CIMAT Centro de Investigación en Matemáticas, A.C.

MODELACIÓN PARA EL ANÁLISIS DE LA EXTINCIÓN CON BASE EN EL REGISTRO FÓSIL

T E S I S

Que para obtener el grado de

Maestra en Ciencias

con Orientación en

Matemáticas Aplicadas

Presenta

Lilian Bárbara Pérez Sosa

Director de Tesis:

Dr. Miguel Nakamura Savoy

Autorización de la versión final

Guanajuato, Gto., 12 de julio de 2021

Centro de Investigación en Matemáticas

Matemáticas Aplicadas

Modelación para el análisis de la extinción con base en el registro fósil

TESIS

QUE PARA OBTENER EL TÍTULO DE:

Maestra en Ciencias con Especialidad en Matemáticas Aplicadas

PRESENTA:

Lilian Bárbara Pérez Sosa

ASESOR:

Dr. Miguel Nakamura Savoy

Guanajuato, Guanajuato, México, 12 de julio de 2021

‘

Lilian Bárbara Pérez Sosa

Fecha

Firma

Agradecimientos

Comenzaré agradeciendo a mis papás, Raquel y Gilberto, por el apoyo que me dan día a día a pesar de la distancia y que siempre han estado junto a mi en cada uno de los pasos que he dado. A mi tío Jesús, mis abuelos y mi hermano les doy las gracias porque han estado siempre preocupados por mi. Agradezco a todos mis amigos y compañeros de maestría que siempre nos hemos apoyado. En especial a Sebas por todo su cariño y a Yamil por su amistad.

Un agradecimiento especial a mi asesor Miguel Nakamura por su apoyo incondicional, disposición para ayudarme siempre y todas las enseñanzas que me ha dado. También quiero agradecer a los profesores de que he tenido durante estos últimos dos años, gracias a ellos considero que he crecido muchísimo académicamente y profesionalmente. Agradezco al Dr. Pablo del Monte por el apoyo en el desarrollo de esta tesis, toda la información que compartió con nosotros y su colaboración como sinodal de mi tesis. Al profesor Enrique Villa por su apoyo y colaboración siendo sinodal de mi tesis. Quiero agradecer al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo otorgado (número 757629) para la realización de mis estudios de la Maestría en Ciencias con especialidad en Matemáticas Aplicadas que realicé en CIMAT .

Resumen

En esta tesis se propone un modelo de regresión de Cox para el análisis de la extinción con base en el registro fósil de especies marinas. A lo largo de los últimos 540 millones de años, la biodiversidad global ha cambiado drásticamente y han habido cinco grandes eventos de extinción masiva. Cada uno de estos eventos variaron en tamaño y causa aniquilando a una abrumadora cantidad de especies que vivían en ese momento, es decir, se ejerció algún efecto sobre su longevidad. En este trabajo se realizó una revisión bibliográfica para identificar dichos factores a fin de incorporar información relevante y disponible en un modelo estadístico paramétrico. Las explicaciones con evidencia consistente acerca de estas extinciones involucran el cambio en el nivel del mar, el clima, variaciones en flujo de rayos cósmicos, entre otras causas. Las características y aptitudes de las especies influyen en su respuesta a las variaciones de las condiciones en las que viven y dado que en los últimos años varios analistas han tratado de apoyar la idea de la evolución mediante el registro fósil se considerará esta variable en el modelo postulado. Se construyó un simulador de tiempos de vida de especies con el objetivo de investigar la plausibilidad de las variables consideradas influyentes en la extinción y la estimación de los parámetros del modelo de regresión de Cox, considerando covariables dependientes del tiempo y habilitando con ello estimación de precisión estadística. El simulador mostró que para la magnitud y estructura de datos contenidos en el registro fósil, es factible realizar estimación sensata de parámetros, dando lugar a interpretaciones en el contexto de extinción. En particular, se estableció compatibilidad con la noción de que la evolución de los organismos mejora su aptitud a lo largo de los años, y se permite analizar la influencia en la extinción de cada uno de los factores identificados.

Índice general

Índice de figuras

Capítulo 1

Extinción en el planeta Tierra

1.1. La extinción e hipótesis referente a una sexta extinción masiva actual

Un evento de extinción masiva es una disminución rápida y generalizada de la biodiversidad de la Tierra. Más del 99 % de todos los organismos que alguna vez han vivido en la Tierra se encuentran extintos. A medida que las nuevas especies se adaptan a nichos ecológicos en constante cambio, las especies más antiguas se desvanecen pero suelen dejar un registro de su existencia en forma de fósiles. Esta tesis versará sobre modelos para comprender dichos registros a lo largo del tiempo geológico, y cómo analizarlos para extraer información válida acerca del fenómeno de extinción.

A lo largo de los últimos 540 millones de años, la biodiversidad global ha cambiado drásticamente y han habido cinco grandes eventos de extinción masiva que aniquilaron a una abrumadora mayoría de especies que vivían en ese momento. Estas cinco extinciones masivas incluyen la extinción masiva del Ordovícico (hace 445 millones de años aproximadamente), la extinción masiva del Devónico (hace 375 MDA aproximadamente), la extinción masiva del Pérmico (hace 252 MDA aproximadamente), la extinción masiva del Triásico-Jurásico (hace 200 MDA aproximadamente) y la extinción masiva del Cretácico-Terciario (hace 66 MDA aproximadamente). Cada uno de estos eventos varió en tamaño y causa, pero todos ellos devastaron sustancialmente la biodiversidad que se encontraba en la Tierra en su época. Las extinciones masivas liberan recursos para que surjan nuevas formas de vida. La extinción más estudiada, que marcó el límite entre los períodos Cretácico y Paleógeno hace unos 66 millones de años, acabó con los dinosaurios y dejó espacio para que los mamíferos y las aves se diversificaran y evolucionaran rápidamente (Dirzo and Raven, 2003). Un asunto muy pertinente de discusión en la literatura científica de paleobiología es cómo es que se arriba a estas conclusiones mediante análisis de datos, especialmente a la luz de que los registros fósiles están llenos de irregularidades y muchas fuentes de incertidumbre. Tomando en cuenta las causas y eventos reconocidos de extinción, en esta tesis se pretende realizar un análisis matemático formal que valide esta información y proporcione las herramientas necesarias para realizar inferencias formales sobre la extinción.

En la siguiente figura se muestran las eras geológicas ya que son el marco de referencia para representar los eventos de la historia de la tierra y de la vida ordenados cronológicamente.

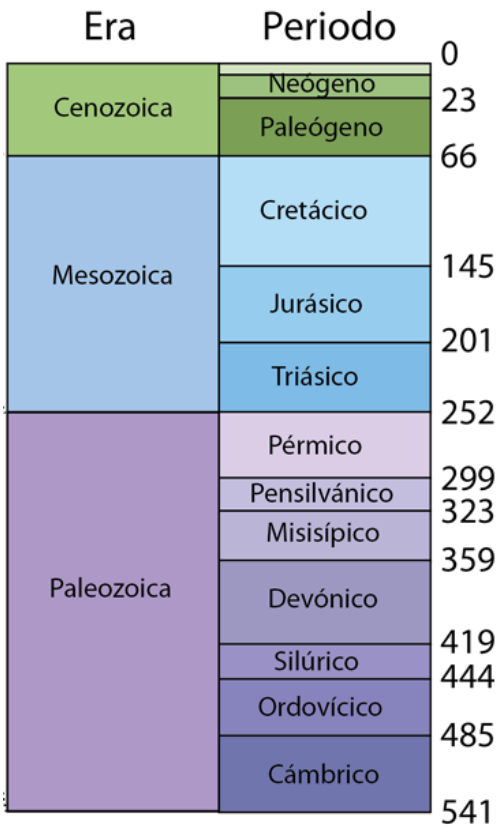

Figura 1.1: Nomenclatura y extensión de eras geológicas en millones de años (MDA) de antigüedad.

Las explicaciones con evidencia más consistente acerca de estas extinciones involucran el impacto de meteoritos contra la tierra, aumento de la actividad volcánica y disminución del nivel medio del mar (Rohde and Muller, 2005). Esta constante sucesión entre extinciones y radiaciones ha dado lugar a un aumento neto de la biodiversidad marina, de tal suerte que en la actualidad existe más variedad de especies que en cualquier otra época pasada. Sin embargo, en los últimos 500 años esta diversidad parece estar desapareciendo a un ritmo comparable con el de los eventos de extinción en masa, y la causa principal se le atribuye a las actividades humanas. Durante los últimos años se ha manejado la hipótesis de que estamos actualmente en camino a una sexta extinción masiva. Para poderlo determinar, será necesario entender primero cómo han sido las tasas naturales de extinción para la historia de la vida en el planeta Tierra, y para luego comparar con las tasas contemporáneas. Esta tarea de determinación de tasas conlleva inevitablemente incertidumbre, y por lo tanto, desde el punto de vista matemático plantea un problema de inferencia estadística formal.

En el planeta, la mayor parte de los fenómenos dejan algún tipo de huella y en el caso de las especies se observa mediante la fosilización. Por lo tanto, los fósiles manifiestan la extinción y con ello, la tasa de extinción. La totalidad de los fósiles identificados y catalogados en los anales de la ciencia se conoce como registro fósil. Un aspecto fundamental para el estudio de los fósiles radica en la estratigrafía, ciencia que estudia, describe e interpreta la disposición de las capas o estratos de rocas y otros materiales en la corteza terrestre (*estratos* refiere a un término empleado para enumerar las capas que puede tener una formación rocosa, como producto de años de sedimentación). Los fósiles marinos

se utilizan principalmente para medir las tasas de extinción debido a su superior registro fósil y rango estratigráfico en comparación con los animales terrestres ya que los animales marinos aparecieron antes que aquellos. El registro fósil, en efecto, revela varios aspectos sobre la vida, pero es crucial reconocer que la ocurrencia de datos que han quedado registrados en el registro fósil no sólo reflejan una cualidad biológica sino que además responden a cierta actividad humana ejercida en su búsqueda e inscripción en una base de datos.

Los paleontólogos recurren al registro para aprender más sobre cómo se formaron las especies de hoy. El descubrimiento de fósiles y sus similitudes con los organismos actuales apoya la idea de una evolución al paso de los años, la cual ha sido ratificada por varios paleontólogos y también se le llama *fitness*. Según Darwin (1859), *fitness*, significa la capacidad de sobrevivir y reproducirse. Se asume que la evolución se ha producido por selección natural, y por lo tanto las estructuras y los comportamientos deben interpretarse en términos de la contribución que hacen a la supervivencia y reproducción de sus poseedores, es decir, a la aptitud darwiniana (Smith, 1978). Darwin creía que apoyaría su teoría de la evolución con el registro fósil; sin embargo, descubrió que estaba (y aún está) incompleto. Esto ratifica la noción de que los datos del registro fósil inevitablemente contienen incertidumbre.

1.2. Motivación y perspectiva de la tesis

Resumiendo, cuando se trata de estudiar la historia y la evolución de la vida, el registro fósil representa una fuente de datos de vital importancia. Es de interés establecer si el cambio de la diversidad que se ha notado en los últimos 500 años se debe a la actividad humana o si es parte de una variación natural. Para esta comparación será necesario analizar el comportamiento de la extinción durante los últimos 540 millones de años, antes de que la especie humana apareciera en la Tierra. Este análisis se complica debido a que existen factores de confusión—los cuales serán detallados posteriormente—en el registro fósil que conllevan a que no se manifieste de manera transparente el proceso de extinción. En esta tesis se reconocerá de entrada que el problema es uno de inferencia estadística, para lo cual es necesario postular modelos probabilísticos que expliquen la variabilidad contenida en el registro fósil como función del información biológica de interés, y tomando en cuenta modelos y métodos que aborden ciertas características distintivas que hay en datos del registro fósil. Con el desarrollo que se propone, se abre el camino para abordar la hipótesis de la la sexta extinción masiva haciendo posible realizar posteriormente una comparación pasado-presente de la extinción marina.

Un ejemplo de característica distintiva es el concepto de dato censurado, pues existen especies en el registro fósil que aun se encuentran vivas. Otros datos son extinciones legítimas, en el sentido de que se sabe que la especie hoy no existe y cuándo aproximadamente se dejaron de producir sus restos fósiles. Por otra parte, se presume que el fenómeno de extinción se asocia con ciertas condiciones imperantes a lo largo de la eras geológicas. Desde el punto de vista matemático, esto significa la noción de covariables, que deben relacionarse con alguna característica indicativa de extinción. La consideración de métodos estadísticos tendrá la importante virtud de poder cuantificar la incertidumbre de inferencias realizadas a partir de datos del registro fósil. Se van a delinear algunos elementos que se deben tomar en cuenta para poder analizar, en un trabajo futuro, el cambio de la diversidad en la actualidad.

En esta tesis, la metodología general a la que se recurrirá será análisis de supervivencia en los últimos 540 millones de años tomando en cuenta las explicaciones más consistentes acerca de las extinciones que han existido, es decir, el cambio en el nivel medio del mar, el clima, los rayos cósmicos, *etc.* Análisis de supervivencia es un enfoque natural, debido a que se reconoce e involucra explícitamente

la censura, y se destina a estudiar los tiempos de longevidad de las especies, concebida ésta como indicio del fenómeno de extinción. Dentro de esta disciplina de supervivencia, el modelo de regresión de Cox está concebido para relacionar la distribución de longevidad como función de covariables. Habilita visión probabilística de datos, estimación de parámetros y cuantificación de incertidumbre, no siendo así en los métodos empleados con anterioridad. Nos basaremos en los datos del Compendio Sepkoski (Sepkoski, 2002), postulando la manera en que participan covariables y parámetros para la especificación de longevidad. Se creará un simulador de tiempos de vida tratando de reproducir la realidad observada con el objetivo de examinar la plausibilidad de la incorporación de covariables en modelo. Como veremos, este modelo contrasta con enfoques anteriormente utilizados para el análisis de la extinción ya que no está encaminado a cuantificar la riqueza existente (el número total de especies vivas en un tiempo geológico determinado), sino que recurre a la caractización de la distribución de longevidad en función de condiciones ambientales a lo largo de la vida de las especies.

1.3. Métodos de la literatura de paleobiología

Esta sección tiene por objeto abordar algunos de los estudios más importantes realizados en el tema de la extinción y para ello es necesario caracterizar los tipos datos disponibles en el registro fósil para poder comprender las diferentes ópticas de los analistas según el tipo de datos que utilizan. Además, es importante remarcar algunos de los principales factores de confusión que se presentan en estos análisis ya que son instrínsecos a los datos.

Una de las bases de registros fósiles más socorrida a nivel mundial es el Compendio Sepkoski, nombrada en honor a su creador, Jack Sepkoski. Esta base está constituida 36 339 géneros de organismos marinos provenientes de distintas partes del mundo que datan desde hace 540 millones de años hasta la era actual. En este registro cada caso representa a un género de animales marinos y contiene la información de la fecha de primera aparición (FA) y última aparición (LA) de un fósil correspondiente a cada género (Tapanila, 2007). Esta estructura nos denota que los datos exhiben censura por ambos lados ya que en realidad no sabemos la verdadera fecha de inicio y fin. Además de la censura por la cuantificación incierta de edades, hay datos censuradas por la derecha ya que existen todavía en la actualidad. Este compendio no contiene información detallada sobre el lugar donde fueron muestreados los fósiles y contiene fósiles observados en una sola ocasión, llamados especies *singletons*, lo cual muestra realidades importantes que privan en el registro. Sin embargo, tiene características deseables por los analistas ya que tiene una estructura sencilla, es un resumen con acceso público sobre la historia de la biodiversidad global y tiene manual de usuario.

Existe una segunda importante fuente de datos, ubicada en `http://fossilworks.org`. *Fossilworks*, creado por John Alroy, proporciona herramientas de consulta, descarga y análisis que utilizan la gran base de datos relacional de *Paleobiology Database* reunida por cientos de paleontólogos de todo el mundo. Incluye una clasificación taxonómica maestra integrada dinámicamente y registros de distribución específicos del sitio. Además, le agrega una escala de tiempo geológico global sintetizada algorítmicamente que habilita sus herramientas de consulta y análisis. El conjunto de datos cubre todas las partes del registro fósil, lo que significa que documenta animales, plantas y microfósiles marinos y terrestres de todas las edades geológicas. Mediante este sitio podemos tener acceso a las ocurrencias taxonómicas, clasificaciones y edades de cada fósil. Permite un involucramiento de información más completa radicada en colecciones a nivel local y por ello una mejor estimación de la riqueza existente. Sin embargo, se aprecian importantes desventajas en esta base de datos ya que no está disponible públicamente en su totalidad, por sus características es muy heterogénea y los datos no están plena-

mente *curados* (término que significa depurados y revisados con plena consideración de su contexto). El proceso de curación es intrincado y de gran dedicación. Además, implementar métodos con estos datos requiere mucho tiempo y da lugar a un desafío de programación; inclusive se ha abordado este tema en varios artículos para intentar facilitar el trabajo (Kocsis et al., 2019). Por otra parte, se tienen unidades de esfuerzo no estándar al incluir numerosos datos de distintas regiones y muestreados de manera totalmente diferente.

Varios efectos de confusión en estos datos provocan que no se manifieste de manera transparente el proceso de extinción en el registro fósil, sino que se encuentra confundida por diversas fuentes de ruido. Estos aspectos provocan una sobrevaloración de la información que contiene el registro fósil y los llamamos sesgos de muestreo. A continuación se enlistan algunos de estos sesgos que se presentan en el registro y nos permitirá tomar en cuenta esta información para el análisis del modelo propuesto en esta tesis.

- Hay filtros taxonómicos o sesgos, relacionados con qué tipos de organismos tienen mayor probabilidad de que se conserven. Es mucho menos probable que se conserven los organismos con cuerpos blandos que los que tienen huesos o conchas. Incluso para organismos con partes duras, las condiciones del lugar de la muerte deben ser las adecuadas para su preservación y mineralización (Benton, 2009).
- El registro fósil está incompleto, pues sólo una pequeña fracción de los individuos son susceptibles de ser fosilizados; de ellos, muy pocos se recopilan e identifican. Por ello, la diversidad biológica que se encuentra registrada en el registro fósil es menor que la diversidad total existente, y los rangos estratigráficos observados de los taxones son más cortos que sus rangos reales. Esto se debe a que la primera aparición de un taxón en el registro fósil ocurre en algún momento después de que realmente se originó, a menos que su primer representante fuera fosilizado y seguidamente muestreado (hecho que es muy poco probable que ocurra). Con esta misma idea, la última aparición de un taxón ocurre en algún momento antes de que realmente se extinga. Las estimaciones con sesgos de los rangos estratigráficos son importantes para el estudio de las tendencias en la biodiversidad porque dichos rangos se han utilizado tradicionalmente para estimar la biodiversidad existente (Bokulich, 2018).
- *Signor Lipps effect* se relaciona con la incompletez ya que la curva de diversidad caerá suavemente a medida que se acerque a una extinción masiva y asimismo, nuevos taxones que aparecen durante la recuperación de una extinción masiva puede tomar algún tiempo para ser muestreados, por lo que una curva se elevará suavemente incluso si la recuperación fue extremadamente rápida (Alroy, 2010a).
- *Pull of the Recent* es un tipo de sesgo que contempla que lo más cercano a la actualidad está mucho mejor muestreado que cualquier otro intervalo en el registro geológico. Este sesgo hace que la curva de la diversidad se eleve a medida que se acerca a la actualidad (Alroy, 2010a).
- El esfuerzo de muestreo es uno de los sesgos más relevantes y difícil de incorporar a los modelos, ya que cuantificarlo es un problema complejo en el estudio de la extinción. Sesgos en la intensidad del muestreo pueden dar lugar a sesgos geográficos ya que la mayoría de los fósiles de hoy se han recolectado en Europa y América del Norte, mientras que otras partes del mundo no están tan bien exploradas (Bokulich, 2018).

Dado que la extinción juega un papel importante en la evolución de la vida, durante muchos años se ha estudiado y procurado llegar a conclusiones que permitan analizar el cambio de la biodiversidad

actual. Algunos estudios se han realizado con datos del Compendio Sepkoski y otros con colecciones y ejemplares extraídas de Fossilworks. Un estudio de los principales análisis realizados nos permitirá obtener una visión panorámica de las fortalezas y debilidades que contienen para poder cambiar o incorporar ideas al modelo desarrollado en esta tesis.

Uno de los grandes analistas del compendio de registros fósiles fue David M. Raup. Ha realizado varios análisis en cuanto a la extinción, uno de los más usados, incluso por otros analistas, es el análisis de Cohort. Raup (1986) introduce términos como *cohort* y *survivorship* en los que se basa para desarrollar un análisis de la extinción. Por *cohort* refiere a un grupo existente en un tiempo geológico definido que se monitorea a través del tiempo geológico para mostrar la desintegración del grupo por extinción de sus especies constituyentes. El término *survivorship* lo utiliza al monitorear cada cohort para ver cuántas de las especies se mantienen vivas. Ajusta una curva utilizando los porcientos de elementos en el grupo que quedan a través del tiempo geológico y a dicha curva la llama *survivorship*. Con estas bases realizan deducciones pero es notable el poco formalismo matemático en el análisis, lo cuál atenta contra la veracidad de los resultados. Sin embargo, obtuvo resultados certeros ya que por ejemplo, con el estudio de cohort remarca las diferencias en las tasas de extinción entre taxones y además concuerda con las extinciones masivas reconocidas. Los estudios de Raup muestran que a pesar de las debilidades del Compendio Sepkoski se pueden obtener buenos resultados con su análisis, lo que nos permitirá desarrollar el modelo propuesto con los datos del Compendio.

Por otra parte, John Alroy, biólogo y paleontólogo creador de Fossilworks, estudia la diversidad y la extinción a través de colecciones y ejemplares. En Alroy (2010a), se abordan métodos para amortiguar los sesgos de muestreo y muestra cómo la elección de un método de conteo resulta ser potencialmente importante. Por método de conteo (o *de facto*, método de selección de subconjuntos) se entiende que, dado el compendio y una fecha geológica dada, T, deben elegirse sólo ciertos fósiles para representar el estado de la biodiversidad en el tiempo T. En general se puede entender que los fósiles seleccionados serán los que figuren en una vecindad de T, pero existen sutilezas respecto a cuáles exactamente deben constituir el subconjunto. Por el momento solo se tomarán en cuenta las ideas pertinentes desarrolladas por Alroy que nos aporte información interesante a nuestro modelo dada la dificultad que conlleva trabajar con este tipo de datos y quedará para un trabajo futuro, contrastar lo obtenido con el Compendio Sepkoski y realizar análisis más profundos.

Uno de los análisis más intuitivos de la extinción recurre a ir monitoreando en cada tiempo geológico, la biodiversidad existente. Como anteriormente mencionamos, en Alroy (2010a) se introducen varios métodos de conteo para realizar una descripción global de los datos. Usualmente se ha utilizado los conteos en BIN, abordados por Alroy, para los datos de ejemplares ya que la ocurrencia de los datos permiten métodos de conteo alternativos, pero vimos que los métodos realizados con datos de ejemplares pueden tener muchas desventajas. A continuación se realizará un breve resumen de estos métodos con el objetivo de mostrar algunas deficiencias que contienen y de cierto modo, justificar la elección de un método de conteo diferente para realizar un análisis descriptivo de los datos en el próximo capítulo. Todos los métodos de conteo mostrados en Alroy (2010a) involucran cinco categorías distintas que se pueden separar usando rangos de edad u ocurrencias:

1. Encontrado antes y después de un intervalo de tiempo, es decir, taxones que abarcan un intervalo y se muestrean dentro de él (N_{r^+}).

2. Taxones que abarcan un intervalo pero no son muestreados dentro del intervalo, es decir, taxones de Lázaro (un taxón que desaparece durante uno o más períodos del registro fósil, y luego vuelve a aparecer) (N_{r^-}).

3. Cruzando la parte inferior del intervalo y se extinguen, es decir, taxones que cruzan el límite solo desde la parte inferior (N_b).
4. Originado dentro de un intervalo y cruzando su límite superior, es decir, taxones primero muestreados en el intervalo y muestreado en cualquier lugar después de él (N_t).
5. Originándose y extinguiéndose inmediatamente, es decir, taxones de intervalo único (N_1)

Los métodos de conteo mencionados en Alroy (2010a) se resumen a continuación:

- RT (*range through*) es el método más común y en donde se cuenta todo lo que está en cualquier lugar del intervalo:

$$N_{r^+} + N_{r^-} + N_b + N_t + N_1.$$

- BC (*boundary crossers*) donde se cuenta todo lo que esta antes y después del inicio del intervalo:

$$N_{r^+} + N_{r^-} + N_b = N_r - N_t - N_1,$$

donde:

N_r: taxones que se identifican en cualquier lugar dentro o a través de un intervalo

- SIB (*sampled in bin*), el más intuitivo de los tres métodos de conteo mediante el cual, cada intervalo contiene todos los taxones muestreados:

$$N_{r^+} + N_b + N_t + N_1 = N_r - N_{r^-}.$$

La diferencia entre RT y SIB radica en que SIB contiene los muestreados, o sea, no contiene a los taxones Lázaro. Se reconocen así a los taxones que desaparecen durante uno o más períodos del registro fósil y después vuelven a aparecer. Las curvas de los métodos RT y BC muestran una desventaja importante ya que caen en sus bordes porque hay muchas oportunidades para muestrear taxones en el medio, pero pocos al principio o al final. La idea del método de conteo propuesto en esta tesis, se fundamenta con ideas muy similares al método SIB pero tiene una estructura más sencilla.

Más adelante, recurriremos a un método para realizar análisis exploratorios globales del compendio de Sepkoski. Este método consiste en contar en una fecha geológica dada, T, los fósiles que están presentes, para así, representar mediante subconjuntos elegidos T x T el estado de la biodiversidad en el tiempo T. Sin embargo, veremos que en la postulación de un modelo de regresión de Cox, no se recurrirá al concepto de subconjunto para describir lo que ocurre en el tiempo T sino que se establecerá que un fósil en el tiempo T estuvo expuesto a covariables propias de ese tiempo, digamos $X(T)$. Esto es, como es usual en un modelo de regresión, para inferir la distribución de supervivencia en un tiempo T no sólo se recurre a datos en una vecindad de T sino que se recurre a todos los individuos, y todos ellos aportan información sobre T a través de una estructura de regresión.

Se ha mencionado también que el esfuerzo de muestreo es un factor importante de confusión cuando se trata de discernir aspectos de extinción a partir de datos registrados en el registro fósil. Este hecho ha sido reconocido en la literatura de paleobiología, y varios trabajos han abordado métodos para hacer correcciones por sesgo de muestreo. Sin embargo, cabe mencionar que estas propuestas están enfocadas a la estimación de riqueza, no a la distribución de longevidad. A continuación se muestra un breve resumen los principales métodos abordados en la literatura de paleobiología para corregir o mejorar las estimaciones de riqueza con base en el registro fósil, tomando en cuenta los sesgos. Este

resumen permitirá mostrar la dificultad que conlleva tratar el sesgo del esfuerzo de muestreo ya que son métodos no triviales y aún así, no eliminan este tipo de sesgo. Estos métodos se dividen principalmente en las siguientes técnicas:

- Submuestreo estandarizado: La idea de este método es corregir sesgos de intensidad de muestreo según varios autores. Se realiza un submuestreo aleatorizado de registros de datos, como ocurrencias taxonómicas individuales o colecciones de fósiles.
- Curvas de acumulación: son curvas que muestran la acumulación de fósiles a lo largo del tiempo histórico. Las curvas de acumulación a menudo contrastan el número de años históricos, años de carrera, publicaciones u observaciones con el número acumulado de taxones encontrados (Alroy, 2010a).
- Rarefacción y extrapolación: el propósito es concebir comparaciones justas entre muestras incompletas. Aunque la rarefacción y extrapolación tradicional basada en el tamaño, en la que las muestras están todas estandarizadas para igual tamaño, proporciona información de muestreo útil, han argumentado que a menudo es más informativo estandarizarlos para que tengan una cobertura igual. Cobertura refiere la proporción del número total de individuos en una comunidad que pertenecen a las especies representadas en la muestra. Se busca comparar muestras de igual calidad e integridad para obtener análisis robustos e inferencias detalladas sobre las comunidades muestreadas (Chao and Jost, 2012).

Con todos los aspectos abordados anteriormente, hemos establecido las bases conceptuales para en el siguiente capítulo presentar la propuesta del modelo elaborado durante desarrollo de esta tesis.

Capítulo 2

Propuesta de modelación basada en regresión de Cox

Al analizar datos de supervivencia, hay dos funciones de interés central, la función de supervivencia y la función de riesgo. Por lo tanto, estas funciones y su relación se definen en la primera parte de este capítulo. En una segunda parte se realizará una descripción gráfica de los datos con fines revelar características relevantes para incorporar al modelo postulado. Luego se construirá un simulador de tiempos de vida de especies con el objetivo de investigar la detección de las variables consideradas influyentes en la extinción y la estimación de los parámetros del modelo de regresión de Cox. Se considerarán algunos de los factores influyentes en la extinción, identificados en la sección de la descripción de los datos y otros identificados en la literatura de paleobiología. Finalmente, se analiza la influencia en la extinción de cada uno de los factores identificados a través del ajuste del modelo de Cox propuesto.

2.1. Conceptos básicos de análisis de supervivencia

En esta sección se resumen conceptos básicos para el análisis de supervivencia con la finalidad de comprender la idea general de nuestro modelo de regresión de Cox posteriormente. El contenido puede ser consultado con mayor lujo de detalle en Collett (2015).

2.1.1. Función de supervivencia

El tiempo de supervivencia de un individuo, t, puede considerarse como el valor de una variable T, que puede tomar cualquier valor no negativo. Llamamos a T la variable aleatoria asociada con el tiempo de supervivencia. Ahora suponemos que la variable aleatoria T tiene una distribución de probabilidad con función de densidad de probabilidad subyacente $f(t)$. La función de distribución de T viene dada por

$$F(t) = P(T < t) = \int_0^t f(u)\, du,$$

y representa la probabilidad de que el tiempo de supervivencia sea menor que algún valor t.

La función de supervivencia, $S(t)$, se define como la probabilidad de que el tiempo de supervivencia sea mayor o igual que t, por lo que

$$S(t) = P(T \geq t) = 1 - F(t).$$

Por tanto, la función de supervivencia se puede utilizar para representar la probabilidad de que un individuo sobrevive desde el origen temporal hasta algún tiempo después de t. En varias referencias que abordan temas de supervivencia se define $S(t) = P(T > t)$ pero por convención, utilizaremos la definición de supervivencia abordada en Collett (2015).

2.1.2. Función de riesgo

La función de riesgo (*función hazard*) es la probabilidad de que un individuo muera en el momento t, condicionado a que haya sobrevivido hasta ese momento. Por tanto, la función de riesgo representa la tasa de muerte instantánea de un individuo que sobrevive hasta el tiempo t. Para obtener una definición formal de la función de riesgo, consideramos la probabilidad que la variable aleatoria asociada con el tiempo de supervivencia de un individuo, T, se encuentra entre t y $t + \delta t$, con la condición de que T sea mayor o igual que t, es decir $P(t \leq T < t + \delta t \mid T \geq t)$. La función de riesgo $h(t)$ es entonces el valor límite de esta probabilidad dividida por el intervalo de tiempo δt, ya que δt tiende a cero, de modo que

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t \mid T \geq t)}{\delta t} \right\}.$$

2.1.3. Relaciones entre función de riesgo y supervivencia

Existen relaciones útiles entre la función de supervivencia y la función hazard. Notemos que

$$P(t \leq T < t + \delta t \mid T \geq t) = \frac{P(t \leq T < t + \delta t)}{P(T \geq t)}$$

y además

$$\frac{P(t \leq T < t + \delta t)}{P(T \geq t)} = \frac{F(t + \delta t) - F(t)}{1 - F(t)}.$$

Tomando en cuenta la definición de $h(t)$, entonces

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} \frac{1}{S(t)},$$

donde

$$\lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\}$$

es la definición de derivada de $F(t)$ con respecto a t. Con ello obtenemos que

$$h(t) = \frac{f(t)}{S(t)}.$$

Por otra parte, $f(t) = -dS(t)/dt$, de donde a partir de la definición de riesgo tenemos que

$$h(t) = -\frac{d \log(S(t))}{dt},$$

o equivalentemente,

$$S(t) = \exp\{-H(t)\},$$

donde

$$H(t) = \int_0^t h(u)\, du.$$

2.1.4. Estimador Kaplan-Meier

El estimador de Kaplan-Meier es un estimador no paramétrico de la función de supervivencia. Hemos ya señalado que en datos de supervivencia, así como en los datos del registro fósil, está presente por contexto el concepto de censura. Para determinar dicha estimación de la función de supervivencia a partir de una muestra de datos de supervivencia censurados, suponemos que tenemos n individuos con tiempos de supervivencia observados $t_1, t_2, \ldots, t_n$ y algunos de estos tiempos estarán censurados. Por lo tanto tenemos r tiempos de muerte con $r \leq n$ y se ordenan ascendentemente, $t_1 < t_2 < \cdots < t_r$. El número de individuos que estaban vivos antes del tiempo t_j, incluyendo aquellos que están a punto de morir en este momento, lo denotamos n_j para $j = 1, 2, \ldots, r$, (en presencia de censura, n_j, es el número de supervivientes menos el número de casos censurados) y d_j el número de individuos que mueren en ese momento. La probabilidad que tiene un individuo de morir durante el intervalo $t_j - \delta$ a t_j, con δ infinitesimal, se estima por d_j/n_j. La probabilidad estimada de supervivencia a través de ese intervalo es entonces $(n_j - d_j)/n_j$. Asumiendo que las muertes de los individuos de la muestra ocurren independientemente una de la otra, la estimación de Kaplan-Meier de la función de supervivencia, viene dada por

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right)$$

$t_k \leq t < t_{k+1}, k = 1, 2, \ldots, r$ con $\hat{S}(t) = 1$ para $t < t_1$ y t_{r+1} se toma como ∞.

2.2. Descripción y representación gráfica con datos del registro fósil

En el compendio Sepkoski tenemos la información de 36 339 géneros distintos pero de ellos, 15 849 solo fueron observadas una vez y se les reconoce como especies *singletons*. Para formarse una idea de las características de estos datos se elaborarán varias gráficas para poder apreciar la información que proporcionan. Este análisis descriptivo será también pertinente para determinar posibles idiosincrasias que ameritan ser incorporadas en las tareas de modelación.

En la Figura 2.1 se muestra la cantidad de fósiles existentes en cada tiempo geológico, a lo cual se le llama riqueza. Las líneas rojas marcan las cinco extinciones masivas reconocidas y se prueba visualmente que los datos del compendio evidencian dichas extinciones. Además, se muestra mucha variación en la riqueza a lo largo del tiempo geológico, y dado que en la vida de las especies un factor importante para su desarrollo y existencia son las condiciones ambientales, sería relevante la incorporación de estas variables a nuestro modelo.

Otro factor importante a considerar lo muestra la vida media de las especies en la Figura 2.2 ya que en los últimos años varios analistas han tratado de apoyar la idea de la evolución mediante el registro fósil. Se evidencia que mientras el tiempo geológico es más cercano a la actualidad, aumenta la vida media de las especies. En esta gráfica, se muestra que en las extinciones masivas (líneas azules) radica un aumento notorio de la vida media de las especies. Este aumento de la vida media en las

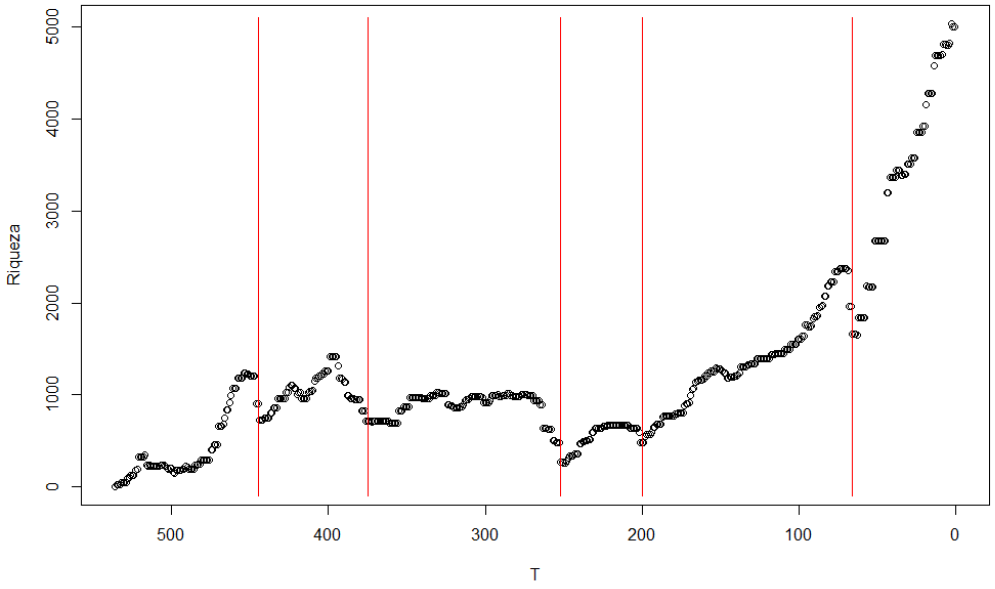

Figura 2.1: Riqueza de géneros de fósiles marinos según el tiempo geológico.

extinciones podría estar justificado con el hecho de que en las extinciones se encuentran menos especies existentes y estas son las más aptas.

A continuación, en la Figura 2.3 se representan los tiempos de vida de todos los géneros que se encuentran en el compendio Sepkoski. Cada segmento se encuentra delimitado por un punto negro al inicio y un punto rojo al final, marcando la primera y última aparición de cada fósil registrado y las estrellas color rosa muestran las especies *singletons*. La altura a la que se encuentra cada línea representa la frecuencia con la que aparece en el compendio el rango graficado. Con esta gráfica se ratifica la riqueza biológica creciente observada en la Figura 2.1 y se aprecia que hace 100 millones de años comienzan a mostrarse más segmentos azules y con una mayor frecuencia. Además, se observan varias especies censuradas ya que según el registro fósil aún no están extintas. De ahí, la importancia de incorporar las censuras al modelo ya que un gran porcentaje de las especies están censuradas.

Utilizaremos la Figura 2.4 principalmente para mostrar nuestro método de selección de fósiles a lo largo del tiempo geológico, aunque también se observa la heterogeneidad de los tiempos de vida de las especies ya que muestra un acercamiento de la Figura 2.3. La línea verde (*cursor*) está ubicada en $T = 335$ y como se puede apreciar, muchos fósiles cruzan a través de ella. Esto es, su primera aparición fue antes de $T = 335$ y su última aparición fue después de $T = 335$. Los fósiles que tienen esta relación respecto a la línea verde se van a tomar como un subconjunto y así va a pasar en cada tiempo geológico con fines de utilizar esta idea para el análisis de la próxima figura. Por otra parte, observando este zoom de la Figura 2.3, vemos que algunas especies viven cientos de millones de años y otras duran menos de cinco millones de años y en el caso de los *singletons* es probable que representen especies que hayan existido poco tiempo y por eso solo se encuentre registrado una sola aparición. Este análisis servirá como punto de partida para nuestro modelo ya que sugiere un modelo que tome en cuenta las heterogeneidades que pueden radicar en el registro y sus posibles causas.

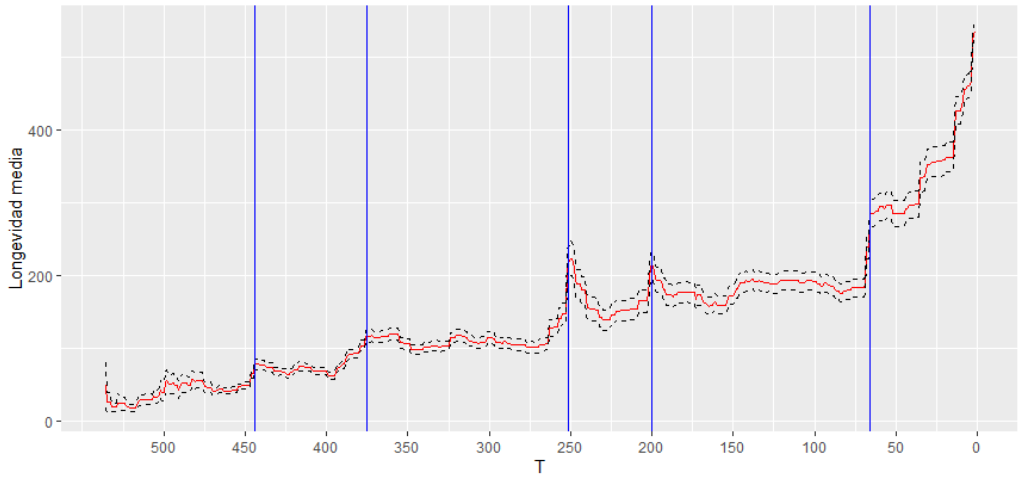

Figura 2.2: Medias de las longevidades.

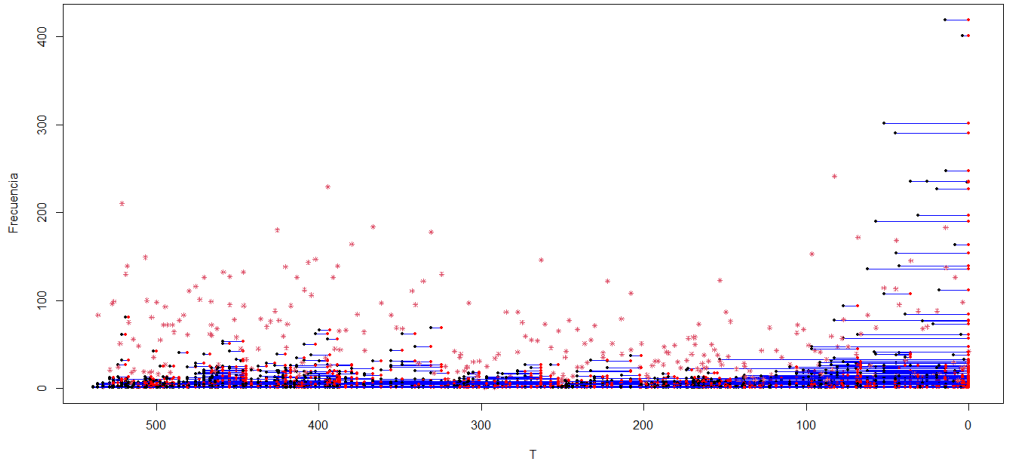

Figura 2.3: Representación del Compendio.

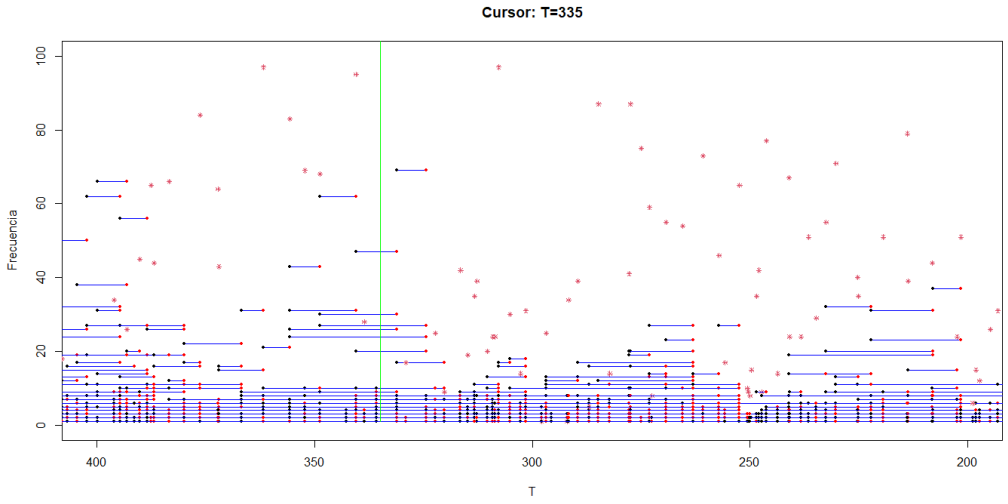

Figura 2.4: Cursor para la selección de los fósiles.

Después de realizar una pequeña discusión de la idea del estimador de la supervivencia de Kaplan-Meier en la sección anterior y abordar la idea del *cursor* para inspeccionar $T \times T$ el estado de la biodiversidad en cada T, sería natural querer revisar cómo va cambiando la supervivencia a lo largo del tiempo geológico. Para ello, se ha creado la Figura 2.5 donde se aprecian dichas supervivencias a lo largo del tiempo geológico a través de una gráfica de contornos. Podemos observar en el eje X el tiempo geológico y en el eje Y longevidades para con los colores de las supervivencias considerar las probabilidades de continuar vivo en el tiempo T teniendo una longevidad t. Notamos que esta figura contiene una fuerte evidencia de una evolución al paso de los años ya que al parecer las especies están más aptas para sobrevivir y por tal razón, al acercarnos a 0 va aumentando la supervivencia en cada T. Esto sugiere que incorporemos está información sobre la evolución a nuestro modelo. Además se corrobora la información de la Figura 2.2 ya que en las extinciones notamos un aumento de la supervivencia.

Con el objetivo de identificar características que deba poseer nuestro modelo, observaremos la información que resalte de los datos mediante estimaciones no paramétricas, al igual que en el caso de las gráficas de contorno con el estimador Kaplan-Meier. En la Figura 2.6 se muestran cuatro gráficas que se construyeron para visualizar mediante el estimador Kernel la densidad de los tiempos de vida en algunos tiempos geológicos. En la mayoría de los casos explorados a través del tiempo geológico se puede apreciar que una gran parte de los datos están concentrados cerca del cero, es decir que las especies no longevas son más comunes que las longevas. Sin embargo, algunas gráficas como la correspondiente a $T = 140$ y $T = 220$ muestran una multimodalidad relevante para analizar. En próximas figuras analizaremos estas multimodalidades que se observan en algunos tiempos geológicos.

En la Figura 2.7 se muestra únicamente la densidad estimada para los tiempos de vida de las especies en $T = 450$ para posteriormente, en la Figura 2.8 seleccionar cinco grandes grupos de taxones reconocidos en Alroy (2010b) y examinar el comportamiento de las densidades de las longevidades en cada grupo. Además de estos cinco grupos, en otra gráfica se muestra la densidad estimada para el resto de los fósiles que no son considerados en estos grupos.

Luego de considerar estas figuras, observamos que a pesar de que el grupo de los *bilvalvia* y

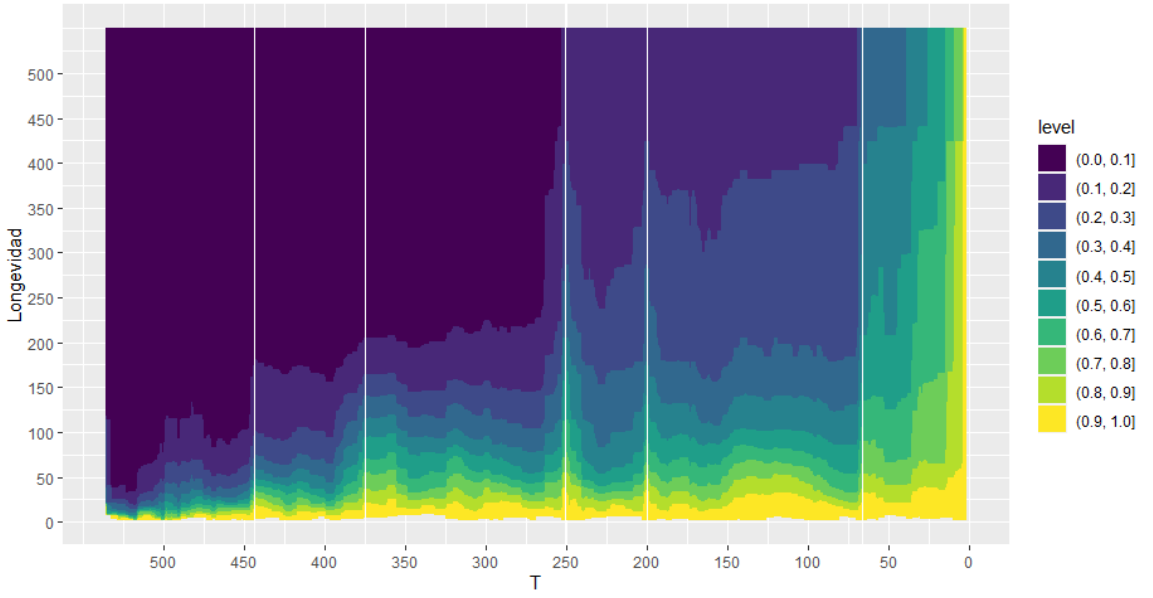

Figura 2.5: Contornos con funciones de supervivencias Kaplan-Meier.

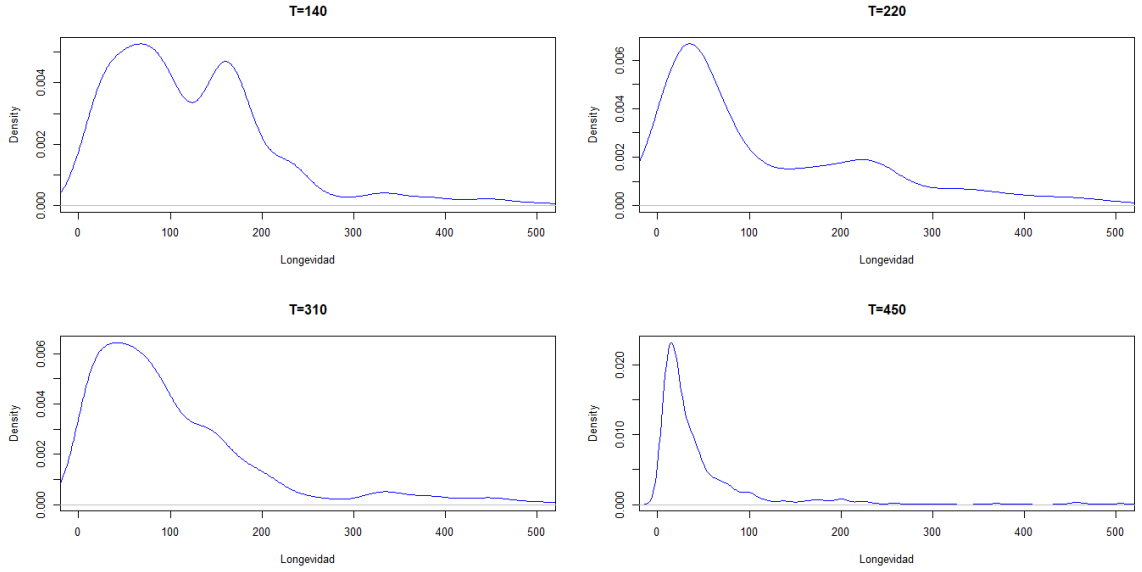

Figura 2.6: Densidad de los tiempos de vida para todas las especies existentes en varios tiempos geológicos.

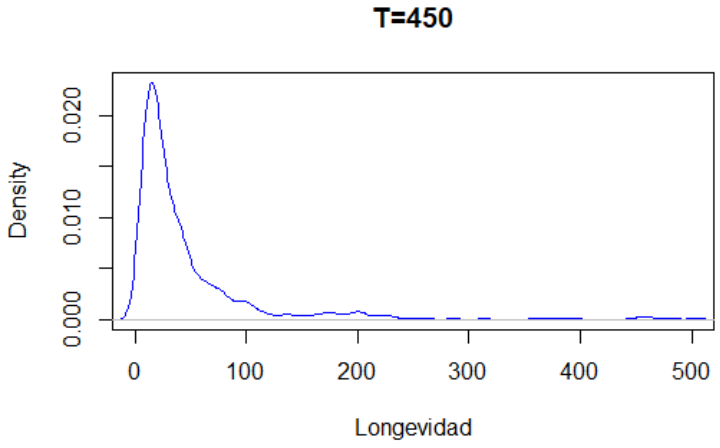

Figura 2.7: Densidad de los tiempos de vida para todas las especies existentes en $T = 450$.

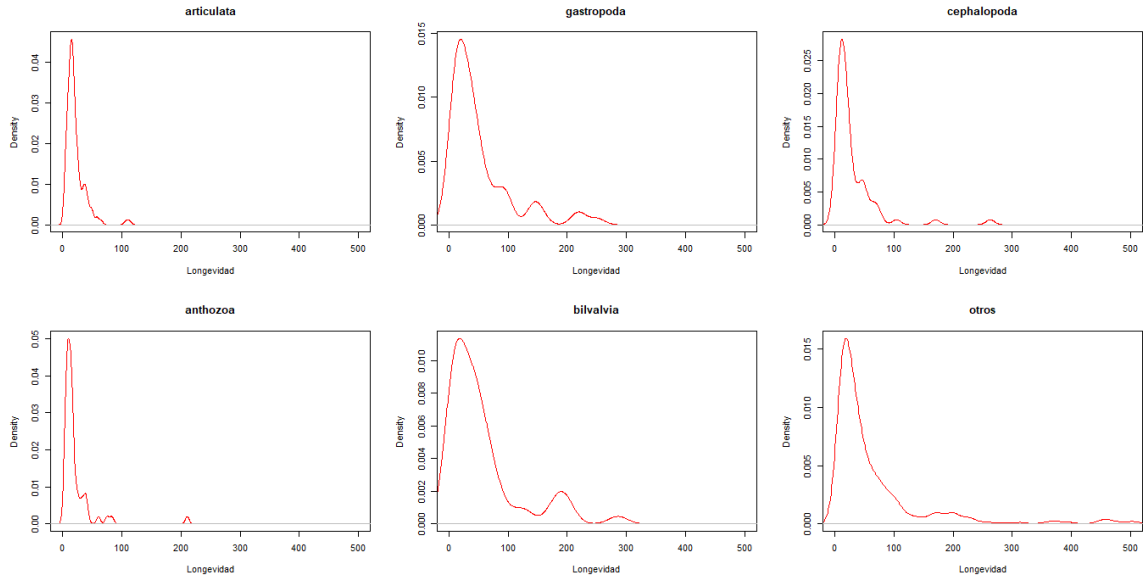

Figura 2.8: Densidad de los tiempos de vida para las especies existentes según distintos filum en $T = 450$.

gastropoda tienen una tendencia a presentar multimodalidad, en la Figura 2.7 de la densidad sin considerar los grupos, no se muestra considerablemente marcada la bimodalidad.

A continuación, en la figura 2.9, realizaremos un análisis similar al anterior pero considerando el tiempo geológico $T = 140$ y posteriormente los las densidades de los grupos considerados por Alroy (2010b).

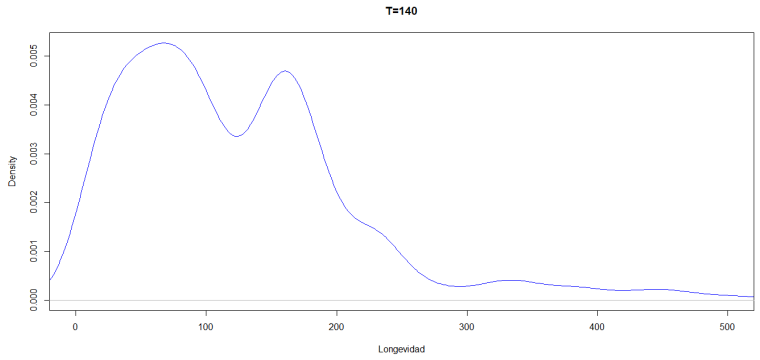

Figura 2.9: Densidad de los tiempos de vida para todas las especies existentes en $T = 140$.

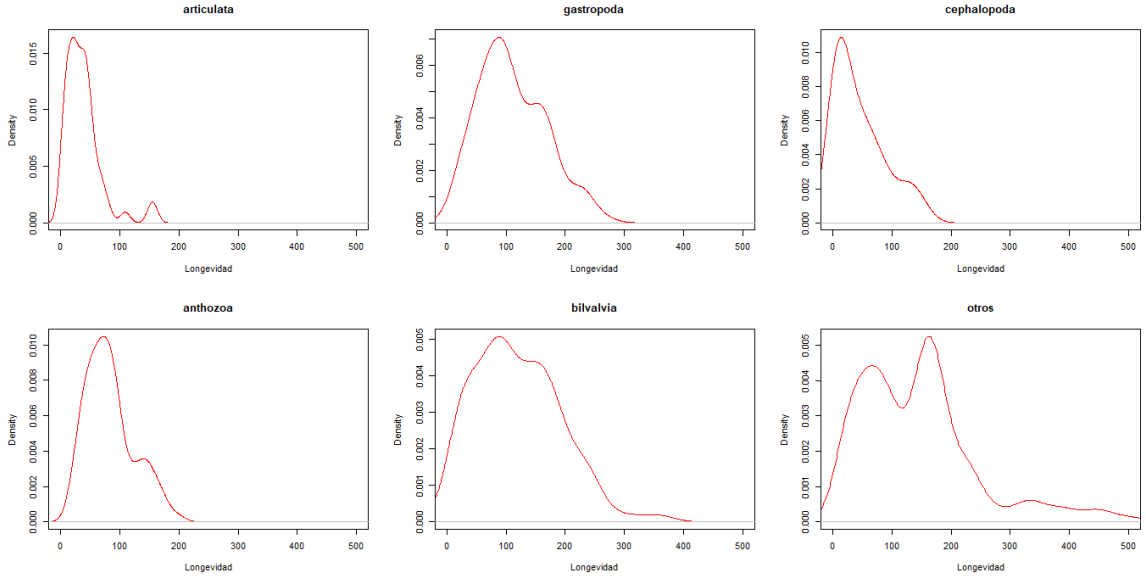

Figura 2.10: Densidad de los tiempos de vida para las especies existentes según distintos filum en $T = 140$.

En este caso, en ninguno de los cinco grupos se muestra tan marcada la bimodalidad como se observa en la Figura 2.9 pero en la gráfica de los restantes taxones (*otros*) sí se aprecia notoriamente esta bimodalidad. Este análisis realizado $T \times T$ con algunos tiempos geológicos nos sugiere que otra variable importante a considerar son los grupos de taxones ya que vimos comportamientos muy variados en cada uno y al considerarlos todos juntos estaríamos desechando la información propia que contiene cada taxón.

Por último, en la Figura 2.11, se muestran gráficas que se construyeron para visualizar la función de riesgo derivada a partir de la función de supervivencia (Figura 2.5) para una selección de tiempos geológicos. Se obtuvieron aproximando la derivada de la función $-\log(S)$. Podemos observar desde otra perspectiva—la función hazard—el comportamiento variado de la función de riesgo según el tiempo geológico, lo que muestra nuevamente la pertinencia de incorporar la mayor cantidad de información relevante a nuestro modelo.

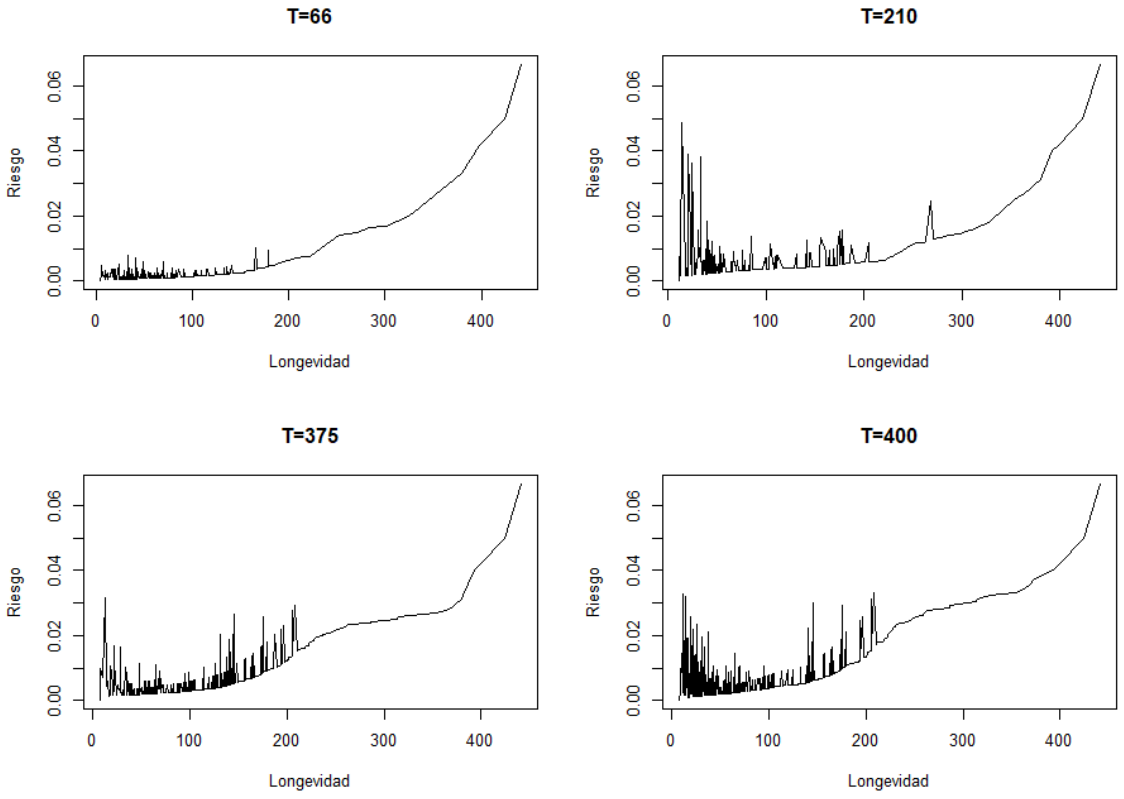

Figura 2.11: Función de riesgo de los tiempos de vida para todas las especies existentes en varios tiempos geológicos.

2.3. Introducción a las covariables

El comportamiento de la tasa de extinción está relacionado con factores que cambien las condiciones del planeta. Un cambio en las condiciones geológicas, atmosféricas, químicas, *etc.*, puede ocasionar cambios en la extinción en relación con la magnitud con la que se produjeron los cambios. En la sección anterior se mostró que existen varias variables que proporcionarían información importante a nuestro modelo. Dichas variables que incorporaremos a nuestros modelo les llamaremos *covariables*. Varios autores han identificado factores importantes que han influido en la extinción y trataremos de incorporarlos a fines de dotar a nuestro modelo con la mayor cantidad de información relevante posible. Dado que existe una relación inherente entre estas variables y la extinción, resulta necesario explorar la naturaleza de esta relación y para ello lo más conveniente es realizar un análisis de regresión.

Durante décadas, los paleobiólogos han estado encontrando ciclos y patrones a gran escala en los registros fósiles. En Rohde and Muller (2005) se muestra un ciclo de 62 millones de años y otro de 140 millones de años en la biodiversidad fósil durante el Fanerozoico, los últimos 540 millones de años de vida en la Tierra. La presencia del ciclo de 140 millones de años se le atribuye a diversos factores como la glaciación, el clima y variaciones en flujo de rayos cósmicos. El ciclo de 62 millones de años lo relacionan con los cambios del nivel del mar ya que estados extremos de nivel del mar corresponden a algunos extremos en este ciclo, pero no de forma consistente. No han encontrado una coincidencia convincente para el ciclo de 62 millones de años a pesar de ser un ciclo evidentemente presente; sin embargo plantean que los registros incompletos y los errores en las escalas de tiempo pueden oscurecer la verdadera periodicidad.

A continuación observaremos la Figura 2.12 extraída de Rohde and Muller (2005) donde muestran la curva de diversidad del Compendio Sepkoski (gráfica **a**) y muestran los ciclos encontrados. En la gráfica en negro en **b**, se muestran los mismos datos vistos en **a**, con una sola ocurrencia y los datos mal fechados eliminados. La línea de tendencia (azul) es un polinomio de tercer orden ajustado a los datos. En **c** muestran los datos con la tendencia azul restada y superponen una onda sinusoidal de 62 millones de años. En **d**, se muestran los datos después de restar el ciclo y con una onda sinusoidal de 140 millones de años superpuesta. Las líneas verticales discontinuas indican los tiempos de las cinco extinciones. Rohde y Muller realizan un análisis espectral para demostrar que estos altibajos oscilan mucho más rítmicamente de lo que se esperaría si fuera un hecho casual (gráfica **e**).

Por lo anteriormente visto, en nuestro modelo consideraremos dos covariables asociadas a estos ciclos ya que con ellas se estarían tomando en cuenta todos los cambios en las condiciones en la tierra que dieron lugar a dichos ciclos. Se considerarán dos sinusoidales, una con un período $\mathcal{T} = 140$ y otra con un período $\mathcal{T} = 62$.

Por otra parte, una de las explicaciones más consistentes en cuanto a la extinción radica en la disminución del nivel medio del nivel del mar (Macleod, 2004). Consideraremos la curva de la variación del nivel del mar a lo largo de los últimos 600 millones de años que se conoce como curva de Exxon (Harrison, 2002).

La curva de Exxon vista en la Figura 2.13 contiene los puntos espaciados cada 0.1 millón de años y en nuestro modelo utilizaremos la Figura 2.14 que sería la tendencia del nivel del mar, idea similar a la utilizada por Rohde y Muller.

Como vimos en la sección anterior, será importante incorporar información sobre la evolución a nuestro modelo. Dado que no se tiene una función que describa la evolución ya que en la mayoría de los estudios solo se aborda de forma cualitativa el problema, se propondrá una función en correspondencia con las características de la evolución. Aunque no sea exactamente la función que mejor se ajusta al proceso de la evolución, permitirá al menos, incorporar una tendencia de aumento en cuanto a la adaptación de los organismos al ambiente a lo largo del tiempo geológico y en un futuro en caso de darse a conocer una mejor aproximación de dicha función, ya se podrán mejorar los resultados.

Para el desarrollo del modelo propuesto debemos distinguir dos tipos de covariables: fijas y temporales. Las covariables fijas se definen como variables cuyos valores no varían a lo largo del tiempo mientras que las covariables temporales son variables cuyos valores varían a lo largo del tiempo. El nivel del mar, las sinusoidales y *fitness* son covariables dependientes del tiempo ya que a lo largo de la vida de una especie van a ir tomando diferentes valores. Además, se analizó en la sección anterior otra covariable importante a tomar en cuenta, los grupos de taxones considerados en Alroy (2010b) ya que

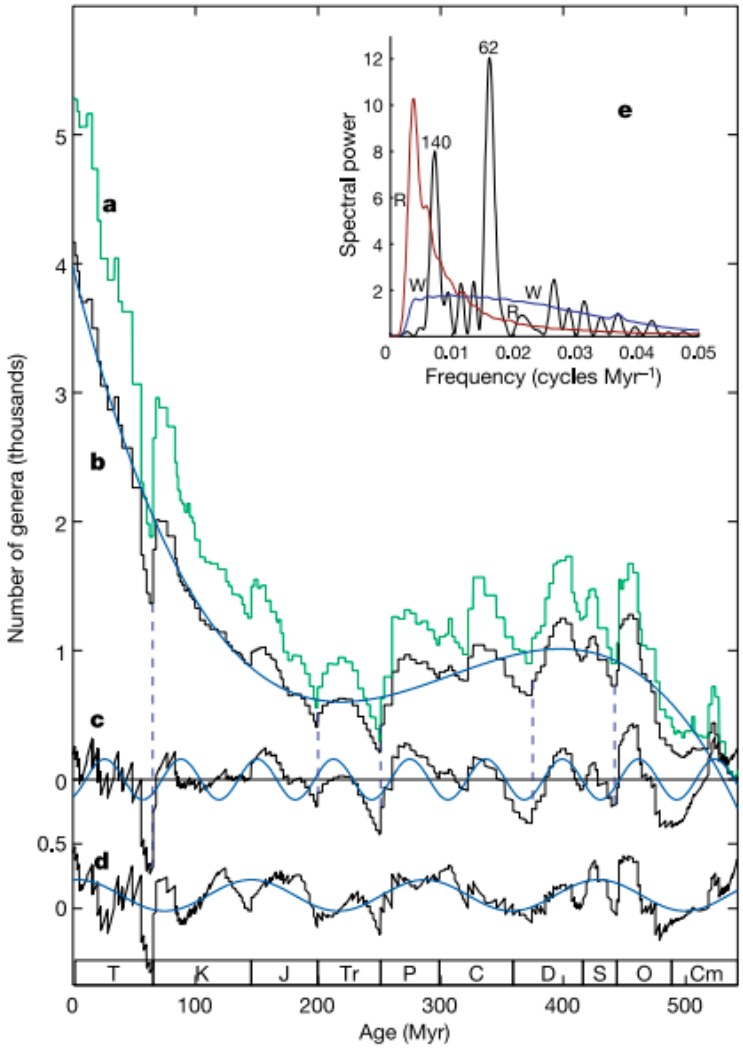

Figura 2.12: Ciclos encontrados en Rohde and Muller (2005).

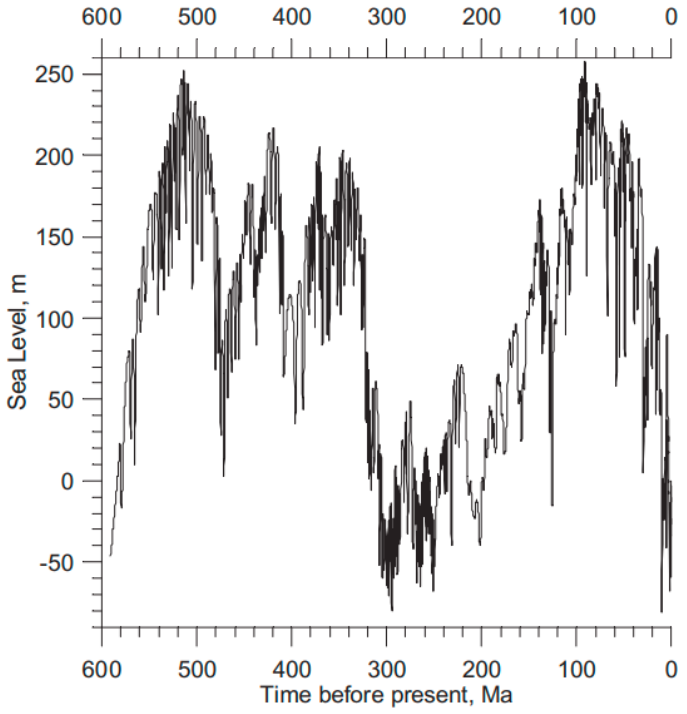

Figura 2.13: Curva del nivel del mar extraída de Harrison (2002).

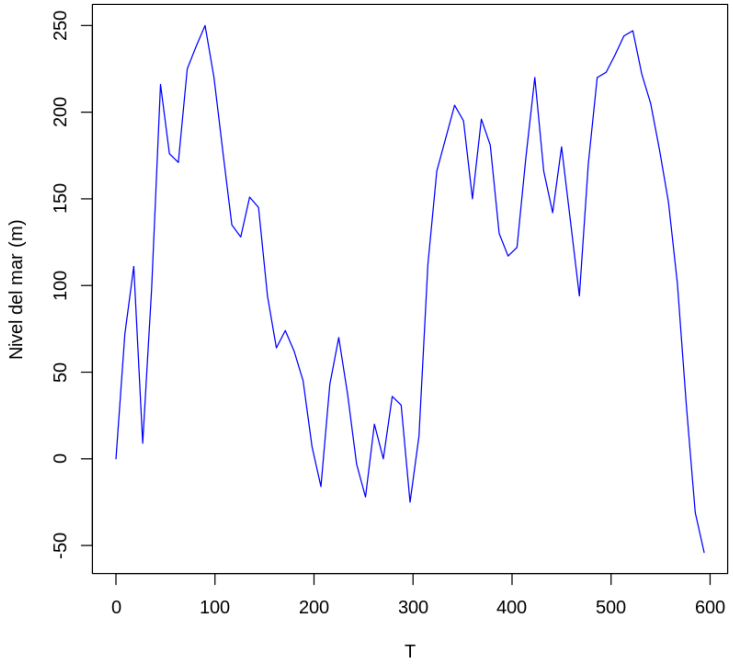

Figura 2.14: Curva de la tendencia del nivel mar.

vimos comportamientos muy variados en cada uno en cuanto al tiempo de vida. La covariable grupos sería una covariable fija y a través de ella considerando solamente seis grupos vistos en Alroy (2010b) obtenemos una gran representación del compendio pues de los 36 339 fósiles, se lograrían representar 19 828 fósiles. Esta distinción en los tipos de covariables tendrá gran importancia en el tratamiento del modelo que propondremos posteriormente.

2.4. Modelo de Cox

El propósito del modelo de Cox es evaluar simultáneamente el efecto de varios factores sobre la supervivencia. En otras palabras, nos permite examinar cómo factores específicos influyen en la ocurrencia de un evento particular. Las variables predictoras (o factores) son las covariables. Con el estimador de Kaplan-Meier no se toman en cuenta estos factores, lo cual, en términos comparativos, es una fortaleza del modelo de Cox ya que permite incorporar información adicional al modelo. Esta sección abordará cuestiones fundamentales para nuestro análisis ya que definimos el modelo de regresión de Cox, algunas de sus particularidades y las herramientas con las que contamos para validar el ajuste realizado. El contenido referente al modelo de regresión de Cox, en su mayoría, puede ser consultado en Collett (2015).

2.4.1. Modelo general de Cox de riesgos proporcionales

El conjunto de valores de las covariables en el modelo estará representado por el vector $\mathbf{X}$, donde $\mathbf{X} = (X_1, X_2, \ldots, X_p)$. La función $h_0(t)$, llamada función de riesgo basal, representará la función de riesgo cuando las covariables que componen al vector $\mathbf{X}$ son cero. Luego, la función de riesgo del modelo de Cox en función del tiempo t y un conjunto de covariables viene dada por

$$h(t, X) = h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j X_j\right).$$

El modelo de esta ecuación, en el que no se realizan suposiciones sobre la forma real de la función de riesgo inicial $h_0(t)$, fue introducida por Cox y ha llegado a conocerse como el modelo de regresión de Cox o el modelo de riesgos proporcionales de Cox. Los coeficientes β_j de las covariables representarán el impacto de las covariables a la función de riesgo según su valor. Un valor positivo del coeficiente β_k implicaría un impacto negativo en la función de riesgo ya que al aumentar el valor de la covariable asociada, aumentaría el riesgo y por el contrario, un valor negativo mostraría una disminución del riesgo aumentando la probabilidad de supervivencia.

Un modelo de riesgos proporcionales refiere a que el riesgo de muerte en un momento dado para un individuo es proporcional al riesgo en ese momento para otro individuo.

Para comprender esta noción definiremos la *razón de riesgos* (Hazard ratio), entre dos sujetos con diferente vector de covariables $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ y $\mathbf{X} = \left(X_1^*, X_2^*, \ldots, X_p^*\right)$ como

$$\mathrm{HR} = \frac{h(t, X^*)}{h(t, X)}$$

donde (Therneau and Grambsch, 2000)

$$\mathrm{HR} = \frac{h_0(t)\exp\left(\sum_{j=1}^{p}\beta_j X_j^*\right)}{h_0(t)\exp\left(\sum_{j=1}^{p}\beta_j X_j\right)} = \exp\left(\sum_{j=1}^{p}\beta_j\left[X_j^* - X_j\right]\right).$$

Esto es, decimos que el modelo es de riesgos proporcionales si el HR es constante en el tiempo. Por lo tanto, en el modelo de Cox se supone la hipótesis de que los riesgos son proporcionales, ya que se suponen covariables no dependientes del tiempo.

2.4.2. Modelo de Cox con covariables dependientes del tiempo

Existe la posibilidad de considerar covariables dependientes del tiempo $\mathbf{X}(\mathbf{t}) = \big(X_1(t), X_2(t), \ldots, X_p(t)\big)$, de manera que la función de riesgo viene dada por

$$h(t, X) = h_0(t)\exp\left(\sum_{j=1}^{p}\beta_j X_j(t)\right).$$

Notemos que en este caso el modelo no es de riesgos proporcionales. El tratamiento de este tipo de modelos se abordará detalladamente más adelante en el capítulo.

2.4.3. Ajuste del modelo de Cox

Ajustar el modelo de Cox a un conjunto observado de datos de supervivencia, implica estimar los coeficientes de las variables explicativas/covariables y es posible que sea necesario estimar la función de riesgo basal. Estos dos componentes del modelo se pueden estimar por separado siendo un resultado importante ya que para hacer inferencias sobre los efectos de las covariables no necesitaríamos especificar la función de riesgo basal. Los coeficientes β en el modelo de regresión de Cox pueden ser estimados usando métodos de máxima verosimilitud pero en este caso la función de verosimilitud será diferente.

Supongamos que tenemos datos de n individuos, de los cuales tenemos r tiempos de muerte y $n - r$ están censurados. Se ordenan los r tiempos de muerte, $t_1 < t_2 < \cdots < t_r$ y el conjunto de individuos que están en riesgo en el tiempo t_j se denota como $R(t_j)$, es decir, es el grupo de individuos que están vivos y sin censura en un momento justo antes de t_j.

A la función de verosimilitud de Cox para el modelo descrito anteriormente se le conoce como función de verosimilitud parcial ya que tiene en cuenta únicamente las probabilidades de los tiempos de muerte y no incluye las probabilidades de los tiempos de datos censurados. Sin embargo, en el cálculo de las probabilidades de los tiempos de muerte toma en cuenta a todos los sujetos (censurados o no) que están en riesgo en cada tiempo de muerte. Esta función está dada por

$$L(\beta) = \prod_{j=1}^{r}\frac{\exp\left(\beta' x_j\right)}{\sum_{l\in R(t_j)}\exp\left(\beta' x_l\right)},$$

donde x_j es el vector de covariables para el individuo que muere en el tiempo ordenado t_j. La suma en el denominador es la suma de los valores de $\exp(\beta' x)$ sobre todos los individuos que están en riesgo en

el tiempo t_j. Los individuos cuyos tiempos de supervivencia se encuentren censurados no contribuyen al numerador de la función de log-verosimilitud pero si contribuyen a la suma sobre los conjuntos de riesgo en los tiempos de muerte que ocurren antes de un tiempo censurado. Posteriormente se calcula la log-verosimilitud parcial y se maximiza para encontrar los coeficientes. A partir de la función de verosimilitud parcial obtenemos una estimación de los coeficientes, $\hat{\beta}$, mediante un método numérico y matriz de covarianzas $\hat{\sum} = I^{-1}(\hat{\beta})$, donde I es la matriz de información observada. Para encontrar una información más detallada y la justificación para considerar esta función de verosimilitud se puede consultar Collett (2015).

Por otra parte, la estimación de la función de riesgo basal suponemos que tenemos r tiempos de muerte distintos y se organizan en orden creciente $t_1 < t_2 < \cdots < t_r$. Además tenemos d_j muertes y n_j individuos en riesgo en el tiempo t_j. La función de riesgo base estimada en el tiempo t_j está dada por

$$\hat{h}\left(t_j\right) = 1 - \hat{\xi}_j,$$

donde $\hat{\xi}_j$ es la solución a la ecuación

$$\sum_{l \in D(t_j)} \frac{\exp\left(\hat{\beta}' x_l\right)}{1 - \hat{\xi}_j^{\exp\left(\hat{\beta}' x_l\right)}} = \sum_{l \in R(t_j)} \exp\left(\hat{\beta}' x_l\right),$$

para $j = 1, 2, \ldots, r$, siendo $D\left(t_j\right)$ el conjunto de los d_j individuos que mueren en t_j y $\hat{\beta}$ el vector de los coeficientes estimados anteriormente.

Cuando no hay covariables esta ecuación se transforma en

$$\frac{d_j}{1 - \hat{\xi}_j} = n_j,$$

obteniéndose

$$\hat{\xi}_j = \frac{n_j - d_j}{n_j}.$$

Luego, la función de riesgo basal en el tiempo t_j sería $\hat{h}_0(t) = d_j/n_j$ y la correspondiente función de supervivencia basal acumulada está dada por

$$\prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right)$$

para $t_k \leq t < t_{k+1}$ con $k = 1, 2, \ldots, r$. Observamos que obtuvimos la función de supervivencia estimada por el estimador Kaplan-Meier. Esto significa que la estimación de la función hazard basal en el caso del modelo de Cox generaliza la estimación de Kaplan-Meier al caso donde la función de riesgo depende de las covariables.

2.4.4. Análisis de residuos

Después de ajustado un modelo a un conjunto observado de datos de supervivencia, la adecuación del modelo se necesita evaluar, siendo ésto una parte esencial del proceso de modelación. El análisis de residuos puede conducir a la identificación de ciertas características importantes, como datos atípicos

y estructuras que ayuden a mejorar el modelo. A continuación se enlistan algunos tipos de residuos útiles en el análisis de supervivencia (Collett, 2015).

Supondremos que tenemos los tiempos de supervivencia de n individuos, donde r de estos son tiempos de muerte y los restantes $n - r$ están censurados por la derecha. Además, suponemos que tenemos un modelo de Cox ajustado a los tiempos de supervivencia y contiene p covariables. La función de riesgo ajustada tomando en cuenta la función de riesgo basal y los coeficientes estimados sería

$$\hat{h}\left(t\right) = \hat{h}_0\left(t\right) \exp\left(\sum_{j=1}^{p} \hat{\beta}_j X_j\left(t\right)\right).$$

- Cox-Snell

  Los residuos de tipo Cox-Snell son ampliamente usados en el análisis de supervivencia. El residuo Cox-Snell para el i-ésimo, $i = 1, 2, \ldots, n$, está dado por

  $$r_{Ci} = \hat{H}_0\left(t_i\right) \exp\left(\sum_{j=1}^{p} \hat{\beta}_j X_j\left(t\right)\right),$$

  donde $\hat{H}_0\left(t_i\right)$ representa la función de riesgo acumulada en el tiempo t_i, el tiempo de supervivencia observado para el individuo. Si el modelo es correcto, los residuos se distribuyen exponencialmente con media 1.

  Notemos que, el residuo de Cox-Snell, r_{Ci}, es el valor de $\hat{H}_i\left(t_i\right) = -\log \hat{S}_i\left(t_i\right)$, donde $\hat{H}_i\left(t_i\right)$ y $\hat{S}_i\left(t_i\right)$ son las funciones de riesgo y supervivencia acumuladas para el i-ésimo individuo en el tiempo t_i.

  Los datos censurados dieron lugar a los residuos Cox-Snell modificados. La idea general de la modificación radica en mantener los residuos de los datos no censurados y a los residuos de los datos censurados sumarle una constante positiva (*exceso residual*) de la siguiente forma:

  $$r'_{Ci} = \begin{cases} r_{Ci} & \text{para observaciones no censuradas} \\ r_{Ci} + \Delta & \text{para observaciones censuradas.} \end{cases}$$

- Martingala

  Los residuos martingala, r_{Mi}, miden la diferencia en $(0, t_i)$ entre el número observado de eventos de interés que experimenta el i-ésimo individuo y el número esperado basado en el modelo ajustado. Se definen como

  $$r_{Mi} = \delta_i - r_{Ci},$$

  donde

  $$\delta_i = \begin{cases} 1 & \text{si se observó el suceso de extinción,} \\ 0 & \text{si la observación está censurada,} \end{cases}$$

  y r_{Ci} es el residuo Cox-Snell.

  Los residuos de martingala toman valores entre $-\infty$ y la unidad y para observaciones censuradas, donde $\delta_i = 0$, son negativos. También se puede demostrar que estos residuos suman cero y, en

muestras grandes, los residuos de martingala no están correlacionados y tienen un valor esperado de cero.

Un residuo martingala negativo y grande indica un individuo de alto riesgo que todavía tenía un tiempo largo de supervivencia. Los residuos de martingala permiten dos usos principales ya que pueden utilizarse para encontrar valores atípicos de individuos que son mal ajustadas por el modelo y para determinar la forma funcional de cada una de las variables en el modelo.

La suma de estos residuos es cero y no están correlacionados, su valor esperado es cero y aun cuando el modelo sea adecuado no se distribuyen de forma simétrica en torno a cero, lo cual dificulta la interpretación de los gráficos.

- Deviance

  Los residuos *deviance* se construyen transformando los residuos martingala de tal manera que produzcan valores simétricos en torno de 0. Luego, están definidos por

$$r_{Di} = \operatorname{sgn}(r_{Mi}) \left[-2 \left\{r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\right\}\right]^{\frac{1}{2}},$$

  donde r_{Mi} es el residuo tipo martingala para el i-ésimo individuo y $\operatorname{sgn}(\cdot)$ es la función signo.

  La desviación de un modelo de regresión es el estadístico que se utiliza para cuantificar hasta qué punto el modelo estimado se aleja de un modelo teórico que se ajustaría perfectamente a nuestros datos (denominado modelo completo o modelo saturado). La estadística viene dada por

$$D = -2 \left\{\log \hat{L}_c - \log \hat{L}_f\right\},$$

  donde $\hat{L}_c$ es la función de verosimilitud para el modelo actual, y $\hat{L}_f$ es la función de verosimilitud para el modelo completo.

  Notemos que cuanto menor sea el valor de la desviación, mejor es el modelo. Las observaciones que corresponden a residuos de desviación relativamente grandes son aquellas que no se ajustan bien al modelo. La suma de los cuadrados de los residuos de desviación corresponde al valor de la desviación del modelo.

- Schoenfeld

  Este residuo difiere de los considerados anteriormente en un aspecto importante ya que no hay un único residuo para cada individuo, sino un conjunto de valores, uno para cada covariable incluida.

  El i-ésimo residuo para la covariable X_j está dado por

$$r_{S_{ji}} = \delta_i \left\{x_{ji} - \hat{a}_{ji}\right\},$$

  donde x_{ji} es el valor de la j-ésima covariable para el i-ésimo individuo en el estudio y además

$$\hat{a}_{ji} = \frac{\sum_{l \in R(t_i)} x_{jl} \exp\left(\hat{\beta}' x_l\right)}{\sum_{l \in R(t_i)} \exp\left(\hat{\beta}' x_l\right)},$$

donde $R(t_i)$ es el conjunto de individuos en riesgo en el tiempo t_i.

Estos residuos presentan la propiedad en muestras grandes de que, el valor esperado de $r_{S_{ji}}$ es cero y no están correlacionados entre sí. Además, su análisis será útil para la realización de pruebas de riesgos proporcionales a un modelo de regresión de Cox ajustado.

2.4.5. Pruebas de riesgos proporcionales

Un supuesto crucial que se hace al utilizar el modelo de regresión de Cox es el de los riesgos proporcionales. Si hay una o más covariables en el modelo cuyos coeficientes varían con el tiempo, o si hay covariables que dependen del tiempo, se violará la suposición de riesgos proporcionales. Por lo tanto, requerimos técnicas que puedan usarse para detectar si existen estas dependencias del tiempo. Existen varias técnicas para evaluar el supuesto de riesgos proporcionales en el modelo de Cox y para un análisis más detallado se puede consultar Collett (2015). A continuación realizaremos un pequeño análisis de una técnica en particular ya que es una de las más usadas y será la que utilizaremos para evaluar el supuesto de riesgos proporcionales en nuestro modelo propuesto.

Los residuos Schoenfeld son particularmente útiles en la evaluación del supuesto de riesgos proporcionales después de ajustar el modelo de regresión de Cox. En Grambsch and Therneau (1994) se mostró que el valor esperado del i-ésimo residuo Schoenfeld $i = 1, 2, \ldots, n$ para la j-ésima covariable en el modelo X_j, $j = 1, 2, \ldots, p$ denotado como $r^*_{S_{ji}}$ está dado por

$$E\left(r^*_{S_{ji}}\right) \approx \beta_j(t_i) - \hat{\beta}_j,$$

donde $\beta_j(t)$ representa el coeficiente variable en el tiempo de X_j, $\beta_j(t_i)$ es el valor de este coeficiente en el tiempo de supervivencia t_i y $\hat{\beta}_j$ es el valor estimado del coeficiente j en el modelo de Cox. Notemos que estos residuos sólo están definidos en los tiempos de muerte.

Este valor esperado sugiere que al graficar $r^*_{S_{ji}} + \hat{\beta}_j$, o equivalentemente $r^*_{S_{ji}}$ contra los tiempos de supervivencia observados debe proporcionar información sobre la forma del coeficiente dependiente del tiempo. Una línea horizontal sugiere que los coeficientes de X_j son constantes y se satisface el supuesto de riesgos proporcionales.

2.5. Simulador de datos de supervivencia con covariables temporales

En esta sección se describirá un simulador de tiempos de supervivencia que fue construido con el objetivo de examinar la plausibilidad de las covariables propuestas que se incorporarán al modelo de Cox en el análisis de la extinción. Como vimos anteriormente, este sistema de covariables responden a hechos reales en el pasado y son fácilmente parametrizables. Por otra parte, será fundamental determinar la corrección de la estimación de los parámetros en el sentido de que intervalos de confianza poseen la cobertura especificada. Para ello se implementa un cuantificador mediante el cual validaremos este aspecto. Con el cuantificador estaremos monitoreando en 100 simulaciones, cuántas veces al realizar la estimación de los parámetros para los tiempos de supervivencia simulados, se logra obtener los parámetros que se utilizaron para simular.

Una de las fortalezas del modelo de Cox es su capacidad para abarcar covariables que cambian con en tiempo. La razón práctica por la que funcionan las covariables dependientes del tiempo se basa en la forma subyacente en la que funciona el modelo de Cox donde en cada momento del evento, se compara

los valores de covariables actuales del sujeto que tuvo el evento con los valores actuales de todos los demás que estaban en riesgo en ese momento. Una forma sencilla de codificar covariables dependientes del tiempo consiste en utilizar intervalos de tiempo (Therneau, 2021). La idea general radica en tomar en cuenta cuántas veces durante la vida de un individuo, cambian de valor las covariables. Después de haber indentificado en qué momentos las covariables cambian su valor, se delimitan los intervalos de tiempo y a cada intervalo se le va a asignar el valor fijo que presentan dichas covariables en el intervalo. Después de realizado este procedimiento, tendremos por cada individuo varios vectores indicando en cada uno de ellos los valores actuales de las covariables hasta el próximo cambio, para así tomar en cuenta sus efectos a lo largo de toda su vida. Para realizar la estimación correspondiente del modelo nos apoyamos en RStudio, donde se podrá indicar cuáles vectores estarían asociados a un mismo individuo.

A continuación describiremos el proceso para simular los datos de supervivencia. Para la simulación utilizaremos como covariables la curva de la tendencia del nivel del mar, y dos sinusoidales con períodos $\mathcal{T} = 140$ y $\mathcal{T} = 62$ respectivamente. Como función de evolución utilizaremos $\exp(-0.01T + 7)$ y además, los grupos de taxones considerados por Alroy descritos anteriormente. La función de riesgo basal empleada para la generación de tiempos de vida resulta de la estimación de Kaplan-Meier para los datos del registro fósil ya que se trató de identificar con modelos paramétricos estándar y no fue evidente la relación entre ambos. El procedimiento para la simulación se encuentra basado en Therneau (2021).

Como INPUTS en el simulador consideramos:

- Los valores de las covariables temporales: función que para un tiempo geológico T entrega k valores de covariables temporales en T.
- Una plantilla que contenga varias columnas para determinar cuáles fósiles específicos simular. La columna 1 será una codificación donde se enumerará cada fósil y se le conocerá como el *id* del fósil. La columna 2 será el tiempo de primera aparición de fósiles seleccionados en el Compendio Sepkoski. La columnas restantes corresponderán a las covariables fijas que corresponden al fósil.
- La función de distribución en el tiempo de longevidad t bajo modelo de Cox.
- Un vector de parámetros que contenga los coeficientes en el modelo de Cox para covariables temporales y fijas (en ese orden).

El simulador construido entrega los siguientes OUTPUTS:

- Una primera matriz que contiene renglones seriados y repetidos para fósiles con varias columnas. En la columna 1 se encuentra el id del fósil. En la columna 2 el tiempo inferior del intervalo donde se encuentra un cambio en los valores de las covariables. En la columna 3 el tiempo superior que sería dónde van a volver a cambiar los valores de las covariables. En nuestro caso los cambios en las covariables se producen cada millón de años. La columna 4 contiene la información sobre el estatus de la especie, es decir, 0 si se encuentra viva y 1 si murió. Por último, en las restantes columnas se tendrán los valores de las covariables fijas y temporales, en ese orden, correspondientes a cada intervalo de tiempo.
- Una segunda matriz de dos columnas que contiene en la columna 1 la longevidad obtenida para cada fósil simulado y en la columna 2 la información sobre si continúa viva la especie (censurada) o si se extinguió.

Para describir el procedimiento de la simulación se enlistan los pasos:

1. Crearemos dos parámetros para controlar el proceso de simulación: `estatus` y `moving_T`. En `estatus` tendremos el valor 0 o el valor 1 en dependencia si la especie sigue viva o no. En `moving_T` se guarda el tiempo geológico en el cuál aún se encuentra viva la especie si `estatus` $= 0$ o el tiempo geológico donde la especie muere si `estatus` $= 1$. Luego se procede a generar los tiempos de supervivencia para cada especie de la plantilla. Mientras `estatus` $= 0$ y `moving_T` > 0 se realizarán los próximos pasos.

2. Comenzando por el valor de `moving_T` igual a la primera aparición asociada a la especie i en la plantilla, calculamos el producto interno de las covariables asociadas a esa especie en el tiempo geológico `moving_T` con los valores definidos en el vector de parámetros que contiene los coeficientes de las covariables. Este valor obtenido corresponde al predictor lineal en el modelo de Cox, es decir, la combinacion lineal entre coeficientes y covariables presente en la fórmula estudiada para la función de riesgo asociada a este modelo.

3. Evaluamos la función de distribución bajo modelo de Cox en $t = 0$ y $t = 1$, es decir, en los extremos del intervalo correspondiente al primer millón de años vividos de la especie en cuestión y utilizando el predictor lineal calculado.

4. Calculamos la probabilidad condicional `conditional_prob` $= (F2 - F1)/(1 - F1)$ donde $F2$ es la función de distribución bajo modelo de Cox en $t = 1$ y $F1$ es la función de distribución bajo modelo de Cox en $t = 0$, ambas utilizando el predictor lineal calculado.

5. Asignamos a `estatus` el valor 1 si generando un número aleatorio de la distribución uniforme se cumple que el número generado sea menor que la probabilidad condicional calculada y se le asignara el valor 0 en caso contrario.

6. Iteramos hasta que se deje de cumplir que `estatus` $= 0$ o `moving_T` > 0 y en cada iteración habría transcurrido un millón de años más de vida para la especie i. Por lo anterior, se actualiza `moving_T` y los valores de t ya que tomaríamos en cuenta el intervalo correspondiente al próximo millón de años vividos.

Al realizar este procedimiento para cada especie de la plantilla, finalmente obtenemos los tiempos de supervivencias simulados.

Luego de creado el simulador, validaremos la plausibilidad de las covariables utilizadas y la viabilidad de la estimación de los parámetros. Realizaremos 100 simulaciones de datos de supervivencia y ajustaremos un modelo de Cox a los datos con las covariables utilizadas. Evaluaremos si se logra estimar cada coeficiente que se utilizó para la simulación. Contaremos la cantidad de veces que el intervalo de confianza obtenido en la salida del ajuste del modelo de Cox contiene al verdadero valor con el que se realizó la simulación. Además, se contará la cantidad de veces que se produce una sobreestimación, es decir, el verdadero valor no está contenido en el intervalo de confianza y se encuentra por debajo del intervalo, en caso contrario diremos se se produjo una subestimación. Estas estimaciones no deben ser exactamente iguales a los coeficientes con que se generaron los datos pero considerando la desviación estándar que contiene la estimación, el intervalo de confianza debe contener al verdadero valor de los coeficientes con que se generaron los datos. Nos apoyamos de la función `coxph` del paquete `survival` en RStudio y la codificación para el tratamiento de las covariables dependientes del tiempo. Posteriormente se analizará detalladamente cómo utilizar esta función en nuestro caso de análisis y cómo interpretar la salida que arroja. Por ahora solo abordamos las generalidades y mostraremos los resultados obtenidos por el contador a fines de validar la utilización de las covariables y las estimaciones que realizaremos para el modelo propuesto en el análisis de extinción.

	nivel_mar	sinusoidal_62	sinusoidal_140	evolución	g_brachiopoda	g_gastropoda	g_cephalopoda	g_anthozoa	g_bilvalvia
contenidos	97	95	94	98	97	94	98	95	97
subestimados	1	2	3	1	2	3	0	2	1
sobreestimados	2	3	3	1	1	3	2	3	2

Figura 2.15: Cuantificador que muestra en 100 simulaciones cuántas veces se logra cubrir con intervalos de confianza nominalmente de 95 % los parámetros con los que se simularon los datos.

Con la Figura 2.15 notamos que para la magnitud y estructura de datos contenidos en el registro fósil, es factible realizar estimación sensata de parámetros, dando lugar a interpretaciones en el contexto de extinción ya que en el 95 % de las simulaciones se obtuvieron resultados aceptables al lograrse recuperar los parámetros de la simulación.

2.6. Aplicación a datos del registro fósil

Hasta este momento se han establecido las bases para poder proponer un modelo de regresión de Cox para el análisis de la extinción. A continuación incorporaremos al modelo las covariables identificadas y posteriormente realizaremos las estimaciones apoyándonos en la función `coxph` de RStudio. Posteriormente procedemos a validar el modelo mediante pruebas de riesgos proporcionales y un análisis de residuos. Finalmente, se estará en condiciones para analizar la influencia en la extinción de cada uno de los factores identificados.

```
  n= 289056, number of events= 8386

                          coef  exp(coef)   se(coef)  robust se       z Pr(>|z|)
t_dep$nivel_mar       1.033e-03  1.001e+00  1.618e-04  1.523e-04   6.781 1.19e-11 ***
t_dep$sinusoidal_62  -4.806e-03  9.952e-01  7.855e-04  7.732e-04  -6.215 5.13e-10 ***
t_dep$sinusoidal_140  3.469e-03  1.003e+00  8.012e-04  8.039e-04   4.316 1.59e-05 ***
t_dep$evolución      -7.790e-04  9.992e-01  4.470e-05  4.463e-05 -17.455  < 2e-16 ***
t_dep$g_trilobita     1.275e+00  3.577e+00  4.567e-02  4.731e-02  26.942  < 2e-16 ***
t_dep$g_brachiopoda   8.541e-01  2.349e+00  3.983e-02  3.973e-02  21.497  < 2e-16 ***
t_dep$g_gastropoda   -6.374e-02  9.383e-01  3.877e-02  3.748e-02  -1.700    0.089 .
t_dep$g_cephalopoda   1.316e+00  3.730e+00  4.187e-02  5.043e-02  26.103  < 2e-16 ***
t_dep$g_anthozoa      4.165e-01  1.517e+00  4.354e-02  4.210e-02   9.895  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                     exp(coef) exp(-coef) lower .95 upper .95
t_dep$nivel_mar         1.0010     0.9990    1.0007    1.0013
t_dep$sinusoidal_62     0.9952     1.0048    0.9937    0.9967
t_dep$sinusoidal_140    1.0035     0.9965    1.0019    1.0051
t_dep$evolución         0.9992     1.0008    0.9991    0.9993
t_dep$g_trilobita       3.5773     0.2795    3.2605    3.9249
t_dep$g_brachiopoda     2.3492     0.4257    2.1732    2.5394
t_dep$g_gastropoda      0.9383     1.0658    0.8718    1.0098
t_dep$g_cephalopoda     3.7302     0.2681    3.3791    4.1178
t_dep$g_anthozoa        1.5167     0.6593    1.3966    1.6472

Concordance= 0.721  (se = 0.003 )
Likelihood ratio test= 4110  on 9 df,   p=<2e-16
Wald test            = 3273  on 9 df,   p=<2e-16
Score (logrank) test = 4553  on 9 df,   p=<2e-16,   Robust = 4477  p=<2e-16
```

Figura 2.16: Estimaciones de los coeficientes del modelo de Cox para los datos del registro fósil.

En la Figura 2.16 se muestran las estimaciones obtenidas al ajustar el modelo. El valor de z es una medida que se calcula como $z =$ `coef/robust_se` y se le conoce como estadístico de Wald. Esta

medida nos permitirá resolver la hipótesis $H_0 : \beta_j = 0$ *versus* $H_1 : \beta_j \neq 0$. Mientras z esté lejos de cero indicaría que se rechaza la hipótesis nula, es decir, se declararía que existe una relación entre la covariable X_j y el riesgo. En nuestro ajuste, los valores del estadístico z para cada una de las covariables son relativamente grandes lo que indicaría que existe una fuerte relación entre el riesgo y las covariables seleccionadas. Un valor pequeño de `Pr(>|z|)` indicaría que el valor de z es grande, o sea, existe una fuerte relación entre la variable explicativa y la variable de respuesta. Cada uno de los tests que se realizan (Wald, Score, Razón de verosimilitud) utilizan diferentes estadísticos y cada uno sigue una distribución Chi cuadrada con p grados de libertad, todos arrojando los mismos resultados.

Luego de ajustado el modelo debemos comprobar que se cumple la hipótesis de riesgos proporcionales. Para el análisis de la hipótesis de riesgos proporcionales del modelo ajustado nos apoyaremos en la función `cox.zph()` del paquete `survival` de RStudio, la cuál utiliza la idea abordada anteriormente en la sección de las pruebas de riesgos proporcionales utilizando los residuos Schoenfeld.

```
                        chisq df       p
t_dep$nivel_mar       62.9127  1 2.2e-15
t_dep$sinusoidal_62    0.9189  1  0.3378
t_dep$sinusoidal_140   0.0532  1  0.8175
t_dep$evolución       26.2521  1 3.0e-07
t_dep$g_trilobita     23.9138  1 1.0e-06
t_dep$g_brachiopoda   16.5733  1 4.7e-05
t_dep$g_gastropoda    41.0713  1 1.5e-10
t_dep$g_cephalopoda  177.0228  1 < 2e-16
t_dep$g_anthozoa      10.4762  1  0.0012
GLOBAL               295.4936  9 < 2e-16
```

Figura 2.17: Prueba de riesgos proporcionales.

En este caso se contrasta: coeficientes no dependen del tiempo *versus* coeficientes dependen del tiempo. Como podemos percibir la mayoría de los p-valores muestran que no se cumple la hipótesis de los riesgos proporcionales en el modelo. Hemos modelado el problema tomando en cuenta que algunas covariables como nivel del mar, las dos sinusoidales y la evolución dependen del tiempo. Vale la pena aclarar que aunque solo las covariables mencionadas dependen del tiempo y la covariable asociada a los grupos de taxones son covariables fijas se necesita realizar la prueba de riesgos proporcionales a todas las covariables. Por ejemplo, el riesgo asociado con el hecho de ser una especie perteneciente a los *trilobita* a lo largo del estudio no debe necesariamente ser constante a pesar de tratarse de una covariable fija ya que hipotéticamente, puede ser que los trilobita al ser más jóvenes tengan mayor riesgo de morir que al alcanzar una edad avanzada.

Con la prueba de riesgos proporcionales realizada con la función `cox.zph()` se identifican coeficientes dependientes del tiempo en el modelo de Cox. A continuación mostraremos estimaciones del coeficiente $\beta(t)$ dependiente del tiempo para cada una de las covariables. Si se cumpliera el supuesto de riesgos proporcionales, entonces la función $\beta(t)$ sería una línea horizontal.

Luego de haber obtenido que el modelo ajustado no cumple la hipótesis de riesgos proporcionales debemos realizar una modificación. En Therneau (2021) se propone una solución a este problema y consiste en variar el valor de los coeficientes durante ciertos intervalos de tiempo. Una forma sencilla de modificar el modelo sería utilizando la función `survSplit` para dividir el conjunto de datos en partes dependientes del tiempo en el que haya alguna variación de los coeficientes. Intuitivamente consiste en identificar a través de las gráficas de los $\beta(t)$, los millones de años de vida en se encuentran variaciones en esta función de los coeficientes y a partir de ahí dividir en conjuntos los datos, tales que tengamos coeficientes constantes en la función de riesgo durante toda la vida de las especies. Por ejemplo, en la

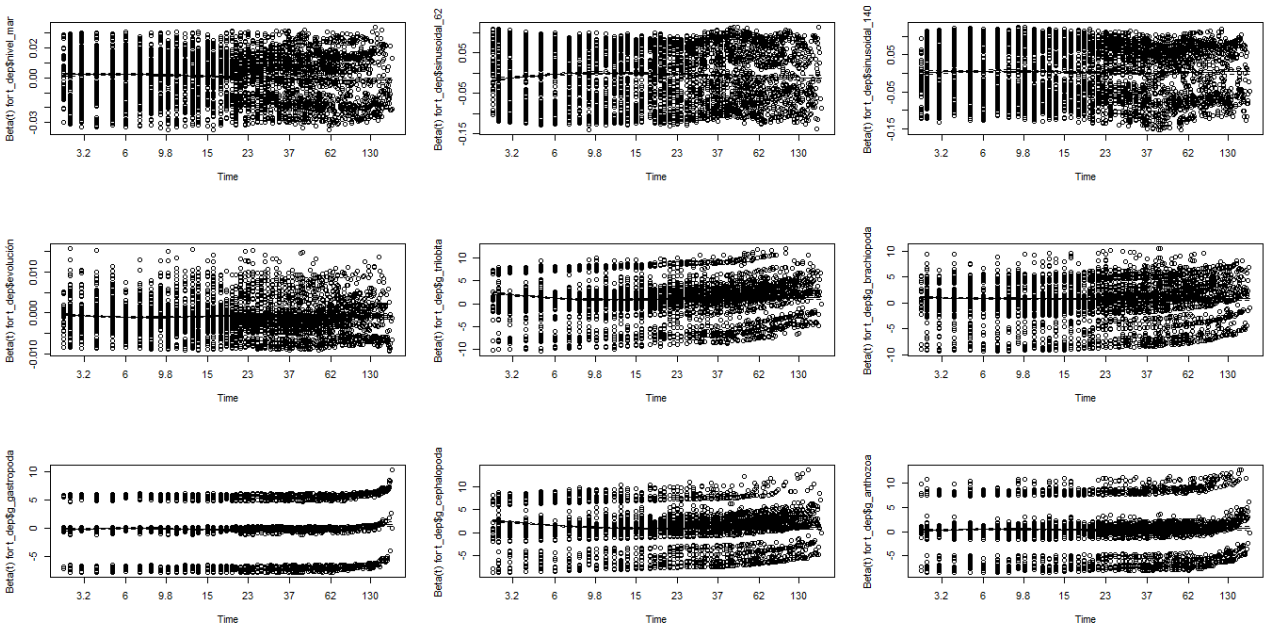

Figura 2.18: Estimaciones de los coeficientes $\beta(t)$ dependientes del tiempo.

primera gráfica del coeficiente correspondiente a la covariable del nivel del mar realizaremos un corte en $t = 9$ y otro en $t = 30$ a fines de contar con tres intervalos de tiempo con coeficientes constantes en cada uno.

En la Figura 2.19 se muestran las estimaciones obtenidas después de modificar el modelo.

Luego de realizar una segunda estimación tras modificar el modelo necesitamos comprobar nuevamente la hipótesis de riesgos proporcionales y obtenemos los resultados anotados en la Figura 2.20. En dicha figura apreciamos una mejoría considerable en todos los p-valores a pesar de que el p-valor global aún muestra que no se cumple la hipótesis de riesgos proporcionales. El método utilizado para modificar el modelo tiene la desventaja de necesitar realizarse una partición *arbitraria* en el conjunto de los coeficientes para intentar lograr cumplir la hipótesis de riesgos proporcionales. En nuestro modelo se intentan estimar nueve parámetros y también complica un poco las condiciones para lograr seleccionar particiones correctas para cada gráfica de β ya que son varios parámetros interactuando. Consideraremos suficiente los resultados obtenidos después de la modificación tras observarse que solo dos covariables están incumpliendo la hipótesis, grupo *trilobita* y grupo *gastropoda*, habiendo resultado esta última covariable la única poco significativa en el modelo.

Para validar el modelo un aspecto importante resulta realizar un análisis de residuos. En la Figura 2.21 y la Figura 2.22 se muestran las gráficas obtenidas para los residuos deviance y martingala.

En la Figura 2.21 observamos que parece estar bien ajustado el modelo ya que no se encuentran patrones sospechosos en la distribución de los residuos, es decir están distribuidos relativamente cerca del cero y sin agruparse dando lugar a una estructura específica. Además no se evidencian residuos relativamente grandes consecuentes de un mal ajuste. Se consideraría un mal ajuste si existieran varios datos con residuos grandes en valor absoluto de tal forma que no estuvieran los puntos cercanos al cero

	coef	exp(coef)	se(coef)	z	p
t_dep$sinusoidal_62	-4.850e-03	9.952e-01	7.872e-04	-6.162	7.20e-10
t_dep$sinusoidal_140	3.337e-03	1.003e+00	8.015e-04	4.163	3.14e-05
t_dep$g_brachiopoda	8.695e-01	2.386e+00	4.054e-02	21.450	< 2e-16
t_dep$nivel_mar:strata(tgroup)tgroup=1	2.552e-03	1.003e+00	2.730e-04	9.350	< 2e-16
t_dep$nivel_mar:strata(tgroup)tgroup=2	7.765e-04	1.001e+00	2.633e-04	2.949	0.00319
t_dep$nivel_mar:strata(tgroup)tgroup=3	-6.057e-04	9.994e-01	3.052e-04	-1.985	0.04720
t_dep$evolución:strata(cut_evo[, 14])cut_evo[, 14]=1	-9.331e-04	9.991e-01	5.990e-05	-15.578	< 2e-16
t_dep$evolución:strata(cut_evo[, 14])cut_evo[, 14]=2	-6.310e-04	9.994e-01	5.921e-05	-10.657	< 2e-16
t_dep$g_trilobita:strata(cut_g1[, 14])cut_g1[, 14]=1	1.778e+00	5.920e+00	7.201e-02	24.695	< 2e-16
t_dep$g_trilobita:strata(cut_g1[, 14])cut_g1[, 14]=2	1.137e+00	3.116e+00	9.701e-02	11.717	< 2e-16
t_dep$g_trilobita:strata(cut_g1[, 14])cut_g1[, 14]=3	1.053e+00	2.865e+00	8.180e-02	12.868	< 2e-16
t_dep$g_trilobita:strata(cut_g1[, 14])cut_g1[, 14]=4	1.096e+00	2.992e+00	6.266e-02	17.493	< 2e-16
t_dep$g_gastropoda:strata(cut_g3[, 14])cut_g3[, 14]=1	-1.027e-01	9.024e-01	4.492e-02	-2.285	0.02229
t_dep$g_gastropoda:strata(cut_g3[, 14])cut_g3[, 14]=2	3.007e-03	1.003e+00	6.116e-02	0.049	0.96079
t_dep$g_cephalopoda:strata(cut_g4[, 14])cut_g4[, 14]=1	2.211e+00	9.125e+00	7.566e-02	29.222	< 2e-16
t_dep$g_cephalopoda:strata(cut_g4[, 14])cut_g4[, 14]=2	1.613e+00	5.016e+00	8.459e-02	19.065	< 2e-16
t_dep$g_cephalopoda:strata(cut_g4[, 14])cut_g4[, 14]=3	1.414e+00	4.114e+00	9.357e-02	15.116	< 2e-16
t_dep$g_cephalopoda:strata(cut_g4[, 14])cut_g4[, 14]=4	1.102e+00	3.010e+00	6.446e-02	17.097	< 2e-16
t_dep$g_cephalopoda:strata(cut_g4[, 14])cut_g4[, 14]=5	7.702e-01	2.160e+00	7.629e-02	10.096	< 2e-16
t_dep$g_anthozoa:strata(cut_g5[, 14])cut_g5[, 14]=1	4.342e-01	1.544e+00	4.900e-02	8.861	< 2e-16
t_dep$g_anthozoa:strata(cut_g5[, 14])cut_g5[, 14]=2	3.812e-01	1.464e+00	7.720e-02	4.937	7.92e-07

Figura 2.19: Estimaciones de los coeficientes del modelo modificado de Cox para los datos del registro fósil.

	chisq	df	p
t_dep$sinusoidal_62	0.784	1	0.37580
t_dep$sinusoidal_140	1.295	1	0.25506
t_dep$g_brachiopoda	2.896	1	0.08877
t_dep$nivel_mar:strata(tgroup)	0.954	3	0.81243
t_dep$evolución:strata(cut_Cox_evo[, 14])	2.524	2	0.28307
t_dep$g_trilobita:strata(cut_Cox_g1[, 14])	12.981	4	0.01137
t_dep$g_gastropoda:strata(cut_Cox_g3[, 14])	8.185	2	0.01670
t_dep$g_cephalopoda:strata(cut_Cox_g4[, 14])	10.968	5	0.05202
t_dep$g_anthozoa:strata(cut_Cox_g5[, 14])	4.209	2	0.12188
GLOBAL	50.411	21	0.00032

Figura 2.20: Prueba de riesgos proporcionales después de modificado el modelo.

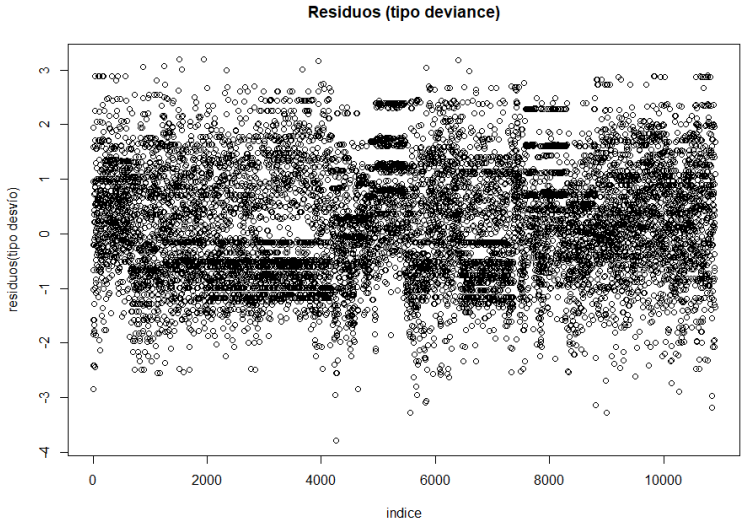

Figura 2.21: Gráfica de residuos Deviance para el modelo modificado considerando coeficientes y covariables variantes en el tiempo.

ya que indicaría que el modelo estimado se diferencia mucho del modelo completo o en otro caso, que no se encuentren arbitrariamente dispersos.

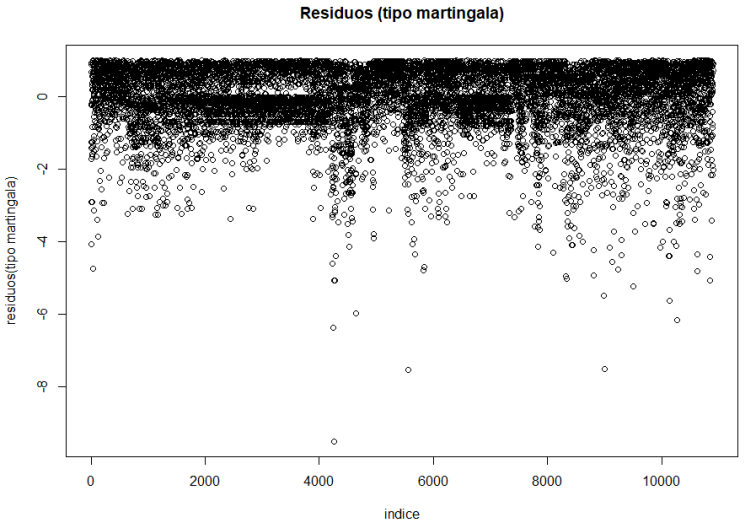

Figura 2.22: Gráfica de residuos Martingala para el modelo modificado considerando coeficientes y covariables variantes en el tiempo.

Podemos observar en la Figura 2.22 que no se determina alguna forma funcional de las variables en el modelo y parece que éste se encuentra bien ajustado. Como habíamos mencionado, estos residuos no se distribuyen de forma simétrica alrededor de cero, lo cual dificulta la interpretación del gráfico. En cambio, son útiles para identificar datos atípicos que, en este caso, parecen ser pocos.

Los residuos de Cox-Snell poseen una distribución exponencial con media unitaria, si el modelo ajustado es correcto. Por lo tanto, tienen una media y una varianza de la unidad y están distribuidas asimétricamente con respecto a la media. Esto significa que los gráficos simples de los residuos, como los gráficos de los residuos frente al número de observación, conocidos como gráficos de índice, no darán lugar a una visualización simétrica. Por ello, usualmente se grafica el riesgo acumulado de los residuos Cox-Snell frente a los residuos Cox-Snell y si el modelo es correcto, los datos se deben ajustar a una recta a través del origen con pendiente unitaria. Esta gráfica se basa en el hecho de que si una variable aleatoria $\mathbb{T}$ tiene una distribución exponencial con media unitaria, entonces la función de supervivencia de $\mathbb{T}$ es $\exp(-t)$. En la Figura 2.23 se evidencia que el modelo ajustado es correcto.

Tras validar las estimaciones obtenidas por el modelo propuesto, nos encontramos en condiciones para poder interpretar los resultados obtenidos en el contexto de la extinción. En la Figura 2.19 se evidencia que todas las covariables identificadas son significativas para la función de riesgo, con excepción del grupo *gastropoda*. Una pregunta en el ámbito de biología sería investigar si hay alguna razón por la cual *gastropoda* es esencialmente insensible a las covariables consideradas. Esto puede constituir una característica esperada, o bien un descubrimiento derivado de la modelación realizada.

En la covariable del nivel del mar se tomaron en cuenta 3 particiones, es decir, se realizó un corte en $t = 9$ y otro en $t = 30$, obteniéndose en los coeficientes estimados en los dos primeros intervalos (`strata` $= 1$ y `strata` $= 2$) valores de signo positivo y en el último (`strata` $= 3$) un valor de signo negativo. Podemos notar que al presentar signo positivo en los dos primeros intervalos nos indica que

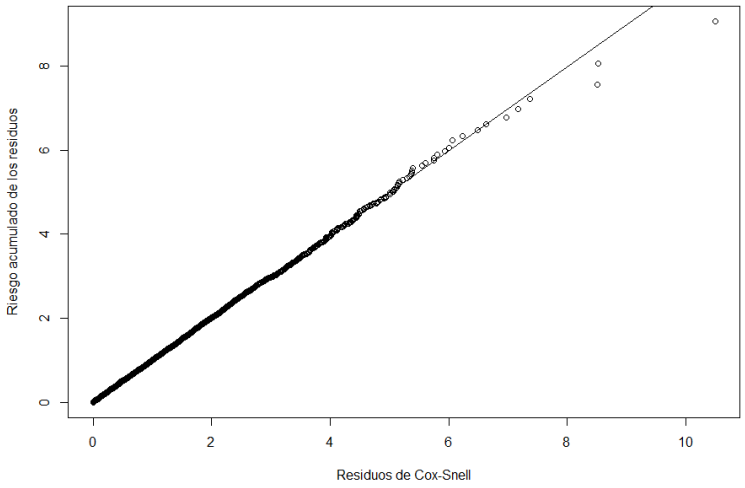

Figura 2.23: Gráfica del riesgo acumulado de los residuos de Cox-Snell.

en las especies jóvenes se tiene un pronóstico desfavorable al aumentar el nivel del mar, es decir, al aumentar el nivel del mar aumenta el riesgo de muerte y disminuye la longevidad. En cambio, en las especies con edades avanzadas al obtenerse un coeficiente negativo, un aumento en el nivel del mar disminuye el riesgo por lo que aumenta la longevidad.

Con relación a la covariable asociada a la sinusoidal con período $\mathcal{T} = 140$, obtenemos que el coeficiente estimado es positivo por lo que a través de ella podemos concluir que un aumento en los valores de la sinusoidal consecuencia de la variación en el clima y el flujo de rayos cósmicos provocarán un aumento del riesgo de extinción. Sucede lo contrario en la sinusoidal de período $\mathcal{T} = 62$ ya que el signo del coeficiente es negativo. El ciclo asociado a esta sinusoidal no contenía causas evidentes comprobadas pero en este estudio inferimos que los factores que pudieran determinarla no indican un pronóstico desfavorable en la longevidad de las especies.

Observando el comportamiento de los coeficientes de la función evolución se establece compatibilidad con la noción de que la evolución de los organismos mejora su aptitud a lo largo de los años ya que se obtienen coeficientes negativos. Esto indica que al pasar el tiempo los organismos disminuyen el riesgo de muerte aumentando la longevidad. Ratificando así la idea de Darwin que intentó probar con los datos del registro fósil sobre la evolución y que además ha sido interés de muchos paleontólogos. Uno de los puntos importantes que esto posee es que la función de evolución es otro de los factores reconocidos que contribuye a confundir la señal que puedan poseer los datos respecto al fenómeno primordial de interés, la extinción en sí misma.

En cuanto a los grupos de taxones notamos que los *trilobita* y *cephalopoda* tenían mucho mayor riesgo de extinguirse que los *anthozoa* y *brachiopodas*. Los coeficientes de los *gastropodas* indican que tenían mayor riesgo de muerte en etapa adulta y menor riesgo en etapa joven pero no son influyentes en el estudio de la evolución. Tomando en cuenta estos aspectos hemos evidenciado la influencia en la extinción de cada uno de los factores identificados.

Capítulo 3

Discusión y perspectiva

Desde la postulación del problema abordado en la tesis, se argumentó que en el análisis de la extinción con base en el registro fósil existen varios aspectos que actúan simultáneamente sobre la longevidad de las especies. Se estableció que se deberían tomar en cuenta para separar efectos concomitantes de aquellos que son de interés primordial, aquellos que tienen que ver directamente con la extinción. El análisis realizado presenta la dificultad de depender del conocimiento previo que se tenga del problema y la información que se encuentre disponible. Consideramos que la aportación principal de este trabajo al análisis de la extinción radica en la propuesta de un modelo con una óptica diferente a la utilizada en estudios previos, que permite incorporar explícitamente mayor información al estudio así como tomar en cuenta los datos censurados en lugar de omitirlos. Sin embargo, es preciso conocer que el modelo postulado contiene algunas limitantes debido a que algunos aspectos no se han tomado en consideración. Al tomar en cuenta nuevos factores en el modelo, consecuentemente nos enfrentamos a nuevos retos de modelación. Por lo anterior, en este capítulo nos dedicaremos principalmente a realizar una discusión del análisis propuesto y diversos elementos que se pueden incorporar para refinar el modelo postulado. Como se mencionó en la introducción, con esta tesis se pretendió sentar las bases para poder analizar posteriormente el cambio en la biodiversidad actual y por ello, al final del capítulo se aborda la pertinencia de nuestro trabajo para posteriores análisis sobre la sexta extinción masiva.

3.1. Resumen de logros

El enfoque de modelación estadística que se utilizó permite introducir información relevante adicional a la contenida en la base de datos del Compendio Sepkoski a través de las covariables y tomar en cuenta que algunos datos están censurados. Mediante el modelo postulado se logra caracterizar probabilísticamente la longevidad de las especies, no siendo así en trabajos anteriores. Esto nos permite cuantificar la incertidumbre asociada a la solución que obtenemos mediante intervalos de confianza de la estimación y con la utilización de pruebas de hipótesis se logra realizar análisis de significancia de las covariables que consideramos influyentes en la extinción. Es decir, se postula un modelo con el cual se toman en consideración márgenes de error en la solución y se pueden realizar inferencias sobre factores influyentes en la extinción. Una de las utilidades primordiales del análisis postulado radica en que se puede validar formalmente mediante pruebas estadísticas la calidad de la propuesta realizada. Con la codificación de los datos a fin de incorporar covariables dependientes del tiempo logramos realizar un

ajuste en el cual se toma en cuenta los valores del nivel del mar y demás covariables a lo largo de la vida de las especies.

Con el modelo postulado se logra convalidar que la variación en el nivel del mar, el clima, la glaciación, el flujo de rayos cósmicos, la adecuación biológica y los factores relacionados con el ciclo de período $\mathcal{T} = 62$ constituyen variables influyentes significativamente en las extinciones ocurridas en los últimos 540 millones de años. Como habíamos visto, diversos estudios se han enfocado en ratificar la idea de Darwin relacionada con la mejora en la aptitud de los organismos al pasar el tiempo y cambiar las condiciones en la Tierra y con este trabajo se logra comprobar empíricamente esta noción.

Por otra parte, se concluye que existen comportamientos muy variados en los grupos de taxones (por ejemplo, grupos *trilobita* y *anthozoa*) del registro fósil, sugiriendo que en efecto, se debe distinguir entre ellos en cualquier análisis que se realice ya que el comportamiento y por consiguiente la información que aporta cada uno es diferente. También se obtuvo la sugerencia de que por lo menos en un grupo (*gastropodas*) no se encuentran las mismas relaciones respecto a covariables temporales, y ello conduce naturalmente a buscar la interpretación biológica que ello conlleva. Por lo anterior, se validan ciertos puntos importantes que ya se conocían sobre la extinción como el nivel del mar, el clima, *etc.*, dando lugar a perfeccionar y modificar el modelo para realizar inferencias sobre la hipótesis de la sexta extinción masiva actual.

3.2. Limitaciones y retos de modelación

Hay ciertos elementos que el modelo desarrollado en la tesis no ha tomado en consideración, no obstante han sido identificados en el análisis del contexto. Así, los resultados obtenidos deben tomarse como resultados ilustrativos. Muestran que el concepto del modelo estadístico invocado—aun sobresimplificado—es útil en principio y que contiene elementos capaces de involucrar componentes pertinentes con resultados interpretables. A continuación se enuncian y comentan algunos de los temas que no han sido explícitamente incorporados en la solución propuesta.

Incertidumbre en fechamientos

En el análisis con los datos del registro fósil hemos mencionado que una característica intrínseca de los datos radica en que en realidad no sabemos a ciencia cierta la fecha de inicio y fin del avistamiento de las especies. Las fechas de primera y última aparición participan en dos aspectos cruciales en el modelo de Cox. Por una parte, definen directamente la longevidad, es decir, la variable de respuesta de un modelo de regresión. Por otra, definen un rango sobre el cual actúan sobre la vida de la especie ciertas covariables que dependen así mismo del tiempo geológico, T. En ambas instancias, se ha actuado como que estas fechas han sido observadas sin error. En realidad, estas fechas para una especie han sido inferidas a partir de una colección de ejemplares, que puede variar en número y homogeneidad. Más aun, el proceso mismo de fechamiento está sujeto a errores, con algunos parámetros propios de precisión y exactitud, y este método de fechamiento pudo no haber sido el mismo para todos los ejemplares. El Compendio de Sepkoski no da cuenta de estas características, que serían *de facto* cuantificaciones adicionales que se deberían incorporar, de ser posible.

Un reto consiste en involucrar la incertidumbre asociada a la longevidad de los fósiles y el efecto que esto tiene sobre las inferencias realizadas acerca de su comportamiento probabilístico. La cuantificación de incertidumbre en una colección tan vasta y variada de ejemplares, con muy

distintos orígenes históricos y varias metodologías utilizadas para su procesamiento, no parece ser una tarea fácil. La base de datos *Fossilworks* contiene información que puede ser más detallada en cuanto a fechamientos, pero no necesariamente con la misma extensión y calidad para todo el registro fósil.

De existir alguna manera de cuantificar la calidad global de un intervalo de tiempo asociado a cada especie, una posibilidad podría ser recurrir a la noción de pesaje (*weighting*) implementada en muchos procedimientos de estimación en modelos de regresión. La idea es que se aplica un mayor peso en la verosimilitud a observaciones que tienen mayor calidad o certeza en los fechamientos. Se trata de modificar el procedimiento de estimación, no de modificar el modelo estadístico propiamente. Por ejemplo, la función `coxph` empleada en R para realizar estimación en el modelo de Cox, considera un argumento llamada `weights` que permite incorporar este esquema de pesaje, noción documentada en Therneau (2021).

En lo que se concibe una manera de involucrar formalmente este aspecto de imprecisión en las fechas de primera y última aparición, se podría abordar el tema a través del modelo de simulación desarrollado en la tesis. En efecto, se podrían simular muy fácilmente datos incorporando errores de medición en fecha al azar con diversos grados de magnitud, para fines de evaluar la robustez de las estimaciones obtenidas por la metodología de regresión de Cox. En otro caso, como una alternativa para abordar la incertidumbre en fechamientos, se podría considerar que los datos están censurados por la izquierda y la derecha e incorporarlos con esta característica al modelo de Cox.

Validación

Se mostró en el Capítulo 2 que por procesos iterativos de modelación y contraste con datos del compendio, que el modelo de Cox requirió de ciertos elementos que introdujeron mayor complejidad. Específicamente, con el método propuesto en Therneau (2021) para incorporar covariables dependientes del tiempo logramos realizar un ajuste más sensato, en el cual se toma en cuenta los valores del nivel del mar y demás covariables a lo largo de la vida de las especies. El método utilizado para la incorporación de covariables dependientes del tiempo resulta útil para nuestros objetivos pero se paga un costo metodológico, relativo a la validación. En un análisis de residuos usualmente podemos detectar insuficiencias en el ajuste graficando la correspondencia de variables explicativas con los residuos correspondientes pero en este caso para cada especie tenemos múltiples valores en cada covariable ya que varían en el tiempo. No es claro qué valor de covariable utilizar para tal fin.

La metodología para validación de modelos de Cox cuando han sido complicados por coeficientes que dependen de t así como covariables que dependen de T, parece ser un problema abierto (Therneau and Grambsch, 2000, página 113).

Esfuerzo de muestreo, y otras variables

Es bien sabido que en ecología, el esfuerzo de muestreo es un factor determinante para analizar datos de riqueza de especies siendo que la dedicación por parte de humanos para buscarlas es esporádica y heterogénea. Los datos en general no han resultado de un esquema de diseño muestral, y se habla de sesgos debido a muestreo. Por ejemplo, cuando se mapean puntos geo-referenciados de especímenes de mamíferos o mariposas, se descubre que los puntos se dispersan de manera muy concentrada sobre las carreteras de una región (Fernández and Nakamura, 2015).

Cuantificar el esfuerzo de muestreo constituye un problema particularmente complejo en el estudio de la extinción. No se trata de una covariable en el mismo sentido que ha sido referido para implementar el modelo de regresión de Cox postulado. Aunque sí existiera una manera clara de cuantificar el esfuerzo de muestreo, no se puede incorporar directamente como covariable ya que no es un factor influyente en la longevidad de una especie, sino una característica del proceso implementado por paleobiólogos para buscar sus restos fósiles. Incorporar el esfuerzo de muestreo constituye uno de los retos más difíciles y sería importante crear un mecanismo en el que se tome en cuenta ya que es uno de los factores que aporta más incertidumbre al análisis.

Desde el punto de vista técnico, el problema puede identificarse como un problema denominado *encounter sampling*, o muestreo encontrado (o circunstancial) (Patil, 1991; Patil and Rao, 1978), en el sentido de que el paleobiólogo se limita a muestrear capas geológicas que le son asequibles en términos de accesibilidad, costo y tiempo, registrando fósiles que circunstancialmente se va encontrando. Si X con densidad $f(x)$ representa la longevidad de un fósil, y $w(x)$ representa la probabilidad de encontrar o registrar un fósil de longevidad x, entonces lo que uno realmente está observando es una densidad mezcla dada por

$$f^w(x) = \frac{w(x)f(x)}{\int w(u)f(u)\,du}.$$

Esto es, en la práctica uno cuenta con observaciones de la densidad f^w cuando uno realmente está interesado en la densidad $f(x)$. La componente de búsqueda por parte de humanos se refleja en $w(x)$, mientras que el objeto de interés biológico radica en $f(x)$. El reto técnico es primero, identificar un *proxy* para esfuerzo de muestreo que pueda representarse a través de algún $w(x)$, y segundo, incorporar ese concepto de muestreo encontrado en el modelo de regresión de Cox para separar los efectos de $w(x)$. En Benton (2009) se plantea que el estudio de ciertos mapas de geología se podría utilizar como un *proxy* para incorporar el esfuerzo de muestreo y sería relevante intentar utilizar esta idea en un trabajo futuro. La idea es que las capas geológicas más fácilmente accesibles serían las que naturalmente, han sido muestreadas con mayor intensidad.

Por aparte de esfuerzo de muestreo, también es posible que haya otras variables dignas de ser consideradas como covariables en el planteamiento de regresión de Cox. Por ejemplo, se ha asumido durante mucho tiempo que un mayor tamaño corporal se correlaciona en gran manera con un mayor riesgo de extinción. En Payne and Heim (2020) se estudia la relación entre la masa de las especies y su extinción. Resultaría interesante incorporar como covariable masa de la especie al modelo propuesto a fin de estudiar su influencia en la extinción, dato que no se encuentra registrado más que para un subconjunto muy pequeño del Compendio de Sepkoski.

Ejemplares únicos (*singletons*)

Los ejemplares que sólo han aparecido una sola vez en el registro fósil *de facto* fueron desechados para fines del análisis aquí presentado. Las especies *singletons* tendrían longevidad 0 y por lo tanto no son susceptibles de ser incorporadas en un estudio de supervivencia. Esto equivale a decir que el análisis realizado es condicional a que la longevidad es positiva. Sin embargo, el número de *singletons* es considerable (15 849 en el Compendio de Sepkoski) y alguna información deberían ser capaces de impartir en sí mismos. Si bien el modelo de Cox está preparado para albergar de manera sencilla censura por la derecha como se ha hecho ya (especies que hoy día permanecen existentes), la censura doble por izquierda y por derecha es un asunto de mayor complejidad.

Si acaso la ocurrencia misma de *singletons* tuviera algún significado, algo que pudiera implementarse fácilmente sería regresión logística para modelar su probabilidad de ocurrencia como función de covariables y tiempo geológico. Sin embargo, esto plantearía una pregunta al margen del proceso de extinción. En el lenguaje de esfuerzo de muestreo y muestreo encontrado, también podría resultar que los *singletons* correspondan a especies sumamente difíciles de encontrar, esto es, que su $w(x)$ en la notación del punto anterior sea minúscula y que por ello no han sido encontradas más que una vez. Es decir, es posible que informen sobre la naturaleza de $w(x)$, e indirectamente sobre la distribución de la longevidad.

La forma funcional del *fitness*

El *fitness* fue introducido como concepto para albergar cualitativamente la idea de que al transcurrir el tiempo, la aptitud de supervivencia mejora. Para fines meramente experimentales, en el trabajo se asumió una forma funcional específica que es monótona ($\exp\{-0.01T+7\}$) que no obedeció a justificación biológica alguna. Posteriormente al haber realizado este ejercicio, se encontró un trabajo (Sibani et al., 1995) que sugiere que hay razones para postular con una combinación de argumentos teóricos y empíricos que la aptitud crece como una potencia de T. Siendo esto así, el parámetro de potencia podría formularse como un parámetro p a estimar introduciendo como covariable $\log(T)$ de manera que en el predictor lineal del modelo de Cox figure el término $p\log(T)$. No sólo esta función concreta podría incorporarse de esta manera, sino también otras familias parametrizadas de funciones monótonas que se pudieran conjeturar.

3.3. Pertinencia para posteriores análisis sobre la sexta extinción masiva

El propósito primordial del trabajo en el Capítulo 2 ha sido caracterizar la distribución de longevidad condicionada a covariables y tiempo geológico, T, donde T corre de 0 a 540 MDA. Esto representa el comportamiento natural de la extinción en el planeta Tierra. Como habíamos mencionado en la introducción, durante los últimos años se ha formulado la hipótesis de que estamos actualmente en camino a una sexta extinción masiva, posiblemente detonada por la misma actividad humana.

Para abordar esta pregunta, ulteriormente habría que comparar con las tasas de extinción contemporáneas, o bien con la distribución de longevidades de especies que hoy existen o que muy recientemente se han extinguido. La mayor de las dificultades aparentes tiene que ver con un escalamiento del tiempo por un factor 10^6: la actualidad representa datos de extinción durante los últimos 500 años, mientras que la extinción en escala geológica abarca 500 millones de años. Más aun, la resolución con la cual figuran las fechas de primeros y últimos avistamientos para cada especie de fósil es mucho más burda que 500 años.

Estamos interesados en comparar distribuciones de fenómenos que están ocurriendo a dos escalas de tiempo órdenes de magnitud diferentes entre sí. Esto parecería sugerir, o que se extrapolan patrones de supervivencia de la actualidad a 500 millones de años, o que se interpolan patrones históricos de supervivencia a 500 años.

No vemos otra manera para lograr esto de no ser por tomar en cuenta el grado de imprecisión que tienen los fechamientos en el registro fósil. Un primer intento podría radicar en simulación de Monte Carlo, utilizando distribuciones de longevidad que la regresión de Cox ha estipulado, incorporando micro variación a escala de 500 años en la escala geológica. Se obtendrían así longevidades comparables

con la actualidad, en el sentido de ocurrir a la misma resolución de tiempo.

Existe un catálogo mundial de especies marinas (`WoRMS`) en `http://www.marinespecies.org/index.php` que proporciona una lista autorizada y completa de nombres de organismos marinos con el cual se puede acceder a información sobre especies actuales existentes y especies que aún no se habían extinguido según la información disponible en la base de datos de Sepkoski. Ello permitiría determinar cuáles y cuántas especies se han extinguido durante los últimos 500 años. Constituiría con ello una muestra de longevidades asociadas a la época presente.

Para el presente, presuntamente podrían conocerse los valores de las covariables involucradas en la modelación del Capítulo 2. El análisis estadístico concluiría con una prueba de hipótesis formal comparando dos distribuciones predichas con valores de covariables por vía de un modelo de regresión: alguna del pasado geológico reciente, y la distribución actual. Es presumible que el número de especies extintas en los últimos 500 años sea órdenes de magnitud menor al que se contabiliza como extintas en el Compendio de Sepkoski, pero una prueba de significancia estadística formal debería tomar esto en cuenta como parte intrínseca de su evaluación. La lógica indicaría a todas luces que las longevidades de las especies en la actualidad están recibiendo grandes embates como consecuencia de actividades humanas (sobreexplotación, destrucción de hábitats, cambio climático). Sin embargo, el asunto de fondo en el análisis de datos es examinar si la evidencia en términos de datos duros así lo determina. Si la incertidumbre para extraer esa conclusión es enorme, por el mismo análisis podría entenderse qué datos será necesario recolectar en el futuro para poder monitorear las tasas de extinción en el planeta con mayor grado de certeza.

Bibliografía

Alroy, J. (2010a). Fair Sampling of Taxonomic Richness and Unbiased Estimation of Origination and Extinction Rates. *The Paleontological Society Papers*, 16:55–80.

Alroy, J. (2010b). The Shifting Balance of Diversity Among Major Marine Animal Groups. *Science*, 329(5996):1191–1194.

Benton, M. (2009). The completeness of the fossil record. *Significance*, 6(3):117–121.

Bokulich, A. (2018). Using models to correct data: paleodiversity and the fossil record. *Synthese*, 1(1).

Chao, A. and Jost, L. (2012). Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, 93(12):2533–2547.

Collett, D. (2015). *Modelling survival data in medical research*. Chapman & Hall/CRC texts in statistical science series. CRC Press, Taylor & Francis Group, Boca Raton, third edition edition.

Dirzo, R. and Raven, P. H. (2003). Global State of Biodiversity and Loss. *Annual Review of Environment and Resources*, 28(1):137–167.

Fernández, D. and Nakamura, M. (2015). Estimation of spatial sampling effort based on presence-only data and accessibility. *Ecological Modelling*, 299:147–155.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526.

Harrison, C. G. A. (2002). Power spectrum of sea level change over fifteen decades of frequency: POWER SPECTRUM OF SEA LEVEL CHANGE. *Geochemistry, Geophysics, Geosystems*, 3(8):1–17.

Kleinbaum, D. G. and Klein, M. (2005). *Survival analysis: a self-learning text*. Statistics for biology and health. Springer, New York, NY, 2nd ed edition.

Kocsis, A. T., Reddin, C. J., Alroy, J., and Kiessling, W. (2019). The R package divDyn for quantifying diversity dynamics using fossil sampling data. *Methods in Ecology and Evolution*, 10(5):735–743.

Macleod, N. (2004). Identifying Phanerozoic extinction controls: statistical considerations and preliminary results. *Geological Society, London, Special Publications*, 230(1):11–33.

Patil, G. and Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, 34:179.

Patil, G. P. (1991). Encountered data, statistical ecology, environmental statistics, and weighted distribution methods. *Environmetrics*, 2(4):377–423.

Payne, J. L. and Heim, N. A. (2020). Body size, sampling completeness, and extinction risk in the marine fossil record. *Paleobiology*, 46(1):23–40.

Raup, D. (1986). Biological extinction in earth history. *Science*, 231(4745):1528–1533.

Rohde, R. A. and Muller, R. A. (2005). Cycles in fossil diversity. *Nature*, 434(7030):208–210.

Sepkoski, J. J. (2002). *A compendium of marine fossil genera*. Paleontological Research Institution, Ithaca, NY. OCLC: 494278457.

Sibani, P., Schmidt, M. R., and Alstrøm, P. (1995). Fitness Optimization and Decay of Extinction Rate Through Biological Evolution. *Physical Review Letters*, 75(10):2055–2058.

Smith, J. M. (1978). Optimization Theory in Evolution. *Annual Review of Ecology and Systematics*, 9(1):31–56.

Tapanila, L. (2007). FossilPlot, an Excel-based Computer Application for Teaching Stratigraphic Paleontology Using the Sepkoski Compendium of Fossil Marine Genera. *Journal of Geoscience Education*, 55(2):133–137.

Therneau, Crowson, A. (2021). Using time dependent covariates and time dependent coefficients in the cox model. `https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf`.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer New York, New York, NY.