Reglas de asociación en ataques de tiburones

Franco Betteo, Francisco Valentini

**Resumen.** En el presente trabajo se realiza un análisis de ataques de tiburones a humanos a partir de una base de casos históricos ocurridos a lo largo del mundo. El objetivo general es encontrar reglas de asociación entre los distintos atributos que definen a cada ataque de modo de identificar patrones interesantes y no triviales. Para tal fin, los datos fueron sometidos a un análisis exploratorio, a tareas de preprocesamiento y por último al algoritmo apriori para encontrar itemsets frecuentes y así generar reglas de asociación. Se utiliza el lift como métrica de evaluación principal.

**1 Introducción**

En el presente trabajo se realiza un análisis de ataques de tiburones a humanos a partir de un dataset cuya fuente original es *sharkattackfile.net* y que recolecta casos históricos a lo largo del mundo y de la historia con atributos que los definen. El objetivo general es encontrar reglas de asociación entre distintas variables relevantes del problema y así descubrir coocurrencias de factores interesantes y no triviales. El análisis presentado está conformado por las siguientes etapas:

* Selección de los atributos cuya información y formato resultan útiles para la aplicación de reglas de asociación.
* Preprocesamiento de los datos y construcción de atributos nuevos.
* Análisis exploratorio de las variables, tanto de manera individual como conjunta.
* Definición de dos configuraciones posibles para el set de datos a analizar.
* Aplicación del algoritmo apriori sobre ambas configuraciones, caracterización de las reglas obtenidas y selección de reglas consideradas robustas e interesantes.
* Identificación de características que hacen que un ataque sea fatal

**2 Materiales y Métodos**

El dataset consiste en 5.897 ataques de tiburones registrados a lo largo de la historia en distintas partes del mundo. Fueron omitidos del análisis aquellos casos muy antiguos o de dudosa ocurrencia (año de ocurrencia menor a 500 d.C.), de modo que se realizó el análisis exploratorio con los 5.772 casos restantes, con ataques acontecidos entre los años 1543 y 2016. Todos los datos fueron procesadas usando el lenguaje R [4]. Los gráficos y cuadros presentados se realizaron con las librerías ggplot2 [5], sjPlot [3] y arulesViz [1].

Múltiples atributos del dataset original fueron omitidos por su irrelevancia para el estudio de patrones de asociación o porque su alta granularidad volvió muy costoso el procesamiento para extraer información útil. Entre este último motivo se destacan atributos que indican tipo de lesión generada, la especie del tiburón involucrado y el distrito donde ocurrió el ataque. Asimismo, se decidió omitir la actividad realizada por las personas en el momento del ataque, debido a que el solapamiento entre un gran número de niveles de la variable impidió construir de forma eficiente niveles mutuamente excluyentes.

En base a la información restante disponible, se construyeron las siguientes variables:

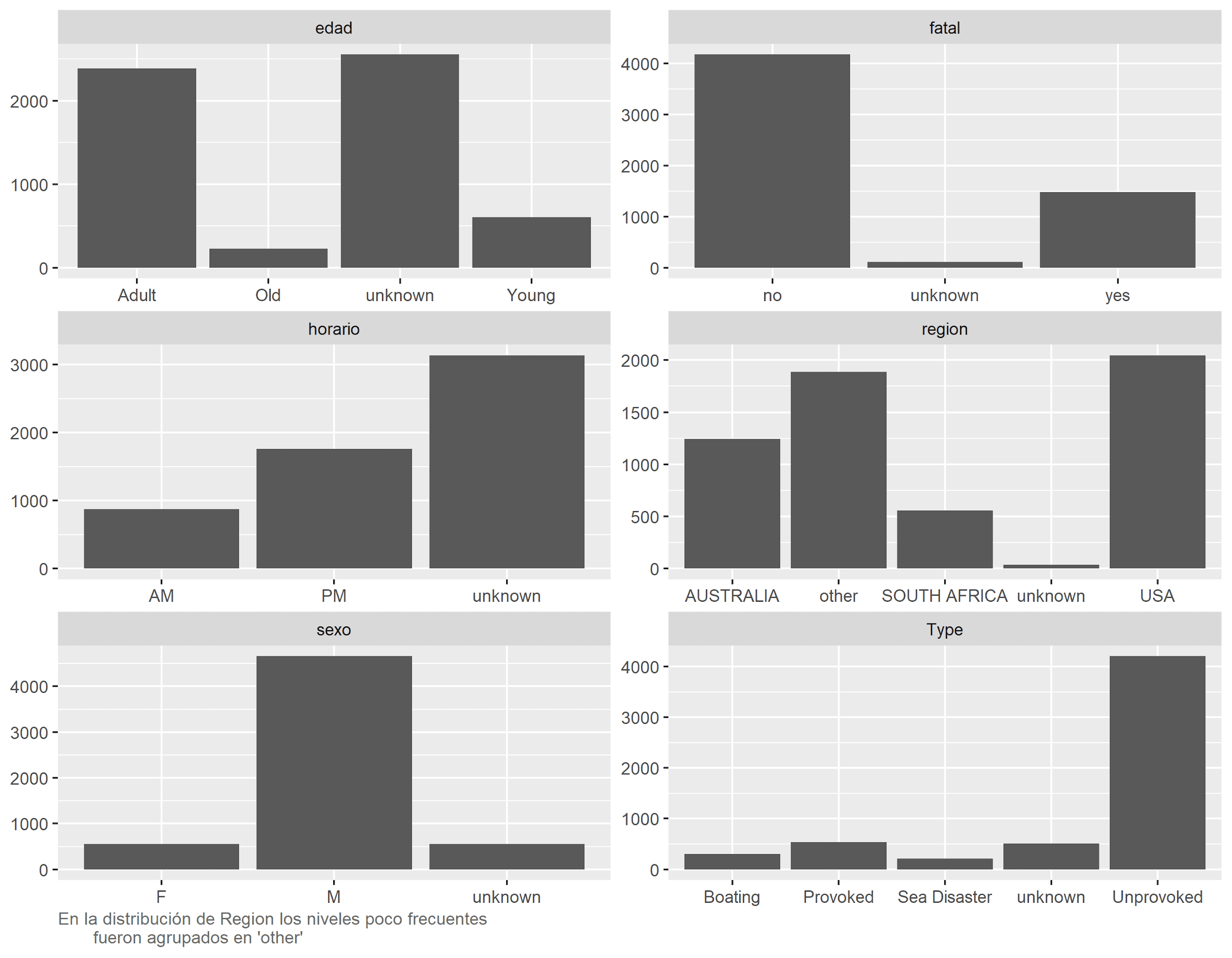
* Region. Se construyó seleccionando las 29 regiones con mayor frecuencia absoluta. Las regiones restantes fueron codificadas como “otros”, mientras que el nivel “unknown” indica la ausencia de la región de ocurrencia del ataque.
* Horario. En base a la información disponible acerca del horario -- cuya especificidad en la base original era muy variable -- se generó una variable de tres niveles: AM, PM y unknown. Debido al alto costo de procesamiento estimado, los niveles con muy baja frecuencia relativa fueron catalogados como “unknown”. Se estimó que la cantidad máxima de casos erróneamente etiquetados como unknown es 159.
* Edad. Se construyó siguiendo un criterio análogo al utilizado para la variable “horario”. Está constituida por cuatro niveles: Young (menores de 16 años), Adult (entre 16 y 50 años), Old (mayores de 50 años) y unknown (edad desconocida). Se estimó que la cantidad máxima de registros erróneamente etiquetados como “unknown” es 13.

Las variables finalmente utilizadas en el análisis se describen en el cuadro 1.

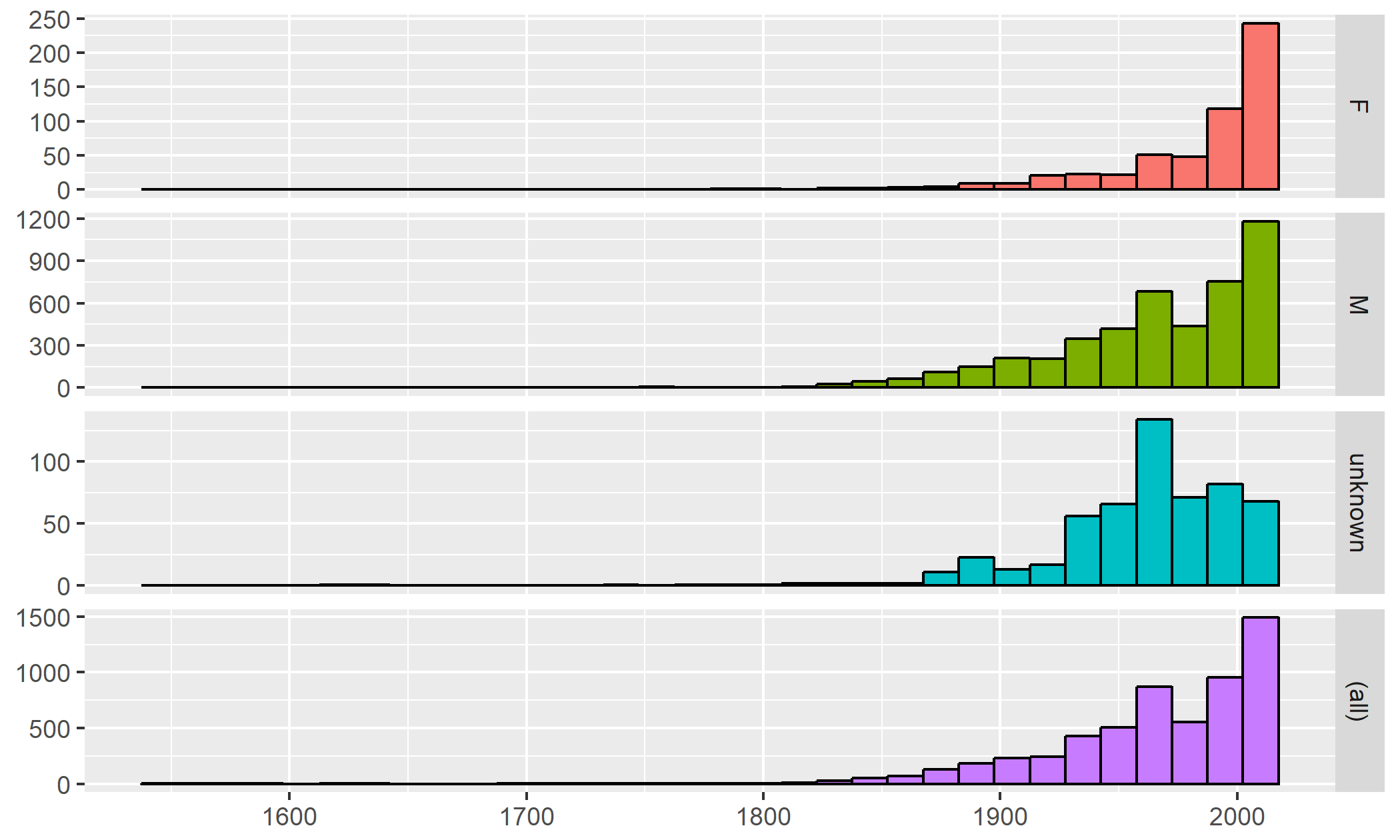
**Cuadro 1. Descripción de los atributos del set de datos**

|  |  |  |  |
| --- | --- | --- | --- |
| *Variable* | *Tipo* | *Descripción* | *Estadística* |
| Año | Numérica | Año de ocurrencia del episodio | Min. :1543 / 1st Qu.:1944 / Median :1975 / Mean :1966 / 3rd Qu.:2003 / Max. :2016 |
| Type | Nominal | Motivo del ataque | Boating: 302 / Unknown: 512 / Provoked: 539 / Sea Disaster: 214 / Unprovoked :4205 |
| sexo | Nominal | Género de la víctima | F: 557 / M :4660 / unknown: 555 |
| region | Nominal | Región de ocurrencia del ataque | AUSTRALIA :1246 / SOUTH AFRICA: 555 / unknown: 37 / USA :2046 / (otros):1888 |
| fatal | Nominal | El ataque fue fatal para la víctima | no: 4174 / unknown: 118 / yes:1480 |
| horario | Nominal | Horario del ataque | AM: 876 / PM :1758 / unknown:3138 |
| edad | Nominal | Edad de la víctima | Adult :2386 / Old: 228 / unknown:2553 / Young: 605 |

En la figura 1 se analiza la distribución de las variables categóricas, equivalente a la información indicada en la columna *Estadística* del cuadro 1.

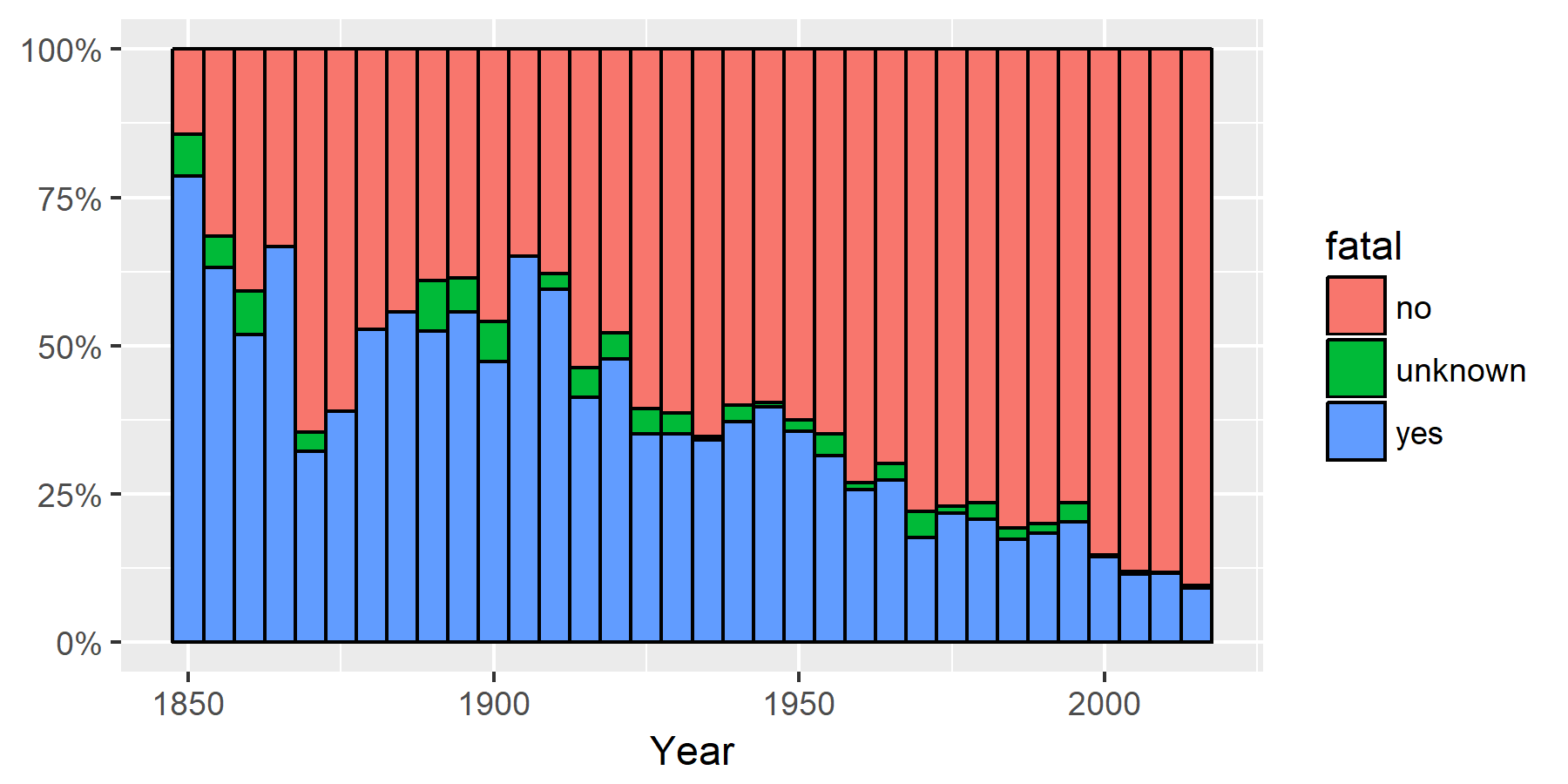
**Figura 1. Distribución de frecuencias de las variables nominales**

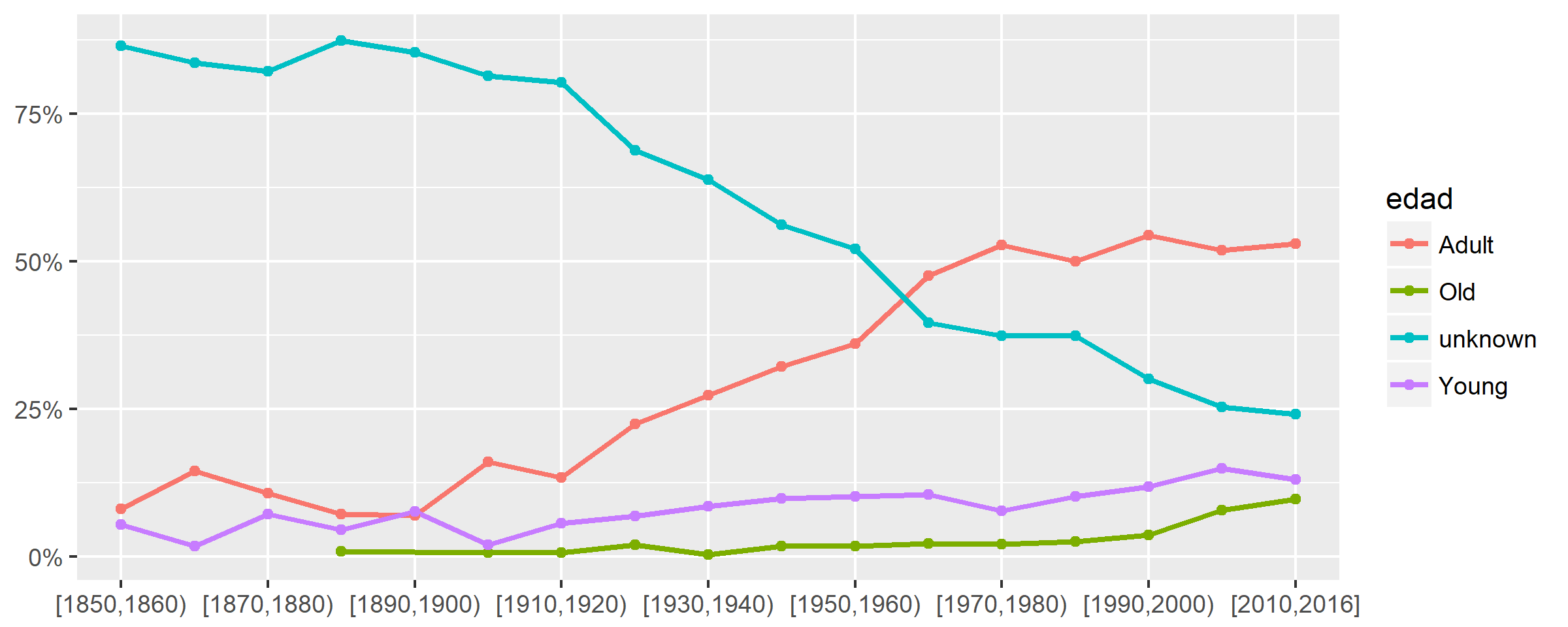
En las distribuciones de las variables edad y horario se destaca un marcado predominio de los valores desconocidos (“unknown”), llegando a representar más de 2.000 y 3.000 casos, respectivamente. En cambio, en fatal, region, sexo y type la relevancia es mucho menor. Omitiendo estos valores, se puede observar que la mayor parte de los casos ocurrieron en Estados Unidos, el rango etario mayoritario es “adulto”, la mayor parte de los ataques no fueron fatales, y las víctimas son en su gran mayoría hombres.

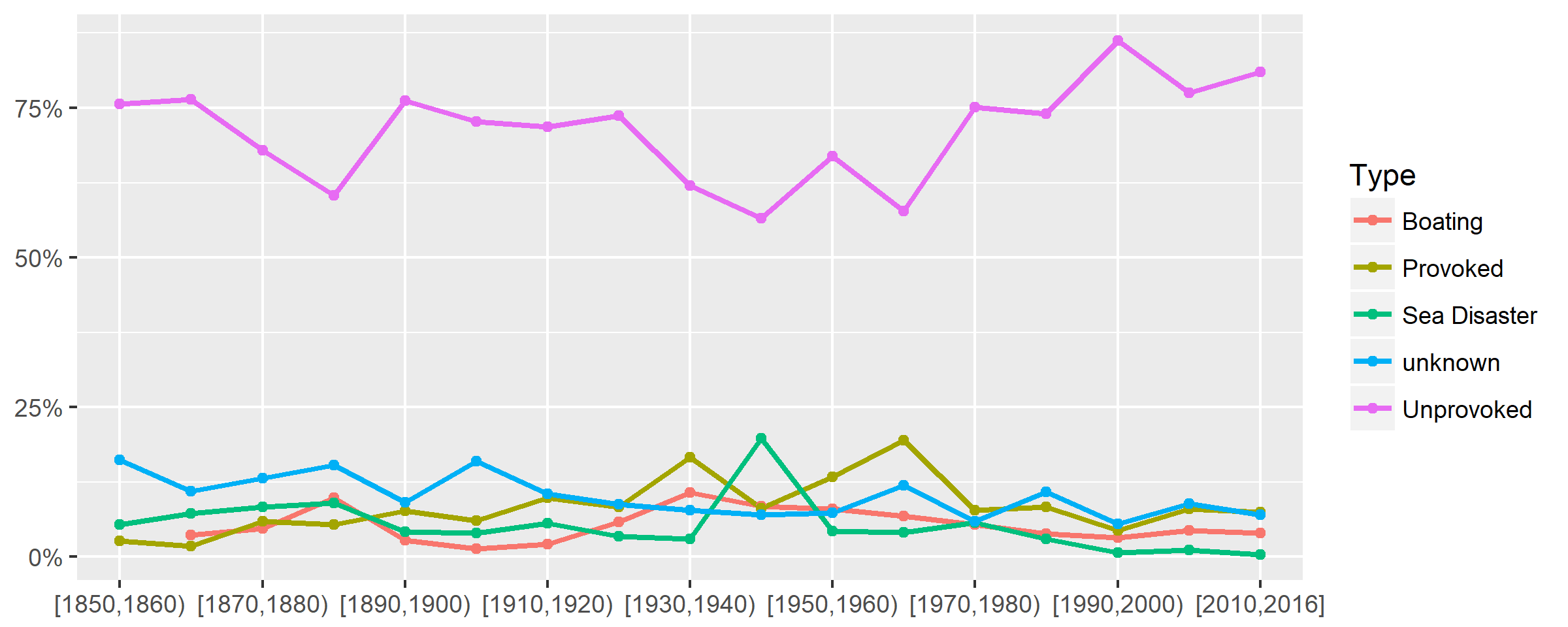
**Figura 2. Distribución de frecuencias -marginal y condicionada por sexo- de la variable Año (intervalos de 15 años)**

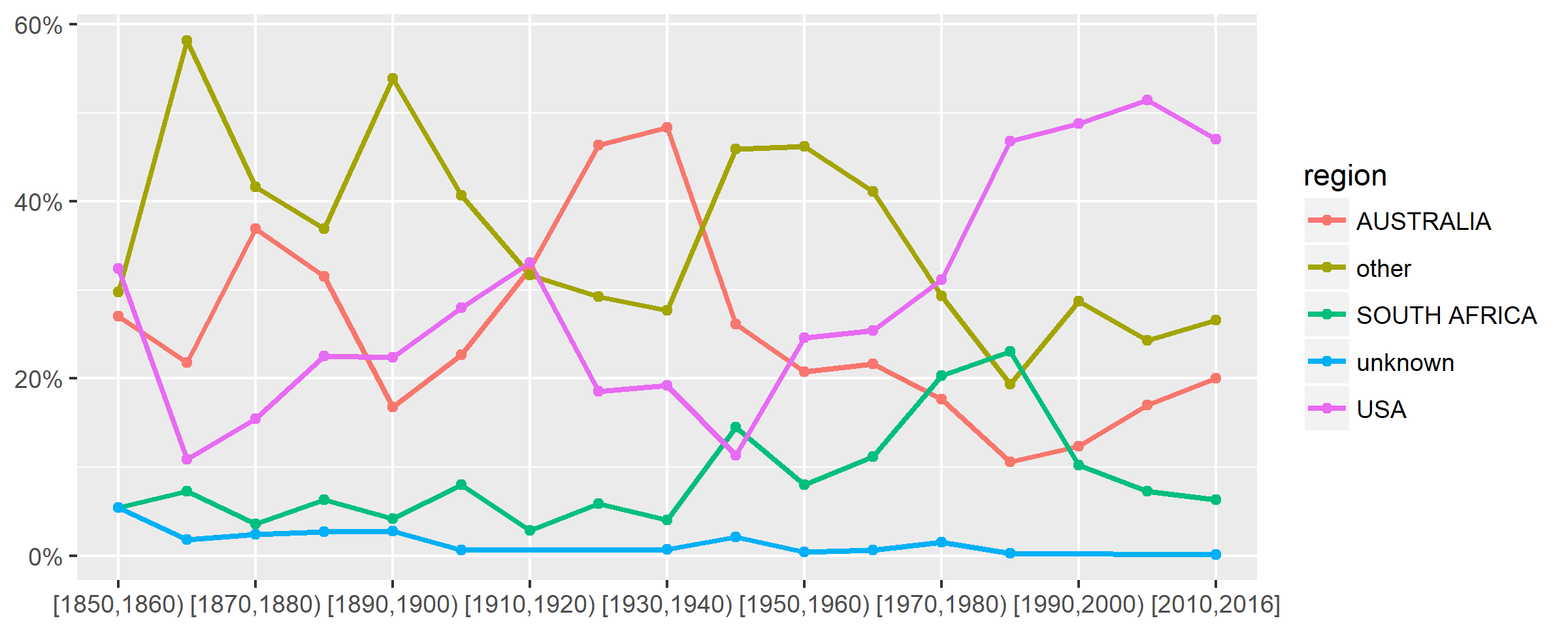
En la distribución de frecuencias de la variable año observada en el extremo inferior de la figura 2 se evidencia que el número de casos es creciente en el tiempo, registrando un pico local de ataques en el período 1959-1974 por episodios ocurridos en los años 1959 y 1960. En los años previos a 1850 la cantidad de casos por año es poco significativa. La distribución de los años condicionada por sexo muestra que la tendencia creciente se verifica para ambos sexos -- masculino y femenino -- a la vez que la cantidad de casos donde el género es desconocido se mantiene constante desde mediados del siglo XX, a excepción del pico registrado en 1959-1974.

**Figura 3. Casos fatales desde 1850 (como % de los casos totales)**

**(intervalos de 5 años)**

**Figura 4. Ataques según edad desde 1850 (como % de los casos totales) (intervalos de 10 años)**

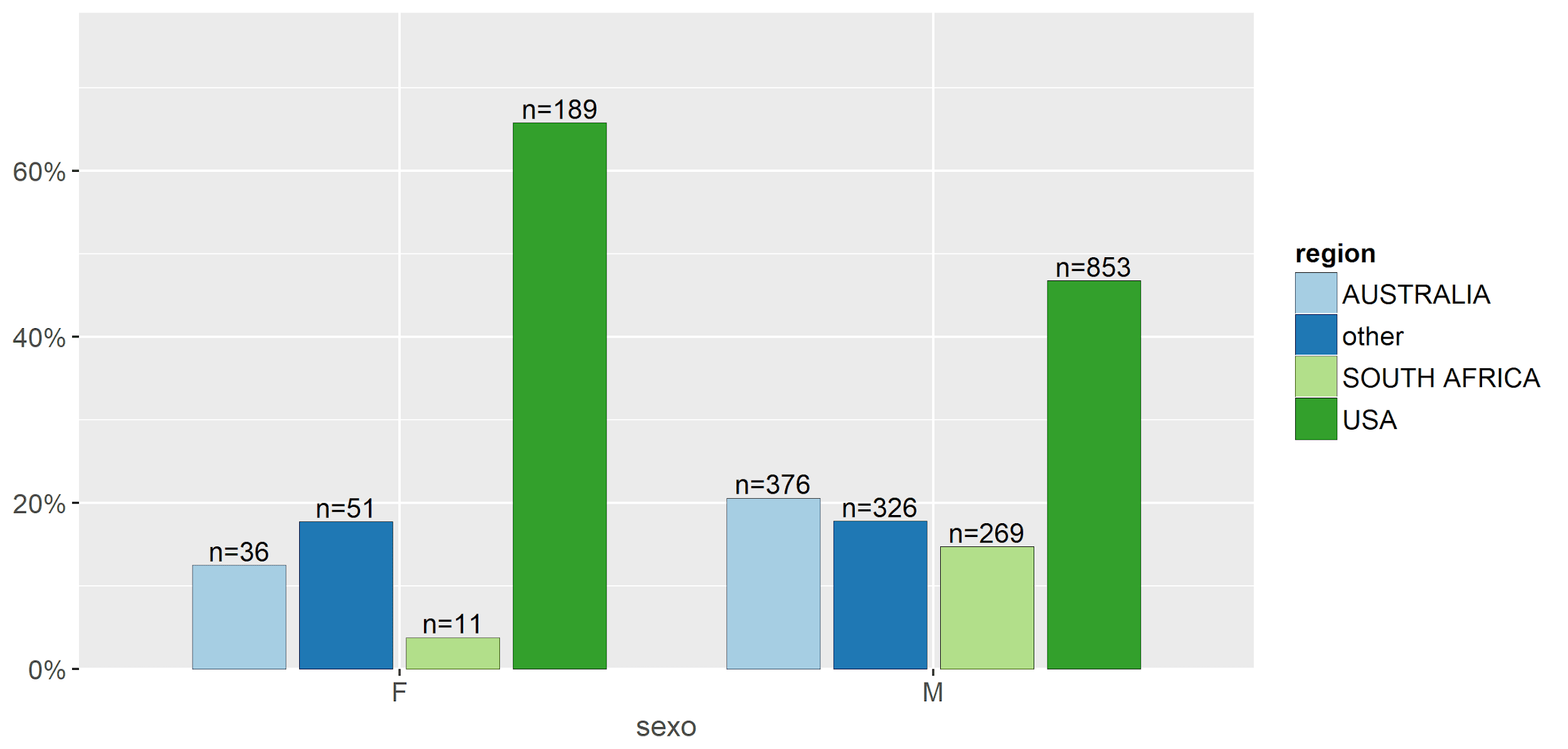
**Figura 5. Ataques según tipo desde 1850 (como % de los casos totales) (intervalos de 10 años)**

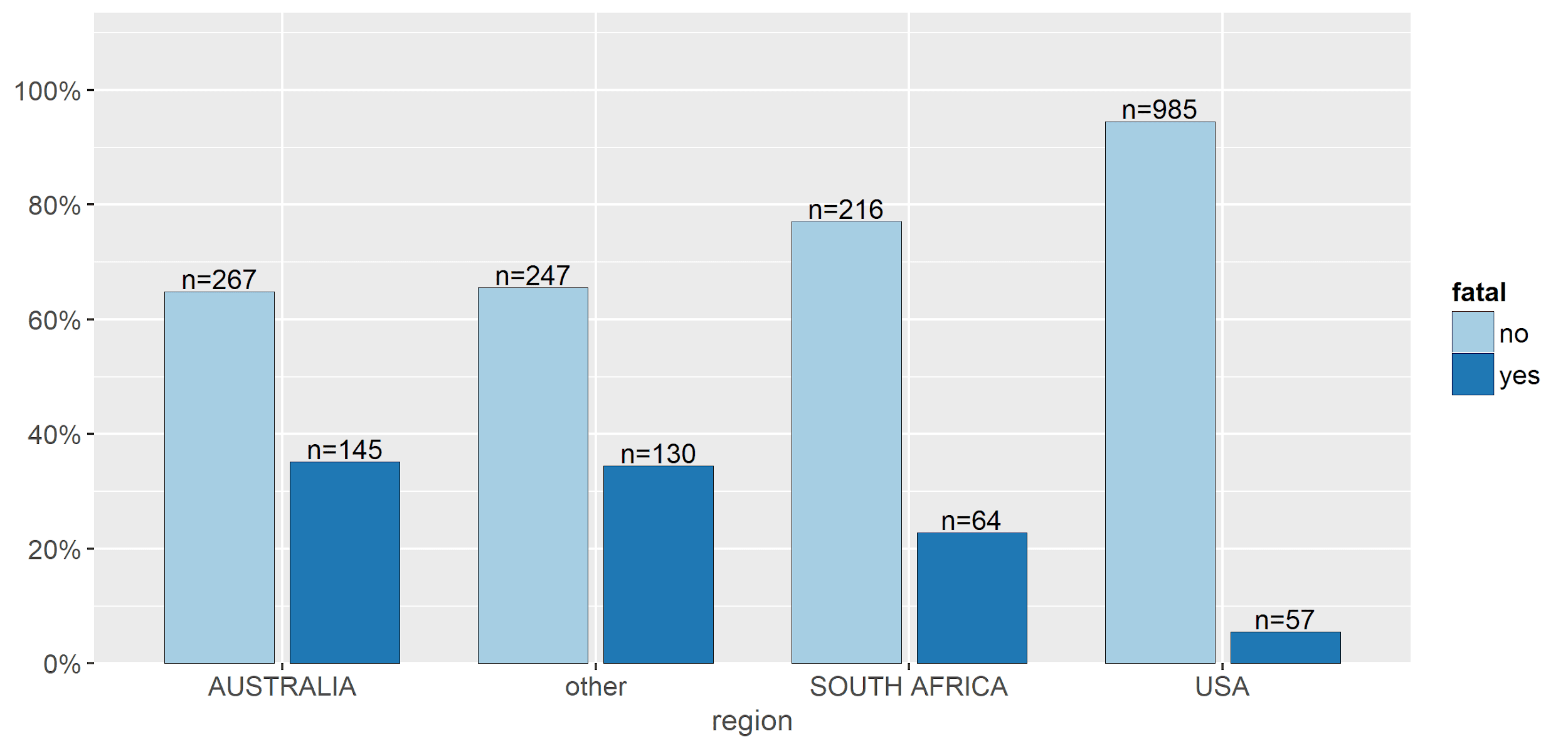
**Figura 6. Ataques según region desde 1850 (como % de los casos totales) (intervalos de 10 años)**

Las figuras 3, 4, 5 y 6 representan la evolución temporal de la frecuencia relativa de los ataques discriminados por fatalidad, edad, tipo y region. En todos los casos se excluyen los episodios previos a 1850 debido a la baja cantidad de casos y a la alta variabilidad de las proporciones estimadas. Del análisis de estas figuras se desprende que:

* La relevancia de los casos fatales tiene una tendencia claramente decreciente a lo largo del tiempo.
* La incidencia de los ataques es aproximadamente similar para jóvenes y adultos a principios de siglo XX. A partir de ese punto, la proporción de casos que involucran a adultos crece notablemente -- sin embargo, esto se explica por la caída de la importancia de los casos donde la edad es desconocida. Hipotetizamos que la incidencia en adultos es siempre más alta que la de jóvenes a lo largo del tiempo, pero en los episodios antiguos la información no estaba disponible.
* Los ataques de tipo “unprovoked” son predominantes a lo largo del tiempo. Lo que es más, ganan aún más peso en los últimos 60 años en detrimento del resto de los tipos -- “boating”, “sea disaster” y “provoked”. Se destaca un máximo local en la proporción de “Sea disaster” en el período 1940-1950.
* Durante los últimos 70 años los casos ocurridos en EEUU reemplazan a los ocurridos en Australia y otras regiones como origen principal de los ataques. Entre 1930 y 1990 se registra además un salto de nivel sensible en la proporción de episodios ocurridos en Sudáfrica, que retorna en los últimos años a lo niveles históricamente normales.
* En términos generales, la proporción de casos sin información disponible (“unknown”) es decreciente a lo largo del tiempo, pasando a ser despreciable en los atributos region y fatal.

Para estudiar las relaciones entre las variables categóricas del dataset se construyeron tablas de contingencia entre las 15 combinaciones posibles de pares de variables. En todos los casos se observó un resultado significativo al realizar la prueba chi-cuadrado (p-valor menor a 0.01). Sin embargo, esto se debe a la inclusión de los valores faltantes (“unknown”) de cada atributo. Al omitir las observaciones faltantes se registran solo dos combinaciones con ausencia de independencia estadísticamente significativa y perfiles claramente distintos entre grupos. En las figuras 7 y 8 se presentan estos resultados gráficamente, condicionando las frecuencias por la variable representada por el color. Las etiquetas de cada barra indican la frecuencia conjunta de cada combinación de niveles. Cabe destacar que la pérdida de información por usar solo los casos completos es significativa: sólo el 36.5% de los 5.722 casos originales cumplen con esta condición.

**Figura 7. Proporción de casos según region (condicionado por sexo)**

**Figura 8. Proporción de casos según fatalidad (condicionado por region)**

En primer lugar, la figura 7 muestra que la incidencia de los casos ocurridos en EEUU es significativamente mayor dentro del grupo de las mujeres que entre los hombres (65.9% vs 46.8%). En cambio, dentro los hombres la relevancia de Sudáfrica como origen es mayor que dentro de los episodios que involucran a mujeres (14.7% vs 3.8%). La figura 8, por su parte, indica que la probabilidad de que un ataque no sea fatal es sensiblemente mayor si ocurre en Estados Unidos que en el resto del mundo: el 5.5% de los casos de EEUU son fatales, guarismo inferior a los observados en Australia (35.2%), Sudáfrica (22.9%) y otros lugares del mundo (34.5%).

Para construir reglas de asociación acerca de las características de los ataques de tiburones, los niveles de las variables descritas en el cuadro 1 fueron considerados como ítems. Como se observa en dicho cuadro, la variable “año” es la única variable numérica, por lo cual debió ser tratada para estudiar los patrones de asociación relativos al año de ocurrencia de los ataques. En particular, la variable fue discretizada construyeron tres bins para los años anteriores a 1800 -- cada uno de un siglo de longitud -- y bins de 10 años de longitud para los años posteriores a 1800. La asimetría en el método elegido se debe al fuerte patrón asimétrico en esta variable, observado en la figura 2.

Para identificar itemsets frecuentes se utilizó el algoritmo apriori, particularmente la implementación disponible en el paquete arules [2]. Se consideraron “frecuentes” todos aquellos itemsets con soporte mayor o igual a 1%, mientras que solo se tuvieron en cuenta reglas con una confianza mayor o igual a 30%. El algoritmo fue aplicado a dos configuraciones del dataset:

* En la configuración “A” se tuvieron en cuenta todas las variables del cuadro 1
* En la configuración “B” se omitió la variable que indica el año y se consideraron únicamente los ataques del período 1990-2016. El objetivo es neutralizar el efecto del tiempo y de eventos históricos muy particulares en los patrones hallados. Asimismo, este un período con una alta frecuencia de casos y con una baja proporción de datos faltantes.

En ambas configuraciones no se consideraron las reglas que tuvieran presencia de valores “unknown” en cualquiera de las variables, en la medida en que refleja ignorancia de alguna/s característica/s del ataque. Asimismo, se eliminaron las reglas redundantes, es decir, todas aquellas reglas para las cuales existen reglas más generales -- con igual consecuente y un precedente más general -- con confianza al menos igual.

**3 Resultados obtenidos**

**Configuración A**

Bajo esta configuración del dataset se obtuvieron 1.198 reglas de asociación de acuerdo a los criterios especificados en la sección anterior. Su composición por tamaño -- cantidad de ítems en cada regla -- se registra en el cuadro 2:

**Cuadro 2. Cantidad de reglas según tamaño (configuración A)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *sizes* | 2 | 3 | 4 | 5 | 6 | 7 |
| *Frequency* | 164 | 390 | 414 | 182 | 35 | 2 |

En el cuadro 2 se observa que alrededor del 67% de las reglas generadas contiene entre 3 y 4 ítems, mientras que se destaca la dificultad para hallar reglas “grandes”: solo el 3% de las reglas tiene entre 6 y 7 items.

**Cuadro 3. Diez reglas con mayor soporte (configuración A)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *rules* | *support* | *confidence* | *lift* | *count* |
| {Type=Unprovoked} => {sexo=M} | 0.619 | 0.849 | 1.052 | 3572 |
| {sexo=M} => {Type=Unprovoked} | 0.619 | 0.767 | 1.052 | 3572 |
| {fatal=no} => {sexo=M} | 0.584 | 0.808 | 1.001 | 3373 |
| {sexo=M} => {fatal=no} | 0.584 | 0.724 | 1.001 | 3373 |
| {fatal=no} => {Type=Unprovoked} | 0.534 | 0.738 | 1.013 | 3080 |
| {Type=Unprovoked} => {fatal=no} | 0.534 | 0.732 | 1.013 | 3080 |
| {sexo=M,fatal=no} => {Type=Unprovoked} | 0.453 | 0.775 | 1.063 | 2613 |
| {edad=Adult} => {sexo=M} | 0.369 | 0.892 | 1.105 | 2128 |
| {sexo=M} => {edad=Adult} | 0.369 | 0.457 | 1.105 | 2128 |
| {edad=Adult} => {Type=Unprovoked} | 0.341 | 0.825 | 1.132 | 1968 |

**Cuadro 4. Diez reglas con mayor lift (configuración A)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *rules* | *support* | *confidence* | *lift* | *count* |
| {sexo=M,edad=Adult,year\_bins=[1980,1990)} => {region=SOUTH AFRICA} | 0.011 | 0.328 | 3.412 | 62 |
| {Type=Sea Disaster} => {region=other} | 0.013 | 0.36 | 3.361 | 77 |
| {Type=Unprovoked,sexo=F,region=USA,fatal=no} => {edad=Young} | 0.014 | 0.347 | 3.313 | 83 |
| {Type=Unprovoked,sexo=F,region=USA} => {edad=Young} | 0.015 | 0.335 | 3.193 | 85 |
| {sexo=F,region=USA,fatal=no} => {edad=Young} | 0.015 | 0.33 | 3.144 | 88 |
| {sexo=F,fatal=no,horario=PM} => {edad=Young} | 0.012 | 0.325 | 3.103 | 67 |
| {sexo=F,region=USA} => {edad=Young} | 0.016 | 0.307 | 2.931 | 90 |
| {sexo=F,horario=PM} => {edad=Young} | 0.012 | 0.304 | 2.904 | 70 |
| {fatal=no,edad=Old} => {year\_bins=[2010,2016]} | 0.011 | 0.356 | 2.746 | 62 |
| {Type=Unprovoked,edad=Old} => {year\_bins=[2010,2016]} | 0.011 | 0.356 | 2.746 | 62 |

En los cuadros 3 y 4 se registran las primeras diez reglas ordenadas por soporte y lift, respectivamente.

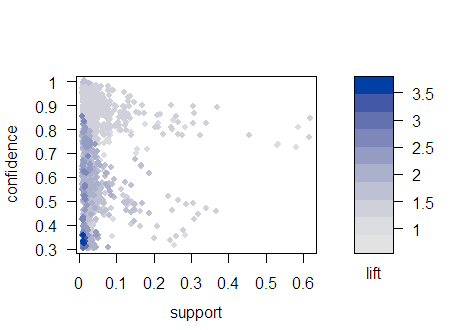
Si bien el ranking por soporte (cuadro 3) no es de gran ayuda para encontrar relaciones interesantes -- ya que solo indica frecuencia de cada regla en tanto itemsets --, sí es útil para ilustrar la composición de los datos. Las tres combinaciones de ítems más frecuentes en los ataques son tipo “unprovoked” y víctimas hombres, accidentes no fatales y a hombres, y accidentes no fatales y de tipo “unprovoked”, reuniendo cada uno de estos itemsets soportes de 61,9%, 58,4% y 53,4%, respectivamente. Desde luego, este resultado se corresponde con lo observado en la figura 1 del análisis exploratorio.

Para poder estudiar la coocurrencia de factores en un sentido condicional es posible usar una métrica de confianza o probabilidad condicional estimada. Sin embargo, esta medida no garantiza la generación de reglas robustas ya que es probable obtener reglas cuyas confianzas son altas “por casualidad”, es decir, únicamente porque sus consecuentes tienen un soporte alto. Por consiguiente, si bien las tablas presentadas incluyen la confianza de cada regla, evaluamos los resultados en términos de la métrica de lift.

En el cuadro 4 se encuentran las diez reglas con mayor lift. El lift es la relación entre el soporte observado de la regla y el que se observaría si antecedente y consecuente fueran independientes, de modo que un lift mayor que 1 indica dependencia entre precedente y consecuente. Esta medida es un mejor indicador de asociaciones no espurias o no triviales ya que tiene en cuenta el soporte de la coocurrencia de antecedente y consecuente controlado por la probabilidad de suceso de cada uno (de manera independiente). Cuánto mayor es el lift, mayor es la capacidad predictiva del precedente con respecto al consecuente.

La regla de mayor lift indica una fuerte dependencia entre ataques a hombres adultos en el período 1980-1990 y que la región del ataque sea Sudáfrica. Dado que estamos en presencia de un lift alto podemos suponer que en la región hubo particularmente más ataques que lo esperado para hombres adultos durante esa década comparado con otras regiones y con otros períodos de la historia. El resultado puede deberse a la presencia de algún episodio particular ocurrido en la región durante la época.

Las reglas 3, 4, 5, 6, 7 y 8 del cuadro 4 indican en su conjunto patrones relevantes para determinar las características de los ataques en los que la víctima es joven: en términos generales, estos ataques ocurren en EEUU, no son provocados, no fueron fatales y afectan a las mujeres. Es llamativo que este tipo de ataques sucedan con mayor probabilidad en Estados Unidos que el resto de las regiones y nos permite suponer que hay mayor exposición a situaciones potencialmente peligrosas para niños y/o mujeres.

**Figura 9. Reglas según soporte, confianza y lift (configuración A)**

En la figura 9 se observa que las reglas no triviales presentadas en el cuadro 5 -- con lift alto -- tienden a tener bajo soporte y baja confianza con relación al resto de las reglas generadas. En otras palabras, estas reglas tienen un poder predictivo alto, pero son aplicables a pocos casos del dataset. Como se observa en la figura, ninguna regla tiene simultáneamente lift alto y soporte alto.

**Cuadro 5. Diez reglas con mayor lift y “fatal=yes” en el consecuente (configuración A)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *rules* | *support* | *confidence* | *lift* | *count* |
| {Type=Sea Disaster} => {fatal=yes} | 0.025 | 0.687 | 2.679 | 147 |
| {Type=Unprovoked,year\_bins=[1900,1910)} => {fatal=yes} | 0.012 | 0.624 | 2.433 | 68 |
| {year\_bins=[1900,1910)} => {fatal=yes} | 0.016 | 0.613 | 2.392 | 92 |
| {Type=Unprovoked,year\_bins=[1890,1900)} => {fatal=yes} | 0.01 | 0.55 | 2.147 | 60 |
| {year\_bins=[1880,1890)} => {fatal=yes} | 0.011 | 0.55 | 2.143 | 61 |
| {sexo=M,year\_bins=[1890,1900)} => {fatal=yes} | 0.011 | 0.529 | 2.063 | 64 |
| {year\_bins=[1890,1900)} => {fatal=yes} | 0.013 | 0.524 | 2.045 | 75 |
| {year\_bins=[1910,1920)} => {fatal=yes} | 0.011 | 0.437 | 1.703 | 62 |
| {Type=Unprovoked,sexo=M,region=AUSTRALIA,horario=PM,edad=Adult} => {fatal=yes} | 0.012 | 0.406 | 1.584 | 67 |
| {Type=Unprovoked,region=AUSTRALIA,horario=PM,edad=Adult} => {fatal=yes} | 0.012 | 0.404 | 1.578 | 72 |

En el cuadro 6 se observa que el principal factor asociado eventos fatales -- evaluado según lift -- es “Sea Disaster”. Este resultado confirma la intuición de que ante una emergencia marítima los medios para evitar un ataque -- o salvarse luego de uno -- son mucho más limitados que en situaciones normales.

Gran parte de las reglas restantes en las que el consecuente es fatal y el lift es alto incluye al período de ocurrencia en su precedente. En particular, los ataques de fines del siglo XIX y principio del XX tuvieron mayor tasa de fatalidad que lo que se esperaría dadas sus probabilidades individuales. Esto mismo se ve reflejado en la figura 3 del análisis exploratorio, donde se observa que la tendencia de casos fatales es a la baja a lo largo de la historia pero en el período mencionado hay un repunte pronunciado.

Por último, se destaca con un lift de 1.58 la relación entre ataques fatales y ataques a hombres adultos en Australia luego del mediodía, independientemente de la época. Sería interesante estudiar la relación entre este resultado y otros atributos relevantes no considerados, como las especies de tiburones de la región o las actividades desarrolladas.

**Configuración B**

Para la configuración B del dataset -- casos entre 1990 y 2016 -- se generaron 413 reglas con los criterios especificados previamente. Se presentan las principales reglas según soporte y lift en los cuadros 6 y 7, respectivamente.

**Cuadro 6. Diez reglas con mayor soporte (configuración B)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *rules* | *support* | *confidence* | *lift* | *count* |
| {Type=Unprovoked} => {fatal=no} | 0.703 | 0.872 | 1.013 | 1615 |
| {fatal=no} => {Type=Unprovoked} | 0.703 | 0.816 | 1.013 | 1615 |
| {sexo=M} => {fatal=no} | 0.677 | 0.859 | 0.998 | 1555 |
| {fatal=no} => {sexo=M} | 0.677 | 0.786 | 0.998 | 1555 |
| {sexo=M} => {Type=Unprovoked} | 0.648 | 0.822 | 1.02 | 1488 |
| {Type=Unprovoked} => {sexo=M} | 0.648 | 0.803 | 1.02 | 1488 |
| {Type=Unprovoked,sexo=M} => {fatal=no} | 0.565 | 0.873 | 1.014 | 1299 |
| {sexo=M,fatal=no} => {Type=Unprovoked} | 0.565 | 0.835 | 1.036 | 1299 |
| {Type=Unprovoked,fatal=no} => {sexo=M} | 0.565 | 0.804 | 1.021 | 1299 |
| {region=USA} => {fatal=no} | 0.47 | 0.953 | 1.107 | 1080 |

**Cuadro 7. Diez reglas con mayor lift (configuración B)**

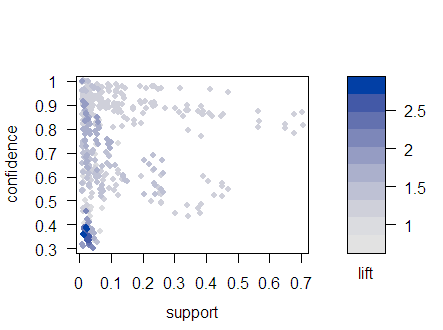
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *rules* | *support* | *confidence* | *lift* | *count* |
| {Type=Unprovoked,sexo=F,region=USA,fatal=no,horario=PM} => {edad=Young} | 0.02 | 0.39 | 2.826 | 46 |
| {sexo=F,region=USA,fatal=no,horario=PM} => {edad=Young} | 0.021 | 0.386 | 2.797 | 49 |
| {Type=Unprovoked,sexo=F,region=USA,horario=PM} => {edad=Young} | 0.02 | 0.383 | 2.779 | 46 |
| {sexo=F,region=USA,horario=PM} => {edad=Young} | 0.021 | 0.38 | 2.754 | 49 |
| {sexo=M,region=BRAZIL} => {fatal=yes} | 0.012 | 0.359 | 2.723 | 28 |
| {region=BRAZIL} => {fatal=yes} | 0.014 | 0.356 | 2.697 | 32 |
| {Type=Unprovoked,sexo=F,region=USA,fatal=no} => {edad=Young} | 0.029 | 0.351 | 2.543 | 67 |
| {sexo=F,fatal=no,horario=PM} => {edad=Young} | 0.025 | 0.35 | 2.535 | 57 |
| {Type=Unprovoked,sexo=F,region=USA} => {edad=Young} | 0.03 | 0.343 | 2.49 | 68 |
| {sexo=F,horario=PM} => {edad=Young} | 0.026 | 0.337 | 2.444 | 59 |

Como se observa en el cuadro 6, las reglas de mayor soporte son similares a las obtenidas bajo la configuración A. Los itemsets compuestos por ataques no fatales, no provocados y/o dirigidos a hombres son los más frecuentes. Las posibles combinaciones de a pares de los mismos registran alrededor de un 75% frecuencia relativa, como se observa en las primeras 6 reglas del cuadro. Se observa a su vez que las reglas que contienen a este itemset registran una confianza relativamente alta -- superior al 80%. Sin embargo, su lift se encuentra alrededor de 1, lo cual indica que probablemente las asociaciones implicadas sean más bien casuales, originadas por el alto soporte de los consecuentes.

Analizando el cuadro 8 se observa que las primeras cuatro reglas con mayor lift indican que la probabilidad de que una víctima sea joven es relativamente alta cuando los ataques son no provocados, no fatales, en EEUU, dirigidos a niñas y/o después del mediodía. Estas reglas también registraban alto lift en la configuración A, lo cual ilustra su robustez.

Por otra parte, se observa otro set de reglas interesantes (reglas 5 y 6 en el cuadro 9) que indican que los ataques en Brasil y dirigidos a los hombres tienen una tasa de fatalidad significativamente alta. Más allá de la baja frecuencia relativa de este itemset (1,4%), los ataques de tiburones que causan la muerte de la víctima en Brasil son más frecuentes de lo que esperado dadas las probabilidades individuales del antecedente y consecuente por separado.

Lo que es más interesante es que las únicas dos reglas generadas según los criterios de este trabajo para la configuración B que incluyen “fatal” en el consecuente son las mismas reglas descritas en el párrafo inmediatamente anterior. Esto refuerza aún más la observación de que en los últimos 25 años la cantidad de accidentes en Brasil es atípicamente alta, y en particular para la población masculina. Sería interesante analizar las características de la costa brasileña, las actividades que allí se desarrollan y de las distintas especies de tiburones que habitan en su cercanía ya que pueden ser los factores que causan esta asociación fuerte.

**Figura 10. Reglas según soporte, confianza y lift (configuración B)**

En la figura 10 vemos que el patrón de reglas es muy similar al de la figura 9, confeccionado con la configuración A del dataset: cuando las reglas tienen lift alto, el soporte es bajo. Por otra parte, en este caso la confianza es un poco más alta para las reglas más robustas que lo observado en la configuración anterior -- de todos modos sus valores apenas superan el 50%.

**Referencias**

[1] Hahsler, M. (2017). *arulesViz: Interactive Visualization of Association Rules with R.* The R Journal Vol 9/2.

[2] Hahsler, M. et al (2018) *arules: Mining Association Rules and Frequent Itemsets*. [Online] Disponible en: https://cran.r-project.org/web/packages/arules/index.html.

[3] Lüdecke, D. & Schwemmer, C. (2018). *sjPlot: Data Visualization for Statistics in Social Science*. [Online]. Disponible en: https://cran.r-project.org/web/packages/sjPlot/index.html.

[4] R Core Team (2017). R: A Language and Environment for Statistical Computing.

[5] Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.