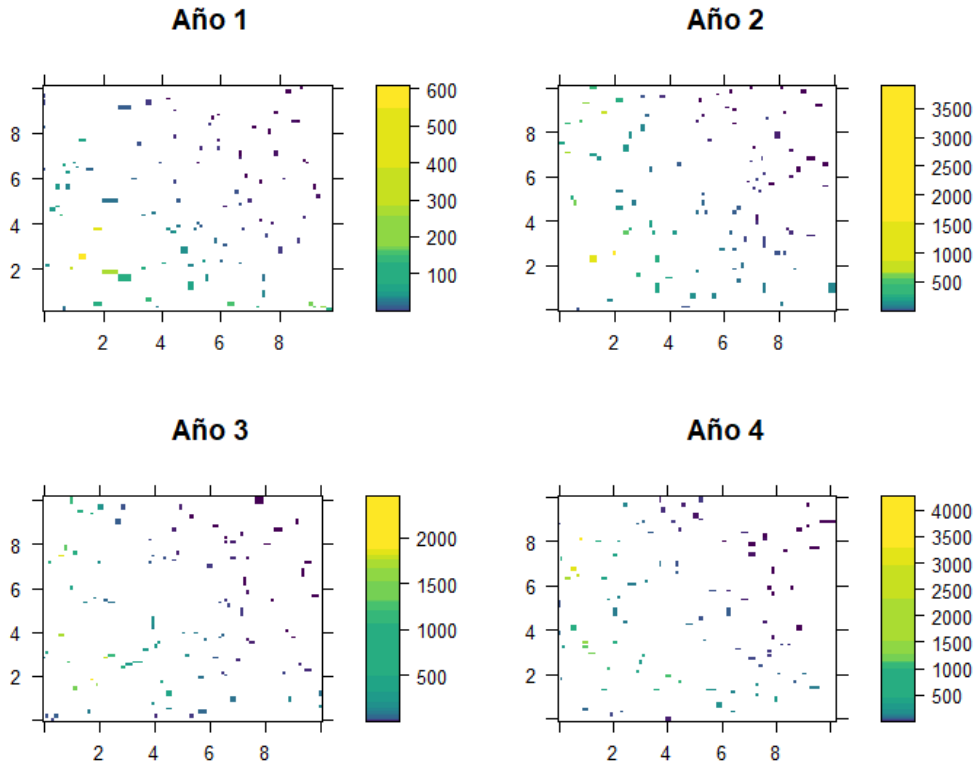

Figura 3.5: Muestreo aleatorio (independiente de la pesca). La escala de color se corresponde con los valores de biomasa muestreados.

3.2.2. Índices de biomasa relativa o de CPUE y capturas

A continuación, gracias a la reproducción de los muestreos podemos imitar los dos *inputs* clásicos en los modelos de evaluación del tipo SPMs: (1) la serie de capturas, la cual contempla la totalidad de registros sobre la captura de una especie y (2) los índices de biomasa relativa o de CPUE, que pretenden ser una serie representativa de la biomasa real del *stock*.

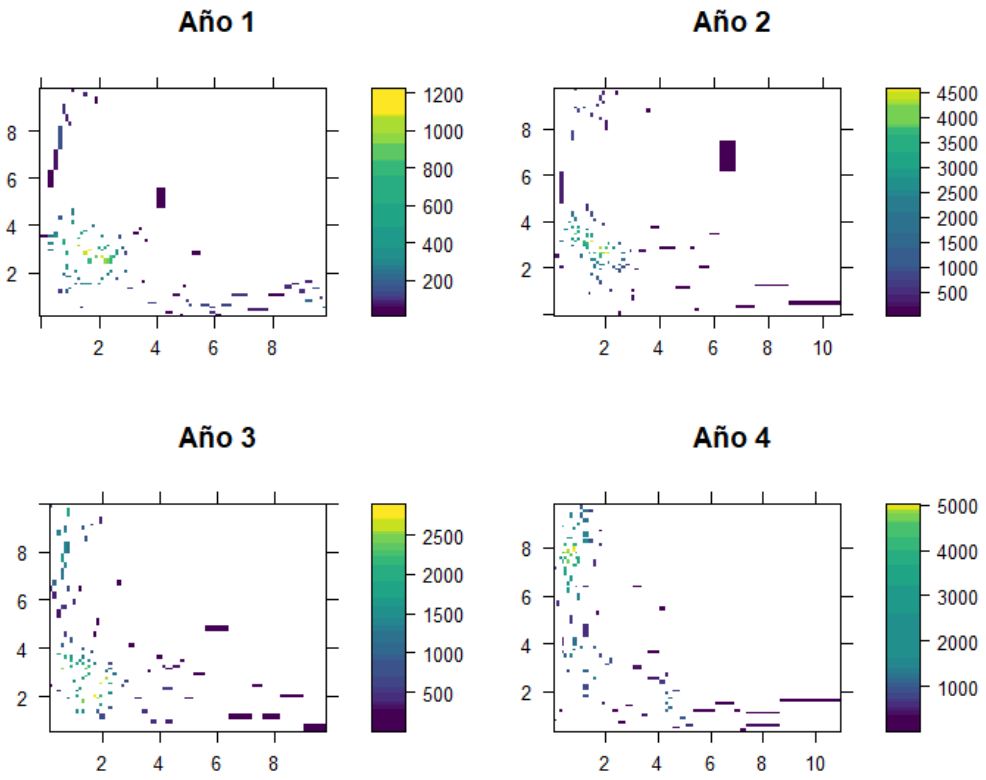

Figura 3.6: Muestreo preferencial (dependiente de la pesca). La escala de color se corresponde con los valores de biomasa muestreados.

Capturas

En primer lugar, gracias a que conocemos la biomasa en toda nuestra región de estudio, podemos aproximar la serie de capturas a partir de la formulación de un modelo SPMs (2.25 y 2.27), asumiendo una curva de Schaefer (1954) en la función de producción $f(B_t)$:

$$\begin{aligned} B_{t+1} &= B_t + f(B_t) - C_t, \\ C_t &= B_t - B_{t+1} + f(B_t), \\ C_t &= B_t - B_{t+1} + rB_t\left(1 - \frac{B_t}{K}\right), \end{aligned} \tag{3.2}$$

donde t es el año correspondiente, $t+1$ el año siguiente, C_t hace referencia a las capturas en el año t, B_t es la biomasa simulada en el año t, $f(B_t)$ es la función de producción, la cual se relaciona con un parámetro r que es la tasa de crecimiento y K que se corresponde con la capacidad de carga del *stock*.

En base a la ecuación anterior (3.2) y con el fin de conocer las capturas para cada año, damos un valor a los parámetros r y K, que junto al valor real de la biomasa simulada en el año t, nos permite calcular las capturas para todos los años, a excepción del último año. Por consiguiente, para poder establecer las capturas globales del *stock* en el último año de la serie, como no conocemos el valor de la biomasa real para el año que sigue al último, hemos asumido un equilibrio, es decir, las biomasa en el año t, siendo este el último de la serie, es la misma que en $t+1$, por tanto, solo afecta la función de producción $f(B_t)$.

Índices de biomasa relativa o de CPUE

Como resultado, al reproducir los muestreos obtenemos una serie de valores de biomasa en las localizaciones seleccionadas. Sin embargo, como ya se ha comentado, los muestreos en pesquerías no obtienen un valor absoluto de biomasa, ya que no es posible pescar toda la biomasa de una especie, sino que, se captura una proporción de la población. De ahí que, el muestreo aleatorio proporcione una serie de índices de biomasa relativa y el muestreo preferencial facilite una serie de índices de captura por unidad de esfuerzo (CPUE).

Ambos índices son proporcionales a la biomasa, por ende, para generar estos índices se cogerá la biomasa seleccionada en los puntos de muestreo y se multiplicará por una constante q a la que denominamos **constante de capturabilidad**. El coeficiente q se define como una medida que relaciona la abundancia del recurso con el esfuerzo pesquero (Arreguín-Sánchez, 1996).

Por otro lado, un detalle a tener en cuenta es el esfuerzo que se ha realizado al pescar. En el muestreo aleatorio el tiempo de pesca es constante para cada lance, por tanto, dichas capturas pueden entenderse como un índice de biomasa relativa. En contraposición, para el muestreo preferencial el esfuerzo no es constante, es decir, en pesquerías cada pescador decide el tiempo que el arte de pesca permanece activo, lo cual puede influir en la capturabilidad de la especie. En consecuencia, se ha de simular una variable de esfuerzo lineal, [4] de manera que se estandarizan las capturas dividendo por el esfuerzo de pesca, y así, obtenemos los índices de CPUE habituales en pesquerías:

- Muestreo aleatorio (campaña oceanográfica):

$$\text{Índice de Biomasa relativa} = \text{Biomasa simulada x } q. \tag{3.3}$$

[4] En nuestro trabajo, hemos supuesto un esfuerzo lineal, sin embargo, el esfuerzo no tiene porque tener un comportamiento lineal, muchas veces la red de pesca se satura y llega un momento que se estabiliza la curva presentando la biomasa un comportamiento logístico frente al esfuerzo.

- Muestreo preferencial (pesquerías):

$$\text{Índice de CPUE} = \frac{\text{Biomasa simulada} \times q}{\text{esfuerzo}}. \tag{3.4}$$

En definitiva, para cada muestreo hemos recreado un banco de datos con los índices de biomasa relativa o de CPUE asociados. Como resultado, estos índices se utilizarán como variable respuesta para estimar y predecir una serie temporal de índices de biomasa relativa o de CPUE con distintos modelos estadísticos utilizados en gestión de pesquerías (GLMs, GAMs y modelos geoestadísticos).

3.3. Modelización de los índices

Gracias al proceso de simulación y muestreo, hemos generado dos bancos de datos que reproducen dos de las principales fuentes de información relacionadas con la evaluación del estado del *stock*. Así pues, se van a plantear una serie de modelos sobre los índices de biomasa relativa y de CPUE con el fin de estimar y predecir en toda la zona de estudio y en un periodo concreto dichos índices.

La modelización de los índices de biomasa relativa o de CPUE ha ido evolucionando de modelos más simples a modelos más complejos a lo largo de las últimas décadas (Stock *et al.*, 2019). En general, se utilizaban GLMMs o métodos similares, incluso algunos GAMs, a fin de estimar una serie temporal de estos índices. Sin embargo, en algunas ocasiones, estos modelos pueden derivar en sesgos a la hora de estimar la biomasa en el modelo de evaluación del *stock*, sobretodo, en series de índices de CPUE derivados de pesquerías (Maunder *et al.*, 2020). De ahí que, poco a poco modelos más complejos como los geoestadísticos hayan ido ganando relevancia (Maunder *et al.*, 2020), resultando ser muy útiles en la toma de decisiones sobre el estado del *stock*.

Por ello, en este trabajo se han planteado un total de 7 modelos divididos en dos bloques: (1) modelos relacionados con la variable respuesta obtenida en el muestreo aleatorio **índices de biomasa relativa** y (2) modelos relacionados con la variable respuesta obtenida en el muestreo preferencial **índices de CPUE**. Para cada una de las variables, se han explorado modelos más simples como puede ser un GLM hasta modelos más complejos como puede ser un modelo geoestadístico.

3.3.1. Modelización índices de biomasa relativa: muestreo aleatorio

Modelo lineal generalizado (GLM)

Uno de los modelos más sencillos que podríamos proponer para modelizar los índices de biomasa relativa es un GLM, de manera que incluimos el tiempo como un factor fijo. El principal problema de estos modelos es que no contemplan la variabilidad espacial [5].

Por un lado, plasmamos la verosimilitud del modelo:

$$\begin{aligned} \mathrm{BR}_{ij} &\sim \mathrm{Gamma}(\mu_{ij}, \phi) \quad i = 1, ..., n, \quad j = 1, .., k\,, \\ log(\mu_{ij}) &= \beta_0 + \beta_1 \mathrm{Batimetría}_{ij} + \beta_2 \mathrm{Batimetría}_{ij}^2 + \alpha_{ij} \mathrm{tiempo}_{ij}\,, \end{aligned} \tag{3.5}$$

donde los subíndices i y j hacen referencia al número de observaciones y años, respectivamente. BR_{ij} es la variable respuesta índice de biomasa relativa, definida por una distribución Gamma donde ϕ es la dispersión y μ_{ij} la media de la distribución. μ_{ij} está enlazada al predictor lineal por la función de enlace logaritmo, β_0 hace referencia al intercepto, β_1 y β_2 se corresponden con los coeficiente asociados a la covariable batimetría ($\mathrm{Batimetría}_{ij}$), y $\alpha_{ij}\mathrm{tiempo}_{ij}$ es un factor fijo relacionado con la variable categórica años.

A continuación, en el contexto bayesiano, necesitaríamos incorporar una priori a cada uno de los parámetros y a los hiperparámetros a estimar del modelo. Además, si resolvemos el modelo con INLA las distribuciones a priori de los parámetros que componente el *latent field* tienen que ser Normales.

Distribuciones a priori para el *Latent Gaussian Field*:

$$\{\beta_0, \beta_1\,, \beta_2\,, \alpha_j\} \sim \mathrm{N}(0, \tau = 0{,}001), \tag{3.6}$$

donde β_0, β_1, β_2 y α_j, siendo j el número de años, son los parámetros del *latent gaussian field* y están definidos con una distribución de probabilidad Normal de media cero y precisión τ conocida.

Hiperparámetros:

$$\phi \sim \mathrm{F}(\phi), \tag{3.7}$$

donde ϕ es un hiperparámetro asociado a la dispersión de la distribución. En el caso de los hiperparámetros se puede suponer una distribución previa F que no sea Normal.

[5] A veces, los investigadores añaden algún factor fijo para representar la componente espacial, por ejemplo, el puerto donde se descarga el producto.

Modelo aditivo generalizado (GAM)

El siguiente escalón sería modelizar un GAM, donde ya podríamos asumir una relación no lineal entre el índice de biomasa relativa y la batimetría. También, podemos contemplar el término espacial añadiendo un suavizado bivariante o un tensor y la tendencia temporal añadiendo el tiempo como factor fijo.

En este caso, en el contexto bayesiano, para ajustar el modelo con el suavizado bivariante hemos utilizado la librería `R2BayesX`, la cual está basada en métodos MCMC para aproximar las distribuciones a posteriori de los parámetros.

Por un lado, plasmamos la verosimilitud del modelo:

$$\begin{aligned} \text{BR}_{ij} &\sim \text{Gamma}(\mu_{ij}, \phi) \quad i = 1, ..., n, \quad j = 1, .., k\,, \\ log(\mu_{ij}) &= \beta_0 + f(\text{Batimetría})_{ij} + f(x, y)_{ij} + \alpha_{ij}\text{tiempo}_{ij}\,, \end{aligned} \tag{3.8}$$

donde i y j hacen referencia al número de observaciones y años, respectivamente. BR_{ij} es la variable respuesta índice de biomasa relativa, definida por una distribución Gamma donde ϕ es la dispersión y μ_{ij} la media de la distribución. μ_{ij} está enlazada al predictor lineal por la función de enlace logaritmo, β_0 hace referencia al intercepto, $f(\text{Batimetría})_{ij}$ se corresponde con una función suave aplicada sobre la covariable batimetría, $f(x, y)_{ij}$ representa un suavizado bivariante [6] para las localizaciones y $\alpha_{ij}\text{tiempo}_{ij}$ se corresponde con un factor fijo para la variable categórica tiempo.

En el caso del paquete `R2BayesX`, las distribuciones a priori de los coeficientes del predictor no es necesario que sean distribuciones Normales, ya que `R2BayesX` trabaja con métodos MCMC. Si se quiere profundizar más sobre cómo funciona la aproximación del paquete `RBayesX`, sus distribuciones a priori por defecto para coeficientes e hiperparámetros del modelo, se nombran las siguientes referencias Osei *et al.* (2012) y Umlauf *et al.* (2012).

Modelo geoestadístico-temporal

Por último, un modelo más complejo con el que podemos modelizar la variable índice de biomasa relativa es un modelo geoestadístico autorregresivo, al que se le añade una tendencia temporal.

[6] Se contempla el efecto conjunto de dos covariables continuas a partir de una función suave.

Función de verosimilitud:

$$\begin{aligned} \mathrm{BR}(s,t) &\sim \mathrm{Gamma}(\mu(s,t),\phi)\,, \\ log(\mu(s,t)) &= \beta_0 + f(\text{Batimetría})(s,t) + f(\text{tiempo})(s,t) + v(s,t)\,, \\ v(s,t) &= \rho \times v(s,t-1) + \mathrm{U}(s,t)\,, \\ \mathrm{U}(s,t) &\sim \mathrm{GMRF}(0,\Sigma)\,, \end{aligned} \tag{3.9}$$

donde $\mathrm{BR}(s,t)$ es la variable respuesta índice de biomasa relativa definida con una distribución Gamma donde ϕ es la dispersión y $\mu(s,t)$ la media de la distribución. $\mu(s,t)$ está enlazada al predictor lineal por la función de enlace logaritmo, β_0 hace referencia al intercepto, $f(\text{Batimetría})(s,t)$ se corresponde con una función suave aplicada sobre la covariable batimetría y $v(s,t)$ es el efecto espacial correlacionado. El espacio $U(s,t)$ se modela como un *Gaussian Markov Random Field* de media cero y matriz de covarianza Σ. Este mismo efecto espacial está relacionado en el tiempo como un modelo autorregresivo de orden 1. Por último, $f(\text{tiempo})(s,t)$ es una función suave para representar la tendencia temporal de la biomasa.

Distribuciones a priori para el *Latent Gaussian Field*:

$$\begin{aligned} \{\beta_0\} &\sim \mathrm{N}(0, \tau = 0{,}001), \\ U(s,t) &\sim \mathrm{N}(0, \Sigma(\sigma_U, \kappa)), \end{aligned} \tag{3.10}$$

donde β_0 es el intercepto y $U(s,t)$ hace referencia al efecto espacial distribuido como una normal de media 0 y matriz de covarianza Σ, siendo σ_{U} y κ los hiperparámetros a estimar de la matriz.

Hiperparámetros:

$$\{\phi, \tau_{\text{Batimetría}}, \tau_{\text{tiempo}}, \rho, \sigma_{\mathrm{U}}, \kappa\} \sim \mathrm{F}(\{\phi, \tau_{\text{Batimetría}}, \tau_{\text{tiempo}}, \rho, \sigma_{\mathrm{U}}, \kappa\}), \tag{3.11}$$

donde ϕ hace referencia a la dispersión de la distribución de la variable, ρ es la correlación temporal del efecto espacial, $\tau_{\text{Batimetría}}$ es el hiperparámetro de un *random walk* de orden 2 ($rw2$) para la batimetría, τ_{tiempo} es el hiperparámetro de un *random walk* de orden 2 ($rw2$) para capturar la tendencia temporal y σ_{U} y κ los hiperparámetros a estimar de la matriz de covarianza Σ.

3.3.2. Modelización índices de CPUE: muestreo preferencial

Modelo lineal generalizado GLM

Función de verosimilitud:

$$\begin{aligned} \mathrm{CPUE}_{ij} &\sim \mathrm{Gamma}(\mu_{ij}, \phi) \quad i = 1,..,n, \quad j = 1,..,k\,, \\ log(\mu_{ij}) &= \beta_0 + \beta_1 \text{Batimetría}_{ij} + \beta_2 \text{Batimetría}_{ij}^2 + \alpha_{ij} \text{tiempo}_{ij}\,, \end{aligned} \tag{3.12}$$

donde los subíndices i y j hacen referencia al número de observaciones y años, respectivamente. CPUE_{ij} es la variable respuesta, se trata de una variable continua y positiva, por tanto, puede venir definida por una distribución de probabilidad Gamma donde ϕ es la dispersión y μ_{ij} la media de la distribución. μ_{ij} está enlazada al predictor lineal por la función de enlace logaritmo, β_0 hace referencia al intercepto, β_1 y β_2 se corresponden con los coeficientes asociados a la covariable batimetría (Batimetría_{ij}) y αtiempo_{ij} es un efecto fijo para la variable categórica años.

Distribuciones a priori para el *Latent Gaussian Field*:

$$\{\beta_0, \beta_1, \beta_2, \alpha_j\} \sim \text{N}(0, \tau = 0{,}001). \tag{3.13}$$

Hiperparámetros:

$$\phi \sim \text{F}(\phi). \tag{3.14}$$

Modelo aditivo generalizado (GAM)

Al igual que con los índices de biomasa relativa, para la perspectiva bayesiana hemos utilizado `R2BayesX` para ajustar un GAM.

Plasmamos la verosimilitud del modelo:

$$\begin{aligned} \text{CPUE}_{ij} &\sim \text{Gamma}(\mu_{ij}, \phi) \quad i = 1, ..., n, \quad j = 1, .., k\,, \\ log(\mu_{ij}) &= \beta_0 + f(\text{Batimetría})_{ij} + f(x, y)_{ij} + \alpha_{ij}\text{tiempo}_{ij}\,, \end{aligned} \tag{3.15}$$

donde i y j hacen referencia al número de observaciones y años, respectivamente. CPUE_{ij} es la variable respuesta índice de CPUE, definida por una distribución Gamma donde ϕ es la dispersión y μ_{ij} la media de la distribución. μ_{ij} está enlazada al predictor lineal por la función de enlace logaritmo, β_0 hace referencia al intercepto, $f(\text{Batimetría})_{ij}$ corresponde con una función suave aplicada sobre la covariable batimetría, $f(x, y)_{ij}$ representa un suavizado bivariante y $\alpha_{ij}\text{tiempo}_{ij}$ se corresponde con un factor fijo para la variable categórica tiempo.

Modelo geoestadístico-temporal

Función de verosimilitud:

$$\begin{aligned} \text{CPUE}(s, t) &\sim \text{Gamma}(\mu(s, t), \phi)\,, \\ log(\mu(s, t)) &= \beta_0 + f(\text{Batimetría})(s, t) + f(\text{tiempo})(s, t) + v(s, t)\,, \\ v(s, t) &= \rho \times v(s, t-1) + \text{U}(s, t)\,, \\ \text{U}(s, t) &\sim \text{GMRF}(0, \Sigma)\,, \end{aligned} \tag{3.16}$$

donde CPUE(s,t) es la variable respuesta índice de CPUE definido con una distribución Gamma donde ϕ es la dispersión y $\mu(s,t)$ la media de la distribución. $\mu(s,t)$ está enlazada al predictor lineal por la función de enlace logaritmo, β_0 hace referencia al intercepto, f(Batimetría) se corresponde con una función suave aplicada sobre la covariable batimetría y $v(s,t)$ es el efecto espacial correlacionado. El espacio se modela como un *Gaussian Markov Random Field* de media cero y matriz de covarianza Σ. Este mismo efecto espacial está relacionado en el tiempo como un modelo autorregresivo de orden 1. Por último, $f(\text{tiempo}(s,t))$ es una función suave para representar la tendencia temporal de la biomasa, esta se modela con un $rw2$.

Distribuciones a priori para el *Latent Gaussian Field*:

$$\begin{aligned} \{\beta_0\} &\sim \mathrm{N}(0, \tau = 0{,}001), \\ U(s,t) &\sim \mathrm{N}(0, \Sigma(\sigma_U, \kappa)), \end{aligned} \tag{3.17}$$

donde β_0 es el intercepto y $U(s,t)$ hace referencia al efecto espacial distribuido como una normal de media 0 y matriz de covarianza Σ, siendo σ_U y κ los hiperparámetros a estimar de la matriz.

Hiperparámetros:

$$\{\phi, \tau_{\text{Batimetría}}, \tau_{\text{tiempo}}, \rho, \sigma_{\text{U}}, \kappa\} \sim \mathrm{F}(\{\phi, \tau_{\text{Batimetría}}, \tau_{\text{tiempo}}, \rho, \sigma_{\text{U}}, \kappa\}), \tag{3.18}$$

donde ϕ hace referencia a la dispersión de la distribución de la variable, ρ es la correlación temporal del efecto espacial, $\tau_{\text{Batimetría}}$ es el hiperparámetro de un *random walk* de orden 2 ($rw2$) para la batimetría, τ_{tiempo} es el hiperparámetro de un *random walk* de orden 2 ($rw2$) para capturar la tendencia temporal y σ_U y κ los hiperparámetros a estimar de la matriz de covarianza Σ.

Modelo geoestadístico-temporal preferencial

Pennino *et al.* (2019) demostraban las consecuencias de no modelizar el patrón puntual asociado a la variable respuesta índice de CPUE. Recordamos, que los índices de CPUE no se consiguen mediante un muestreo aleatorio de la región de estudio, sino que, son derivados de las pesquerías y estás suelen generar un patrón de puntos agrupado, ya que el pescador conoce los caladeros en los que la biomasa de la especie es alta. Por lo tanto, en este trabajo se plantea la modelización del patrón puntual junto a la variable continua CPUE para intentar paliar los efectos de la dependencia en el muestreo.

Función de verosimilitud:

$$\begin{aligned} \text{CPUE}(s,t) &\sim \text{Gamma}(\mu(s,t),\phi)\,, \\ log(\mu(s,t)) &= \beta_{0\text{CPUE}} + f(\text{tiempo})(s,t) + v(s,t)\,, \end{aligned} \tag{3.19}$$

$$\begin{aligned} \text{PP}(s,t) &\sim \text{LGCP}(\lambda(s,t)), \\ log(\lambda(s,t)) &= \beta_{0\text{PP}} + \alpha v(s,t) \end{aligned}$$

$$\begin{aligned} v(s,t) &= \rho \times v(s,t-1) + \text{U}(s,t)\,, \\ \text{U}(s,t) &\sim \text{GMRF}(0,\Sigma)\,, \end{aligned}$$

donde $\text{CPUE}(s,t)$ es la variable respuesta definida por una distribución Gamma donde ϕ es la dispersión y $\mu(s,t)$ la media de la distribución. $\mu(s,t)$ está enlazada al predictor lineal por la función de enlace logaritmo, β_0 hace referencia al intercepto, $f(tiempo)(s,t)$ es una función suave para la tendencia temporal y $v(s,t)$ es el término espacial correlado en el tiempo. Además, la variable respuesta del patrón puntual $\text{PP}(s,t)$ presenta las mismas componentes en el predictor que la variable índice de $\text{CPUE}(s,t)$, excepto la tendencia temporal, ya que esta es únicamente sobre la media de la variable respuesta. Así mismo, el efecto espacial correlado se relaciona a través del parámetro de escalado α.

Distribuciones a priori para el *Latent Gaussian Field* LGF:

$$\begin{aligned} \{\beta_{0\text{CPUE}}, \beta_{0\text{PP}}, \alpha\} &\sim \text{N}(0, \tau = 0{,}001), \\ U(s,t) &\sim \text{N}(0, \Sigma(\sigma_U, \kappa)), \end{aligned} \tag{3.20}$$

donde $\beta_{0\text{CPUE}}$ y $\beta_{0\text{PP}}$ hacen referencia a los efectos fijos (interceptos) de la verosimilitud, incluyendo α. $U(s,t)$ se corresponde con el efecto espacial distribuido como una normal de media 0 y matriz de covarianza Σ, siendo σ_U y κ los hiperparámetros a estimar de la matriz.

Hiperparámetros:

$$\{\phi, \tau_{\text{tiempo}}, \rho, \sigma_\text{U}, \kappa\} \sim \text{F}(\{\phi, \tau_{\text{tiempo}}, \rho, \sigma_\text{U}, \kappa\}), \tag{3.21}$$

donde ϕ hace referencia a la dispersión de la distribución de la variable, ρ es la correlación temporal, τ_{tiempo} es el hiperparámetro de un *random walk* de orden 2 ($rw2$), y σ_U y κ los hiperparámetros a estimar de la matriz de covarianza.

Después de proponer y detallar los distintos modelos, hemos de conseguir, con la ayuda de herramientas estadísticas, estimar y predecir sus parámetros, es decir, vamos a inferir y predecir sobre los modelos. En este trabajo algunos de los modelos se resolverán desde la perspectiva frecuentista, intentado reproducir como suelen resolverse estos modelos por la comunidad

científico-pesquera. Por el contrario, todos los modelos se resolverán en el contexto bayesiano utilizando como herramienta `R2BayesX`, `R-INLA` e `inlabru`.

Por último, al inferir y predecir sobre los parámetros de los distintos modelos, podremos obtener las series de índices predichas en el tiempo, que funcionan como *inputs* en modelos de evaluación del *stock*, p.ej. SPiCT. En consecuencia, para determinar que serie ha conseguido capturar mejor el comportamiento de la biomasa simulada, se propone comparar entre la serie predicha de índices y la biomasa simulada, calculando dos medidas de error RMSE y MAPE con las librerías de `R` (Team, 2013) `Metrics` y `LMetrics`, respectivamente:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\text{Índice}_i – \text{Biomasa}_i)^2}{n}}, \tag{3.22}$$

$$MAPE = \frac{100}{n} \times \sum_{i=1}^{N} \left| \frac{\text{Biomasa}_i - \text{Índice}_i}{\text{Biomasa}_i} \right|,$$

donde i es el número de observaciones y n coincide con el número de años.

Capítulo 4

Resultados

En el siguiente capítulo se desarrolla un ejemplo de cómo podemos aplicar el protocolo elaborado en este trabajo. En primer lugar, escogemos unos parámetros para la simulación de un escenario de biomasa y, posteriormente, reproducimos un muestreo aleatorio emulando una campaña oceanográfica y un muestreo preferencial imitando la actividad pesquera. A partir de estos dos muestreos obtendremos dos de las principales fuentes de información en la gestión de *stocks*, los índices de biomasa relativa, derivados de campañas oceanográficas y los índices de CPUE derivados de la actividad pesquera.

Una vez disponemos de los distintos índices de biomasa relativa y de CPUE, ya podemos inferir y predecir sobre los modelos propuestos en el capítulo 3 con las herramientas descritas en el capítulo 2. De estos modelos podemos obtener distintos resultados, por ejemplo, en el caso de los modelos geoestadísticos, es posible, presentar mapas de distribución de la biomasa. En todo caso, el *output* en el que nos centraremos es la serie temporal predicha de índices de biomasa relativa o de CPUE, para así, poder compararla con la serie de tiempo de la biomasa real que hemos simulado.

Para finalizar, las series predichas de índices de biomasa relativa o de CPUE se compararán con la biomasa simulada, calculando una serie de medidas de error (RMSE y MAPE), que nos permitirán concluir qué modelización ha conseguido capturar mejor el comportamiento de la biomasa simulada. El modelo que menor RMSE y MAPE obtengan en la serie para el índice de biomasa relativa y para el índice de CPUE, se utilizará como *input* en un modelo de evaluación del *stock* (SPiCT), del que mostraremos los *outputs* más relevantes en el campo de la gestión pesquera.

En el caso de que se quiera profundizar en la confección de la simulación y el ajuste de los modelos, es posible acceder a los *scripts* de este trabajo desde los materiales complementarios.

4.1. Simulación

Para comenzar, hemos simulado un escenario de biomasa con sus respectivos muestreos. La biomasa se simula a partir de un modelo (3.1) compuesto por un intercepto, la batimetría, una tendencia temporal y un efecto espacial autoregresivo. En consecuencia, se ha de dar valor a cada uno de los parámetros del modelo, y así, conseguir simular la biomasa.

En el *script* elaborado para la simulación de la biomasa y la reproducción de fuentes de datos pesqueros, disponible en el material complementario, se han de fijar los siguientes parámetros para su correcto funcionamiento:

- **Coordenadas**: en este caso la biomasa la hemos simulado en un cuadrado, por lo tanto, debemos darle las coordenadas de sus vértices.
- **Parámetros del efecto espacial autoregresivo**: para el efecto espacial damos un valor a la varianza $\sigma^2_{U(s,t)}$ y a κ que se relaciona con el rango tal que $r = \frac{\sqrt{8}}{\kappa}$. Además, para que el efecto espacial esté correlado en el tiempo damos un valor a un parámetro ρ.
- **Número de años**: fijamos una k que será el número de años que queramos estudiar la biomasa.
- **Coeficientes del predictor**: una vez simulamos el efecto espacial y la batimetría, debemos darle valores a los coeficientes del predictor β_0, β_1, β_2 y al vector con el coeficiente de la tendencia temporal para cada año.
- **Parámetro β de la distribución Gamma**: la distribución Gamma en `R` viene dada en términos de α y β, por lo que, reparametrizamos α relacionándolo con la media de la distribución μ y le damos un valor fijo a β.
- **Parámetros de los muestreos**: tenemos un parámetros m que será el número de puntos que muestreamos de la simulación de biomasa y dos parámetros q_{random} y q_{pref} que serán los coeficientes de capturabilidad para relacionar los índices relativos con la biomasa.

Con el valor de los parámetros mencionados ya escogidos, gracias a las distintas funciones que componen el *script* de la simulación, estaríamos en disposición de obtener una simulación de la biomasa con sus respectivos bancos de datos con los índices de biomasa relativa y de CPUE. Por consiguiente, vamos a ilustrar los distintos resultados que se pueden extraer de una simulación.

En primer lugar, se muestra la simulación del efecto espacial. En la Figura 4.1 podemos observar un efecto en el espacio con un comportamiento autoregresivo en un periodo de 10

años. Si nos fijamos de nuevo en la Figura 2.4, la simulación de nuestro efecto espacial tiene un comportamiento similar, de manera que las zonas con valores más altos presentan una fuerte correlación entre los puntos, mientras que, aquellas zonas donde el efecto espacial es bajo están débilmente relacionadas.

Por otro lado, además del efecto espacial autoregresivo, se ha simulado la batimetría, esta se trata de una variable cuantitativa, continua y positiva. Igualmente, hemos supuesto que la batimetría no cambia a lo largo del periodo de estudio y que está relacionada con los ejes cartesianos (Figura 4.2). Además, hemos relacionado la biomasa con la batimetría, de manera que la relación no es lineal (se simula mediante un polinomio de grado dos).

Después de fijar los parámetros y tener simuladas las componentes del predictor, podemos construir la variable respuesta biomasa. Para ello, vamos a utilizar la función *rgamma*, dicha función viene dada por los parámetros α y β de una distribución Gamma, por lo tanto, reparametrizamos α tal que $\alpha = \frac{\mu}{\beta}$, donde μ es la media que hemos simulado a través del predictor lineal, y a β le hemos dado un valor fijo.

En la Figura 4.3, podemos apreciar el escenario de biomasa simulado a lo largo del espacio y el tiempo, presentando un comportamiento autorregresivo. Si nos fijamos en las Figuras correspondientes al efecto espacial y la batimetría, vemos que la biomasa se está distribuyendo acorde al efecto espacial autoregresivo y su relación (no lineal) con la batimetría (Figuras 4.1 y 4.2).

Por otra parte, como el objetivo final de nuestro trabajo reside en recuperar la tendencia temporal de la biomasa a lo largo del periodo de estudio comparando entre los resultados de distintas modelizaciones (GLM, GAM, modelos geoestadísticos, etc.), mostramos en la Figura 4.4 la mediana de la serie temporal de la biomasa simulada. Como bien se observa en la Figura 4.4, se ha simulado una tendencia en la media de la biomasa a lo largo de los años, y es esta la tendencia que queremos recuperar.

Hasta aquí, ya hemos conseguido nuestro escenario de biomasa simulado, por lo tanto, podríamos proceder a reproducir las fuentes principales de datos en la gestión pesquera: los índices de biomasa relativa y los índices de CPUE. Para ello, lo primero que debemos hacer es muestrear del escenario de biomasa, primeramente, muestrearemos de forma aleatoria, imitando una campaña oceanográfica y, a continuación, muestrearemos de forma preferencial, imitando la actividad de las pesquerías. Los resultados del muestreo aleatorio y preferencial se aprecian en las Figuras 4.5 y 4.6, respectivamente.

Acto seguido, con los puntos de biomasa seleccionados en ambos muestreos aplicamos el coeficiente de capturabilidad q a cada valor de biomasa y ya tenemos los índices de biomasa relativa. Para los índices de CPUE además de multiplicar q, simulamos un efecto lineal al que llamamos esfuerzo y que irá dividiendo a la biomasa, y ya tenemos los índices de CPUE (3.3 y

3.4). En la Tabla 4.1, se recogen los estadísticos resumen de la mediana, el cuantil 0.025 y el cuantil 0.975, para estos índices.

Índices	***Mediana***	***Cuantil 2.5 %***	***Cuantil 97.5 %***
Biomasa relativa			
Año 1	6.53	$4.06 \cdot 10^{-20}$	64.71
Año 2	6.3	$6.46 \cdot 10^{-14}$	114.03
Año 3	19.57	$2.24 \cdot 10^{-5}$	104.37
Año 4	21.712	$4.54 \cdot 10^{-3}$	134.77
Año 5	39.13	$1.50 \cdot 10^{-6}$	378.53
Año 6	12.12	$3.23 \cdot 10^{-10}$	318.02
Año 7	26.28	$2.54 \cdot 10^{-4}$	461.90
Año 8	14.58	$3.83 \cdot 10^{-8}$	581.36
Año 9	16.07	$8.85 \cdot 10^{-9}$	334.38
Año 10	10.83	$1.25 \cdot 10^{-4}$	256.81
CPUE			
Año 1	2.33	0.33	4.55
Año 2	4.24	1.02	7.62
Año 3	3.55	1.15	7.64
Año 4	4.70	0.77	17.92
Año 5	11.13	1.52	23.62
Año 6	6.87	1.36	21.57
Año 7	10.12	1.76	32.76
Año 8	7.84	0.81	27.81
Año 9	10.29	0.69	34.25
Año 10	4.02	0.64	11.35

Tabla 4.1: Estadísticos resumen para los índices de biomasa relativa y de CPUE.

Por último, se ha elaborado un Anexo, Anexo I, donde es posible acceder a otros resultados que podrían ser de interés, como es la relación entre la batimetría y los índices, histogramas con las variables, etc.

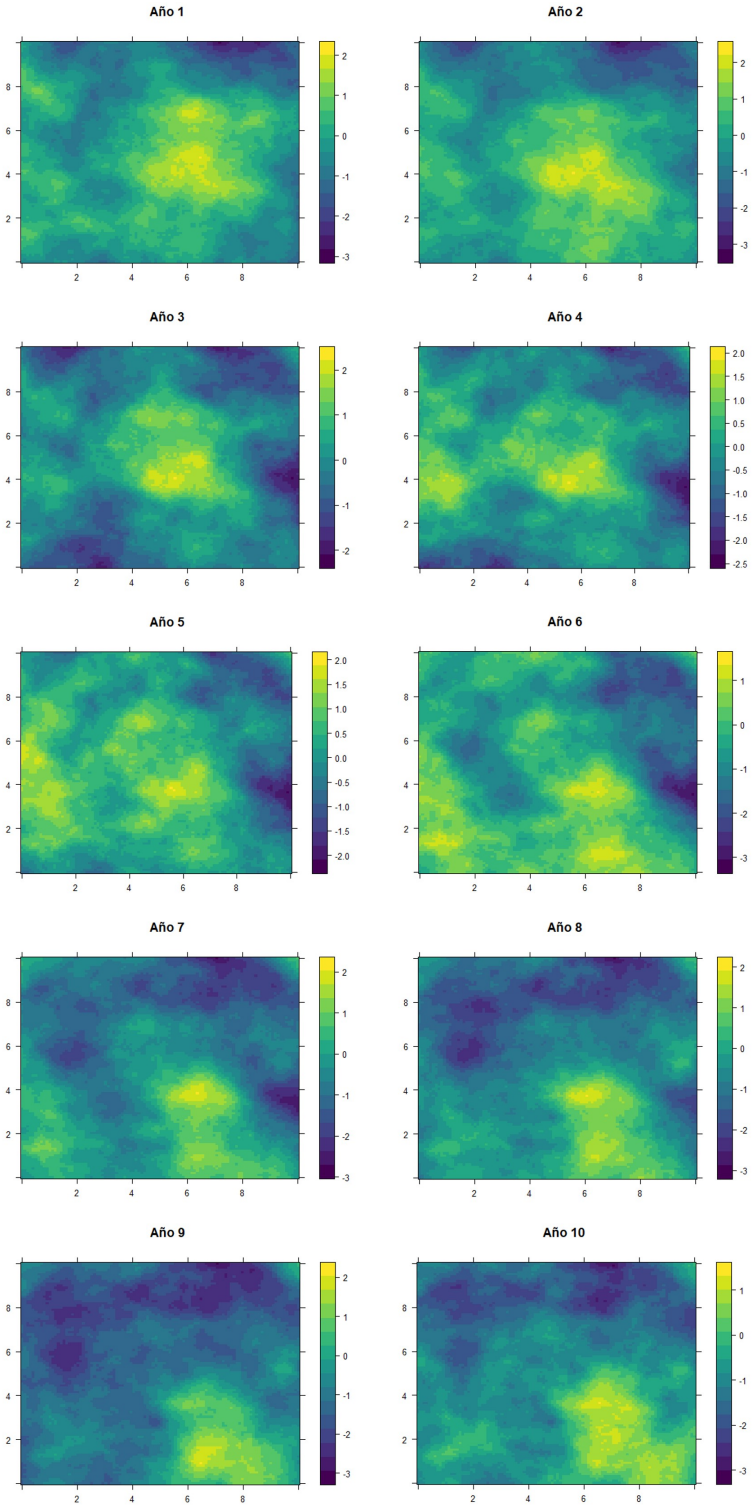

Figura 4.1: Simulación del efecto espacial autoregresivo.

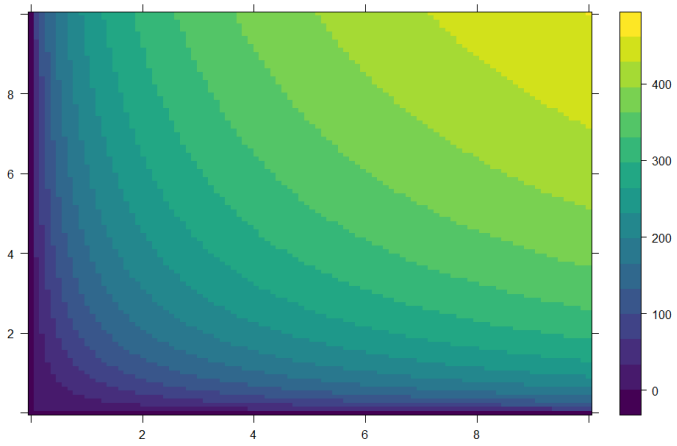

Figura 4.2: Simulación de la batimetría.

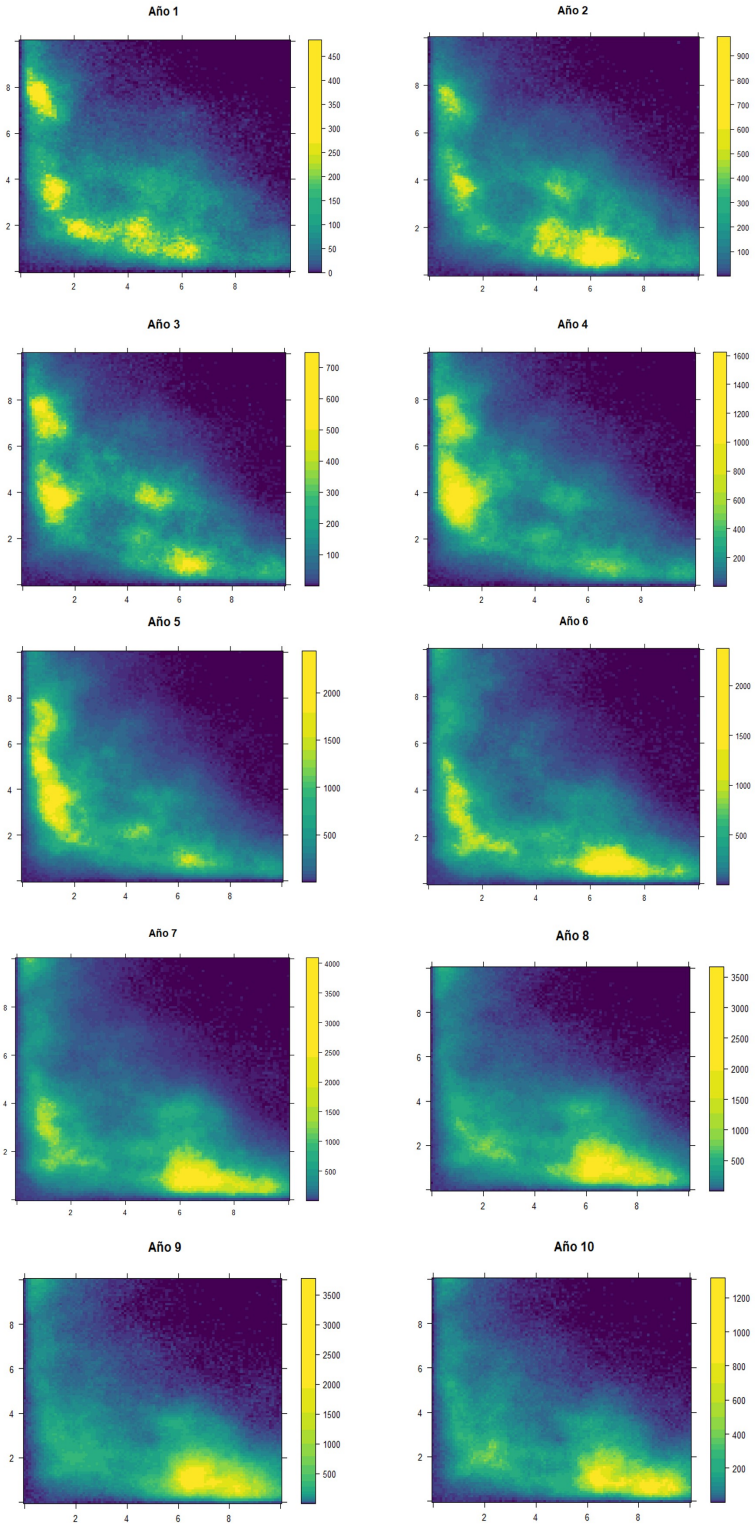

Figura 4.3: Simulación de la biomasa.

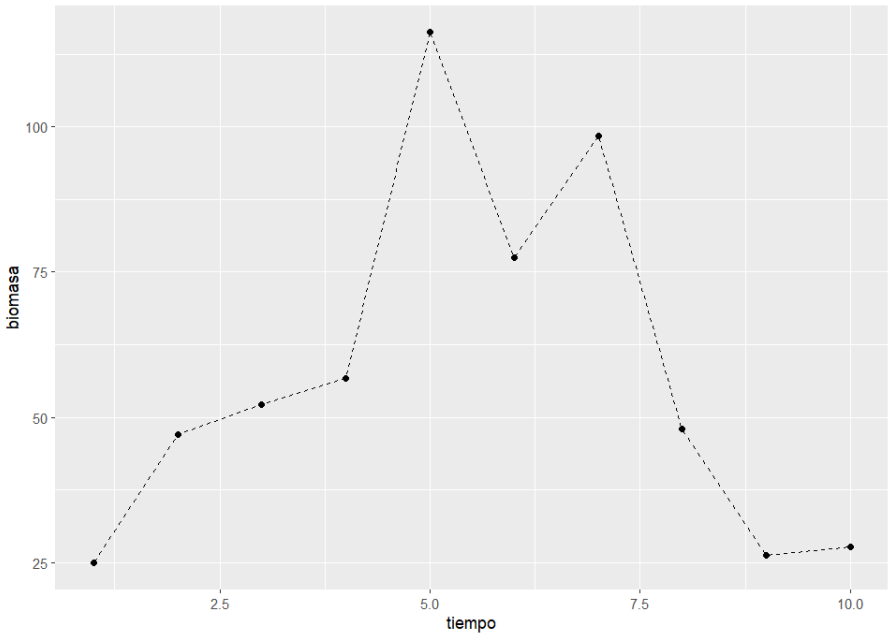

Figura 4.4: Mediana de la biomasa a lo largo del periodo de estudio.

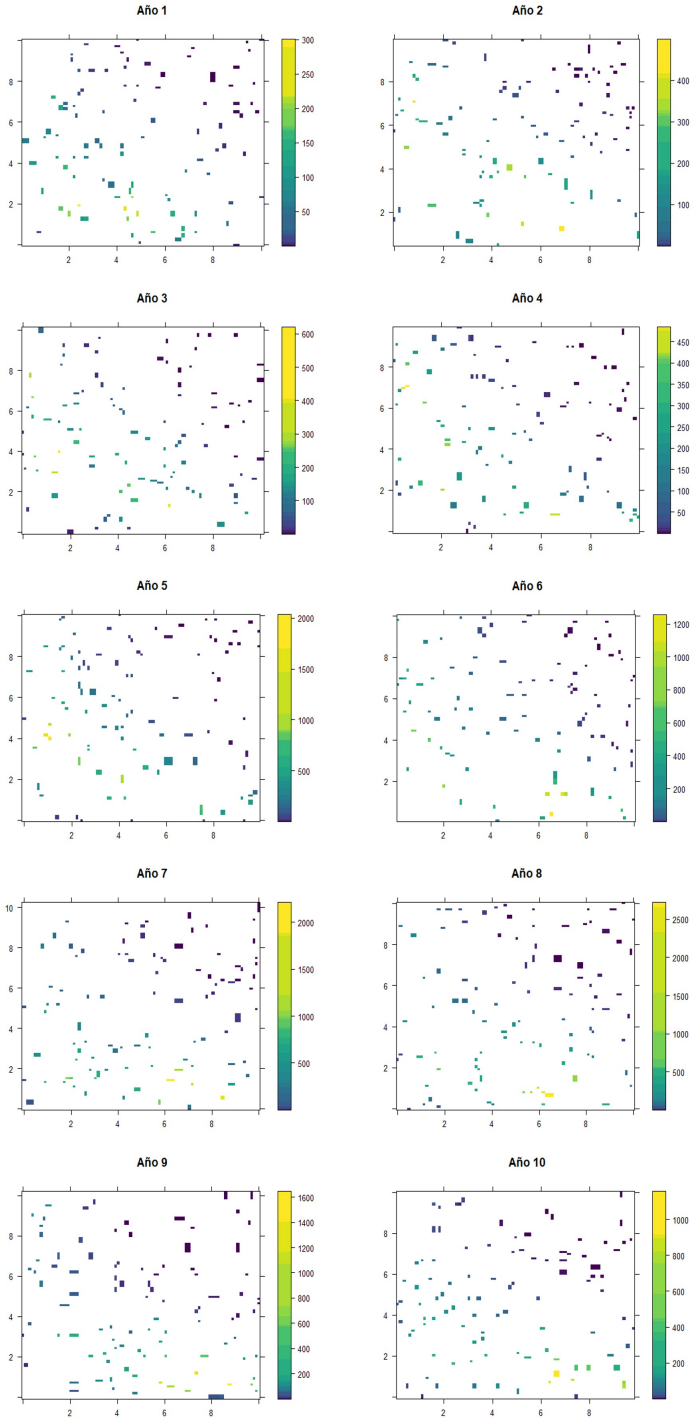

Figura 4.5: Muestreo aleatorio de la biomasa (independiente de la pesca). La escala de color se corresponde con el valor de biomasa simulado.

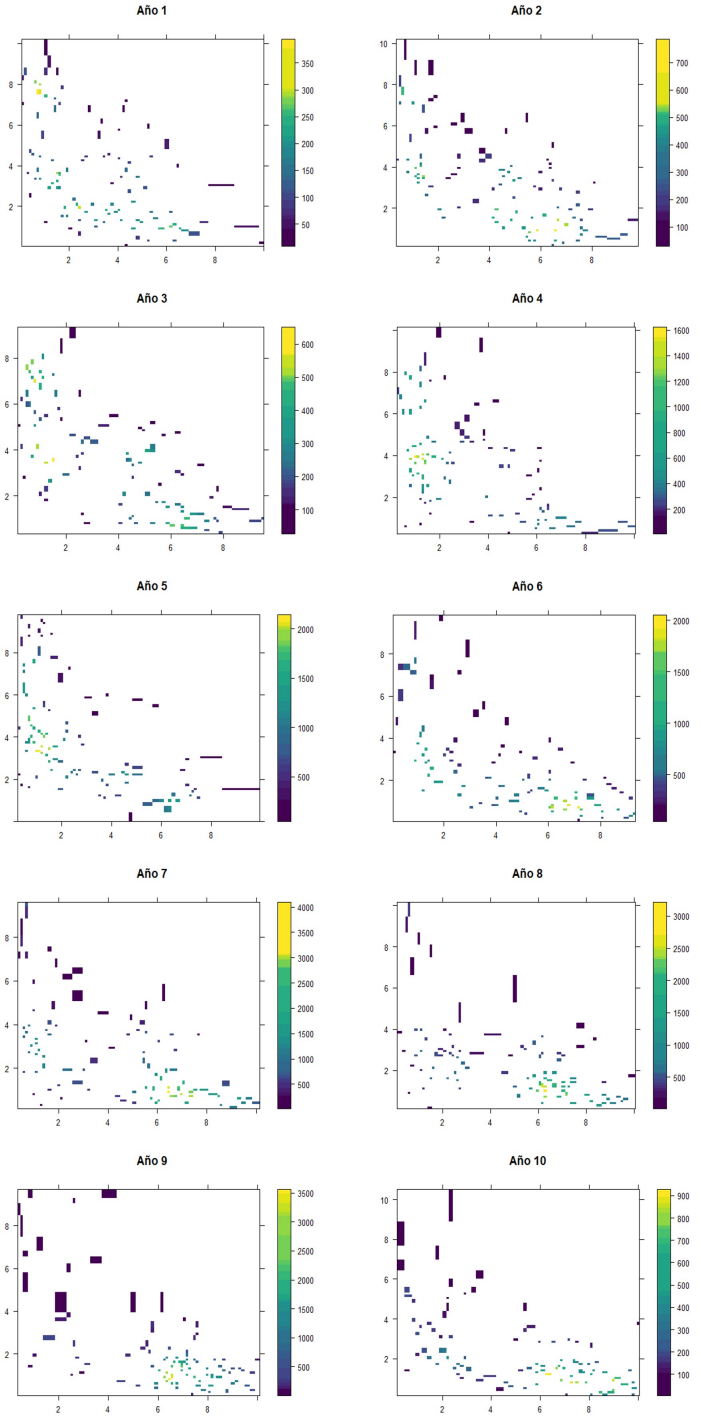

Figura 4.6: Muestreo preferencial de la biomasa (dependiente de la pesca). La escala de color se corresponde con el valor de biomasa simulado.

4.2. Resultados: Inferencia y predicción

En el siguiente apartado, a partir de la simulación escogida se van a ajustar los modelos propuestos en el capítulo anterior, diferenciando entre los modelos para un muestreo aleatorio (índices de biomasa relativa derivados de campañas oceanográficas) y para un muestreo preferencial (índices de CPUE derivados de las pesquerías). De cada uno de los modelos expuestos, se predice un valor de biomasa por año, con el fin de obtener una serie de biomasa relativa o de CPUE a comparar con la biomasa real. Por último, aquel modelo que de mejores resultados, es decir, consiga la serie de índices más representativa de la biomasa real (menor RMSE y MAPE), se utilizará como *input* en un modelo de evaluación del *stock* (SPiCT).

4.2.1. Índices de biomasa relativa: muestreo aleatorio

Para los índices de biomasa relativa derivados de un muestreo aleatorio hemos ajustado un total de cinco modelos, recordamos que cuatro de los modelos son un GLM y un GAM desde la perspectiva frecuentista y otro GLM y otro GAM desde la bayesiana y, el quinto es un modelo geoestadístico bayesiano resuelto mediante la aproximación INLA. Todos los modelos presentan como variable respuesta el índice de biomasa relativa, de manera que, una vez hemos predicho la serie temporal de biomasa relativa, podríamos recuperar la biomasa simulada a través del coeficiente de capturabilidad q asignado en la simulación (3.3).

Así pues, en la Figura 4.7 se muestran las series de tiempo predichas y divididas por q para cada uno de los modelos. Dichas series, representan la mediana del valor de biomasa predicho para cada año. En general, parece que todos los modelos han conseguido capturar mejor o peor la tendencia en el tiempo de la biomasa. Sin embargo, puede observarse como el modelo geoestadístico parece ajustar mejor la tendencia temporal real, aunque tiende a sobreestimar el valor de biomasa si se compara con la biomasa simulada. En cuanto a los GLMs y los GAMs destaca que, en los años con valores más altos de biomasa el modelo sobreestima y en los años con valores más bajos de biomasa el modelo infraestima.

Por otro lado, en la Tabla 4.2 se recogen los resultados de las medidas de error MAPE y RMSE para cada una de las series predichas. De entre todas las series de biomasa relativa predichas con los distintos modelos, aquella que ha obtenido un menor RMSE y MAPE ha sido la serie resultante del modelo geoestadístico, con un RMSE de 12.45 y un MAPE de 0.19. Por lo tanto, puede considerarse el índice que mejor representa la biomasa simulada. Después del modelo geoestadístico, la modelización del índice con menor RMSE y MAPE, 21.94 y 0.45 respectivamente, ha sido el GAM frecuentista, seguido del GAM bayesiano y, por último, los GLMs frecuentista y bayesiano, cuyos resultados han sido muy similares.

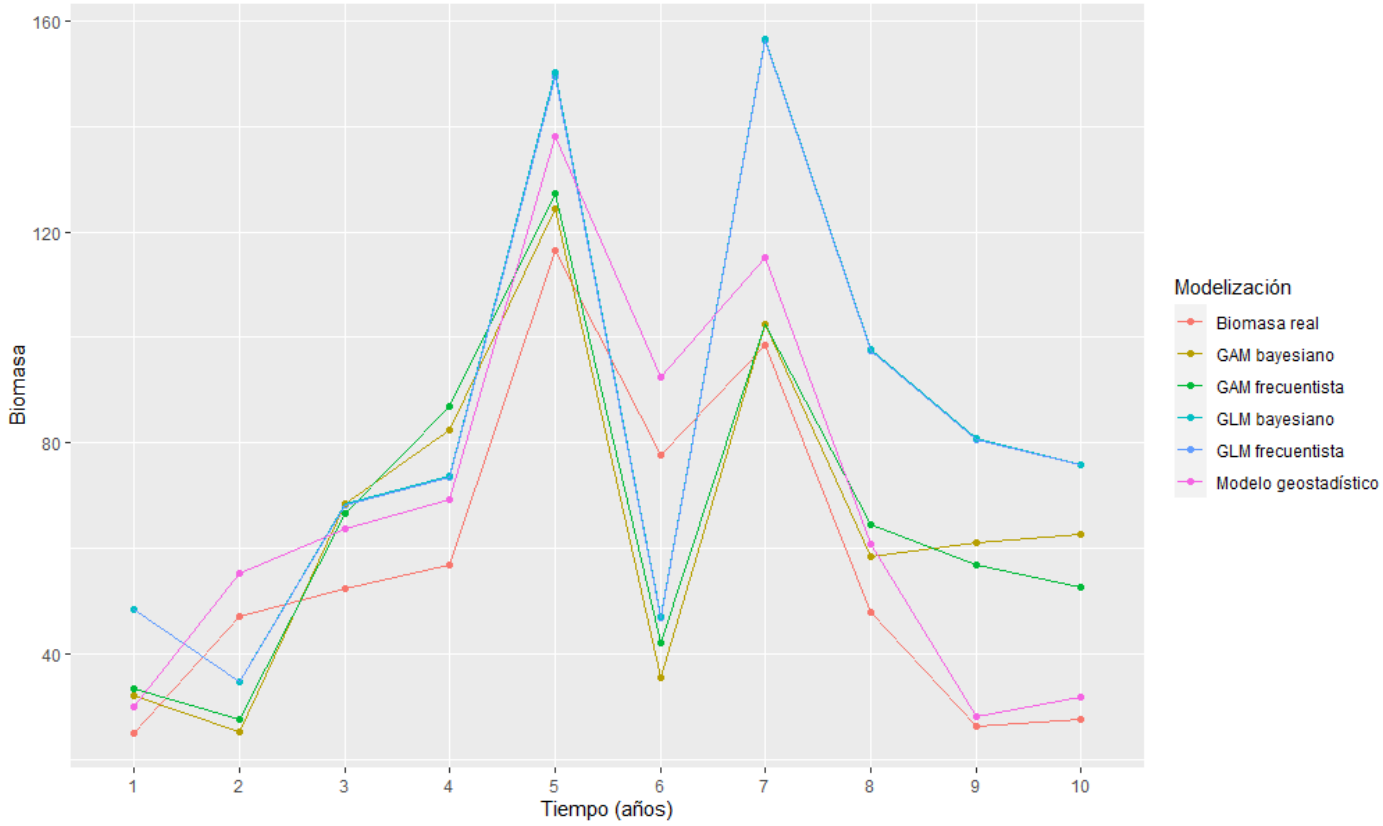

Figura 4.7: Series de biomasa simulada frente a series predichas con los distintos modelos para el muestreo aleatorio.

Modelo	***RMSE***	***MAPE***
Bayesiano		
GLM	35.52	0.74
GAM	24.20	0.49
Geoestadístico	**12.45**	**0.19**
Frecuentista		
GLM	35.40	0.74
GAM	21.94	0.45

Tabla 4.2: Medidas de error para comparar la biomasa simulada y las series de tiempo predichas en un muestreo aleatorio.

En vista a que el modelo geoestadístico ha presentado los mejores resultados, procedemos a ilustrar algunos de los mapas de distribución de la biomasa que derivan de este tipo de modelos (Figura 4.8). Además, en el Anexo II, se han recopilado otros *outputs* del modelo geoestadístico, como pueden ser la incertidumbre asociada a cada punto, los cuantiles, etc.

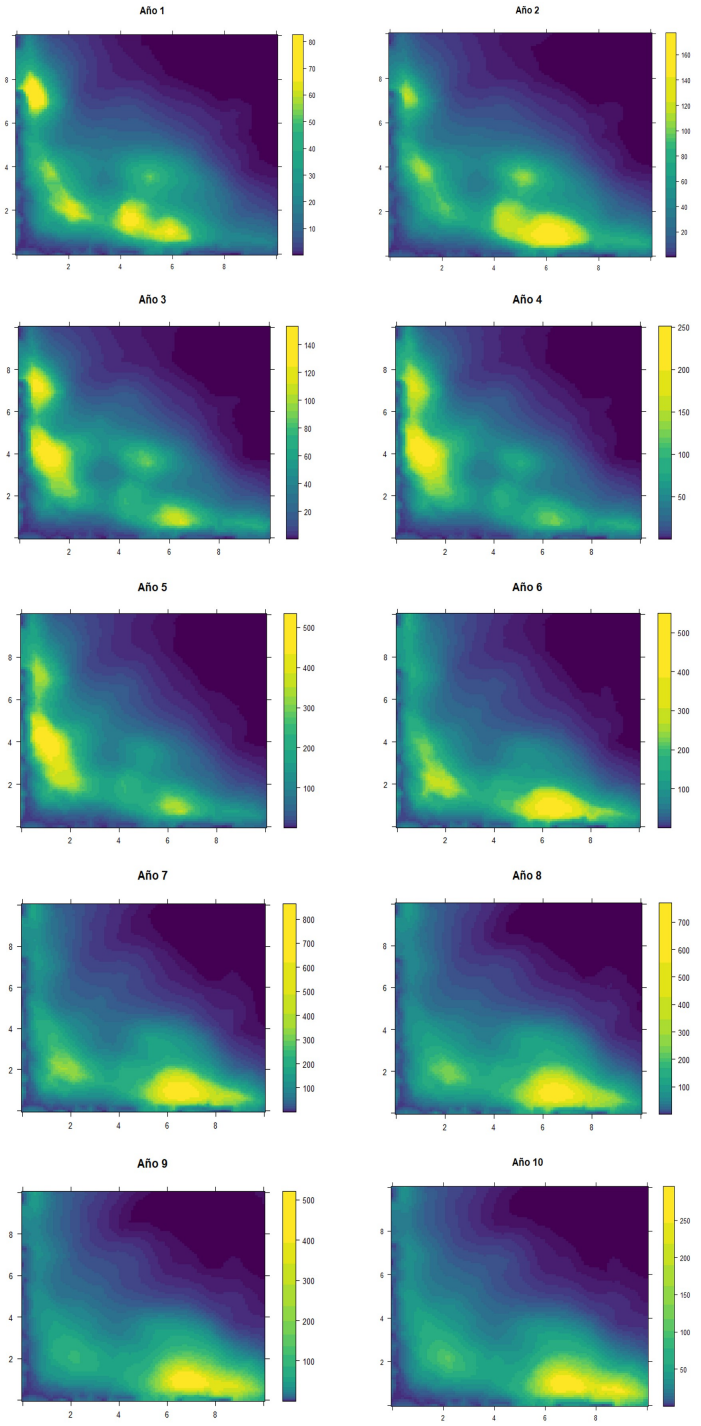

Figura 4.8: Mediana de la distribución predictiva a posteriori para el índice de biomasa relativa (muestreo aleatorio).

4.2.2. Índices de CPUE: muestreo preferencial

En pesquerías los datos pueden provenir de campañas oceanográficas (muestreo aleatorio) o de la actividad pesquera (muestreo preferencial). En este apartado, nos centraremos en los resultados de los modelos que se han ajustados para los índices de CPUE derivados de un muestreo preferencial (dependiente de la pesca). En total, hemos propuesto y resuelto seis modelos, dos de ellos GLMs (frecuentista y bayesiano), otros dos GAMs (frecuentista y bayesiano), un modelo geoestadístico y un patrón puntual marcado (combinación de un modelo geoestadístico para la variable CPUE y un modelo LGCP para el patrón de puntos asociado). Como la variable respuesta es el índice de CPUE, para recuperar la biomasa, una vez ajustado el modelo, dividimos por el coeficiente de capturabilidad q asignado en la simulación y multiplicamos el esfuerzo (3.4).

En primer lugar, en la Figura 4.9 se observan las serie temporales predichas de biomasa para cada uno de los modelos propuestos. En sí, la Figura 4.9 representa la mediana del valor de biomasa para cada año en un total de 10 años. Destaca que todos los modelos están sobrestimando si comparamos con la biomasa simulada. No obstante, queda reflejado como utilizar técnicas geoestadísticas ha ido disminuyendo la sobreestimación, siendo el modelo preferencial (patrón puntual marcado) el modelo que está obteniendo mejores resultados.

Por otra parta, además de la Figura 4.9, en la Tabla 4.3 se recogen las medidas de error RMSE y MAPE para cada una de las series predichas. De entre todas las series de índices de CPUE predichas con distintas modelizaciones, el modelo que menor RMSE y MAPE ha conseguido, 16.84 y 0.31 respectivamente, ha sido el modelo preferencial (patrón puntual marcado). En consecuencia, podemos considerarlo como el índice que mejor representa la biomasa simulada para el muestreo preferencial. El siguiente modelo con menor RMSE y MAPE ha sido el modelo geostadísitco, 67.96 y 1.18 respectivamente, seguido del GAM bayesiano y frecuentista y, por último, los GLMs los cuales obtuvieron resultados muy similares.

Modelo	***RMSE***	***MAPE***
Bayesiano		
GLM	613.89	10.90
GAM	498.37	8.91
Geoestadístico	67.96	1.18
Preferencial	**16.84**	**0.31**
Frecuentista		
GLM	612.67	10.88
GAM	511.14	9.09

Tabla 4.3: Medidas de error para comparar la biomasa simulada y las series de tiempo predichas en un muestreo preferencial.

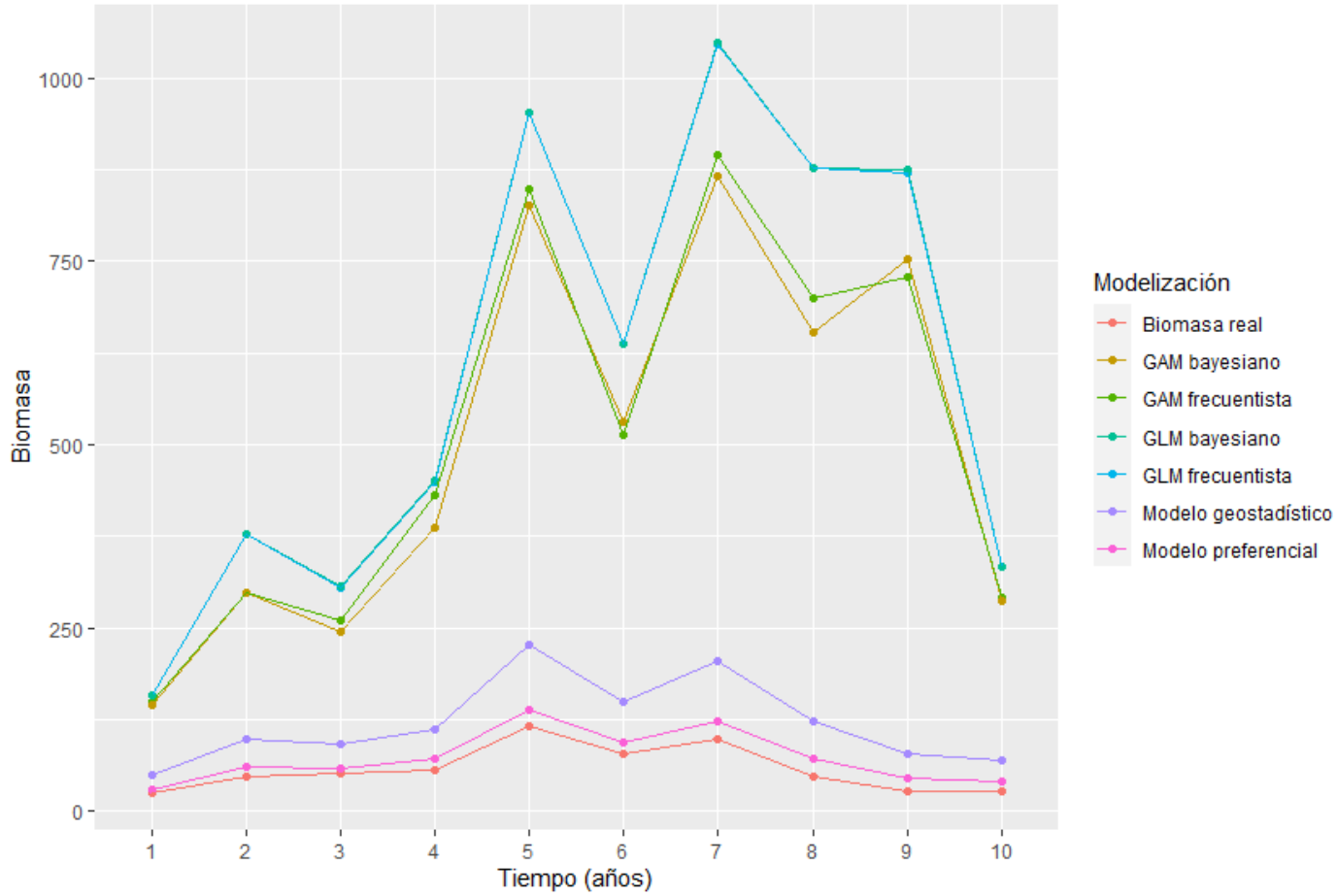

Figura 4.9: Series de biomasa simulada frente a series predichas con los distintos modelos para el muestreo preferencial.

Al igual que con los resultados del muestreo aleatorio, vamos a proceder a mostrar el mapa de predicción para los índices de CPUE, que hemos conseguido con el modelo que mejor (menor RMSE y MAPE), es decir, el patrón puntual marcado o modelo geoestadístico preferencial. El modelo preferencial, ha sido ajustado con el paquete `inlabru`, el cual por defecto tiene implementadas una serie de funciones que nos permiten extraer directamente el mapa con la distribución de los índices.

En la Figura 4.10, es posible observar el mapa con la mediana de la distribución predictiva a posteriori de los índices de CPUE para cada año. Se aprecia que los mapas han conseguido captar la tendencia en el espacio y en el tiempo si los comparamos con la simulación de la biomasa (Figura 4.3). Así mismo, en el Anexo II, se han recopilado otros *outputs* del patrón puntual marcado que podrían ser de interés.

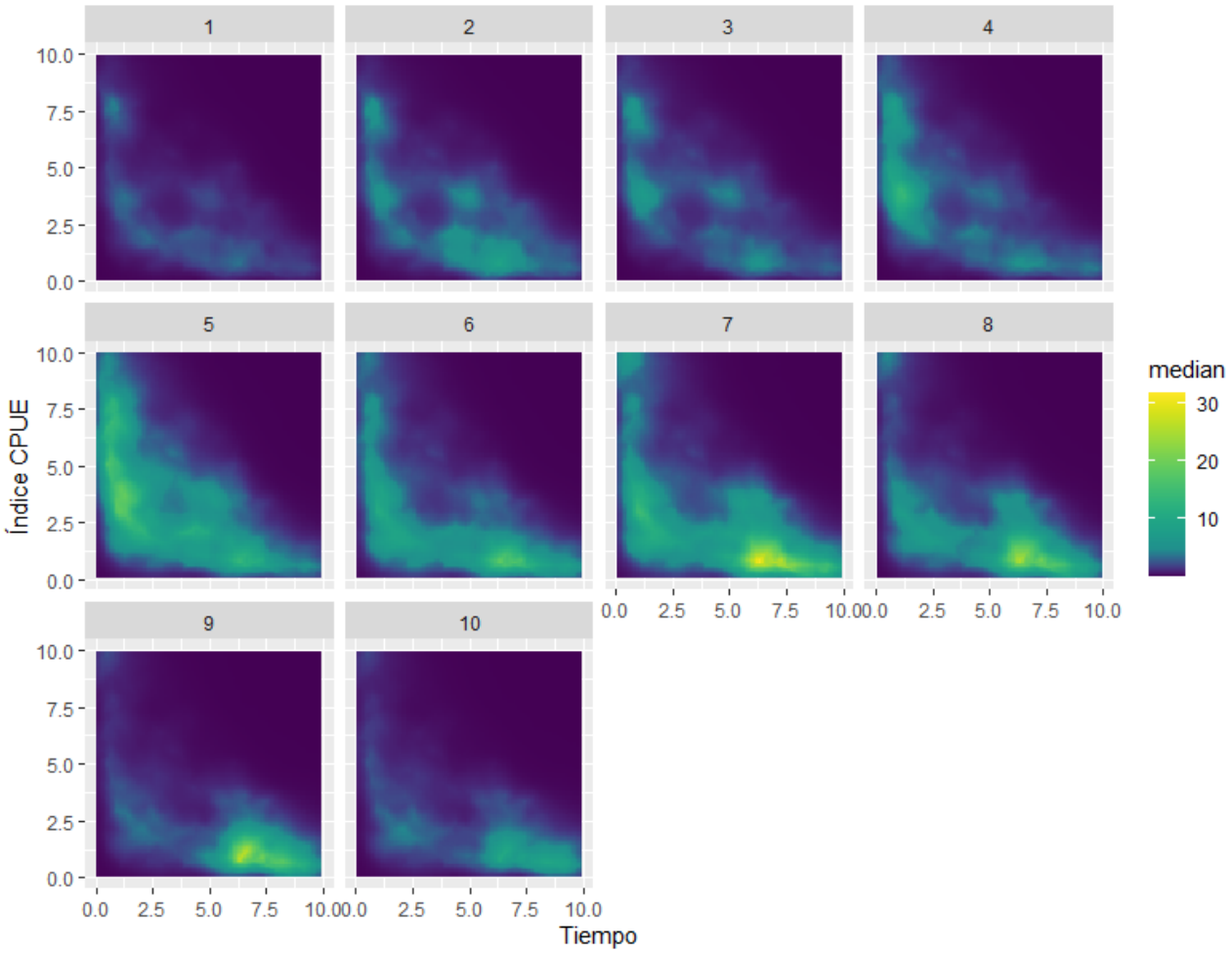

Figura 4.10: Mediana de la distribución predictiva a posteriori para el índice de biomasa relativa en un muestreo aleatorio.

4.3. Evaluación del *stock*: SPiCT

Los índices de biomasa relativa o de CPUE, se utilizan para alimentar Modelos de producción excedentaria, como por ejemplo, SPiCT. En consecuencia, un índice de biomasa relativa que sea representativo de la biomasa conseguirá buenos resultados en la evaluación del *stock*. Por consiguiente, vamos a utilizar las series predichas del índice de biomasa relativa y de CPUE, que menor RMSE y MAPE presenten, para ajustar un model SPiCT, y así, apreciar cuáles son los *outputs* típicos de un modelo de evaluación del *stock* y su utilidad en la gestión de pesquerías.

4.3.1. Índices de biomasa relativa: muestreo aleatorio

En primer lugar, recordamos que un modelo de producción excedentaria (SPiCT), necesita de dos *inputs*: (1) la serie de capturas que hemos aproximado (3.2) y una serie de índices de biomasa relativa o de CPUE con una valor para cada año. En la Figura 4.11, se observan los dos *inputs* que hemos utilizado para ajustar un modelo SPiCT, en el primer gráfico se observan las capturas aproximadas y en el segundo gráfico la serie del índice de biomasa relativa ajustado con un modelo geoestadístico para el muestreo aleatorio.

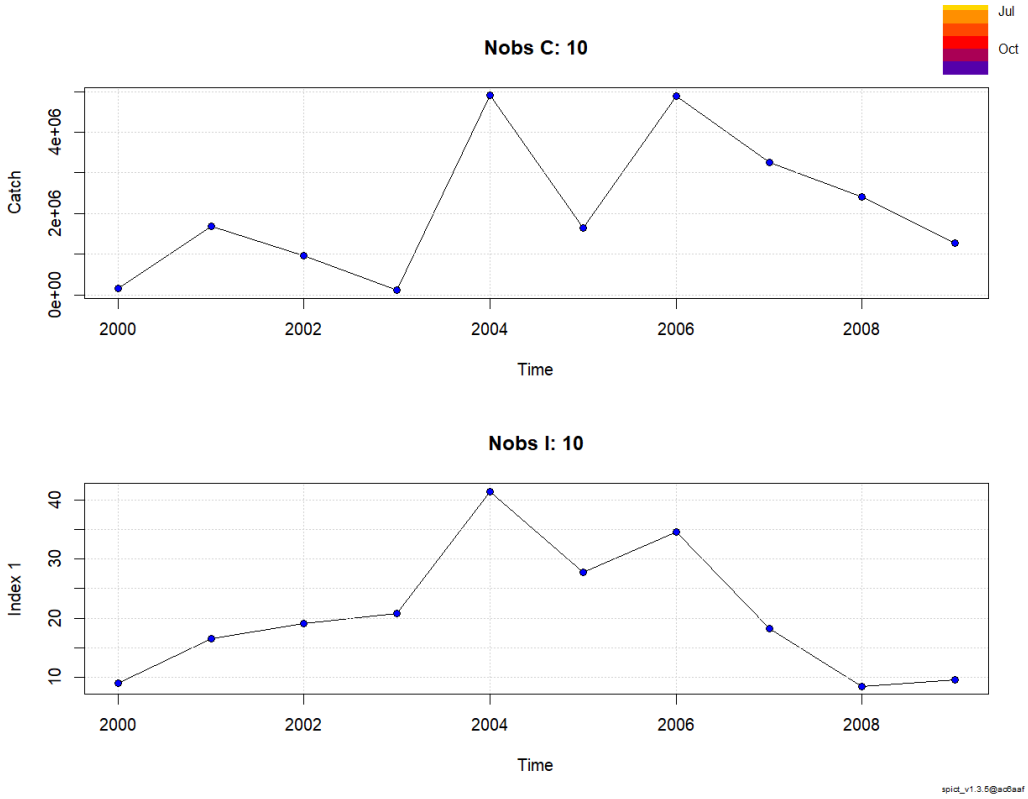

Figura 4.11: *Inputs* SPiCT: serie de capturas simulada (Nobs C:10) y serie de índices de biomasa relativa predichos (Nobs I:10).

A continuación, la Figura 4.12 recoge una selección de algunos de los resultados más relevantes que ha obtenido el modelo de SPiCT con los *inputs* mostrados en la Figura 4.11. En primer lugar, la Figura 4.12 muestra la estimación de la biomasa absoluta del *stock* (*Absolute biomass*). En segundo lugar, se observa la estimación de la mortalidad por pesca (*Absolute fishing mortality*) y las capturas (*Catch*). Por último, podemos observar como en los primeros años de la serie el *stock* ha tenido una producción de biomasa positiva, estando el mismo dentro de los límites sostenibles. Por el contrario, en los últimos años la población está teniendo unos niveles de producción de biomasa muy bajos que ponen en peligro la conservación del *stock*.

En resumen, un modelo de producción excedentaria como SPiCT, es capaz de devolvernos una evaluación completa sobre el estado del *stock*, de forma que estima la biomasa absoluta, la mortalidad por pesca, los límites de explotación sostenibles o puntos de referencia, etc. Para la gestión del *stock* uno de los *outputs* más relevantes son los puntos de referencia o límites de explotación sostenibles, ya que en base a ellos se marcan los límites de pesca necesarios para evitar el colapso del sistema.

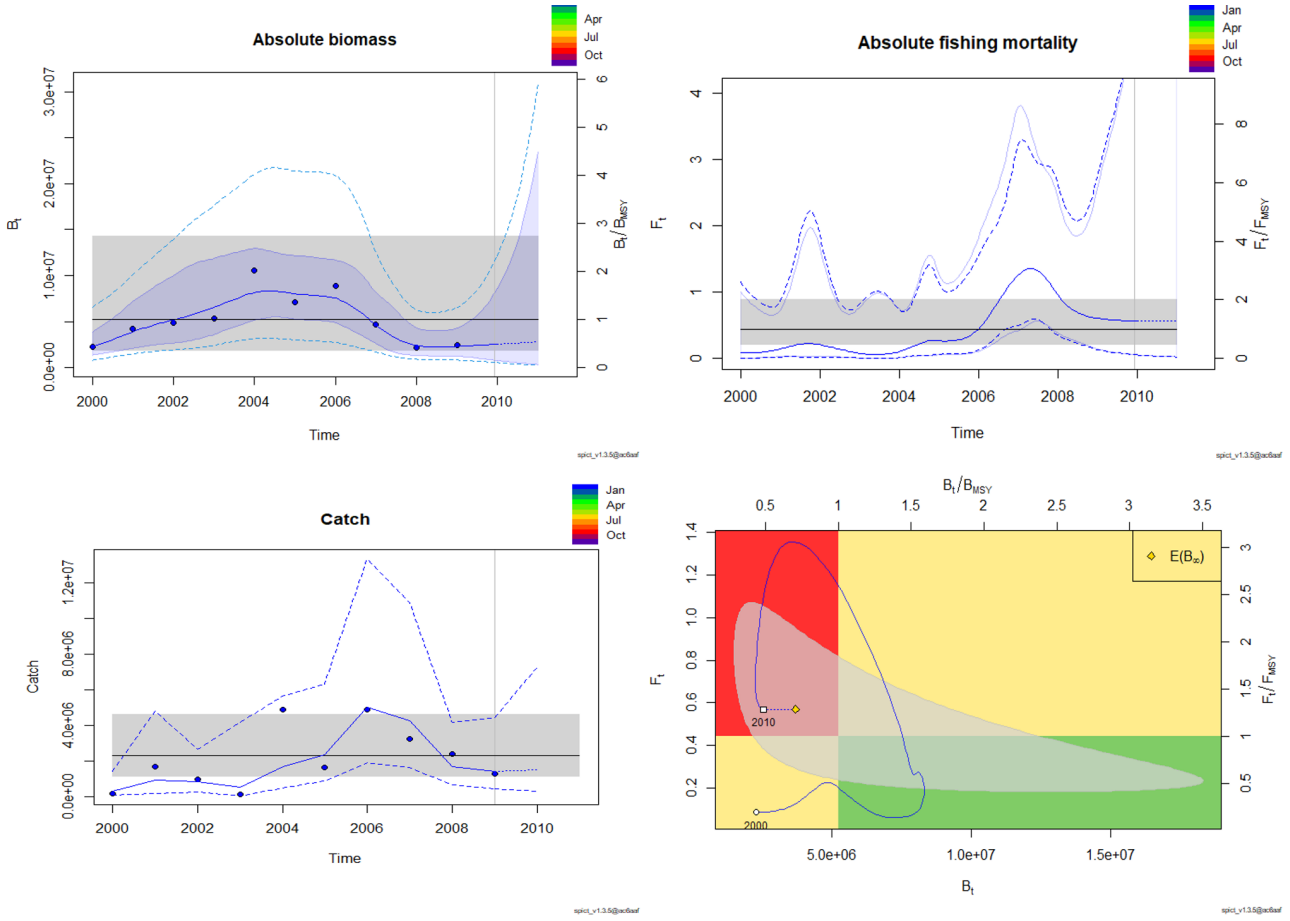

Figura 4.12: Resultados modelo SPiCT para los índices de biomasa relativa (muestreo aleatorio).

Para finalizar, se ha de comprobar que todos los parámetros han convergido correctamente y que la estimación es oportuna (Figura 4.13). SPiCT devuelve un diagnóstico por defecto, a través de una de las funciones implementadas en el paquete. En este caso, los p-valores del diagnóstico para la calidad del modelo, para las capturas y los índices, no son significativos, por lo tanto, no parece haber ningún problema en el ajuste del modelo y los resultados son fiables.

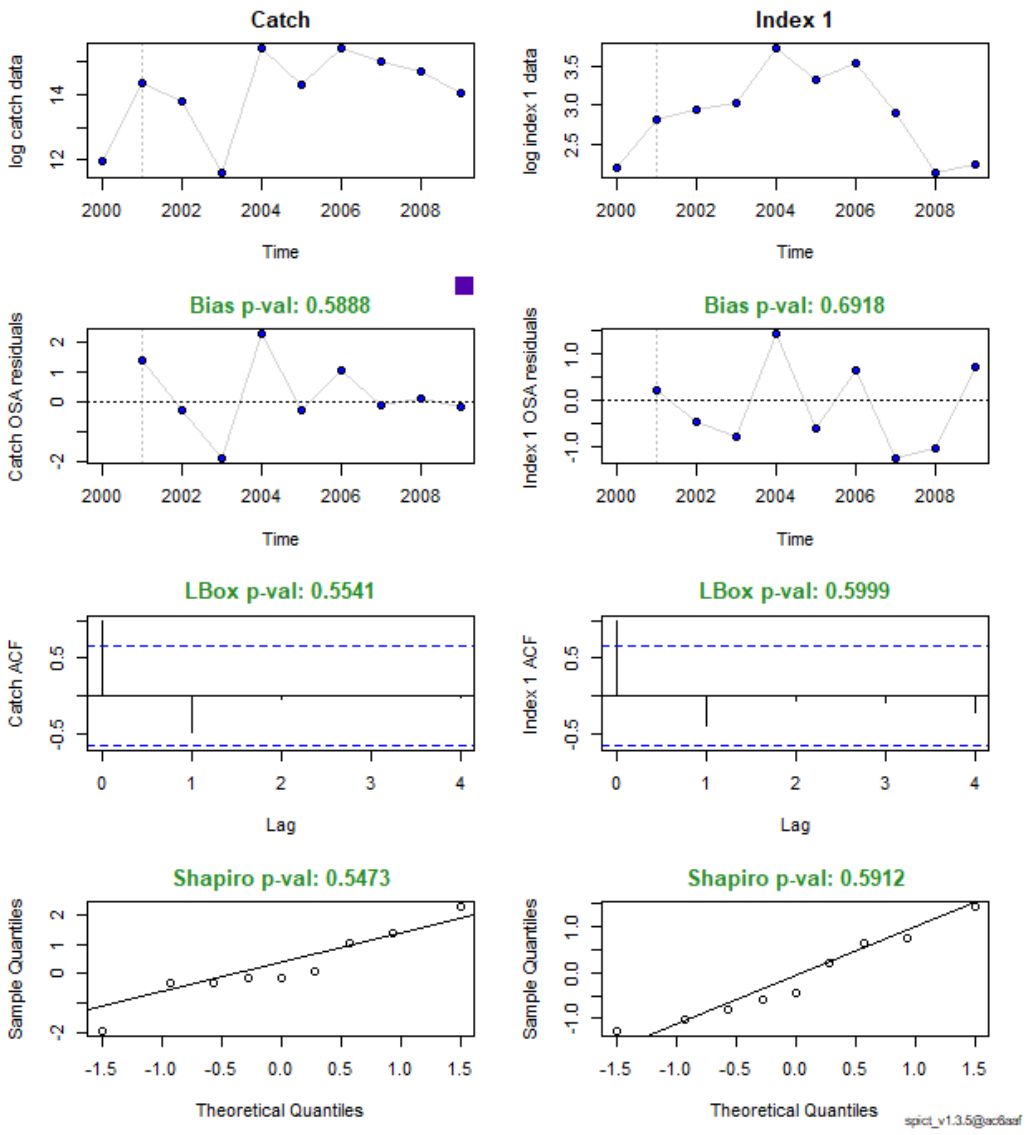

Figura 4.13: Diagnóstico del modelo SPiCT para el índice de biomasa relativa (muestreo aleatorio).

4.3.2. Índices de CPUE: muestreo preferencial

En este último apartado, vamos a reproducir los pasos para aplicar un modelo SPiCT, de igual modo que hemos hecho con el índice de biomasa relativa del muestreo aleatorio, pero, utilizando como *input* la serie de índices de CPUE predicha con el modelo preferencial (patrón puntual marcado) para el muestreo preferencial (dependiente de la pesca).

Al igual que con el modelo de evaluación anterior, en la Figura 4.14, mostramos los *inputs* que hemos utilizado para el SPiCT, de manera que, por un lado, tenemos la serie de índices de CPUE predichos con un modelo preferencial y, por otro lado, la serie de capturas que hemos aproximado. Esta última coincide con la serie de capturas de la Figura 4.11.

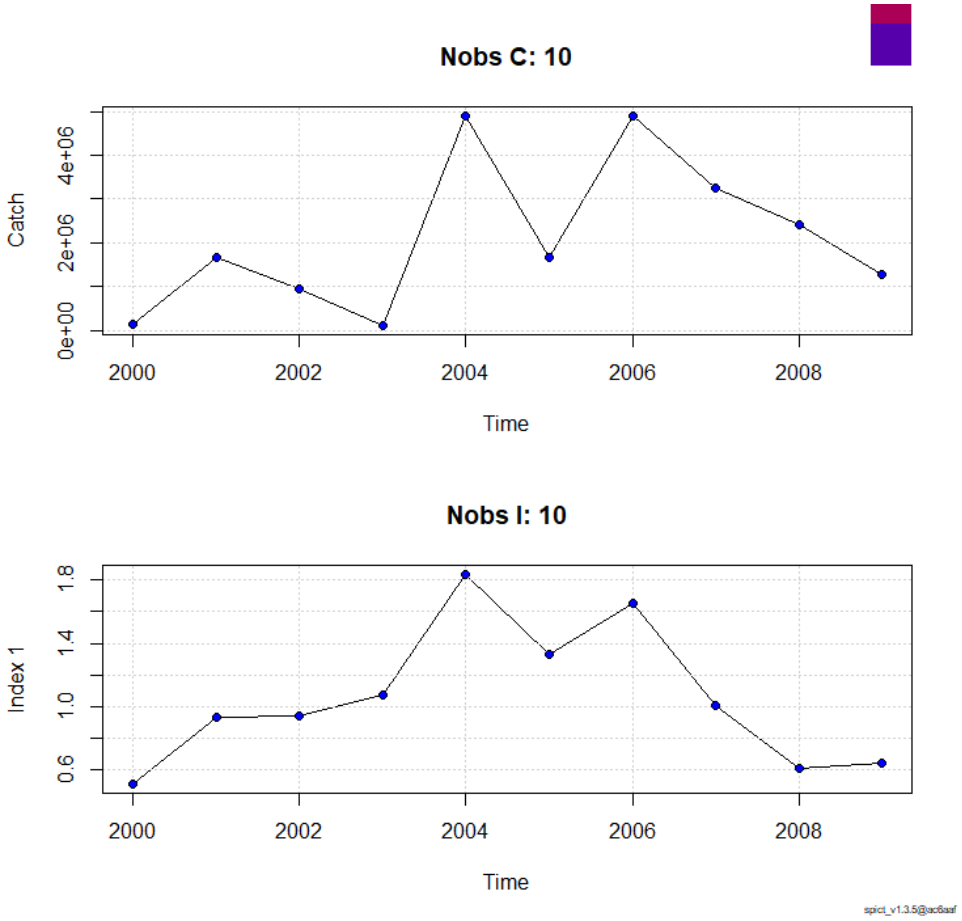

Figura 4.14: *Inputs* SPiCT: serie de capturas simulada (Nobs C:10) y serie de índices de CPUE predichos (Nobs I:10).

Una vez introducimos los *inputs* y ajustamos el modelo, en la Figura 4.15, mostramos algunos de los *outputs* para el SPiCT ajustado con los índices de CPUE. Si comparamos la Figura 4.15 y la Figura 4.12, los resultados de ambos modelos de evaluación (SPiCT) son muy parecidos.

Por último, en la Figura 4.16, se recoge el diagnóstico del modelo, como puede verse todos los p-valores son no significativos, así que, no parece haber ningún problema en el ajuste del modelo, siendo los resultados fiables.

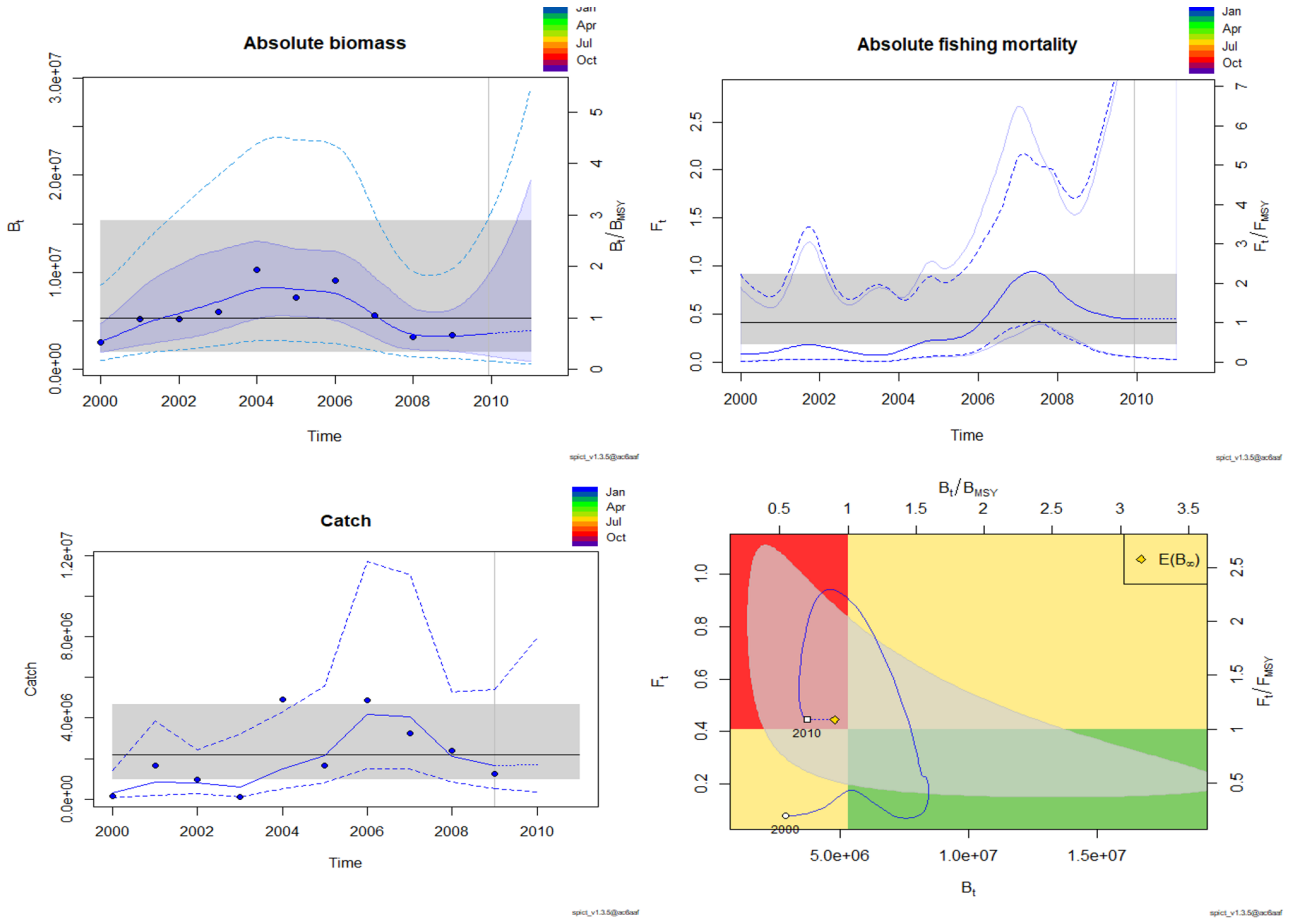

Figura 4.15: Resultados modelo SPiCT para los índices de CPUE (muestreo preferencial).

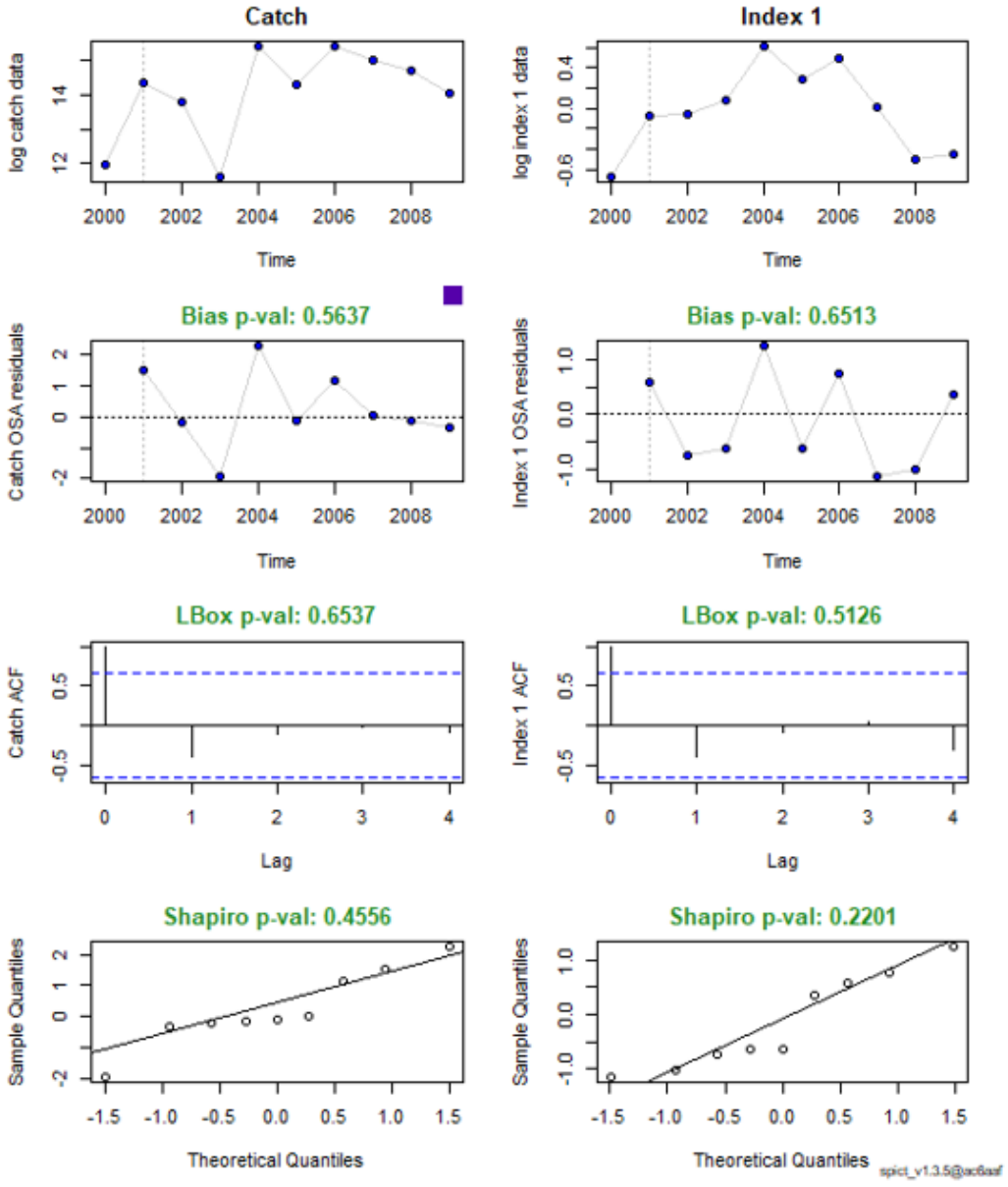

Figura 4.16: Diagnóstico del modelo SPiCT para el índice de CPUE (muestreo preferencial).

Capítulo 5

Discusión

En el siguiente capítulo, se discuten los resultados obtenidos para cada una de las partes de este trabajo, comenzando por los resultados de la modelización hasta proponer una serie de conclusiones y líneas futuras.

5.1. Comparativa modelizaciones

En lo referente a los resultados obtenidos para las distintas modelizaciones de los índices de biomasa relativa y de CPUE, ambos análisis han concluido que es conveniente la modelización del proceso espacio-temporal subyacente para poder capturar, de manera precisa, el comportamiento de la biomasa real del *stock*. Este resultado, pone en manifiesto la necesidad de emplear modelos más complejos como son los geoestadísticos, en lugar de utilizar GLMs o GAMs para predecir las series de índices que posteriormente alimentarán los modelos de evaluación del *stock*.

Por consiguiente, los resultados que hemos extraído de este trabajo para la modelización de los índices de biomasa relativa y de CPUE, en el contexto de la evaluación de un *stock*, coinciden con otras investigaciones. Entre algunos de los trabajos destacan Maunder *et al.* (2020), donde recalcaban la necesidad de modelos espacio-temporales, como son los geoestadísticos, para poder capturar el comportamiento de la biomasa a través de índices de biomasa relativa y de CPUE. Del mismo modo, Stock *et al.* (2019) y Stock *et al.* (2020), también manifestaban que los modelos geoestadísticos resultan en estimaciones más precisas en comparación con GLMs o GAMs.

En resumen, realizar una predicción en toda la zona de estudio, en la que se tiene en cuenta un proceso espacial subyacente sobre los índices de biomasa relativa o de CPUE, consigue

capturar mejor el comportamiento de la biomasa del *stock*. Por lo tanto, dichas series predichas de índices deberían funcionar mejor como *input* en los modelos de producción excedentaria, como, por ejemplo, la aproximación SPiCT.

Por otro lado, cabe matizar en algunos aspectos relacionados con los resultados para las series predichas de índices de CPUE. Estos índices derivados de la actividad pesquera, a diferencia de los índices derivados de campañas oceanográficas, presentan una dependencia en el muestreo. En otros términos, las observaciones de las que dispone el investigador en el espacio están sesgadas, ya que los pescadores siempre recurren a los parches de máxima biomasa al conocer la posición de dichos parches.

En base a esta dependencia en el muestreo, asumir un modelo geoestadístico puede seguir derivando en sesgos a la hora de recuperar el comportamiento de la biomasa del *stock*. Por ello, Pennino *et al.* (2019) proponen modelizar un patrón puntual marcado o modelo preferencial para inferir y predecir sobre los índices de CPUE. Un modelo preferencial consiste en combinar un modelo geoestadístico para el proceso continuo con un modelo LGCP para el proceso del patrón puntual.

En vista a los resultados obtenidos por Pennino *et al.* (2019), en este trabajo, hemos modelizado un patrón puntual marcado para las CPUE, que además ha sido el modelo que mejores resultados ha obtenido a la hora de recuperar el comportamiento de la biomasa simulada. No obstante, Ducharme-Barth *et al.* (2022) llevan a cabo la simulación de diversos escenarios con distintos grados de preferencialidad, concluyendo que, en el caso de que la preferencialidad no sea muy marcada, un modelo geoestadístico puede conseguir resultados satisfactorios a la hora de captar el comportamiento de la biomasa real del *stock*.

Finalmente, se ha utilizado la serie de índices de biomasa relativa predicha con un modelo geoestadístico y la serie de índices de CPUE predicha con un modelo preferencial (patrón puntual marcado) para ajustar un modelo de producción excedentaria del tipo SPiCT. Estos modelos llevaban a cabo una evaluación completa del estado del *stock* en base a los niveles de biomasa. Los resultados de ambos modelos SPiCT convergen, de forma que ambos *inputs* (serie de índices de biomasa relativa y de CPUE) han conseguido una evaluación muy parecida para el *stock*, lo que puede ser un indicativo de que los índices han sido modelados correctamente.

5.2. Protocolo de simulación

En el siguiente apartado profundizaremos en el protocolo de simulación que hemos elaborado para conocer la biomasa del *stock* y reproducir las fuentes de información más características en la evaluación de pesquerías (índices de biomasa relativa y de CPUE).

Para comenzar, en lo referente a la simulación de la biomasa del *stock*, recalcar que dicha simulación se ha hecho bajo el supuesto de un modelo de biomasa en el que existe un proceso espacial correlado subyacente (3.1). Los motivos principales por los que asumimos dicho supuesto son:

1. Suponer que no existe un efecto espacial correlacionado en el tiempo en la columna de agua resulta prácticamente inverosímil, puesto que, las poblaciones de peces están en constante interacción con el medio.

2. En la mayoría de los artículos presentes en la literatura, en los que se asume un proceso espacial subyacente, las conclusiones extraídas de los modelos mejoran notablemente (Paradinas *et al.*, 2017; Martínez-Minaya *et al.*, 2018; Pennino *et al.*, 2019; Stock *et al.*, 2019; Maunder *et al.*, 2020; Stock *et al.*, 2020; Izquierdo *et al.*, 2021; Pennino *et al.*, 2022).

3. Resulta muy interesante observar las consecuencias de utilizar modelos más sencillos (GLMs o GAMs), sabiendo que existe un proceso espacial subyacente, para así, confirmar si es necesario el uso de modelos más complejos como los geoestadísticos o es suficiente con los GLMs o GAMs.

Por otro lado, la reproducción de las principales fuentes de información, como son los índices de biomasa relativa derivados de campañas oceanográficas (muestreo aleatorio) o los índices de CPUE derivados de la actividad pesquera (muestreo preferencial), ha sido posible gracias a la reproducción de distintos escenarios de muestreo y la introducción de un coeficiente de capturabilidad q, que relacionada la biomasa del *stock* con los índices de biomasa relativa o de CPUE (Cousido-Rocha *et al.*, 2022). Además, para los índices de CPUE, se ha simulado una variable a la que denominamos esfuerzo de pesca. Esta se basa en el tiempo que el pescador mantiene el arte activo, de forma que hemos asumido una relación lineal entre las capturas y el esfuerzo de pesca. Esta asunción lineal del esfuerzo de pesca no siempre ocurre, p.ej. en las pesquerías de arrastre la red puede llegar a saturarse, de manera que por mucho que se invierta más tiempo en la pesca, el arte ya está saturado, es decir, ha llegado a su límite de carga.

5.3. Limitaciones del trabajo

Limitaciones con SPiCT

La mayor limitación de este trabajo, ha sido a la hora de aplicar el modelo de producción excedentaria SPiCT. El motivo principal, ha sido la naturaleza de la simulación, es decir, recordamos que un modelo de producción excedentaria necesita dos *inputs*: (1) una serie de

capturas y (2) una serie de índices de biomasa relativa o de CPUE. La simulación se enfocó para obtener la serie de índices de biomasa relativa o de CPUE, puesto que, el fin de este trabajo es modelar estos índices para que sean representativos de la biomasa. Sin embargo, nuestra simulación no contempla simular la serie de capturas, sino que, se aproximan a través de las ecuaciones base de un modelo SPMs (3.2).

Por consiguiente, obtener la serie de capturas por una vía ajena a la simulación desarrollada, puede estar afectando a la hora de ajustar el modelo de producción excedentaria SPiCT. Igualmente, las series de índices de biomasa relativa o de CPUE con las que estamos alimentando a SPiCT contienen un total de 10 años, un número que puede considerarse insuficiente para conseguir una buena estimación de la biomasa Cousido-Rocha *et al.* (2022).

En definitiva, para poder llevar a cabo una buena evaluación del *stock*, a través de un modelo de producción excedentaria (SPiCT), necesitamos profundizar más en la formulación del modelo y ligar la simulación elaborada al mismo. Todo ello, con el fin de conseguir la simulación de un escenario de biomasa que tenga un proceso espacial subyacente correlado en el tiempo y, a su vez, siga una dinámica basada en estos modelos de producción.

Limitaciones en la modelización de los índices

Por otra parte, hemos tenido alguna problemática a la hora de ajustar y predecir sobre los parámetros de los modelos para los índices de biomasa relativa y de CPUE. Por un lado, para los índices de biomasa relativa, es cierto que el modelo geoestadístico ha sido el que menor RMSE y MAPE ha conseguido. No obstante, el modelo sobreestima sistemáticamente la biomasa en todos los años. Suponemos que puede ser un problema en las previas de los parámetros o hiperparámetros del modelo o una cuestión del propio muestreo.

Por otro lado, para los índices de CPUE, también tenemos una sobreestimación constante de la biomasa. Este patrón suele estar relacionado con la preferencialidad del muestreo, ya que únicamente se cubre el rango de biomasas más alto. Sin embargo, el modelo preferencial consigue paliar notablemente los efectos del muestreo preferenciado, provocando una reducción en el valor de las medidas de error (RMSE y MAPE). Profundizando en el modelo preferencial, en la literatura, no se había ajustado previamente un modelo preferencial o patrón puntual marcado en el espacio y tiempo. Por lo tanto, hemos tenido algunas limitaciones a la hora de la inferencia y la predicción con `inlabru`, de forma que no hemos conseguido introducir la covariable batimetría con certeza. Como consecuencia, el modelo preferencial únicamente contempla el intercepto, el efecto espacial correlado y la tendencia temporal.

Para concluir, muchas de las limitaciones con las que nos hemos encontrado en la modelización de los índices de biomasa relativa o de CPUE, pueden estar relacionadas con los parámetros

fijados para la simulación. Por ende, sería conveniente preparar y correr una secuencia de simulaciones y comprobar si llegamos a las mismas conclusiones.

5.4. Líneas futuras

De acuerdo con las limitaciones del trabajo y mirando hacia el futuro, se plantean algunas líneas futuras, con las que pretendemos completar el presente trabajo y que podrían considerarse de interés científico:

- Conseguir una simulación en la que combinemos los modelos de evaluación del *stock*, como son los modelos de producción excedentaria, con los modelos espacio-temporales basados en geoestadística. Eso permitiría tener la capacidad de comparar las distintas estimaciones de biomasa que consigue SPiCT, en función de la calidad de los *inputs* (serie de capturas y serie de índices de biomasa relativa o de CPUE).
- Aumentar el número de años en la simulación, de forma que obtengamos una serie predicha de índices con un número considerable de años y, así conseguir que SPiCT funcione correctamente.
- Generar una batería de simulaciones, de manera que ajustemos los modelos para los índices de biomasa relativa y de CPUE en diferentes condiciones de muestreo, de biomasa simulada, etc., y así afianzar los resultados que obtengamos.
- Introducir covariables en un modelo preferencial en el espacio-tiempo.

Capítulo 6

Conclusiones

Para concluir con el trabajo, procedemos a enumerar las conclusiones principales que hemos obtenido gracias al protocolo elaborado:

- El protocolo desarrollado para la simulación de un escenario espacio-temporal de biomasa y la reproducción de distintos escenarios de muestreo con sus fuentes de información asociadas, ha resultado satisfactorio para evaluar qué modelización de los índices de biomasa relativa y de CPUE consigue capturar mejor el comportamiento de la biomasa.
- No tener en cuenta el proceso espacial subyacente en la modelización de los índices de biomasa relativa y de CPUE puede generar cierto sesgo en la inferencia y predicción de los parámetros del modelo.
- Para los índices de biomasa relativa derivados de campañas oceanográficas (muestreo aleatorio), la modelización que mejor ha conseguido capturar el comportamiento de la biomasa simulada, presentando menor RMSE y MAPE, ha sido el modelo geoestadístico resuelto mediante la aproximación INLA, implementada en `R` a través del paquete `R-INLA`.
- Para los índices de CPUE derivados de la actividad pesquera (muestreo preferencial), la modelización que mejor ha conseguido capturar el comportamiento de la biomasa simulada, presentando menor RMSE y MAPE, ha sido el modelo preferencial (patrón puntual marcado) resuelto mediante la aproximación INLA, implementada en `R` a través del paquete `inlabru`.
- El modelo que mejor resultado ha obtenido, en términos de RMSE y MAPE, de la totalidad de modelos implementados (englobando índices de biomasa relativa y de CPUE) ha sido el modelo geoestadístico para el muestreo aleatorio, con un RMSE y MAPE de 12.45 y 0.19 respectivamente.

- Los resultados de la evaluación del estado del *stock* obtenidos mediante dos modelos SPiCT (uno de ellos, con el índice de biomasa relativa derivado del modelo geoestadístico como *input* y, el otro, con el índice de CPUE derivado del modelo preferencial como *input*) convergen a conclusiones simulares.

Anexo I: simulación

En el siguiente anexo incluimos una serie de gráficas descriptivas de la simulación empleada en este trabajo. En concreto, los histogramas para la biomasa simulada, para los índices de biomasa relativa y para los índices de CPUE y, un gráfico representando la relación de la batimetría con la biomasa simulada.

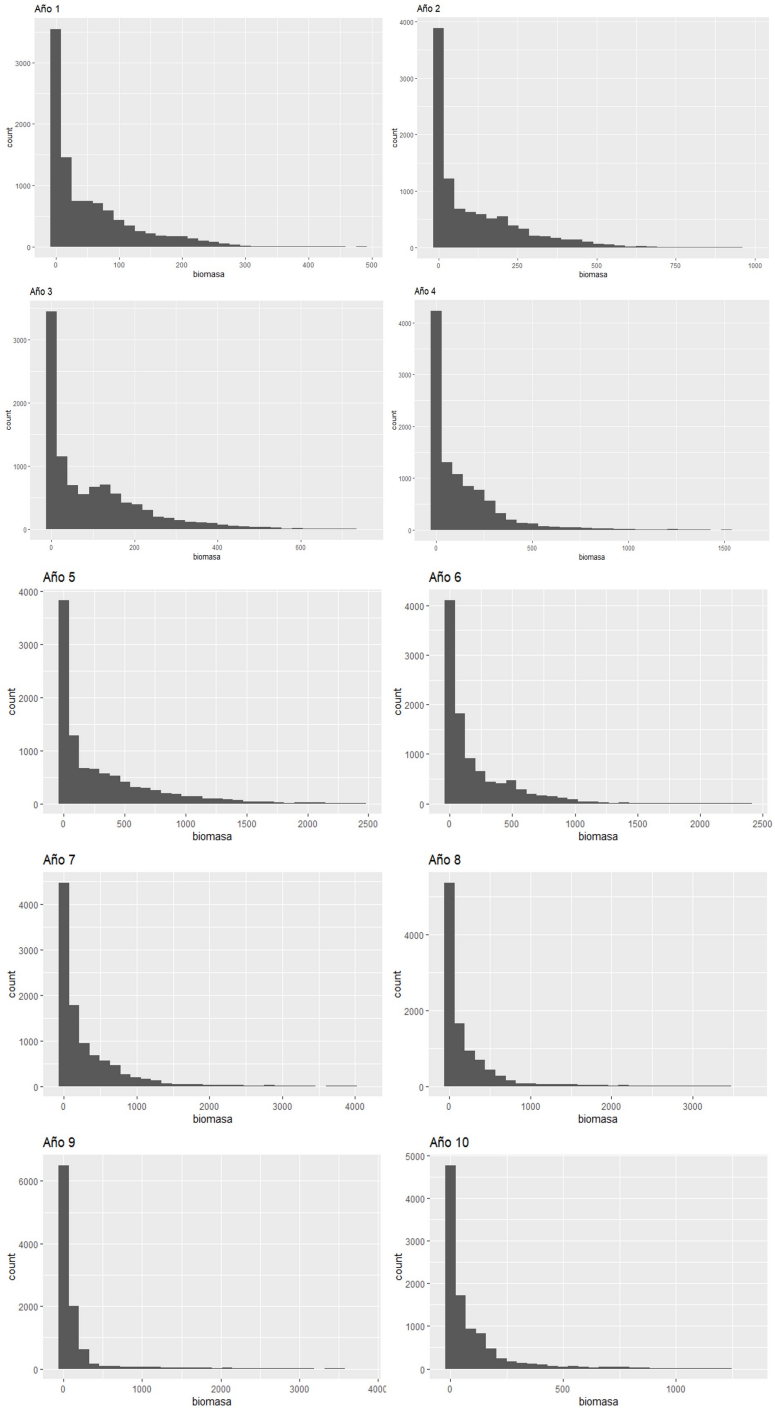

Figura 6.1: Histograma de la biomasa simulada para cada año.

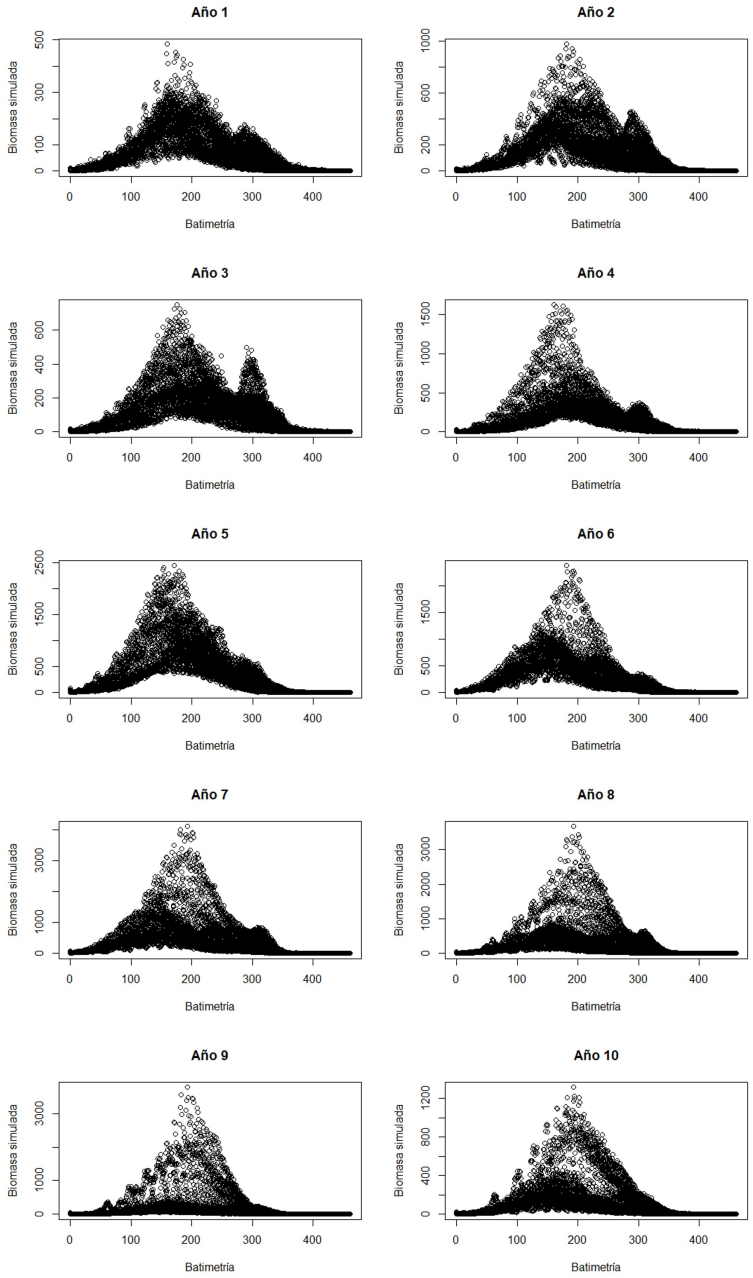

Figura 6.2: Relación entre la biomasa simulada y la covariable batimetría.

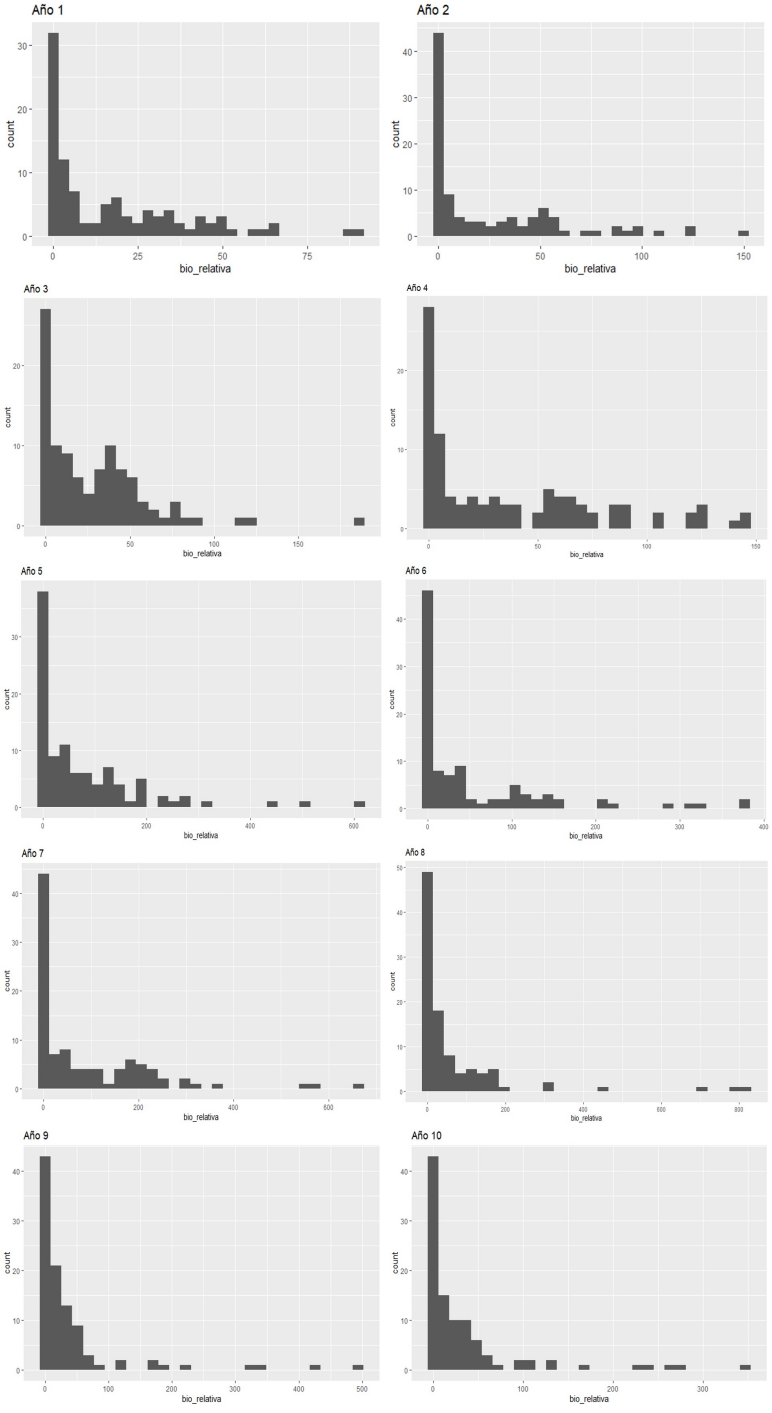

Figura 6.3: Histograma índice de biomasa relativa para cada año.

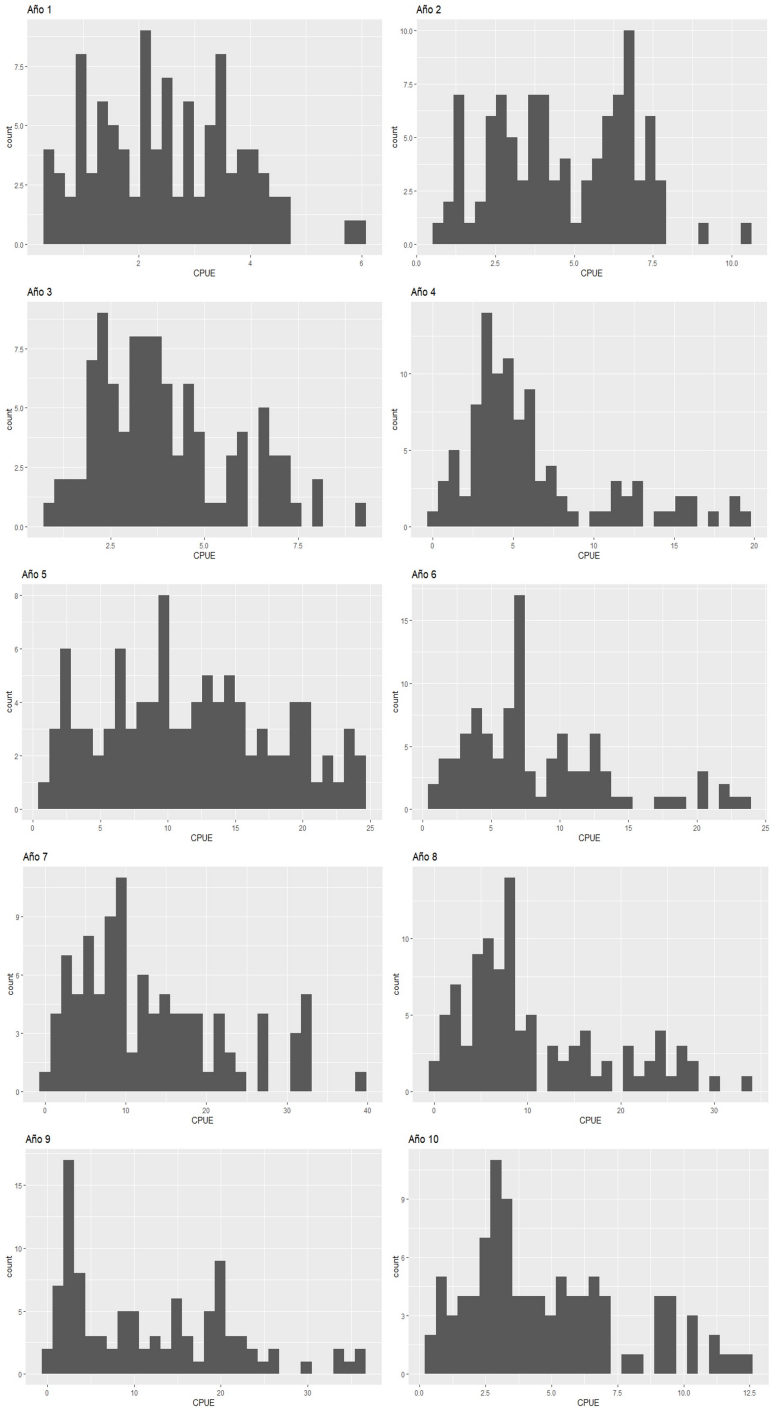

Figura 6.4: Histograma índice de CPUE para cada año.

Anexo II: mapas

En el siguiente anexo se presentan algunos mapas complementarios de interés. En concreto, el mapa de la desviación típica y de los cuantiles (2.5 % y 97.5 %) para el modelo geoestadísticos asociado índice de biomasa relativa y, el mapa de la desviación típica y de los cuantiles (2.5 % y 97.5 %) para el modelo preferencial asociado al índice de CPUE.

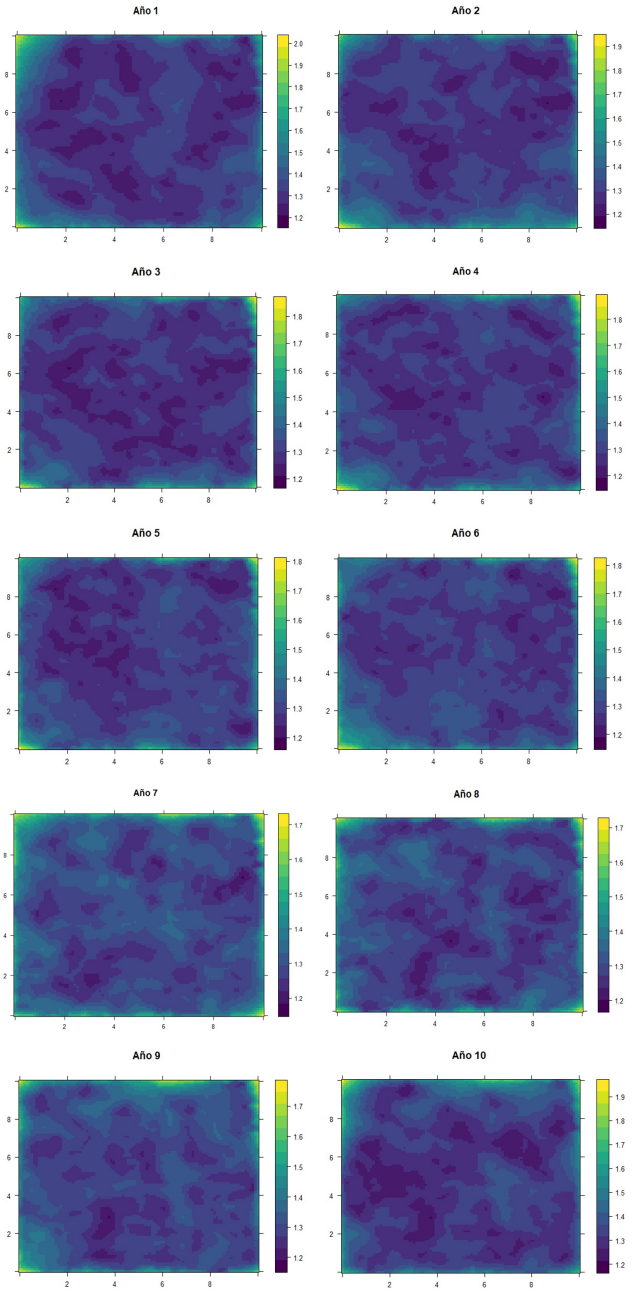

Figura 6.5: Desviación típica de la distribución predictiva a posteriori del índice de biomasa relativa con un modelo geoestadístico.

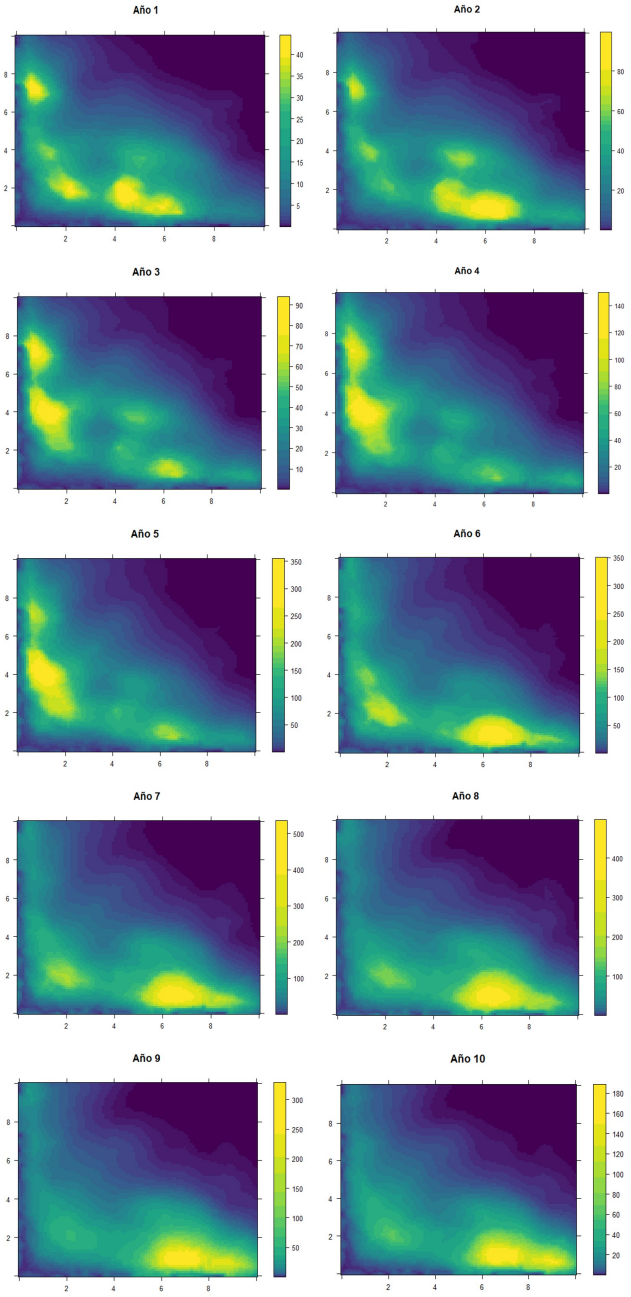

Figura 6.6: Cuantil 2.5 % de la distribución predictiva a posteriori del índice de biomasa relativa con un modelo geoestadístico.

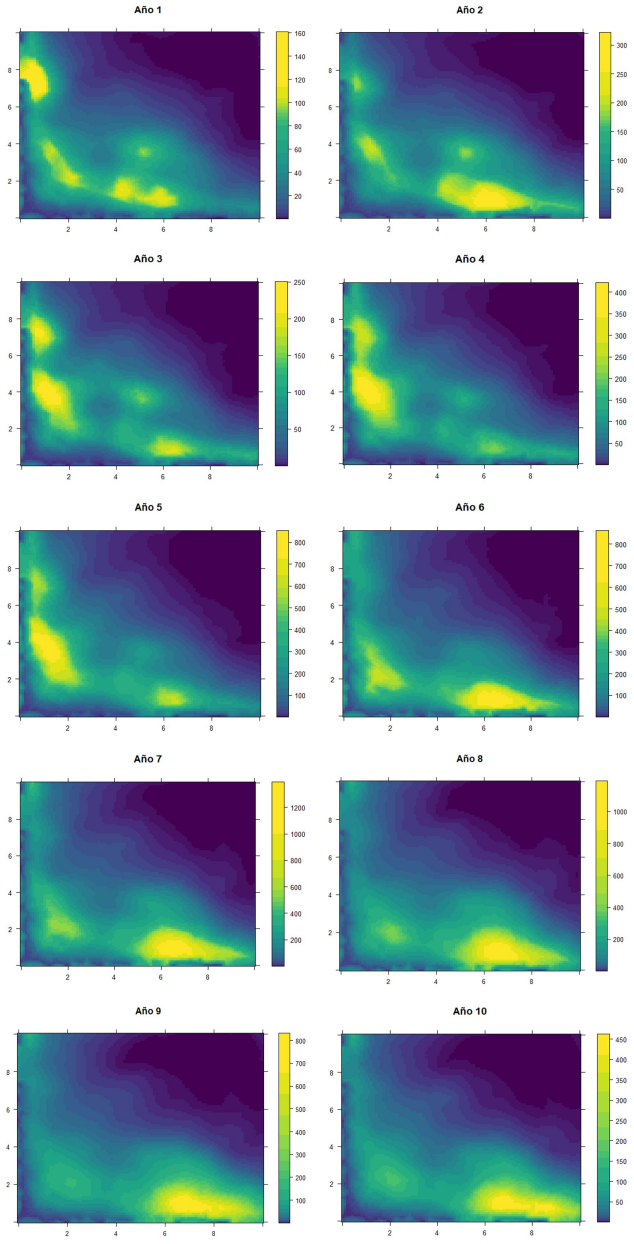

Figura 6.7: Cuantil 97.5 % de la distribución predictiva a posteriori del índice de biomasa relativa con un modelo geoestadístico.

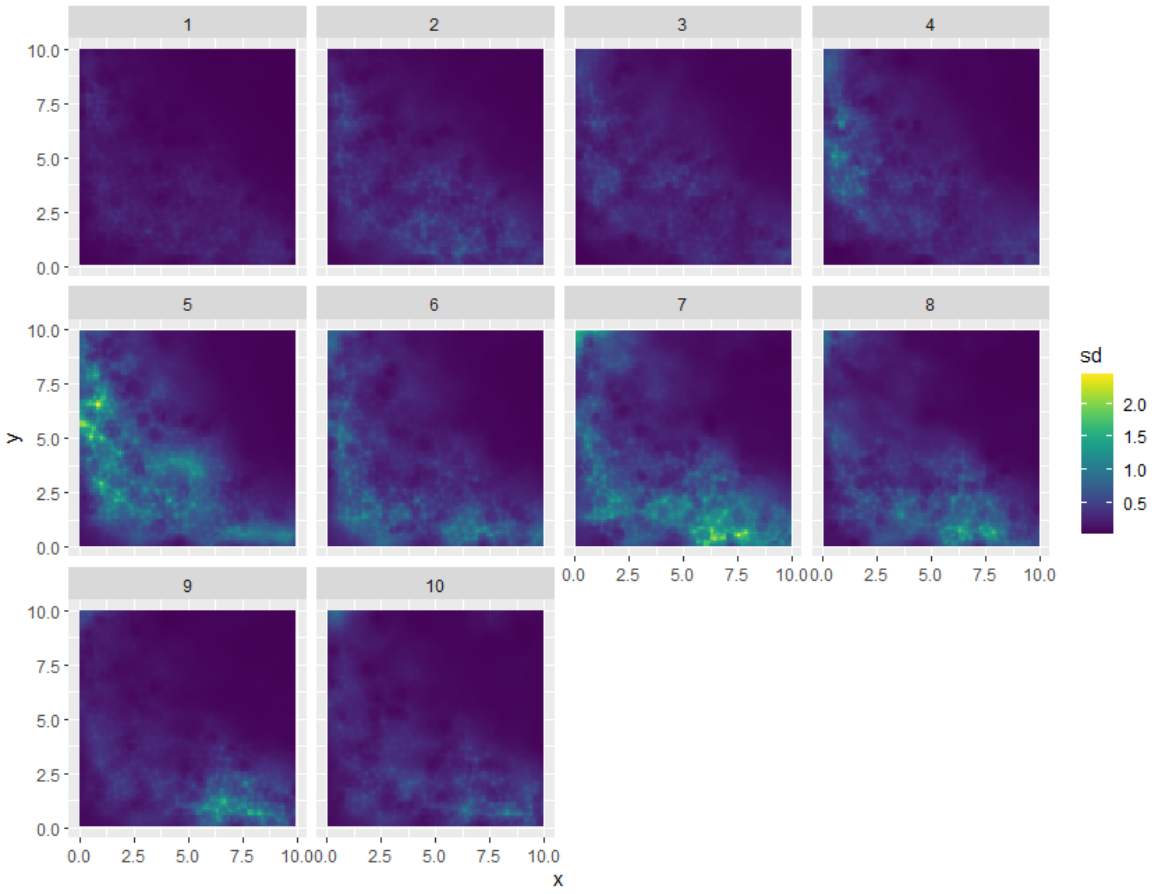

Figura 6.8: Desviación típica de la distribución predictiva a posteriori del índice de CPUE con un modelo preferencial.

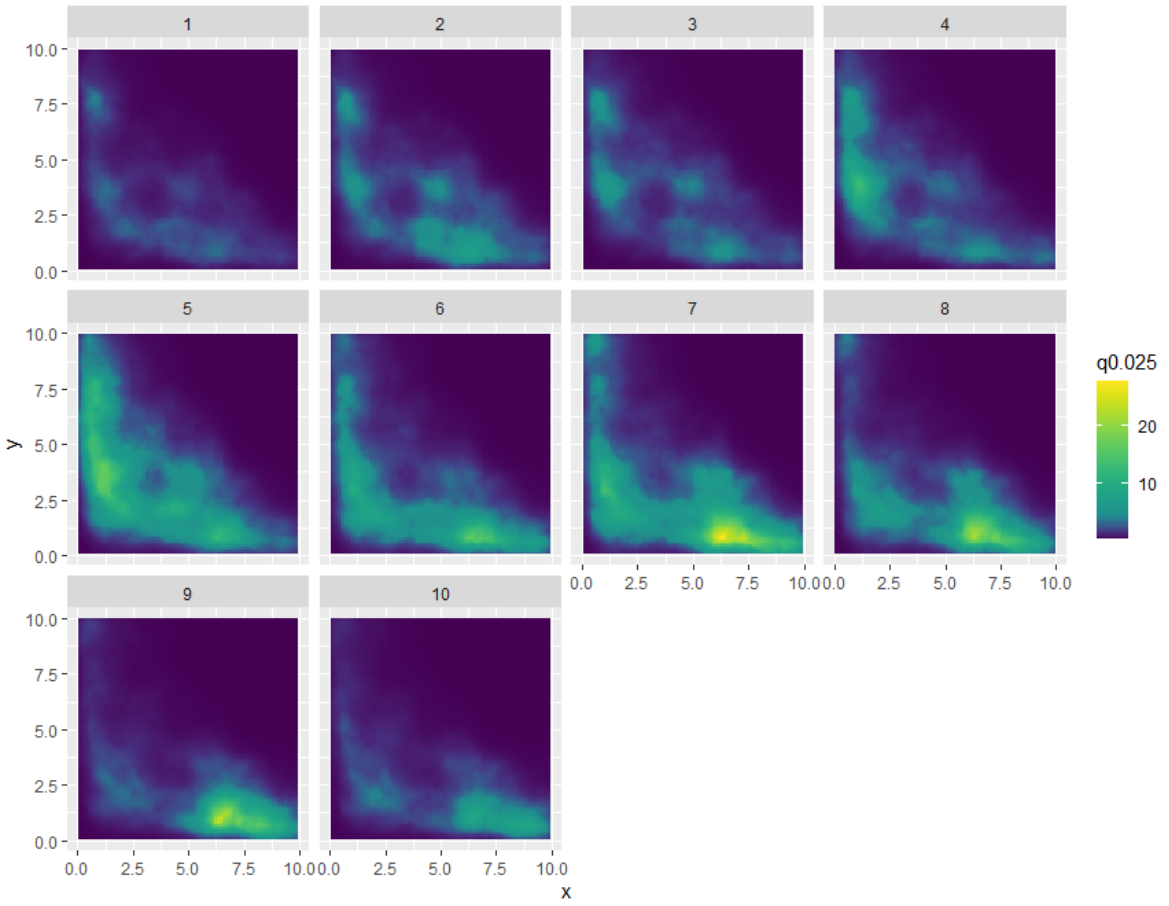

Figura 6.9: Cuantil 2.5 % de la distribución predictiva a posteriori del índice de CPUE con un modelo preferencial.

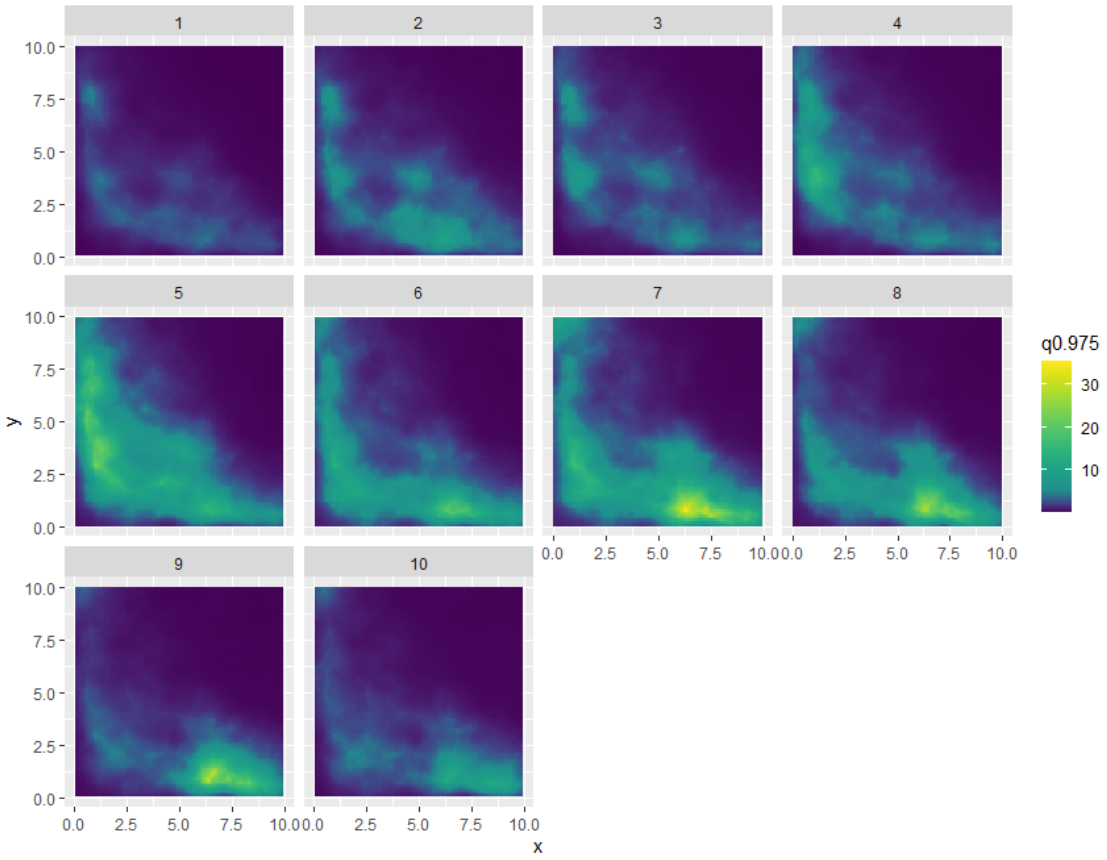

Figura 6.10: Cuantil 97.5 % de la distribución predictiva a posteriori del índice de CPUE con un modelo preferencial.

Material complementario.

Por un lado, para el acceso y descarga a los datos simulados empleados en este trabajo, adjuntamos el siguiente *link*: datos simulados.

Por otro lado, para el acceso y descarga del código empleado en este trabajo, adjuntamos el *link* a un repositorio GitHub: código TFM Alba

Referencias

Arreguín-Sánchez, F. (1996). Catchability: a key parameter for fish stock assessment. *Reviews in fish biology and fisheries*, 6(2):221–242.

Bachl, F. E., Lindgren, F., Borchers, D. L., e Illian, J. B. (2019). `inlabru`: an `R` package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10(6):760–766.

Baddeley, A., Bárány, I., y Schneider, R. (2007). Spatial point processes and their applications. *Stochastic Geometry: Lectures Given at the CIME Summer School Held in Martina Franca, Italy, September 13–18, 2004*, pp. 1–75.

Banerjee, S. (2016). Spatial data analysis. *Annual review of public health*, 37:47–60.

Bivand, R. S., Pebesma, E. J., y Gómez-Rubio, V. (2008). Areal data and spatial autocorrelation. *Applied spatial data analysis with R*, pp. 237–272.

Björndal, T., Lane, D. E., y Weintraub, A. (2004). Operational research models and the management of fisheries and aquaculture: a review. *European Journal of Operational Research*, 156(3):533–540.

Box George, E., Jenkins Gwilym, M., Reinsel Gregory, C., y Ljung Greta, M. (1976). *Time series analysis: forecasting and control*. San Francisco: Holden Bay.

Cavieres, J. y Nicolis, O. (2018). Using a spatio-temporal Bayesian approach to estimate the relative abundance index of yellow squat lobster (*Cervimunida johni*) off Chile. *Fisheries research*, 208:97–104.

Cerviño, S. (2004). Estudio de la incertidumbre asociada a los métodos de evaluación de las poblaciones de peces. Tesis de máster, Universidad de Vigo.

Chatfield, C. (2003). *The analysis of time series: an introduction*. Chapman and hall/CRC.

Clark, C. W. (2006). Fisheries bioeconomics: why is it so widely misunderstood? *Population Ecology*, 48(2):95–98.

Cousido-Rocha, M., Pennino, M., Izquierdo, F., Paz, A., Lojo, D., A, T., Zanni, M., y Cerviño, S. (2022). Surplus Production models: a practical review of recent approaches. *Reviews in Fish Biology and Fisheries*.

Depaoli, S., Clifton, J. P., y Cobb, P. R. (2016). Just another Gibbs sampler (JAGS) flexible software for MCMC implementation. *Journal of Educational and Behavioral Statistics*, 41(6):628–649.

Ducharme-Barth, N. D., Grüss, A., Vincent, M. T., Kiyofuji, H., Aoki, Y., Pilling, G., Hampton, J., y Thorson, J. T. (2022). Impacts of fisheries-dependent spatial sampling patterns on catch-per-unit-effort standardization: A simulation study and fishery application. *Fisheries Research*, 246:106169.

FAO (2020). *The State of World Fisheries and Aquaculture (SOFIA)*. FAO.

García, F. M. (2004). Aplicación de la geoestadística en las ciencias ambientales. *Ecosistemas*, 13(1).

Gómez-Rubio, V. (2020). *Bayesian inference with INLA*. CRC Press.

Grüss, A. y Thorson, J. T. (2019). Developing spatio-temporal models using multiple data types for evaluating population trends and habitat usage. *ICES Journal of Marine Science*, 76(6):1748–1761.

Guerra Sierra, A. y Sánchez Lizaso, J. L. (1998). *Fundamentos de explotación de recursos vivos marinos*. ACRIBIA, S.A.

Hilborn, R. y Walters, C. J. (2013). *Quantitative fisheries stock assessment: choice, dynamics and uncertainty*. Springer Science & Business Media.

Hinton, M. G. y Maunder, M. N. (2004). Methods for standardizing CPUE and how to select among them. *Col. Vol. Sci. Pap. ICCAT*, 56(1):169–177.

Hjort, J., Jahn, G., y Ottestad, P. (1933). The optimum catch. *Hvalradets Skrifter*, 7:92–127.

Hutchings, J. A. (1996). Spatial and temporal variation in the density of northern cod and a review of hypotheses for the stock's collapse. *Canadian Journal of Fisheries and Aquatic Sciences*, 53(5):943–962.

Iversen, E. S. (1996). *Living marine resources: their utilization and management*. Springer Science & Business Media.

Izquierdo, F., Paradinas, I., Cerviño, S., Conesa, D., Alonso-Fernández, A., Velasco, F., Preciado, I., Punzón, A., Saborido-Rey, F., y Pennino, M. G. (2021). Spatio-temporal assessment of the European hake (*Merluccius merluccius*) recruits in the northern Iberian Peninsula. *Frontiers in Marine Science*, 8:1.

Jentzen, A. y Kloeden, P. E. (2009). The numerical approximation of stochastic partial differential equations. *Milan Journal of Mathematics*, 77(1):205–244.

Kai, M. (2019). Spatio-temporal changes in catch rates of pelagic sharks caught by Japanese research and training vessels in the western and central north Pacific. *Fisheries Research*, 216:177–195.

King, M. (2013). *Fisheries biology, assessment and management*. John Wiley & Sons.

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., y Rue, H. (2018). *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman and Hall/CRC.

Larkin, P. A. (1989). Comments on the workshop presentations. *Fishery Science and Management: Objectives and Limitations*, 28:287–289.

Lindgren, L.-E. (2001). Finite element modeling and simulation of welding part 1: increased complexity. *Journal of thermal stresses*, 24(2):141–192.

Lunn, D. J., Thomas, A., Best, N., y Spiegelhalter, D. (2000). Winbugs-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337.

Martínez-Minaya, J., Cameletti, M., Conesa, D., y Pennino, M. G. (2018). Species distribution modeling: a statistical review with focus in spatio-temporal issues. *Stochastic environmental research and risk assessment*, 32(11):3227–3244.

Maunder, M. N., Thorson, J. T., Xu, H., Oliveros-Ramos, R., Hoyle, S. D., Tremblay-Boyer, L., Lee, H. H., Kai, M., Chang, S.-K., Kitakado, T., Christoffer, A. M., Minte-Vera, C., Lennert-Cody, C. E., Aires-da Silva, A. M., y Piner, K. R. (2020). The need for spatio-temporal modeling to determine catch-per-unit effort based indices of abundance and associated composition data for inclusion in stock assessment models. *Fisheries Research*, 229:105–594.

Møller, J. y Waagepetersen, R. P. (2007). Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684.

Myers, R. A. y Worm, B. (2003). Rapid worldwide depletion of predatory fish communities. *Nature*, 423(6937):280–283.

Osei, F. B., Duker, A. A., y Stein, A. (2012). Bayesian structured additive regression modeling of epidemic data: application to cholera. *BMC Medical Research Methodology*, 12(1):1–11.

Paradinas, I., Conesa, D., López-Quílez, A., y Bellido, J. M. (2017). Spatio-temporal model structures with shared components for semi-continuous species distribution modelling. *Spatial Statistics*, 22:434–450.

Pauly, D. (1996). One hundred million tonnes of fish, and fisheries research. *Fisheries research*, 25(1):25–38.

Pauly, D., Christensen, V., Guénette, S., Pitcher, T. J., Sumaila, U. R., Walters, C. J., Watson, R., y Zeller, D. (2002). Towards sustainability in world fisheries. *Nature*, 418(6898):689–695.

Pauly, D. y Zeller, D. (2003). The global fisheries crisis as a rationale for improving the FAO's database of fisheries statistics. *Fisheries Centre Research Reports*, 11(6):1–9.

Pawlowsky-Glahn, V. y Olea, R. A. (2004). *Geostatistical analysis of compositional data*, volumen 7. Oxford University Press.

Pedersen, M. W. y Berg, C. W. (2017). A stochastic Surplus Production model in continuous time. *Fish and Fisheries*, 18(2):226–243.

Pella, J. J. y Tomlinson, P. K. (1969). A generalized stock production model. *Inter-American Tropical Tuna Commission Bulletin*.

Pennino, M. G., Izquierdo, F., Paradinas, I., Cousido, M., Velasco, F., y Cerviño, S. (2022). Identifying persistent biomass areas: The case study of the common sole in the northern Iberian waters. *Fisheries Research*, 248:106196.

Pennino, M. G., Paradinas, I., Illian, J. B., Muñoz, F., Bellido, J. M., López-Quílez, A., y Conesa, D. (2019). Accounting for preferential sampling in species distribution models. *Ecology and evolution*, 9(1):653–663.

Peterman, R. M. (1990). Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Sciences*, 47(1):2–15.

Prager, M. (1992). Aspic: A Surplus Production model incorporating covariates. *Coll. Vol. Sci. Pap., Int. Comm. Conserv. Atl. Tunas (ICCAT)*, 28:218–229.

Prager, M. (1994). A suite of extensions to a nonequilibrium Surplus Production model. *Fish. Bull.*, 92:374–389.

Ripley, B. D. (2005). *Spatial statistics*. John Wiley & Sons.

Rue, H., Martino, S., y Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Russell, E. S. (1931). Some theoretical considerations on the "overfishing" problem. *ICES Journal of Marine Science*, 6(1):3–20.

Schaefer, M. B. (1954). Some aspects of the dynamics of populations important to the management of the commercial marine fisheries. *Bulletin, Inter American Tropical Tuna Commission*, pp. 1:25–56.

Shelton, P. y Lilly, G. (2000). Interpreting the collapse of the northern cod stock from survey and catch data. *Canadian Journal of Fisheries and Aquatic Sciences*, 57(11):2230–2239.

Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H., y Rue, H. (2016). Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103(1):49–70.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., y Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 32(1):1–28.

Stock, B. C., Ward, E. J., Eguchi, T., Jannot, J. E., Thorson, J. T., Feist, B. E., y Semmens, B. X. (2020). Comparing predictions of fisheries bycatch using multiple spatiotemporal species distribution model frameworks. *Canadian Journal of Fisheries and Aquatic Sciences*, 77(1):146–163.

Stock, B. C., Ward, E. J., Thorson, J. T., Jannot, J. E., y Semmens, B. X. (2019). The utility of spatial model-based estimators of unobserved bycatch. *ICES Journal of Marine Science*, 76(1):255–267.

Team, R. C. (2013). `R`: A language and environment for statistical computing.

Thorson, J. T. y Barnett, L. A. (2017). Comparing estimates of abundance trends and distribution shifts using single-and multispecies models of fishes and biogenic habitat. *ICES Journal of Marine Science*, 74(5):1311–1321.

Thorson, J. T., Shelton, A. O., Ward, E. J., y Skaug, H. J. (2015). Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for west coast groundfishes. *ICES Journal of Marine Science*, 72(5):1297–1310.

Tremblay-Boyer, L., Hampton, J., McKechnie, S., y Pilling, G. (2018). Stock assessment of south Pacific albacore tuna. *14th Regular Session of the Scientific Committee of the WCPFC. Busan, Republic of Korea.*

Tremblay-Boyer, L., McKechnie, S., Pilling, G., y Hampton, J. (2017). Geostatistical analyses of operational longline CPUE data. Technical report, Technical Report WCPFC-SC13-2017/SA-WP-03, Rarotonga, Cook Islands, 9-17 August.

Umlauf, N., Adler, D., Kneib, T., Lang, S., y Zeileis, A. (2012). Structured additive regression models: An `R` interface to BayesX. Technical report, Working Papers in Economics and Statistics.

Walters, C. y Maguire, J.-J. (1996). Lessons for stock assessment from the northern cod collapse. *Reviews in fish biology and fisheries*, 6(2):125–137.

Winker, H., Carvalho, F., y Kapur, M. (2018). JABBA: just another Bayesian biomass assessment. *Fisheries Research*, 204:275–288.

Xu, H., Thorson, J. T., Methot, R. D., y Taylor, I. G. (2019). A new semi-parametric method for autocorrelated age-and time-varying selectivity in age-structured assessment models. *Canadian Journal of Fisheries and Aquatic Sciences*, 76(2):268–285.

Yee, T. W. (2015). *Vector generalized linear and additive models: with an implementation in R*, volumen 10. Springer.

Zhou, S., Campbell, R. A., y Hoyle, S. D. (2019). Catch per unit effort standardization using spatio-temporal models for Australia's eastern tuna and billfish fishery. *ICES Journal of Marine Science*, 76(6):1489–1504.

Zuur, A. F., Ieno, E. N., y Saveliev, A. A. (2017). Spatial, temporal and spatial-temporal ecological data analysis with `R-INLA`. *Highland Statistics Ltd*, 1.