Comportamiento de los procesos de baja laboral Modelos probabilísticos

Behaviour of processes of sick leaves Probabilistic Models

Alexandra García Lorenzo

Trabajo de Fin de Grado en Matemáticas

Facultad de Ciencias. Sección de Matemáticas

Universidad de La Laguna

La Laguna, 16 de septiembre de 2014

El Dr. D. **Carlos González Alcón**, con N.I.F. 50.821.751-P profesor Titular de Universidad adscrito al Departamento Matemáticas, Estadística e Investigación Operativa. de la Universidad de La Laguna

C E R T I F I C A

Que la presente memoria titulada:

"*Comportamiento de los procesos de baja laborales*"

ha sido realizada bajo su dirección por D. **Alexandra García Lorenzo**, con N.I.F. 54.064.061-P.

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos firma la presente en La Laguna a 16 de septiembre de 2014

Agradecimientos

A mi familia, y en especial a mis padres por apoyarme siempre en los momentos difíciles.

Al tutor de este trabajo, pues sin él no hubiese sido posible la realización del mismo.

A la Mutua de Accidentes de Canarias, por facilitarme los datos con los que se ha elaborado este trabajo.

Resumen

El objetivo principal de este trabajo es desarrollar una serie de modelos que nos permitan entender mejor el comportamiento de los procesos de baja laboral. Para ello, contamos con una serie de datos proporcionados por la Mutua de Accidentes de Canarias (MAC), en los que se proporciona información sobre los episodios de baja de los trabajadores a los que MAC realizó seguimiento durante los años 2006–2013.

En una primera etapa, dividimos estos datos en dos grupos con el propósito de utilizar uno de ellos, formado por el 60% de los datos, para elaborar los modelos y el otro, con los datos restantes, para comprobar que los resultados obtenidos fuesen correctos. Una vez hecho esto, estudiamos cuales eran las enfermedades y los pares de enfermedades más comunes entre los trabajadores, utilizando para esto un análisis cluster.

Detectadas dichas enfermedades, elaboramos tres modelos apoyándonos en los procesos de cadenas de Markov. Sobre estos modelos se estudiaron aspectos como: la media y la distribución de la duración de los episodios de baja, la proporción de días que cada trabajador permanece de baja, etc. Estos modelos presentan distintos niveles de dificultad, los cuales varían dependiendo del número de estados que estemos considerando en las cadenas.

Finalmente y haciendo uso de los datos restantes, comprobamos la veracidad de los tres modelos desarrollados durante la elaboración de esta memoria.

Palabras Clave: Baja laboral, análisis cluster, cadenas de Markov, matriz de transición.

Abstract

The principal objective of this project is to develop a series of models which helps us to understand the behavior of the processes of work absence for illness better. In order to achieve our goal, we have used the data provided by MAC, Mutua de Accidentes de Canarias, (Benefit Society of Accidents of Canaries). In these data we could find information about the processes of work absence for illness which MAC monitored between the years 2006 and 2013.

Firstly, we divided this information into two groups. One of them, formed by the 60 % of the data, was used to develop the models and the other group, formed by the 40 % of the remaining information, was used to verify that the results we got were correct.

Secondly, we used a cluster analysis in order to establish which were the most frequent illnesses and pairs of illnesses among the workers.

Once we had determined these illnesses, we developed three models using the processes of Markov chains. On these models we studied different aspects such as the mean and the distribution of the duration of the episodes of work absence, the proportion of days on which each worker is absent of work, etc. These models present different levels of difficulty which change depending on the number of states we are considering in the chains.

Finally and using the remaining data, we check the veracity of the three models developed during the elaboration of this memory.

Keywords: *Sick leaves, cluster analysis, Markov chains, transition matrix.*

Índice general

Capítulo 1

Motivación y Objetivos

Como es conocido, una de las asignaturas que forman parte del último curso del Grado en Matemáticas es la de Prácticas Externas, la cual se debe realizar en una entidad colaboradora, ya sea en una empresa o en una institución. En mi caso, desarrollé esta asignatura en la Mutua de Accidentes de Canarias (MAC). Durante esas prácticas estuve realizando un estudio sobre las duraciones medias de los procesos de baja por incapacidad temporal, principal causa de ausencia al trabajo, de los pacientes a los que MAC realizó seguimiento durante los años 2010, 2011 y 2012. Tras comentarle esto al tutor de este trabajo de fin de grado, llegamos a la conclusión de que podía ser muy interesante aprovechar estos datos para realizar algún tipo de estudio complementario en el que se elaborara alguna clase de modelo probabilístico.

Nos fijamos como principal objetivo de este proyecto desarrollar una serie de modelos que nos permitieran obtener resultados más complejos a los obtenidos durante el periodo de prácticas, para entender mejor cómo se comportan los procesos de baja laboral. Antes de desarrollar estos modelos tenemos que tener en cuenta por un lado, el estudio realizado durante las prácticas en MAC, pues este nos aportará información acerca de cuáles son las enfermedades que aparecen con más frecuencia. Por otro lado, será necesario realizar algún tipo de análisis a través del cual podamos detectar los pares de enfermedades que suelen darse de forma simultánea en los pacientes.

Una vez detectadas estas enfermedades, nos proponemos construir los modelos mencionados anteriormente. Pretendemos con todo esto estudiar una serie de aspectos como:

- El número de días que un trabajador que ha iniciado un proceso de baja, por una enfermedad concreta se ausenta al trabajo.
- El número de bajas que de los trabajadores toman al año, ya sean motivadas por una única enfermedad o por varias.
- La probabilidad que existe de que un trabajador que ha iniciado un proceso de baja y se recupera, vuelva a tener un episodio de baja, bien motivado por esta misma enfermedad o por otra distinta. Así mismo, estamos interesados en estudiar

la probabilidad de que un paciente que está enfermo lo siga estando al días siguiente, bien por la misma enfermedad o por otra distinta, o que por el contrario se recupere, es decir, pase a estar sano.

Capítulo 2

Datos

En el desarrollo de este trabajo, y como hemos mencionado anteriormente, se han utilizado los datos proporcionados por MAC. Dichos datos recogen información sobre las bajas laborales para cada uno de los pacientes a los que MAC realizó seguimiento entre los años 2006 y 2013, ambos inclusive. De entre toda la información proporcionada, para cada episodio de baja haremos uso de las siguientes variables:

- **Paciente**
  Código que permite identificar a los trabajadores. Cada trabajador tiene un código de paciente, este siempre es el mismo en todos los procesos de baja. Es de tipo entero.

- **Episodio**
  Código que se le asigna a los trabajadores cuando entran en un proceso de Incapacidad Temporal (IT). Cabe mencionar que cada vez que un paciente inicia un proceso de IT se le asigna un episodio distinto, esto quiere decir que cada baja tiene asociado un código diferente. Esta variable también es de tipo entero.

- **Diagnóstico**
  Patología que origina la baja, de un conjunto de 46 diagnósticos diferentes, entre ellos encontramos: cervicalgia, anomalías congénitas, trastornos mentales, etc. La lista con todas la enfermedades se podrá ver en el capítulo 3 junto con el código que las identifica.

- **Fecha de baja**
  Designamos con esta variable al instante de tiempo en el cual un trabajador inicia un proceso de IT. Esta variable es de tipo fecha, $dd/mm/aaaa$.

- **Fecha de alta**
  Instante de tiempo en el que el paciente finaliza el proceso de IT y se incorpora al trabajo. Es una variable de tipo fecha, $dd/mm//aaaa$.

- **Días de baja**
  Periodo de tiempo durante el cual el trabajador se ausenta en el trabajo. Esta variable

se obtiene de la diferencia de las dos anteriores, es decir, días de baja = fecha de alta - fecha de baja, y es de tipo entero.

- **Años de antigüedad**
  Tiempo en años que el trabajador lleva en el sistema. Si a la fecha en la que se inicia un proceso de IT (fecha de baja) le restamos esta variable, obtenemos el instante de tiempo en el cual el trabajador fue dado de alta en el sistema. Esta variable es real, así pues, que un trabajador tenga una antigüedad de 11,8904109589041 años quiere decir que lleva dado de alta en el sistema $11{,}8904109589041 \cdot 365 = 4340$ días.

Los datos que nos proporcionó MAC estaban en una hoja de cálculo de Microsoft Excel, de manera que cada episodio de baja estaba recogido en una fila y las distintas variables que aportaban información sobre dichos procesos de baja estaban distribuidas en las distintas columnas. Así contamos con 41 variables (41 columnas) y 33.848 episodios diferentes.

Usaremos para nuestro análisis el lenguaje R [4] a través de la interfaz de usuario R-Studio. Para ello, necesitamos exportar los datos originales proporcionados por MAC a un formato de texto delimitado por tabulaciones, con el fin de que R los importe a una estructura de datos.

Una vez hecho esto, creamos dos scripts, los cuales podemos encontrar en las secciones A.1 y A.2 del apéndice A. El primero de ellos, `tomadedatos.R`, permitió importar los datos al software mencionado anteriormente. El segundo, `divisiondatos2grupos.R`, se utilizó para dividir los datos en dos grupos, de forma que uno de ellos estuviese formado por el 60 % de los datos y el otro por el 40 %. Esta división se realizó a partir de un sorteo sobre los distintos trabajadores, teniendo en cuenta el número de bajas que ha tenido cada uno de ellos en el intervalo de tiempo para el que tenemos información (2006–2013). El hecho de hacer el sorteo teniendo en cuenta esto, se debe a que queremos obtener grupos equilibrados, entendiendo por esto grupos en los que la proporción de pacientes con n episodios de bajas, sea la misma aproximadamente para los dos grupos. Los grupos obtenidos recibieron el nombre de `datos.trabajo` y `datos.control`. El primero de ellos formado por el 60 % de los datos será con el que realizaremos nuestro estudio; y el segundo tiene como finalidad la validación de los resultados obtenidos.

Capítulo 3

Relación entre las distintas enfermedades

Una vez divididos los datos en dos grupos, nos propusimos estudiar qué relación existe entre los distintos diagnósticos (enfermedades).

Creamos, para ello, una matriz de correlación entre las distintas enfermedades apoyandonos en el script `CorrelacionEnfermedades.R` (véase en los anexos). En él, dedicamos una parte a pintar la matriz de correlación obtenida, con la finalidad de conseguir una forma visual y más sencilla para interpretar los resultados. Al hacer esto, obtuvimos el siguiente gráfico:

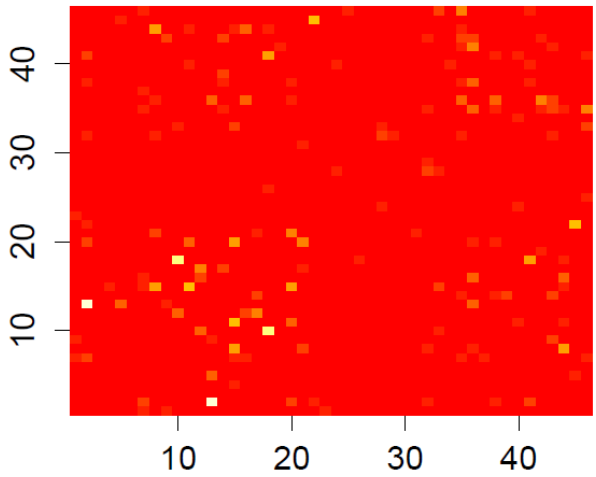


Figura 3.1: Correlación entre los 46 diagnósticos. Los colores más claros indican mayor correlación.

Sabemos que los colores claros representa los pares de enfermedades que están más

relacionados, aunque resulte algo complicado identificarlas a partir de esta representación gráfica, podemos decir que los 10 pares de enfermedades que presentan un índice de correlación más elevado, y por tanto las que parecen estar más relacionadas son las que se muestran en la siguientes tabla:

Fila	**Columna**	**Correlación**
E02	E13	0.08001
E10	E18	0.07195
E11	E15	0.04535
E08	E44	0.03980
E15	E20	0.03758
E08	E15	0.03712
E18	E41	0.03531
E35	E46	0.03276
E20	E21	0.02983
E12	E17	0.02823

Nota 3.1 *Los elementos de las dos primeras columnas de la tabla anterior, son códigos utilizados para identificar a las distintas enfermedades, la relación que existe entre estos códigos y cada uno de los diagnóstico se recoge en un catálogo expuesto más adelante (página 9).*

Lo anterior nos condujo a aplicar un Análisis Clúster o Análisis por Conglomerados, que no es más que una técnica estadística que permite dividir los distintos individuos en grupos maximizando la homogeneidad dentro de cada uno de ellos y minimizándola entre los distintos grupos, de acuerdo con los valores que toman sobre distintas características.

De entre los distintos procedimientos que se han diseñado para llevar a cabo este agrupamiento de datos, nos centraremos en los métodos jerárquicos aglomerativos, también conocidos como métodos ascendentes. Estos comienzan el análisis con tantos clústers como individuos haya. A partir de estas unidades iniciales se van formando grupos de forma ascendente, hasta que al final del proceso todos los casos tratados están englobados en un mismo conglomerado.

El algoritmo utilizado en este tipo de métodos supone que tenemos definida una distancia entre cada par observaciones (o individuos). Además a partir de distancias entre individuos tenemos muchas formas de definir una distancia entre grupos. En la función `hclust` de R esta distancia entre grupos puede elegirse entre enlace simple, enlace compuesto, enlace promedio, distancia promedio o método de Ward. Expliquemos a continuación en que consiste cada una de estas distancias:

- **Enlace Simple**
  La distancia entre grupos se define como la distancia más pequeña entre sus observaciones.

- **Enlace Compuesto**
  La distancia entre grupos se define como la distancia más grande entre sus observaciones.

- **Enlace Promedio o Método del Centroide**
  La distancia entre grupos se define como la distancia entre los centros de cada uno de ellos, siendo dicho centro un punto cuya coordenada se corresponde con la media aritmética de las observaciones que forman cada grupo.

- **Distancia Promedio**
  La distancia entre grupos coincide con la media de las distancias de todos los posibles pares de observaciones, una de cada grupo.

- **Método de Ward**
  La distancia entre grupos equivale al aumento producido en la suma de cuadrados cuando dichos grupos se unen en uno.

Así, dado los grupos $X = \{X_1, \ldots, X_n\}$ e $Y = \{Y_1, \ldots, Y_m\}$ podemos definir las distancias explicadas anteriormente de la siguiente forma:

Distancia entre grupos	**Fórmula**
Enlace Simple	$d(X,Y) = \text{mín}\{d(X_i, Y_j)/X_i \in X, Y_j \in Y\}$
Enlace Compuesto	$d(X,Y) = \text{máx}\{d(X_i, Y_j)/X_i \in X, Y_j \in Y\}$
Método del Centroide	$d(X,Y) = DE(\bar{X}, \bar{Y})$
Distancia Promedio	$d(X,Y) = (\sum_{X_i \in X, Y_j \in Y} d(X_i, Y_j))/n \cdot m$
Método de Ward	$d(X,Y) = \sum_{Z_k \in X,Y} DE(Z_k - \bar{X,Y}) -$ $(\sum_{X_i \in X} DE^2(X_i - \bar{X}) + \sum_{Y_j \in Y} DE^2(Y_j - \bar{Y}))$

Además, el algoritmo que hemos mencionado anteriormente viene descrito como sigue:

- **Paso 0**
  Se calculan las distancias entre las n observaciones $X_1, \ldots, X_n$ construyendo la correspondiente matriz de distancias:

$$D^{(1)} = \begin{array}{c} \\ X_1 \\ \vdots \\ X_n \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & \cdots & X_n \end{array} \\ \begin{pmatrix} & & \\ & & \\ & & \end{pmatrix} \end{array}$$

- **Paso 1**
  Se construyen inicialmente tantos grupos como observaciones haya, de forma que $C_1, C_2, \ldots, C_n$ contenga una sola observación $X_1, X_2, \ldots, X_n$.

- **Paso 2**
  Se identifican las dos observaciones con distancia más pequeña entre ellas, y las unimos formando un clúster con ellas, de modo que ahora tenemos $n - 1$ clústers.

- **Paso 3**
  Se calcula la distancia entre el clúster y el resto de las observaciones no agrupadas, obteniendo así una nueva matriz de distancias:

$$D^{(2)} = \begin{array}{c} \\ C_1C_2 \\ \vdots \\ C_n \end{array} \begin{array}{c} \begin{array}{ccc} C_1C_2 & \cdots & C_n \end{array} \\ \begin{pmatrix} & & \\ & & \\ & & \end{pmatrix} \end{array}$$

  Una vez hecho esto evaluamos la matriz $D^{(2)}$, si obtenemos que la distancia mínima es la correspondiente al clúster C_1C_2 y a la observación C_j , estos se unen formando un nuevo clúster de tres elementos, si esto no sucede, entonces se forma otro clúster de tamaño dos con las observaciones más cercanas.

- **Paso 4**
  Se repite sucesivamente el paso anterior hasta encontrar un clúster que contenga todas las observaciones, o bien, hasta que se supere una distancia prefijada.

El resultado de aplicar este algoritmo lo podemos representar mediante un dendrograma que no es más que un tipo de representación gráfica o diagrama en forma de árbol que organiza los datos en subcategorías que se van dividiendo en otros hasta llegar al nivel de detalle deseado. Gracias a este tipo de representaciones podemos apreciar las relaciones de agrupaciones entre los datos e incluso entre grupos de ellos. Además, observando las sucesivas subdivisiones podemos hacernos una idea sobre los criterios de agrupación de los mismos, la distancia entre los datos según las relaciones establecidas, etc.

Para obtener un dendrograma con nuestros datos que nos permita identificar los grupos de enfermedades más frecuente en los pacientes elaboramos dos scripts, `CorrelacionEnfermedades.R` y `Cluster.R`, los cuales podemos encontrar en el apéndice.
El primero de estos dos archivos de comandos consta de tres partes. En la primera de ellas, creamos una variable a la que denominamos `enfermedades` y que contiene las distintos diagnósticos que han padecido los trabajadores pertenecientes al grupo `datos.trabajo`, durante el periodo de tiempo del que tenemos información. Apoyándonos en esta variable construimos un catálogo en el que a cada una de las enfermedades le asociamos un código con el objetivo de obtener resultados más sencillos de visualizar, pues los nombres de las enfermedades resultan demasiado grandes en comparación con este código. El catálogo obtenido es el siguiente:

```
Codigos   Nº Episodios Descripción Enfermedades
E01       21           ANOMALIAS CONGENITAS
E02       112          ARTROPATIAS
E03       40           CARDIOPATIA
E04       2776         CERVICALGIA
E05       1088         CONTUSION DE TRONCO
```

```
E06     3124        CONTUSION MIEMBRO INFERIOR Y OTROS
E07     1993        CONTUSION MIEMBRO SUPERIOR
E08     278         CUERPO EXTRAÑO EN ORIFICIO
E09     4157        DORSOPATIAS
E10     39          EFECTOS TOXICOS
E11     52          ENFERMEDAD SISTEMA CIRCULATORIO
E12     109         ENFERMEDADES APARATO DIGESTIVO
E13     32          ENFERMEDADES APARATO URINARIO
E14     186         ENFERMEDADES DE LA PIEL
E15     17          ENFERMEDADES DEL OIDO
E16     44          ENFERMEDADES DEL SISTEMA NERVIOSO
E17     80          ENFERMEDADES INFECCIOSAS
E18     46          ENFERMEDADES RESPIRATORIAS
E19     950         ESGUINCE MUÑECA Y MANO
E20     567         ESGUINCE PARTE SUPERIOR DEL BRAZO
E21     534         ESGUINCE ESPALDA
E22     1070        ESGUINCE RODILLA Y PIERNA
E23     3235        ESGUINCE TOBILLO Y PIE
E24     45          FRACTURAS CRANEALES
E25     160         FRACTURAS DE CUELLO Y TRONCO
E26     847         FRACTURAS DE MIEMBRO INFERIOR
E27     967         FRACTURAS DE MIEMBRO SUPERIOR
E28     335         HERIDA ABIERTA CABEZA, CUELLO O TRONCO
E29     379         HERIDA ABIERTA MIEMBRO INFERIOR
E30     1910        HERIDA ABIERTA MIEMBRO SUPERIOR
E31     24          LESION POR APLASTAMIENTO
E32     643         LESION SUPERFICIAL
E33     243         LESIONES INTRACRANEALES
E34     213         LUXACION
E35     59          OSTROPATIAS, CONDROPATIAS, DEFORMIDADES MUSCULOESQUELETICAS
E36     1006        OTRO TRASTORNO DE ARTICULACIÓN
E37     46          OTROS EFECTOS
E38     500         OTROS TRASTORNOS DE SINOVIA, TENDON Y BURSA
E39     491         OTROS TRAUMATISMOS
E40     318         QUEMADURAS
E41     131         SINTOMAS, SIGNOS Y ESTADOS MAL DEFINIDOS
E42     1002        TENDINITIS DE INSERSIONES PERIFERICAS
E43     2793        TRASTORNO DE MUSCULOS, LIGAMENTOS Y FASCIAS
E44     776         TRASTORNO DEL OJO Y ANEXOS
E45     325         TRASTORNO INTERNO DE RODILLA
E46     85          TRASTORNOS MENTALES
```

En la segunda parte, creamos una tabla cuyas filas están formadas por los distintos pacientes y sus columnas por los diagnósticos. De este modo, conseguimos visualizar fácilmente las enfermedades que han padecido cada uno de los trabajadores y el número de veces que las han tenido. Por último, construimos una matriz de correlación (`correlacionE`) entre las enfermedades basándonos en el coeficiente de correlación de Pearson, que no es más que una medida de la relación lineal que existe entre dos variables cuantitativas X e Y. Este coeficiente se define como el cociente de la covarianza de las dos variables entre el producto de las desviaciones típicas de cada una de ellas, esto es:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sqrt{V(X) \cdot V(Y)}}$$

Una vez obtenida la matriz `correlacionE` creamos el script `Cluster.R` (sección A.4 del apéndice), con en el que vamos a realizar el análisis clúster que nos permitirá identificar cuáles son las enfermedades que están más relacionadas, es decir, las que aparecen con más

frecuencia y de forma simultánea en los individuos. Antes de definir la matriz de distancia que nos permita realizar un análisis de este tipo, elaboramos una matriz `corE` dividiendo la matriz `correlacionE`, obtenida anteriormente entre el elemento máximo de la misma, sin tener en cuenta en ningún caso el elemento diagonal 1. Matemáticamente podemos expresar esto a través de la siguiente fórmula:

$$\texttt{corE} = \frac{\texttt{correlacionE} - 1}{\text{máx}\,(\texttt{correlacionE} - 1)}$$

Por definición sabemos que la distancia entre dos puntos siempre es mayor o igual que cero, luego diremos que una matriz es verdaderamente una matriz de distancia si todos sus elementos cumplen esta condición, es decir, que todos ellos sean mayor o igual que 0. Teniendo en cuenta esto y la matriz anterior (`corE`), podemos definir la matriz de distancia para el análisis clúster de la siguiente forma: `as.dist(1 - cor(E))`.

Puesto que no sabíamos la influencia que podría tener en el agrupamiento de los datos el uso de una u otra de las posibles distancias, repetimos el mismo análisis cambiando la distancia utilizada para comprobar los resultados. Tras haber analizado y estudiado los dendrogramas obtenidos al implementar el script `Cluster.R` localizado en la sección A.4, llegamos a la conclusión de que las enfermedades que están más relacionadas, y que podremos observar más adelante en la figura 3.2 son:

1. Enfermedades infecciosas (E17) y Trastorno del ojo y anexos (E44).
2. Esguince parte superior del brazo (E20) y Esguince de espalda (E21).
3. Esguince de rodilla y pierna (E22) y Trastorno interno de rodilla (E45).

La aparición de las enfermedades E20 y E21 en un mismo paciente parece deberse a una tipificación del diagnóstico médico, esto quiere decir a un fallo médico al pasar al ordenador la patología que presenta el paciente. Este hecho nos lleva a no considerar estas enfermedades en un posible estudio posterior. Además de entre los pares que nos quedan podemos decir que E17 y E44 son enfermedades poco comunes que aparecen en muy pocos individuos y de las cuales tenemos poca información en comparación a los diagnósticos E22 y E45, los cuales son más frecuentes. Este hecho se observa fácilmente en la columna `Frecuencias` del catálogo expuesto anteriormente, donde podemos comprobar que de los 33.848 episodios de baja que componen los mismo, 80 son debidos a enfermedades infecciosas (E17), 776 a trastornos del ojo y anexos (E44), 1070 a esguince de rodilla y pierna (E22) y 325 a trastorno interno de rodilla (E45).

En la figura siguiente, se muestra el dendrograma obtenido al aplicar el método de Ward para obtener la distancia entre grupos.

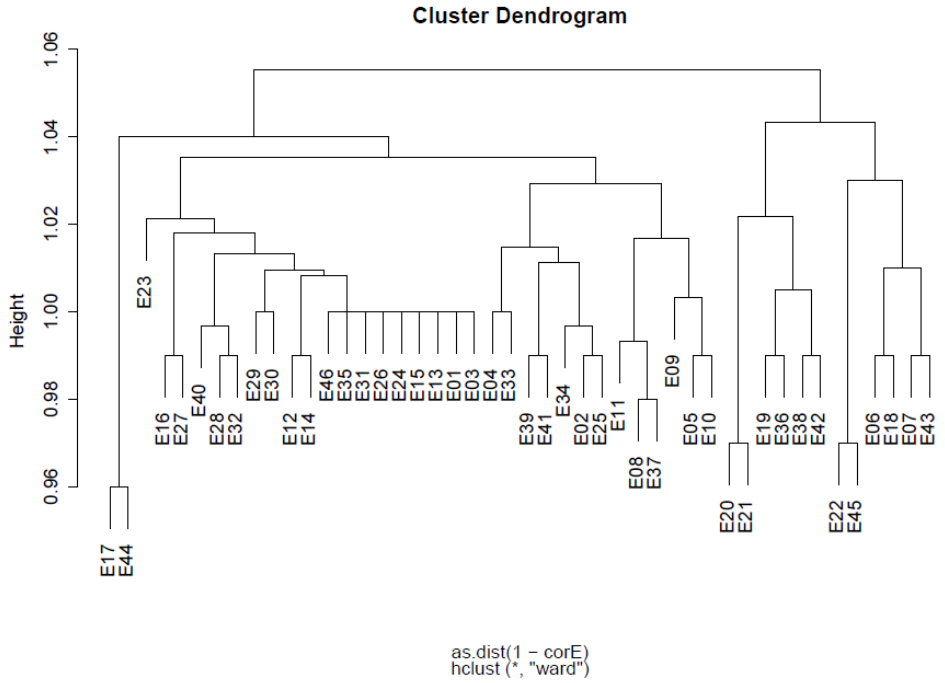


Figura 3.2: Dendograma con la agrupación de las enfermedades utilizando la distancia de Ward.

Capítulo 4

Cadenas de Markov

Las cadenas de Markov son modelos muy ricos y fuertes para capturar el comportamiento dinámico de procesos estocásticos con un gran número de componentes. En lo que sigue, nos centraremos en explicar los procesos en tiempo discreto.

El primer elemento que define una cadena de Markov es un conjunto $S = \{s_1, s_2, \ldots, s_N\}$ de posibles estados. Vamos a considerar que el proceso comienza en uno de estos estados y que se mueve sucesivamente de uno a otro. Este movimiento recibe el nombre de paso. Además, podemos decir que si una cadena está actualmente en el estado s_i, entonces la probabilidad que existe de pasar a un estado s_j en el paso siguiente se denota por $p_{i,j}$. Cabe mencionar también que esta probabilidad no depende de la historia pasada sino únicamente del estado actual en el que se encuentra el proceso (y tal vez del instante considerado), y que a partir de ellas se obtiene la llamada matriz de transición $P = (p_{i,j})$.

Si P es una matriz de transición para una cadena de Markov, entonces $p_{i,j}^{(m)}$ simboliza la probabilidad de que una cadena que está actualmente en el estado s_i esté en el estado s_j al cabo de m pasos. Así, tenemos que $P^{(m)}$ es la matriz de transición en m pasos, y además se cumple que $P^{(m)} = P^m$.

Dentro de los distintos tipos de cadenas de Markov existentes, estamos interesado en estudiar las cadenas ergódicas. Por esta razón, toda la teoría que se desarrollará a continuación busca explicar y definir en qué consiste este tipo de cadenas.

Definición 4.1 *Una cadena de Markov es ergódica si es posible ir (tal vez en más de un paso) desde cualquier estado al resto de ellos. En muchas ocasiones, este tipo de cadenas reciben el nombre de cadenas de Markov irreducibles.*

Definición 4.2 *Diremos que una cadena de Markov es regular si alguna potencia de la matriz de transición tiene todos sus elementos positivos. En otras palabras, una cadena es regular si para algún m, es posible ir de cada estado a cualquier otro en exactamente m pasos.*

Es claro que a partir de estas dos definiciones podemos hacer las siguientes afirmaciones:

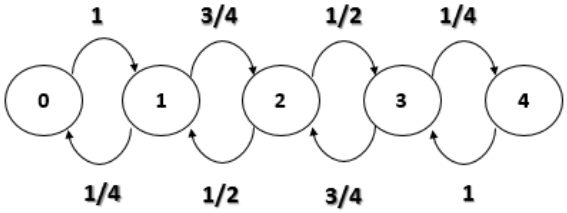
$$\text{Cadena regular} \Rightarrow \text{Cadena ergódica}$$
$$\text{Cadena ergódica} \nRightarrow \text{Cadena regular}$$

Ejemplo 4.1 *Consideremos que tenemos una matriz de transición definida de la siguiente forma:*

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/4 & 0 & 3/4 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 3/4 & 0 & 1/4 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

Es claro en este ejemplo, que podemos acceder desde cualquier estado al resto de ellos. Sin embargo, se puede observar que partiendo del estado 0 se puede llegar a los estados 0, 2 o 4 sólo después de un número par de pasos (m par), y a los estados 1 y 3 después de un número impar de ellos (m impar). En otras palabras, no existe ningún m que permita ir desde el estado 0 a todos los demás. Por lo tanto, teniendo en cuenta las definiciones 4.1 y 4.2 se puede concluir que esta matriz de transición corresponde a una cadena de Markov ergódica no regular.

A continuación, podemos observar un esquema de este tipo de cadena.



Para una cadena de Markov ergódica, podemos afirmar que cualquier vector de probabilidad w que satisfaga la condición $w = wP$ define una distribución de equilibrio de la cadena. Además, la j-ésima componente viene dada por la expresión:

$$w_j = \sum_{i \in S} w_i p_{i,j}. \tag{4.1}$$

La parte izquierda de la ecuación 4.1 representa la probabilidad de estar actualmente (instante de tiempo t) en el estado j, mientras que la parte derecha indica la probabilidad de estar en el estado j en el siguiente instante de tiempo $(t + 1)$. En una distribución de equilibrio tiene que ocurrir que estas dos probabilidades sean iguales. Además, se tiene

que en las cadenas con un número finito de estados siempre existe esta distribución de equilibrio.

En lo que sigue y como ya hemos mencionado anteriormente, desarrollaremos una serie de modelos que nos permitirán entender de forma más clara y precisa cómo se comportan los procesos de baja laboral. En todos ellos usaremos modelos de cadenas de Markov ergódicas con un número finito de estados.

4.1. Modelo 1

Como ya habíamos mencionado anteriormente, los datos sobre los que hemos trabajado para desarrollar este trabajo son datos reales proporcionado por MAC (Mutua de Accidentes de Canarias), lugar en el cual cursé la asignatura Prácticas Externas. Durante la realización de las prácticas, elaboré un estudió que me permitió concluir que una de las enfermedades que aparece con más frecuencia en la mayoría de los trabajadores es la cervicalgia. Por este motivo, dicha enfermedad es la que vamos a utilizar para el desarrollo del modelo 1.

Así este primer modelo que proponemos, consiste en un sistema de cadenas de Markov con dos estados transitorios, estar de baja por Cervicalgia o no estarlo. Por lo tanto, para construir este tipo de cadenas solo tendremos en cuenta aquellos pacientes que han padecido alguna vez esta enfermedad y que están dentro del grupo `datos.trabajo` (que como se explicó en el capítulo 2, este es el grupo utilizado a la hora de estimar nuestro modelo y por tanto, con el que se obtendrán los resultados). En la figura 4.1, podemos ver un esquema en el que se ilustra el modelo comentado.

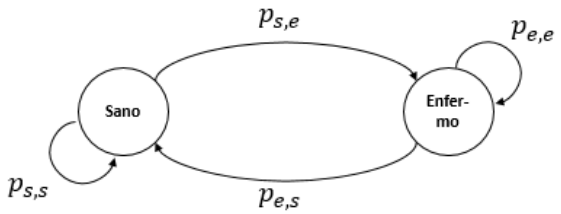


Figura 4.1: Representación de la cadena del Modelo 1

Con todo esto, nuestro objetivo ahora es proponer una matriz de transición, a través de la cual, podremos tener una idea de la probabilidad que existe de que una persona que esté enferma lo siga estando al día siguiente o se recupere, y viceversa. Dicha matriz de transición presenta la siguiente estructura:

$$P = \begin{pmatrix} p_{e,e} & p_{e,s} \\ p_{s,e} & p_{s,s} \end{pmatrix}.$$

Donde:

- $p_{e,e}$ simboliza la probabilidad de que un paciente permanezca enfermo.
- $p_{e,s}$ es la probabilidad de que un trabajador que esté de baja por cervicalgia se incorpore al día siguiente al trabajo.
- $p_{s,e}$ indica la probabilidad que hay de pasar del estado transitorio sano al estado enfermo.
- $p_{s,s}$ representa la probabilidad de que un trabajador que está sano lo siga estando al día siguiente.

A continuación, estamos interesado en estimar los elementos de la matriz de transición basándonos en nuestros datos. Para ello, es necesario tener en cuenta algunas de las variables explicadas en el capítulo 2 como pueden ser: los días de baja, las fechas de baja y de alta de cada episodio (proceso de baja) y los años de antigüedad de cada uno de los trabajadores a los que MAC pasó consulta durante los años 2006–2013. Estas variables permitieron, a su vez, obtener otras nuevas entre las cuales podemos destacar las siguientes:

- **alta** = máx{(**fechas.baja** − **antig**), **fecha1**}.
  Esta variable nos permite saber el tiempo que un trabajador lleva dado de alta en el sistema. Que consideremos el máximo entre la fecha real en la que el paciente fue dado de alta (diferencia entre la fecha de su primer episodio de baja y la antigüedad que tenía en ese momento) y `fecha1` = 01–01–2006 se debe a que los datos que conocemos para cada uno de los pacientes comienzan a partir de dicho instante de tiempo (`fecha1`).
- **dias.enfermo** = **sum**(**dias.baja**).
  Con esta variable podemos saber el total de días que cada paciente ha estado de baja por la enfermedad Cervicalgia, en nuestro caso.
- **dias.sano** = **as.numeric**(**fecha2** − **alta** − **dias.enfermo**, **unit** = "**days**").
  El número de días que un trabajador ha estado sano, dentro del intervalo de tiempo para el que tenemos información (desde `fecha1` = 01–01–2006 hasta `fecha2` = 31–12–2013), se obtiene sin más que restarle a la `fecha2` el día en el que el paciente fue dado de alta en el sistema y los días que este ha estado enfermo. Cabe mencionar que es el comando `as.numeric(argumento, unit = 'days')` el que nos permite obtener como resultado a esta diferencia el número de días que ha estado sano.

Una vez aclarado esto, para obtener los coeficientes de la matriz de transición, hay que tener en cuenta el número de saltos que se producen al pasar del estado transitorio enfermo

al estado sano (`saltos.es`) y viceversa (`saltos.se`). Además, estamos considerando que la unidad de salto es de un día, por lo que los saltos de sano a sano (`saltos.ss`) coinciden con el total de días seguidos que el paciente permanece en este estado, obtenidos mediante la variable `dias.sano`, menos un día que simboliza el paso seguido de sano a enfermo. Así mismo, los saltos de enfermo a enfermo (`saltos.ee`) coinciden con el número de días que el paciente está enfermo, `dias.enfermo`, menos un día (paso de enfermo a sano).

Al implementar el script `Modelo1.R` (véase en la sección A.5 del apéndice), obtenemos los elementos de la matriz de transición.

$$P = \begin{pmatrix} 0{,}97919 & 0{,}02081 \\ 0{,}00044 & 0{,}99956 \end{pmatrix}.$$

Además, podemos observar que se cumple la propiedad:

- $(p_{e,e} = 0{,}97919) + (p_{e,s} = 0{,}02081) = 1$
- $(p_{s,e} = 0{,}00044) + (p_{s,s} = 0{,}99956) = 1$

Lo cual nos lleva a afirmar, según lo comentado al inicio de este capítulo, que la matriz P calculada en base a nuestros datos es una verdadera matriz de transición.

Para finalizar esta sección, reflexionaremos sobre aspectos como: la duración media y la distribución de las longitudes de las bajas y el tiempo que cada trabajador está de baja a lo largo del año.

- **Duración media de las bajas.**
  Teniendo en cuenta las longitudes de las bajas de los trabajadores que han tenido algún episodio de cervicalgia durante los años 2006–2013, queremos estudiar la duración media de las bajas por esta enfermedad. Al estar trabajando con procesos de cadenas de Markov en el desarrollo de nuestro modelo, podemos calcular esta media a través del tiempo medio de recurrencia. Pero, ¿qué es el tiempo medio de recurrencia?

  Definición 4.3 (Tiempo medio de retorno o recurrencia) *Si una cadena de Markov ergódica comienza en un estado s_i, el tiempo que transcurre hasta volver al estado s_i por primera vez es lo que se conoce como tiempo medio de recurrencia para s_i.*

  Una vez entendida esta definición, para calcular el tiempo medio de recurrencia aplicaremos el siguiente resultado.

  Teorema 4.1 *Para una cadena de Markov ergódica, el tiempo medio de recurrencia para el estado s_i viene dado por:*

$$r_i = \frac{1}{w_i}$$

donde w_i es la i-ésima componente del vector de probabilidad fijo de la matriz de transición P.

Teniendo en cuenta que para obtener w_i hay que aplicar el *Teorema fundamental del límite para cadenas regulares*, el cual se enuncia a continuación.

Teorema 4.2 (Teorema fundamental del límite para cadenas regulares) *Si P es una matriz de transición para una cadena de Markov regular, entonces:*

$$\lim_{n\to\infty} P^n = W$$

donde W es una matriz cuyas filas son todas iguales. Además, todas las entradas de dicha matriz W son estrictamente positivas.

Estamos en condiciones ya de obtener la duración media de cada proceso de baja por cervicalgia, calculando para ello el tiempo medio de recurrencia para el estado s_i = sano, no de nuestro modelo original sino del siguiente modificado:

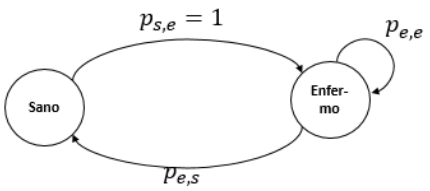


Figura 4.2: Representación de la cadena utilizada para el cálculo del tiempo medio de recurrencia en el Modelo 1.

Cuya matriz de transición es:

$$P_1 = \begin{pmatrix} p_{e,e} & p_{e,s} \\ p_{s,e} & p_{s,s} \end{pmatrix} = \begin{pmatrix} 0{,}97919 & 0{,}02081 \\ 1 & 0 \end{pmatrix}$$

Aplicando ahora el teorema 4.2 y teniendo en cuenta la matriz de transición P_1 expuesta anteriormente, obtenemos el vector $w = (w_e;\ w_s)$ sin más que resolver el siguiente límite:

$$\lim_{n\to\infty} P_1^n = \lim_{n\to\infty} \begin{pmatrix} 0{,}97919 & 0{,}02081 \\ 1 & 0 \end{pmatrix}^n$$

Para facilitar el cálculo de este límite, diagonalizamos en primer lugar la matriz de transición anterior. Esto es:

$$P_1 = \begin{pmatrix} 0{,}70711 & -0{,}02081 \\ 0{,}70711 & 0{,}99979 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -0{,}02081 \end{pmatrix} \begin{pmatrix} 1{,}38538 & 0{,}02883 \\ -0{,}97982 & 0{,}97982 \end{pmatrix}$$

En segundo lugar, calculamos la n-ésima potencia de la matriz.

$$P_1^n = \begin{pmatrix} 0{,}70711 & -0{,}02081 \\ 0{,}70711 & 0{,}99979 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -0{,}02081 \end{pmatrix}^n \begin{pmatrix} 1{,}38538 & 0{,}02883 \\ -0{,}97982 & 0{,}97982 \end{pmatrix}.$$

Por último, tomamos límite en la expresión anterior. Al hacer esto tenemos que:

$$\lim_{n\to\infty} P_1^n = \begin{pmatrix} 0{,}70711 & -0{,}02081 \\ 0{,}70711 & 0{,}99979 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1{,}38538 & 0{,}02883 \\ -0{,}97982 & 0{,}97982 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}97961 & 0{,}02039 \\ 0{,}97961 & 0{,}02039 \end{pmatrix}$$

De aquí concluimos que el vector $w = (w_e;\ w_s) = (0{,}97961;\ 0{,}02039)$ y por tanto, al aplicar el teorema 4.1, el tiempo medio de recurrencia es $r_s = \frac{1}{w_s} = \frac{1}{0{,}02039} = 49{,}0436$ días. Esto nos dice que cada episodio de baja por cervicalgia dura en media unos 49.0436 días.

- **Distribución de las longitudes de las bajas**

Antes de comenzar a estudiar cómo se distribuyen las longitudes de las bajas, es necesario tener presente la siguiente definición:

Definición 4.4 (Distribución Geométrica) *Una variable aleatoria X sigue una distribución geométrica, $X \sim Geo(p)$, si:*

$$P(X = x) = (1-p)^{x-1} \cdot p\,, \quad x = 1, 2, \ldots$$

En otras palabras, la variable X tiene una distribución geométrica si se puede expresar como el número de pruebas necesarias hasta la aparición del primer éxito, en la repetición sucesiva de pruebas de Bernuilli independientes y todas presentando la misma probabilidad de éxito p.

Para estudiar la distribución de las duraciones de las bajas, vamos a representar mediante un histograma las longitudes de los distintos procesos de baja de aquellos trabajadores que han padecido algún episodio de cervicalgia durante el intervalo de tiempo 2006–2013 y que se encuentran en el grupo `datos.trabajo`. Dichas longitudes se encuentran acumuladas en el vector `long.bajas`, el cual podemos encontrar

en el script `Modelo1.R` localizado en la sección A.5 del apéndice. Así mismo, representamos a través de una línea roja la verdadera distribución de la geométrica, para ello y teniendo en cuenta la definición 4.4, tomamos como probabilidad de éxito $p_{e,s}$, ya que estamos interesados en estudiar las longitudes de las bajas, por lo que nuestro proceso finaliza cuando un trabajador que está enfermo se recupera.

Dichas representaciones se harán agrupando las longitudes de las bajas en intervalos de tamaño 5 y representando la distribución de la geométrica según intervalos de tamaño uno, llegando así al histograma que se muestra a continuación:

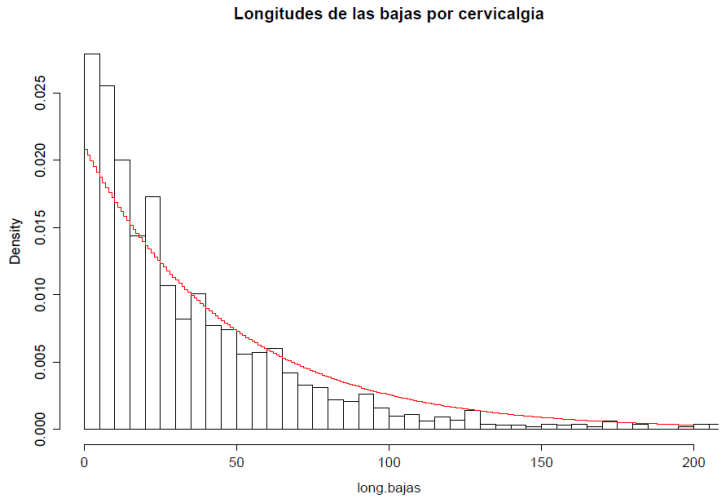


Figura 4.3: Histograma de la duración de las bajas debidas a cervicalgia. En rojo las probabilidades correspondientes a una distribución geométrica de parámetro $p_{e,s} = 0{,}02081$

A la vista podemos concluir que la forma general no dista mucho de una geométrica. Sin embargo, esta geométrica presenta una media mayor que la de los datos, lo cual nos indica que este modelo no aproxima muy bien los datos. Además, podemos observar que los procesos de baja con una duración menor de 35 días siempre están por encima de la línea roja que representa la distribución geométrica mencionada anteriormente. Este hecho junto con el de que el modelo no aproxima bien los datos, hace que nos planteemos la realización de un estudio similar al realizado en esta sección, pero considerando en este caso que nuestro sistema de cadenas de Markov está formado por tres estados transitorios: estar sano, estar enfermo durante un periodo corto (bajas de duración menor o igual a 35 días) o estar enfermo por un periodo largo (bajas de más de 35 días). Este estudio se verá en profundidad en la sección 4.2.

- **Tiempo (proporción) de baja al año.**
  Para obtener el tiempo que cada trabajador está de baja al año por cervicalgia,

basándonos en los que ya han tenido algún episodio de esta enfermedad durante los años 2006–2013, nos apoyamos en el cálculo de las posiciones de equilibrio. Para obtener dichas posiciones basta con aplicar el teorema 4.2, enunciado en el apartado dedicado a la duración media de las bajas, a la matriz P obtenida a partir de este primer modelo. Si recordamos, esta matriz tenía la siguiente forma:

$$P = \begin{pmatrix} 0{,}97919 & 0{,}02081 \\ 0{,}00044 & 0{,}99956 \end{pmatrix}.$$

Al diagonalizar esta matriz (del mismo modo que en apartado dedicado a la duración media de las bajas) tenemos que la potencia n-ésima es:

$$P^n = \begin{pmatrix} -0{,}70711 & -0{,}99978 \\ -0{,}70711 & 0{,}02109 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 0{,}97875 \end{pmatrix}^n \cdot \begin{pmatrix} -0{,}02921 & -1{,}38500 \\ -0{,}97956 & 0{,}97956 \end{pmatrix}$$

Por lo tanto, al tomar límite en esta expresión llegamos a que:

$$\lim_{n\to\infty} P^n = \begin{pmatrix} 0{,}02066 & 0{,}97934 \\ 0{,}02066 & 0{,}97934 \end{pmatrix}$$

Podemos concluir así que las posiciones de equilibrio vienen dadas por el vector $w = (w_e;\ w_s) = (0{,}02066;\ 0{,}97934)$. Esto nos dice que la proporción de tiempo que cada trabajador que inicia un proceso de baja por cervicalgia está de baja al año es $0{,}02066 \cdot 365 = 7{,}5409$ días.

4.2. Modelo 2

El modelo que vamos a desarrollar en esta sección, consiste en una cadena de Markov con tres estados transitorios. El hecho de considerar este tipo de proceso, viene motivado de la ambición de conseguir un mejor ajuste de la distribución de las longitudes de los procesos de baja, que como habíamos visto en la sección anterior, estos estaban siempre por encima de la distribución geométrica de parámetro $p_{e,s}$ (representada en la figura 4.3 en color rojo), cuando dichos procesos presentaban una duración de más de 35 días.

Los estados que vamos a considerar son: estar sano, tener una baja por un periodo corte de tiempo o padecer una baja durante un periodo largo; considerando como periodo corto a todas aquellas bajas que tuvieran una duración no superior a 35 días, y de periodo largo a las que durasen más. A continuación podemos ver un esquema en el que se ilustra el modelo comentado anteriormente.

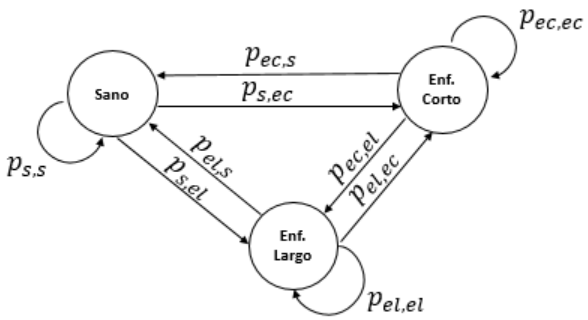


Figura 4.4: Representación de la cadena del Modelo 2.

A continuación, queremos obtener una matriz de transición P que nos permita estudiar la probabilidad que existe de pasar de un estado transitorio a otro o la de permanecer en el mismo. Así, si denotamos por $ec \equiv$ estar enfermo durante un periodo corto de tiempo, por $el \equiv$ estar enfermo durante un periodo largo y por $s \equiv$ estar sano, tenemos que la matriz de transición que estamos buscando se define como:

$$P = \begin{pmatrix} p_{ec,ec} & p_{ec,el} & p_{ec,s} \\ p_{el,ec} & p_{el,el} & p_{el,s} \\ p_{s,ec} & p_{s,el} & p_{s,s} \end{pmatrix}.$$

Pasamos ahora, a obtener los coeficientes de la matriz P. Para ello, creamos para cada uno de los trabajadores con algún episodio de cervicalgia durante el periodo de tiempo 2006–2013 y que están en el grupo `datos.trabajo`, una tabla que conste de dos columnas de modo que la primera de ellas, contenga los distintos estados por los que ha pasado y la segunda los días que permaneció en dichos estados. Además, es necesario tener en cuenta, que estamos considerando que si el número de días que el trabajador estuvo sano entre dos procesos de baja distintos es menor o igual que 30, entonces suponemos que pasó directamente de una enfermedad a la otra, es decir, no tenemos en cuenta esos días en los que estuvo sano. Para hacer todo esto creamos una función, la cual llamamos `extraer.historia` y que podemos encontrar en el script `Modelo2.R` de la sección A.6 del apéndice. Un ejemplo de estas tablas para uno de nuestros pacientes se puede observar a continuación.

```
  Estado Dias.Baja
1      3       653
2      2        63
3      3       626
4      1        11
5      3        77
```

Donde:

- $1 \equiv$ tener una enfermedad por un periodo corto de tiempo.
- $2 \equiv$ padecer la enfermedad durante un periodo largo.
- $3 \equiv$ estar sano.

Mirando este ejemplo, podemos decir que este trabajador ha tenido dos procesos de baja por cervicalgia en el intervalo de tiempo 2006–2013, uno por un periodo largo de tiempo durante 63 días y otro por un periodo corto de 11 días. Además, podemos observar que entre estos dos procesos de baja, el trabajador permaneció sano durante 626 días. Así mismo, este paciente empieza y termina estando sano durante 653 y 77 días, respectivamente.

Para obtener los coeficientes de la matriz P aparte de todo lo anterior, hay que tener en cuenta los saltos que se producen al pasar de un estado transitorio al otro. También, es importante saber que estamos considerando que la unidad de salto es de 5 días cuando estamos en el estado transitorio enfermo durante un periodo corto o en el estado sano, y que dicha unidad de salto es de 35 días si estamos en el estado enfermo por periodo largo. Esto nos lleva a que los saltos de un estado al mismo se obtengan dividiendo entre x los días que el paciente permanece en cada uno de ellos (`días.estado`), siendo $x = 5$ cuando estamos en los estados *ec* o *s* y $x = 35$ si el estado es *el*. Además, tenemos que tener en cuenta lo siguiente:

$$\begin{cases} \text{saltos} = \frac{dias.estado}{x} - 1, & \text{si} \quad \frac{dias.estado}{x} \equiv \text{entera} \\ \text{saltos} = \lfloor \frac{dias.estado}{x} \rfloor, & \text{si} \quad \frac{dias.estado}{x} \equiv \text{no entera} \end{cases}$$

Con todo esto aclarado y teniendo en cuenta que nuestras probabilidades se obtienen a través de los saltos explicados anteriormente, al ejecutar el script recogido en la sección A.6 del apéndice, llegamos a que la matriz de transición estimada es:

$$P = \begin{pmatrix} 0{,}99845 & 0{,}00001 & 0{,}00153 \\ 0 & 0{,}99902 & 0{,}00098 \\ 0{,}00155 & 0{,}00097 & 0{,}99749 \end{pmatrix}$$

Como podemos observar, esta matriz nos dice que la probabilidad de pasar del estado transitorio enfermo largo al estado enfermo corto es 0, esto implica que una vez que un individuo entre en este estado (enfermo largo) o permanece en él o por el contrario, pasa a estar sano. Este hecho hace que el esquema considerado previamente (4.4) se vea modificado de la siguiente forma:

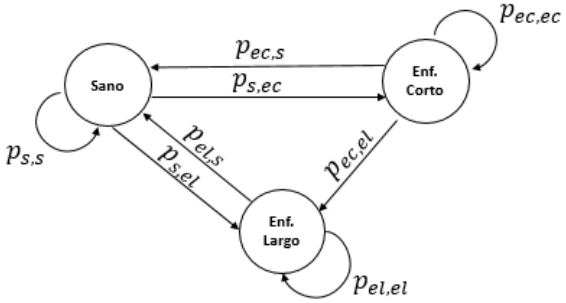


Figura 4.5: Representación de la cadena del Modelo 2, según la matriz P anterior.

Para concluir y al igual que hicimos en la sección 4.1, reflexionaremos sobre una serie de aspectos como pueden ser: la duración media de las bajas y el tiempo que cada trabajador permanece de baja al año.

- **Duración media de las bajas.**
  Basándonos en los trabajadores que presentaron algún episodio por cervicalgia entre los años 2006–2013, podemos obtener información acerca de la duración media de las bajas para un individuo que inicia una suspensión del trabajo por esta enfermedad. Para calcular esta media, tendremos en cuenta el esquema recogido en la figura 4.6, los teoremas 4.1 y 4.2.

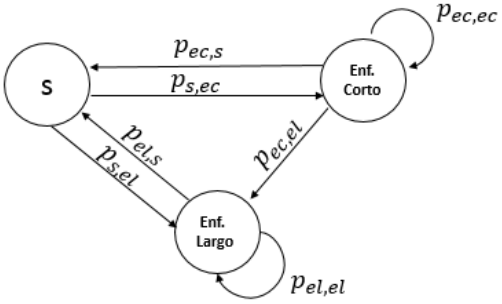


Figura 4.6: Representación de la cadena utilizada para el cálculo del tiempo medio de recurrencia en el Modelo 2.

la matriz de transición asociada a dicho esquema es:

$$P_2 = \begin{pmatrix} p_{ec,ec} & p_{ec,el} & p_{ec,s} \\ p_{el,ec} & p_{el,el} & p_{el,s} \\ p_{s,ec} & p_{s,el} & p_{s,s} \end{pmatrix} = \begin{pmatrix} 0{,}99845 & 0{,}00001 & 0{,}00153 \\ 0 & 0{,}99902 & 0{,}00098 \\ 0{,}61508 & 0{,}38492 & 0 \end{pmatrix}$$

donde $p_{s,ei} = \frac{p_{s,ei}}{p_{s,ec}+p_{s,el}}$, con $i = c, l$. Teniendo en cuenta que las probabilidades son las obtenidas en la matriz P del modelo 2.

Queremos calcular ahora el vector $w = (w_{ec};\ w_{el};\ w_s)$. Para ello basta con resolver el $\lim_{n\to\infty} P_2^n$, según el teorema 4.2. Así, si diagonalizamos y elevamos a la n-ésima potencia tenemos que:

$$\lim_{n\to\infty} P_2^n = \begin{pmatrix} 0{,}4968199 & 0{,}4977698 & 0{,}001247955 \\ 0{,}5009703 & 0{,}5019281 & 0{,}001258380 \\ 0{,}4984199 & 0{,}4993729 & 0{,}001251974 \end{pmatrix} \simeq \begin{pmatrix} 0{,}5 & 0{,}5 & 0 \\ 0{,}5 & 0{,}5 & 0 \\ 0{,}5 & 0{,}5 & 0 \end{pmatrix}$$

Concluimos así que el vector $w = (w_{ec};\ w_{el};\ w_s) = (0{,}5;\ 0{,}5;\ 0)$ y por tanto, al aplicar el teorema 4.1 teniendo en cuenta que $w_i \equiv$ sano, el tiempo medio de recurrencia es $r_s = \frac{1}{w_s} = \frac{1}{0{,}5} = 2$días. Esto nos dice que cada episodio de baja por cervicalgia dura en media unos 2 días.

- **Tiempo (proporción) de baja al año.**
  Para obtener el tiempo que cada trabajador está de baja al año por cervicalgia, basándonos en este segundo modelo nos apoyamos en el cálculo de las posiciones de equilibrio. Para obtener estas posiciones de equilibrio basta con aplicar el teorema 4.2 a la matriz P obtenida a partir de este segundo modelo.

  Repitiendo los mismos pasos que en el primer modelo (diagonalizar la matriz P y elevarla a la n-ésima potencia), obtenemos las posiciones de equilibrio.

$$\lim_{n\to\infty} P^n = \begin{pmatrix} 0{,}33266 & 0{,}33437 & 0{,}33296 \\ 0{,}33266 & 0{,}33437 & 0{,}33296 \\ 0{,}33266 & 0{,}33437 & 0{,}33296 \end{pmatrix}$$

  Podemos concluir así que las posiciones de equilibrio vienen dadas por el vector $w = (w_{ec};\ w_{el};\ w_s) = (0{,}33266;\ 0{,}33437;\ 0{,}33296)$. Esto nos dice que la proporción de tiempo que cada trabajador que inicia un proceso de baja por cervicalgia es de $(0{,}33266 + 0{,}33437) \cdot 365 = 243{,}4660$ días.

4.3. Modelo 3

En el Análisis Clúster realizado en el capítulo 3 concluimos que los pares de enfermedades que se presentaban en el mismo trabajador con más frecuencia eran:

- Enfermedades infecciosas (E17) y Trastorno del ojo y anexos (E44)
- Esguince parte superior del brazo (E20) y Esguince de espalda (E21)
- Esguince de rodilla y pierna (E22) y Trastorno interno de rodilla (E45)

Así mismo decidimos que de entre todos estos pares de enfermedades, estudiaríamos en profundidad el correspondientes a los diagnósticos: esguince de rodilla y pierna (E22) y trastorno interno de rodilla (E45).

Teniendo en cuenta lo anterior es claro que nuestro modelo consta de tres estados transitorios: estar sano, tener la enfermedad esguince de rodilla y pierna o la enfermedad trastorno interno de rodilla. Por lo tanto, sólo nos vamos a centrar en los pacientes que padecieron alguna de estas dos enfermedades (o las dos), en el intervalo de tiempo para el que tenemos información, y que están dentro del grupo `datos.trabajo`, que como ya hemos explicado en numerosas ocasiones, dicho grupo contiene los datos con los que realizamos nuestro estudio. A continuación se presenta un esquema que puede ayudar a entender mejor el modelo que queremos desarrollar.

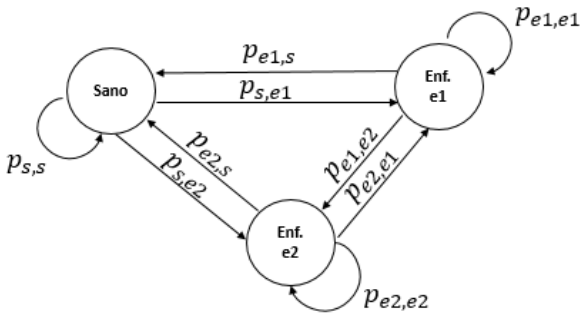


Figura 4.7: Representación de la cadena del Modelo 3.

Nuestro objetivo ahora, es buscar una matriz de transición que nos permita saber la probabilidad que existe de que una persona que está sana lo siga estando al día siguiente, o que por el contrario pase a estar enfermo. Así mismo estamos interesados en conocer la probabilidad que existe de que un trabajador que éste enfermo por alguna de estas dos enfermedades lo siga estando al día siguiente, o pase a tener la otra o se recupere. Además si denotamos por $s \equiv$ estado transitorio estar sano, $e1 \equiv$ tener la enfermedad esguince de rodilla y pierna y por $e2 \equiv$ padecer trastorno interno de rodilla, sabemos que la estructura de esta matriz de transición es:

$$P = \begin{pmatrix} p_{e1,e1} & p_{e1,e2} & p_{e1,s} \\ p_{e2,e1} & p_{e2,e2} & p_{e2,s} \\ p_{s,e1} & p_{s,e2} & p_{s,s} \end{pmatrix}$$

Para obtener los coeficientes de esta matriz tenemos que tener en cuenta las siguientes variables: días de baja, diagnóstico y fechas de baja y de alta de cada episodio, años de antigüedad de cada uno de los trabajadores a los que MAC pasó consulta durante los años 2006–2013, las cuales han sido explicadas en el capítulo 2. A través de ellas, obtenemos la variable $\mathbf{alta} = \mathbf{max}\{(\mathbf{fechas.baja} - \mathbf{antig}), \mathbf{fecha1}\}$ (explicada en la sección 4.1).

En este caso también es necesario crear una tabla similar a la obtenida en la sección 4.2 pero considerando ahora que:

- $1 \equiv$ tener la enfermedad esguince de rodilla y pierna.
- $2 \equiv$ padecer la enfermedad trastorno interno de rodilla.
- $3 \equiv$ estar sano.

Aclarado esto, para obtener los coeficientes de la matriz de transición tenemos que tener en cuenta el número de saltos que se producen al pasar de un estado transitorio al otro. Además, en este caso estamos considerando que la unidad de salto es de 5 días, por lo que los saltos de un estado al mismo se obtienen dividiendo entre 5 los días que el pacientes permanece en cada uno de esos estados (`días.estado`), teniendo en cuenta lo siguiente:

$$\begin{cases} \text{saltos} = \frac{dias.estado}{5} - 1, & \text{si} \quad \frac{dias.estado}{5} \equiv \text{entera} \\ \text{saltos} = \lfloor \frac{dias.estado}{5} \rfloor, & \text{si} \quad \frac{dias.estado}{5} \equiv \text{no entera} \end{cases}$$

Al ejecutar el script `Modelo3.R` (véase en sección A.7 del apéndice), obtenemos la siguiente matriz de transición:

$$P = \begin{pmatrix} 0{,}90417 & 0{,}00059 & 0{,}09524 \\ 0{,}00065 & 0{,}94078 & 0{,}05858 \\ 0{,}00170 & 0{,}00049 & 0{,}99782 \end{pmatrix}$$

Por último y para concluir esta sección, daremos una serie de resultados a modo de conclusión. Para ello, estudiaremos los siguientes aspectos:

- **Duración media de las bajas.**
  Apoyándonos en los trabajadores que han padecido algún episodio de trastorno interno de rodilla y pierna o esguince interno de rodilla, podemos obtener una estimación de la duración media de las bajas producidas por estas enfermedades. Para obtener dicha estimación, basta con aplicar los teoremas 4.1 y 4.2.

  Antes de aplicar los resultados mencionados anteriormente, hay que tener presente que estamos considerando que s_i es el estado estar sano, por lo tanto, la cadena en la que nos apoyamos para estimar la duración media de bajas es ligeramente diferente a la del modelo en cuestión. Dicha cadena se puede observar en la figura 4.8 y tiene asociada la matriz de transición siguiente:

$$P_3 = \begin{pmatrix} p_{e1,e1} & p_{e1,e2} & p_{e1,s} \\ p_{e2,e1} & p_{e2,e2} & p_{e2,s} \\ p_{s,e1} & p_{s,e2} & p_{s,s} \end{pmatrix} = \begin{pmatrix} 0{,}90417 & 0{,}00059 & 0{,}09524 \\ 0{,}00065 & 0{,}94078 & 0{,}05858 \\ 0{,}77626 & 0{,}22374 & 0 \end{pmatrix}$$

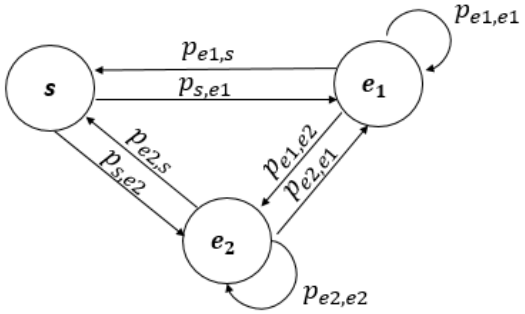


Figura 4.8: Representación de la cadena utilizada para el cálculo del tiempo medio de recurrencia en el Modelo 3.

donde $p_{s,ei} = \frac{p_{s,ei}}{p_{s,e1}+p_{s,e2}}$, con $i = 1, 2$. Teniendo en cuenta que las probabilidades son las obtenidas en la matriz P del modelo 2.

Queremos calcular ahora el vector $w = (w_{e1};\ w_{e2};\ w_s)$. Para ello basta con resolver el $\lim_{n\to\infty} P_3^n$, según el teorema 4.2.

$$\lim_{n\to\infty} P_3^n = \begin{pmatrix} 0{,}62578 & 0{,}29716 & 0{,}07701 \\ 0{,}62588 & 0{,}29721 & 0{,}07702 \\ 0{,}62580 & 0{,}29717 & 0{,}07701 \end{pmatrix} \simeq \begin{pmatrix} 0{,}626 & 0{,}297 & 0{,}077 \\ 0{,}626 & 0{,}297 & 0{,}077 \\ 0{,}626 & 0{,}297 & 0{,}077 \end{pmatrix}$$

De aquí concluimos que el vector es $w = (w_{e1}; w_{e2}; w_s) = (0{,}626; 0{,}297; 0{,}077)$ y por tanto, al aplicar el teorema 4.1, el tiempo medio de recurrencia es $r_s = \frac{1}{w_s} = \frac{1}{0{,}077} =$ 12,9870días. Esto nos dice que cada episodio de baja por esguince de rodilla y pierna o trastorno interno de rodilla dura en media unos 12,9870 días.

- **Distribución de las longitudes de las bajas.**
  Queremos estudiar la distribución de la longitudes de las bajas, teniendo en cuenta los procesos de baja de aquellos trabajadores con algún episodio de esguince de rodilla y pierna o trastorno interno de rodilla durante los años 2006 – 2013. Para ello, representaremos gráficamente a través de histogramas, los vectores `long.bajas1` y `long.bajas2` que podemos encontrar en el script `Modelo3.R` (sección A.7 del apéndice). Dichos vectores recogen de forma acumulada la duración de los procesos de bajas de las enfermedades e1 (esguince de rodilla y pierna) y e2 (trastorno interno de rodilla), respectivamente. Representamos, también, a través de una línea de color rojo la verdadera distribución de la geométrica tomando como probabilidad de éxito, según la definición 4.4, $p_{e1,s} + p_{e1,e2}$, en el caso de considerar el vector `long.bajas1`, y $p_{e2,s} + p_{e2,e1}$ cuando representamos el vector `long.bajas2`. Dichas representaciones se harán agrupando las longitudes de las bajas en intervalos de tamaño 5 y la dis-

tribución de la geométrica en intervalos igual a la unidad. Teniendo en cuenta esto, obtenemos las figuras 4.9 y 4.10 que se muestran a continuación.

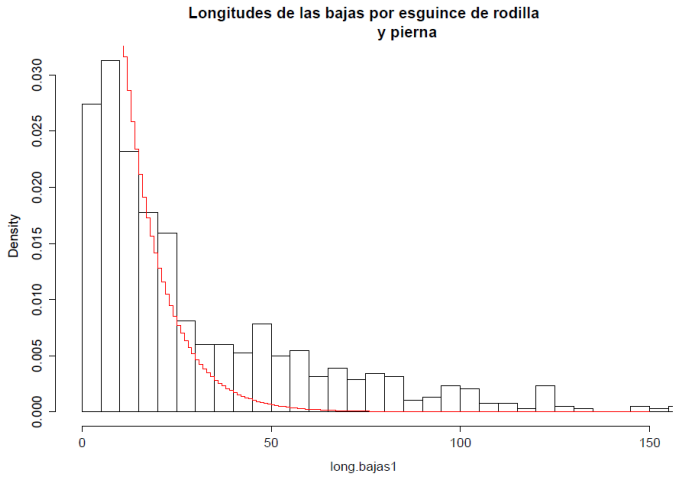


Figura 4.9: Histograma de la duración de las bajas debidas a esguince de rodilla y pierna. En rojo las longitudes de las bajas correspondientes a una distribución geométrica de parámetro $p_{e1,s} + p_{e1,e2} = 0{,}09583$

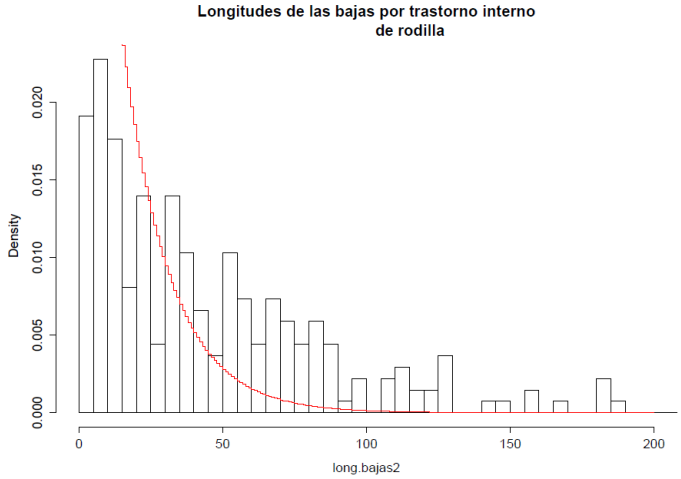


Figura 4.10: Histograma de la duración de las bajas debidas a trastorno interno de rodilla. En rojo las longitudes de las bajas correspondientes a una distribución geométrica de parámetro $p_{e2,s} + p_{e2,e1} = 0{,}05923$

A la vista de estos dos histogramas podemos concluir que la forma general del que representa las longitudes de las bajas por trastorno interno de rodilla (figura 4.10), dista mucho más de la forma de una geométrica que el que representa las longitudes

de las bajas por esguince de rodilla y pierna (figura 4.9). Sin embargo, se puede observar que los procesos de baja por esguince de rodilla y pierna (ver figura 4.9), con una duración de mayor de 20 días están siempre por encima de la línea roja que representa la distribución de la geométrica de parámetro $p_{e1,s}$. Esto mismo sucede con la enfermedad trastorno interno de rodilla (ver figura 4.10), pero en este caso la duración de las bajas está por encima de la distribución de la geométrica para aquellos procesos con una duración superior a 35 días.

- **Tiempo (proporción) de baja al año.**
  Para obtener el tiempo que cada trabajador está de baja al año por esguince de rodilla y pierna o trastorno interno de rodilla, basándonos en aquellos que ya han tenido algún episodio por estos diagnósticos entre los años 2006–2013, nos apoyamos en el cálculo de las posiciones de equilibrio. Esto es posible al estar nuestros modelos fundamentados en los procesos de cadenas de Markov. Para obtener estas posiciones de equilibrio basta con aplicar el teorema 4.2, enunciado en el apartado dedicado a la duración media de las bajas, a la matriz P obtenida a partir de este tercer modelo.

  Repitiendo los mismos pasos que en el primer modelo (diagonalizar la matriz P y elevarla a la n-ésima potencia), obtenemos las posiciones de equilibrio sin más que resolver el siguiente límite:

$$\lim_{n\to\infty} P^n = \begin{pmatrix} 0{,}01729 & 0{,}00818 & 0{,}97452 \\ 0{,}01729 & 0{,}00818 & 0{,}97452 \\ 0{,}01729 & 0{,}00818 & 0{,}97452 \end{pmatrix}$$

  Podemos concluir así que las posiciones de equilibrio vienen dadas por el vector $w = (w_{e1};\ w_{e2};\ w_s) = (0{,}01729;\ 0{,}00818;\ 0{,}97452)$. Esto nos dice que la proporción de tiempo que cada trabajador que inicia un proceso de baja por esguince de rodilla y pierna está de baja al año es $0{,}01729 \cdot 365 = 6{,}3109$ días, y por trastorno interno de rodilla es de $0{,}00818 \cdot 365 = 2{,}2857$ días.

Nota 4.1 *Para ver las demostraciones de todos los teoremas utilizados en este capítulo, remitirse al libro [2].*

Capítulo 5

Validación del modelo

Esta sección tiene como finalidad la comprobación de los resultados obtenidos en el desarrollo de este trabajo. Para ello tenemos que tener presente que inicialmente, en el capítulo 2 dividimos los datos originales que nos proporcionó MAC en dos grupos, de forma que uno de ellos y al que llamamos `datos.trabajo` contuviese el 60 % de los datos y el otro llamado `datos.control` el resto.

El desarrollo de todo el modelo recogido en los capítulos anteriores, se hizo para aquellos trabajadores pertenecientes al primero de estos grupos. Esto nos lleva a utilizar la información restante recogida en el grupo `datos.control`, en la comprobación de los resultados obtenidos y por tanto en la validación del modelo.

5.1. Relación entre las distintas enfermedades

Lo que se pretende conseguir en esta sección es encontrar aquellas enfermedades que aparecen con más frecuencia y de forma simultánea en los individuos, para ello y como ya vimos en el capítulo 5.1 nos apoyamos en los dendrogramas obtenidos tras aplicar un análisis clúster jerárquico.

En el estudio realizado en el capítulo mencionado anteriormente, habíamos concluido que los pares de enfermedades más comunes eran:

1. Enfermedades infecciosas (E17) y Trastorno del ojo y anexos (E44)

2. Esguince parte superior del brazo (E20) y Esguince de espalda (E21)

3. Esguince de rodilla y pierna (E22) y Trastorno interno de rodilla (E45)

Además tras analizar los datos originales, decidimos estudiar en profundidad el par correspondiente a los diagnósticos E22 y E45 .

Lo que pretendemos ahora, es encontrar aquellos pares de enfermedades más frecuentes en los trabajadores pertenecientes al grupo `datos.control`. Para ello basta ejecutar el script `Cluster.R` (sección A.4 del apéndice), sustituyendo `datos.control` por `datos.trabajo`. Al hacer esto, obtenemos el siguiente dendrograma:

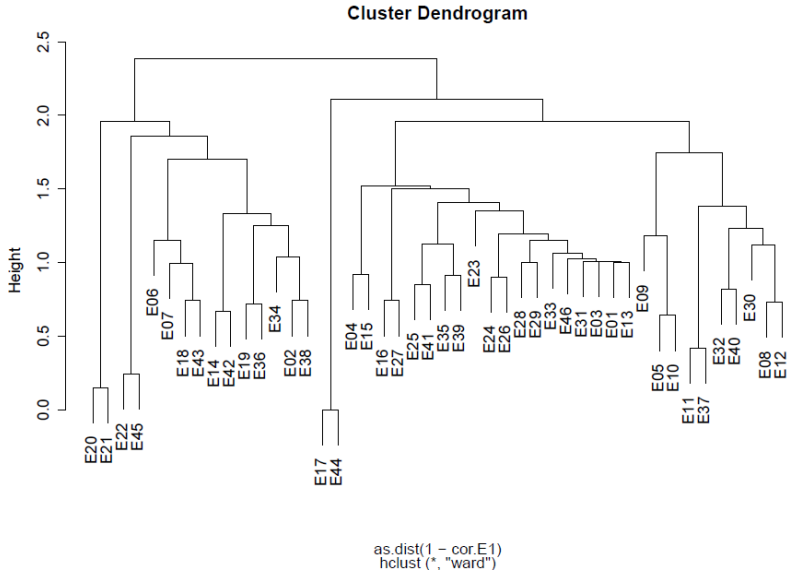


Figura 5.1: Dendrograma con la agrupación de las enfermedades utilizando la distancia de Ward.

Como podemos observar, los pares de enfermedades más relacionados coinciden con los citados anteriormente. Esto nos lleva a pensar que por el momento los resultados obtenidos son correctos.

5.2. Cadenas de Markov

Esta sección está enfocada al desarrollo de modelos que nos permitan entender mejor como se comportan los distintos procesos de baja.

5.2.1. Modelo 1

Este primer modelo consistía en una cadena de Markov con dos estados transitorios (estar sano o enfermo por cervicalgia). Lo que pretendemos ahora, es sacar conclusiones a partir de los datos, para contrastarlas con la obtenida mediante el modelo. Estas conclusiones serán sobre las variables: el número de bajas y de días de baja al año y las longitudes de las mismas.

- **Número de bajas.**
  Estamos interesados en predecir el número medio de bajas al año por cervicalgia,

teniendo en cuenta los trabajadores que han padecido algún episodio por esta enfermedad durante los años 2006–2013. Para ello, necesitamos conocer el número de bajas por cervicalgia que ha presentado cada uno de estos pacientes. Esta información la obtenemos a partir del vector `N.Bajas` (script `Modelo1.R` de la sección A.5 del apéndice). Además, tenemos que conocer los días que cada uno de estos trabajadores lleva dado de alta en el sistema, esto equivale a la suma de los días que cada uno de ellos ha estado enfermo más los que ha estado sano. En este caso, el vector `Dias.Total` nos proporciona esta información.

Vamos a calcular, ahora, la media de bajas al año para cada uno de los grupos de datos que tenemos. Si consideramos `datos.trabajo`, podemos definir una variable `pacientes` formada por aquellos trabajadores que hayan padecido algún episodio por cervicalgía y que pertenezcan a este grupo. Teniendo en cuenta todo esto, podemos obtener dicha media a través de la operación siguiente:

$$\text{Nº de bajas al año} = \frac{\sum_{p \in pacientes} N.Baja(p)}{\sum_{p \in pacientes} Dias.Total(p)} \cdot 365 = 0{,}1919$$

Haciendo lo mismo, pero en este caso considerando que la variable `pacientes` contiene a los pacientes del grupo `datos.control`, llegamos a que:

$$\text{Nº de bajas al año} = 0{,}1930$$

De estas dos operaciones se concluye que el número medio de bajas al año por cervicalgia es de 0,191, cuando implementamos `datos.trabajo`, y de 0,1930 días, cuando usamos `datos.control`.

Sabemos que si multiplicamos el número de bajas que cada trabajador tiene al año, por la duración media de cada una de estas bajas, obtenemos la media de bajas al año. Por lo tanto, el número medio de bajas al año que nos aporta el modelo se obtiene sin más que dividir la media de bajas al año entre la duración de las mismas. Al hacer esto, obtenemos que el número medio de días que cada trabajador que inicia un proceso de baja por cervicalgia está de baja al año es de 0,1538. Teniendo en cuenta el valor de esta media obtenida a través de nuestros datos, podemos concluir que este primer modelo predice un número medio de bajas al año algo inferior al valor real.

- **Duración de las bajas.**
  Estamos interesados en obtener la duración media de las bajas por cervicalgia a partir de nuestros datos, para ello utilizamos el comando `mean(long.bajas)` para obtener la media de las longitudes de las bajas por cervicalgia . Al hacer esto, obtenemos que dicha media es de 40,616 días cuando usamos `datos.trabajo` y es de

40,2641 cuando trabajamos con `datos.control`.

Teniendo en cuenta que a través del primer modelo obteníamos que cada episodio de baja por cervicalgia dura en media unos 49,0436 días, y que según los datos esta media está entorno a los 40 días, podemos concluir que este modelo obtiene una media de las longitudes de las bajas por cervicalgia superior a la real (la que nos proporcionan los datos).

- **Días (proporción) de baja.**
  En este estudiaremos la media de días de baja al año por cervicalgia. Para ello, tenemos que conocer el número de días que cada trabajador ha estado de baja por algún episodio de esta enfermedad. Esta información la obtenemos al partir del vector `Dias.Baja` (script `Modelo1.R` de la sección A.5). Así mismo, necesitamos conocer los días que cada uno de los trabajadores llevan dado de alta en el sistema.

  Teniendo en cuenta que `pacientes` es una variable que contiene a los trabajadores del grupo `datos.trabajo` con algún episodio de cervicaliga.

$$\text{Media de bajas al año} = \frac{\sum_{p \in pacientes} Dias.Baja(p)}{\sum_{p \in pacientes} Dias.Total(p)} \cdot 365 = 7{,}794148$$

  Si considerando ahora que la variable `pacientes` pertenece al grupo `datos.control`, tenemos que:

$$\text{Media de bajas al año} = 7{,}7712$$

  A través de estas dos operaciones llegamos a una media de 7,7941 días de baja por cervicalgia y trabajador al año, al utilizar el grupo `datos.trabajo`, y de 7,7712 días al considerar `datos.control`.

  A través del Modelo 1 concluíamos que la proporción de tiempo que está de baja al año cada trabajador que inicia un procesos de baja por cervicalgia es de 7,5409 días. Fijándonos en las medias anteriores, vemos que tanto a través de los `datos.trabajo` como de los `datos.control` llegamos a una media similar de días de baja al año, por esta misma enfermedad. Por lo tanto, podemos concluir que este primer modelo ajusta bastante bien la proporción de días de baja al año.

- **Distribución de las longitudes de las bajas.**
  Para estudiar la distribución de las longitudes de la bajas por cervicalgia, procedemos del mismo modo que en la sección 4.1 de este mismo apartado, teniendo en cuenta

la definición (4.4) e implementando el script `Modelo1.R` (sección A.5), pero usando ahora los `datos.control`. Al hacer esto obtenemos el histograma que se muestra a continuación en la figura 5.2.

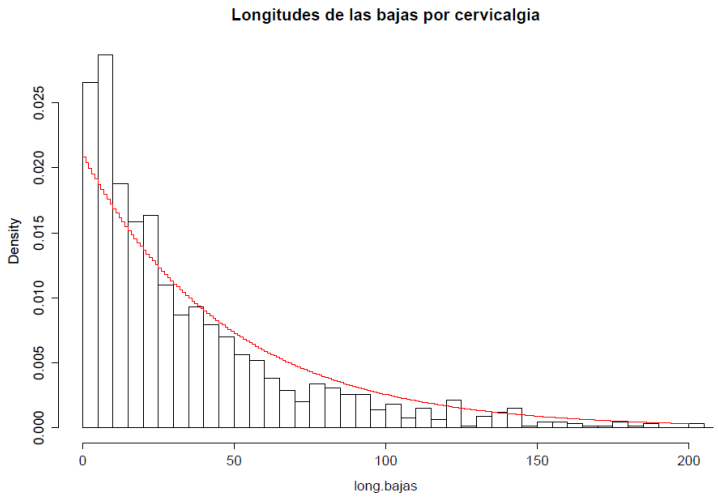


Figura 5.2: Histograma de la duración de las bajas debidas a cervicalgia. En rojo las probabilidades correspondientes a una distribución geométrica de parámetro $p_{e,s} = 0{,}02081$

A la vista del histograma podemos concluir que la forma general no se aleja mucho de la forma de una geométrica, aunque dicha geométrica presenta una media mucho mayor que la de los datos.

5.2.2. Modelo 2

Este segundo modelo consta de en una cadena de Markov con tres estados transitorios (estar sano o enfermo por cervicalgia durante un periodo corto de tiempo o durante un periodo largo). Lo que se pretendemos en esa sección, es obtener conclusiones a partir de los datos, para contrastarlas con las obtenida mediante el modelo. Estas conclusiones serán sobre las variables: el número de bajas y de días de baja al año y las longitudes de las mismas.

En este caso, podemos concluir que tanto el *número medio de bajas* al año por cervicalgia, la *media de días de baja* por esta misma enfermedad y trabajador al año y la *media de las longitudes de las bajas* debidas a este mismo diagnóstico, coinciden con las obtenidas en la subsección 5.4. Esto se debe a que en ambos casos, para obtener esta información, nos basábamos en el mismo conjunto de individuos, formado por aquellos trabajadores que padecieron algún episodio de cervicalgia durante los años 2006–2013, por lo que también estamos considerando la misma enfermedad.

Finalmente concluimos que:

- Sabemos que si multiplicamos el número de bajas que cada trabajador tiene al año, por la duración media de cada una de estas bajas, obtenemos la media de bajas al año. Por lo tanto, el número medio de bajas al año que nos aporta el modelo se obtiene sin más que dividir la media de bajas al año entre la duración de las mismas. Al hacer esto, obtenemos que el número medio de días que cada trabajador que inicia un proceso de baja por cervicalgia está de baja al año es de 121,73. Además el valor de esta media, obtenida a través de nuestros datos, es de 0,19. Por lo tanto, podemos concluir que este segundo modelo predice un número medio de bajas peor que el anterior.

- Teniendo en cuenta que a través del segundo modelo obteníamos que cada episodio de baja por cervicalgia dura en media unos 2 días, y que según los datos esta media está entorno a los 40 días, podemos concluir que este modelo obtiene una media de las longitudes de las bajas por cervicalgia inferior a la real.

- Además este modelo nos permitía concluir que la proporción de tiempo que está de baja al año cada trabajador que inicia un procesos de baja por cervicalgia es de 243,4660 días. Fijándonos en las medias obtenidas a partir de los datos, vemos que este tercer modelo predice una media muy superior a la real.

De este modo llegamos a que el Modelo 2 es peor que el Modelo 1. Por lo tanto, sería necesario buscar otro modelo que ayude a predecir mejor las medias anteriores.

5.2.3. Modelo 3

Este tercer modelo desarrolla una cadena de Markov con tres estados (sano, enfermo por esguince de rodilla y pierna y enfermo por trastorno interno de rodilla). Al igual que hemos hecho en los modelos anteriores, nos proponemos sacar conclusiones a partir de los datos, con el propósito de compararla con la obtenida a partir del modelo y validar así la eficiencia del mismo. Estas conclusiones serán sobre las variables:

- **Número de bajas.**
  Queremos obtener información acerca del número medio de bajas al año por esguince de rodilla y pierna (e1) y por trastorno interno de rodilla (e2), basándonos en la información que conocemos sobre los trabajadores que han padecido algún episodio por estas enfermedades. Por lo tanto, necesitamos conocer los vectores `N.Bajas1` y `N.Bajas2` localizados en el script `Modelo3.R` de la sección A.7, los cuales recogen el número de bajas por e1 y e2, respectivamente. También es necesario conocer el número de días que cada trabajador lleva dado de (`Dias.Total`).

Denotando por `pacientes` a los trabajadores que han tenido algún episodio de esguince de rodilla y pierna o trastorno interno de rodilla (o por los dos) y que pertenecen al grupo `datos.trabajo`, podemos calcular:

$$\text{Nº bajas al año por e1} = \frac{\sum_{p \in pacientes} N.Bajas1(p)}{\sum_{p \in pacientes} Dias.Total1(p)} \cdot 365 = 0{,}13779$$

$$\text{Nº bajas al año por e2} = \frac{\sum_{p \in pacientes} N.Bajas2(p)}{\sum_{p \in pacientes} Dias.Total2(p)} \cdot 365 = 0{,}04887$$

Procediendo de igual forma, pero considerando en este caso que `pacientes` pertenece al grupo `datos.control`, llegamos a que las medias buscadas son:

$$\text{Nº bajas al año por e1} = 0{,}1430$$

$$\text{Nº bajas al año por e2} = 0{,}0440$$

De aquí concluimos que el número medio de bajas al año por esguince de rodilla y pierna es de 0,13779 y por trastorno interno de rodilla es de 0,04887 para `datos.trabajo`. Así mismo para `datos.control` este número medio es de 0,1430 y de 0,0440, respectivamente.

Para obtener el número medio de bajas al año que nos aporta el modelo, hay que dividir la media de bajas al año entre la duración de las mismas. Al hacer esto, obtenemos que el número medio de días que cada trabajador, que inicia un proceso de baja por esguince de rodilla y pierna, está de baja al año es de 0,4859 y es de 0,1760 si la baja es trastorno interno de rodilla. Teniendo en cuenta el valor de estas medias obtenidas a través de nuestros datos, podemos concluir que este tercer modelo predice un número medio de bajas al año mayor al real.

- **Duración de las bajas.**
  Utilizamos los comandos `mean(long.bajas1)` y `mean(long.bajas2)` para obtener la media de las longitudes de las bajas por esguince de rodilla y pierna (e1) y trastorno interno de rodilla (e2), respectivamente. Al hacer esto, obtenemos que dicha media para e1 es de 42,97914 días y para e2 es de 56,09191 cuando utilizamos `datos.trabajo`; y que dicha media es de 49,74494 y es 47,32237 cuando trabajamos con `datos.control`.

Teniendo en cuenta que cada episodio de baja por esguince de rodilla y pierna (e1) o trastorno interno de rodilla (e2) dura en media unos 12,9870 días. A la vista de las medias obtenidas anteriormente, podemos concluir que este tercer modelo se queda muy por debajo de la duración media real de los procesos de baja por estas enfermedades.

- **Días (proporción) de bajas.**
  Para estudiar la media de días de baja por e1 y e2 por trabajador al año, necesitamos conocer el número de días que los estos trabajadores han estado de baja (por padecer algún episodio de los mencionados anteriormente); y el número de días que cada uno de ellos lleva dado de alta en el sistema. Esta información la podemos encontrar el script `Modelo3.R` (sección A.7 del apéndice), a través de los vectores `Dias.Baja1`, `Dias.Baja2` y `Dias.Total`, respectivamente.

  Denotando por `pacientes` a los trabajadores que han presentado algún episodio por e1 o e2, y que pertenecen al grupo `datos.trabajo`, llegamos a que:

$$\text{Media bajas al año (e1)} = \frac{\sum_{p \in pacientes} Dias.Baja1(p)}{\sum_{p \in pacientes} Dias.Total(p)} \cdot 365 = 5{,}92220$$

$$\text{Media bajas al año (e2)} = \frac{\sum_{p \in pacientes} Dias.Baja2(p)}{\sum_{p \in pacientes} Dias.Total(p)} \cdot 365 = 2{,}74094$$

  Si consideramos ahora que `pacientes` contiene a los trabajadores que han padecido alguna de las enfermedades anteriores (o las dos), y que pertenecen al grupo `datos.control`, obtenemos que la media buscada es:

$$\text{Media bajas al año (e1)} = 7{,}11106$$

$$\text{Media bajas al año (e2)} = 2{,}08146$$

  De aquí podemos concluir que, la media es de 5,92 días de baja por esguince de rodilla y pierna al año por trabajador cuando usamos `datos.trabajo` y de 7,11 días cuando tomamos `datos.control`. Así mismo, podemos decir que la media es de 2,74094 días de baja por trastorno interno de rodilla al año por trabajador, cuando trabajamos con `datos.trabajo` y de 2,08 días cuando usamos `datos.control`.

  A través del Modelo 3 concluíamos que la proporción de tiempo que está de baja al año cada trabajador que inicia un procesos de baja por esguince de rodilla y pierna

es de 6,3109 días y por trastorno interno de rodilla de 2,2857 días al año. Fijándonos en lo obtenido anteriormente, vemos que tanto a través de los `datos.trabajo` como de los `datos.control` llegamos aproximadamente al mismo resultado. Por lo tanto, podemos concluir que este tercer modelo ajusta bien la proporción de días de baja al año, para las enfermedades en cuestión que estamos considerando.

- **Distribución de las longitudes de las bajas.**
  Para estudiar la distribución de las longitudes de la bajas por esguince de rodilla y pierna y trastorno interno de rodilla, procedemos del mismo modo que en la sección 5.1. En este caso implementamos el script `Modelo3.R` (sección A.7 del apéndice) sustituyendo `datos.trabajo` por `datos.control`. Al hacer eso, se obtienen los histogramas que se muestran a continuación en las figuras 5.3 y 5.4, donde `long.bajas1` y `long.bajas2` son vectores formados por las duraciones de los procesos de bajas de aquellos trabajadores que padecieron algún episodio por esguince de rodilla y pierna (e1) y trastorno interno de rodilla (e2), respectivamente.

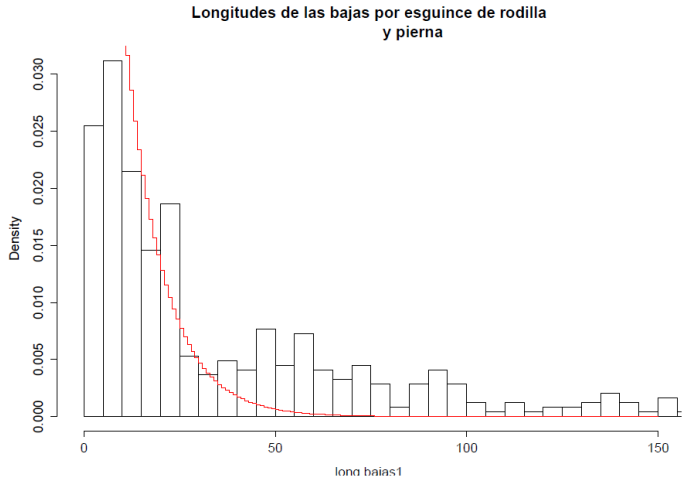


Figura 5.3: Histograma de la duración de las bajas debidas a esguince de rodilla y pierna. En rojo las longitudes de las bajas correspondientes a una distribución geométrica de parámetro $p_{e1,s} + p_{e1,e2} = 0{,}08206$

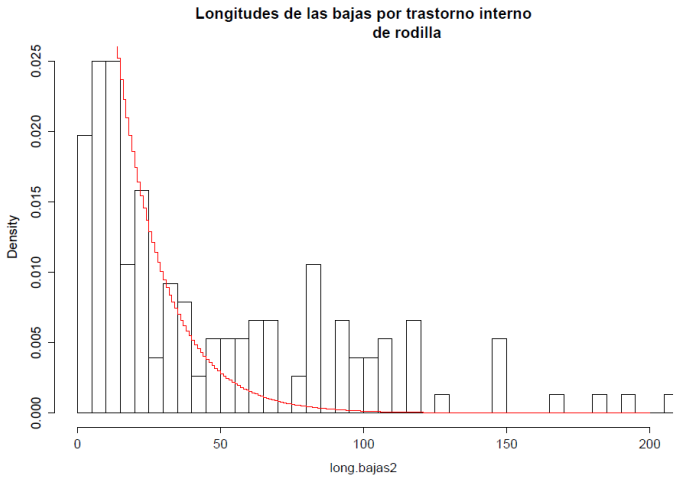


Figura 5.4: Histograma de la duración de las bajas debidas a trastorno interno de rodilla. En rojo las longitudes de las bajas correspondientes a una distribución geométrica de parámetro $p_{e2,s} + p_{e2,e1} = 0{,}06787$

A la vista de estos dos histogramas podemos concluir que la forma general del que representa las longitudes de las bajas por trastorno interno de rodilla (figura 5.4), dista mucho más de la forma de una geométrica que el que representa las longitudes de las bajas por esguince de rodilla y pierna (figura 5.3).

Apéndice A

Scripts

A.1. Importación de los datos

```
#Fichero: TomaDatos.R

#Este script se utiliza para llamar los datos con los que vamos a trabajar
#y que tenemos guardados en el documento Consulta.Arreglo.txt
setwd("C:/Users/Alexandra/Desktop/Proyecto")
load("C:/Users/Alexandra/Desktop/Proyecto/.RData")
datos <- read.table("Consulta.Arreglo.txt", header = TRUE, sep = "\t",
                    dec = ",")
datos
```

A.2. División de los datos en dos grupos

```
#Fichero: DivisionDatos2Grupos.R

#En este script, haremos uso de muestras aleatorias para dividir los datos
#que tenemos  en dos grupos, de  forma que uno de ellos, el que llamaremos
#datos.trabajo y con el que  vamos a realizar nuestro estudio, contenga el
#60% de los datos y el otro, datos.cotrol, el cual usaremos para comprobar
#los resultados obtenidos contendrá el resto.
tfrec <- table(datos$paciente)
df.frec <- data.frame(tfrec)
names(df.frec) <- c('pac','frec')
df.frec$pac <- as.numeric(names(tfrec))
tm <- 0.6  #tasa de muestreo

# sorteo es un vector con los números de paciente que han salido sorteados
sorteo <- numeric()
for(i in 1:max(df.frec$frec)){
```

```
  df.frecuencias <- df.frec$pac[df.frec$frec == i]
  t <- length(df.frecuencias)
  nuevo.sorteo <- sample(c(T,F), t, prob=c(tm,1 - tm), replace=T)
  sorteo <- c(sorteo, df.frecuencias[nuevo.sorteo])
}

#esta.sorteado  es una liste de TRUE O FALSE y de dimension igual a la del
#numero de pacientes, donde TRUE representa a los pacientes que han salido
#en el sorteo y FALSE a los que no.
esta.sorteado <- datos$pac == 0
for (i in sorteo){
  esta.sorteado[datos$pac == i] <- TRUE
}

datos.trabajo <- datos[esta.sorteado,]
datos.control <- datos[!esta.sorteado,]
```

A.3. Correlación entre las enfermedades

```
#Fichero: CorrelacionEnfermedades.R

#Este script consta de tres partes:

#En la  primera, creamos  un catálogo en el  que se recogen las  distintas
#enfermedades y le asignamos a cada una de ellas un código, con  el fin de
#identificarlas de forma mas visual.
enfermedades <- levels(datos.trabajo$descripcion.diag)
codigosE <- c(paste("E0", 1:9, sep = ""), paste("E", 10:46, sep = ""))
catalogo <- data.frame(Desc.Enf = enfermedades, Codigos = codigosE)

#La segunda, se utiliza  para  obtener una tabla que  tenga por  filas los
#distintos  pacientes  y  por columnas las enfermedades,  de forma  que se
#muestre que enfermedad ha tenido  cada  pacientes y cuantas  veces la  ha
#padecido.
tablaP.E <- table(datos.trabajo$paciente, datos.trabajo$descripcion.diag)
dimnames(tablaP.E)[[2]] <- catalogo[,2]

#En la tercera, crearemos una matriz de correlación entre las enfermedades
#y  las  pintamos, con el fin de obtener  información sobre que  variables
#podrían estar relacionas.
correlacionE <- cor(tablaP.E)

#Sustitumos la diagonal de la matriz anterior por el valor 0 con el fin de
```

```
#ver un poco mejor la relación  que pueden existir entre  las enfermedades
#cuando pintamos dicha matriz.
diag(correlacionE) <- 0
image(correlacionE)
```

A.4. Clúster

```
#Fichero: Cluster.R

#En este  Script vamos a  realizar un  análisis cluster con el objetivo de
#identificar que enfermedades están mas relacionadas.

#En primer lugar dividiremos los elementos de la matriz correlacionE entre
#el elemento máximo de la misma, sin tener en cuenta el elemento diagonal 1.
correlacionE1 <- cor(tablaP.E)
corE <- round((correlacionE1 - diag(1, 46, 46)) / max(correlacionE1 -
               - diag(1, 46, 46)), 2)

#A continuación, vamos a realizar el análisis cluster de  la matriz  corE,
#para ello:
#Guardamos en la variable etiquetas las  cabeceras de la matriz corE, las
#cuales se corresponden con los códigos que  identifican a las  distintas
#enfermedades.
etiquetas <- row.names(corE)

#Dibujamos el dendrograma  utilizando  la  distancia enlace complejo como
#diastancia entre grupos.
plot(hclust(as.dist(1 - corE)), main = 'Dendrograma')

#En este caso utilizamos como distancia entre grupos el enlace simple.
plot(hclust(as.dist(1 - corE), method = "single"))

#Por último utilizaremos el método de Ward.
plot(hclust(as.dist(1 - corE), method = "ward"))
```

A.5. Modelo 1

```
#Fichero: Modelo1.R

#Con este script se obtenemos las probabilidades de la matriz de transición
#en  un proceso de cadenas de Markov con dos estados transitorios (sano -
#enfermo por cervicalgia).
```

```
pacientes <- datos.trabajo$paciente[datos.trabajo$descripcion.diag ==
                                      'CERVICALGIA']
episodios.e1ye2 <- datos.trabajo[datos.trabajo$descripcion.diag ==
                                   'CERVICALGIA', ]

#[fecha1, fecha2]  constituye  en intervalo de tiempo en el que se mueven
#nuestros datos
fecha1 <- as.Date('2006/01/01')
fecha2 <- as.Date('2013/12/31')

#Con esta función se pretende calcular una tabla formada por dos columnas,
#la primera de ellas con los estados (sano o enfermo) de cada  uno de los
#pacientes y la segunda con el nº de días de cada uno de los estados.
extraer.historia = function(episodiospaciente){
  dias.baja <- episodiospaciente$dias.baja
  diagnostico <- episodiospaciente$descripcion.diag
  fechas.baja <- as.Date(episodiospaciente$fecha.baja, format = '%d-%m-%y')
  fechas.alta <- as.Date(episodiospaciente$fecha.alta, format = '%d-%m-%y')
  antig <- round(episodiospaciente$anios.antig * 365)
  #Tiempo que el trabajador lleva en el sistema
  alta <- max((fechas.baja[1] - antig[1]), fecha1)
  dias.enfermo <- sum(dias.baja)
  dias.sano <- as.numeric(fecha2 - alta - dias.enfermo, unit='days')
  estado <- 2 #Simboliza el estado estar sano
  dias <- as.numeric(fechas.baja[1] - alta, unit = 'days')
  for(i in 1:length(diagnostico)){
    if(i > 1){
      dias.sano[i] <- as.numeric(fechas.baja[i] - fechas.baja[i-1], unit =
                                   'days')
      estado[length(estado) + 1] <- 2
      dias[length(dias) + 1] <- dias.sano[i] - sum(dias.baja[i-1])
    }
    if(diagnostico[i] == 'CERVICALGIA'){
      estado[length(estado) + 1] <- 1
      dias[length(dias) + 1] <- sum(dias.baja[i])
    }
  }
  estado[length(estado) + 1] <- 2
  dias[length(dias) + 1] <- as.numeric(fecha2 - fechas.alta[i], unit =
                                         'days')
  matriz <- data.frame(Estado = estado, Dias.Baja = dias)
  return(matriz)
}
```

```
#Creamos una matriz de saltos formada por dos filas y dos columnas.
saltos <- matrix(0, nrow = 2, ncol = 2)
fila <- c('e', 's')
columna <- c('e','s')
dimnames(saltos) <- list(fila, columna)

#Definimos los siguientes vectores vacíos.
long.bajas <- c()
N.Bajas <- c()
Dias.Baja <- c()
Dias.Total <- c()

for(p in pacientes){
  historia <- extraer.historia(episodios.e1ye2[episodios.e1ye2$paciente ==
                                                 p,])
  #Acumulamos las longitudes de los distintos procesos de baja.
  long.bajas <- append(long.bajas, historia$Dias.Baja[historia$Estado ==
                                                        1])
  #Guardamos el número de bajas de cada paciente.
  N.Bajas <- append(N.Bajas, sum(historia$Estado == 1))
  #Sumamos los días de bajas de cada paciente.
  Dias.Baja <- append(Dias.Baja, sum(historia$Dias.Baja[historia$Estado ==
                                                          1]))
  #Contamos los días que cada paciente lleva en el sistema.
  Dias.Total <- append(Dias.Total, sum(historia$Dias.Baja))
  #Este bucle será usado para obtener los saltos de un estado a él mismo.
  for(i in 1:length(historia[, 2])){
    for(j in 1:2){
      if(historia[i, 1] == j){
        saltos[j, j] <- saltos[j, j] + (historia[i, 2] - 1)
      }
    }
  }
  Estado <- historia$Estado
  #Calculamos a continuación los saltos de un estado a otro distinto.
  for(i in 1:(length(historia.enfermoi[, 2]) - 1)){
    saltos[Estado[i], Estado[i+1]] <- saltos[Estado[i], Estado[i+1]] + 1
  }
}

#La matriz M es una matriz de probabilidades obtenida a partir de los saltos.
M <- matrix(0, nrow = 2, ncol = 2)
filas <- c('e1', 's')
columnas <- c('e1', 's')
```

```
dimnames(M) <- list(filas, columnas)

M[1,] <- saltos[1,] / sum(saltos[1, ])
M[2,] <- saltos[2,] / sum(saltos[2, ])

#Creamos un data.frame con los vectores definidos anteriormente.
tabla <- data.frame(N.Bajas = N.Bajas, Dias.Baja = Dias.Baja, Dias.Total =
                      Dias.Total)

#Nos  apoyamos  en los histogramas para representar la distribución  de las
#longitudes de las bajas.
histograma1 <- hist(long.bajas, breaks = c(seq(0, 250), 750),
                    xlim = c(0,200), plot = 'true',
                    main = 'Longitudes de las bajas por cervicalgia')

#En rojo representamos la verdadera distribución de la  geométrica  tomando
#como probabilidad de exito M.es.
lines(0:200, dgeom(0:200, prob = M[1,2]), type = "s", col = "red")

#Repetiremos  el  mismo procedimiento anterior, pero agrupando en este caso
#las longitudes de las bajas en intervalos de tamaño 5.
histograma2 <- hist(long.bajas, breaks = c(seq(0, 250, by = 5), 750),
                    xlim = c(0,200), plot = 'true',
                    main = 'Longitudes de las bajas por cervicalgia')
lines(0:200, dgeom(0:200, prob = M[1,2]), type = "s", col = "red")
```

A.6. Modelo 2

```
#Fichero: Modelo2.R

#Este  script se utiliza  para hallar las probabilidades  de la  matriz de
#transición  en  un   proceso  de  cadenas  de  Markov  con  tres  estados
#transitorios (sano - enfermo periodo corto - enfermo periodo largo).

pacientes <- datos.trabajo$paciente[datos.trabajo$descripcion.diag ==
                                      'CERVICALGIA']
episodios.e1ye2 <- datos.trabajo[datos.trabajo$descripcion.diag ==
                                   'CERVICALGIA', ]

# [fecha1,fecha2] constituye en  intervalo de tiempo en el  que  se mueven
#nuestros datos
fecha1 <- as.Date('2006/01/01')
```

```
fecha2 <- as.Date('2013/12/31')

ec <- 'enfermo.corto'
el <- 'enfermo.largo'

#Con esta función se pretende calcular una tabla formada por dos columnas,
#la primera de ella  con  los  estados  (sano, enfermo e1 o enfermo e2) de
#cada  uno  de los pacientes y la segunda con el nº de días de cada uno de
#los estados.
extraer.historia = function(episodiospaciente){
  dias.baja <- episodiospaciente$dias.baja
  diagnostico <- episodiospaciente$descripcion.diag
  fechas.baja <- as.Date(episodiospaciente$fecha.baja, format = '%d-%m-%y')
  fechas.alta <- as.Date(episodiospaciente$fecha.alta, format = '%d-%m-%y')
  antig <- round(episodiospaciente$anios.antig * 365)
  #Tiempo que el trabajador lleva en el sistema
  alta <- max((fechas.baja[1] - antig[1]), fecha1)
  dias.enfermo <- sum(dias.baja)
  dias.sano <- as.numeric(fecha2 - alta - dias.enfermo, unit='days')
  estado <- 3
  dias <- as.numeric(fechas.baja[1] - alta, unit = 'days')
  for(i in 1:length(diagnostico)){
    if(i > 1){
      dias.sano[i] <- as.numeric(fechas.baja[i] - fechas.baja[i-1], unit
                                 = 'days')
      if(dias.sano[i] > 30){#Con esto consideramos saltos  de  enfermo  a
                            #sano sólo cuando entre dos procesos de  baja,
                            #el trabajador está sano más de 30 días
        estado[length(estado) + 1] <- 3
        dias[length(dias) + 1] <- dias.sano[i] - sum(dias.baja[i-1])
      }
    }
    if(dias.baja[i] <= 35){
      estado[length(estado) + 1] <- 1
      dias[length(dias) + 1] <- sum(dias.baja[i])
    }
    else{
      estado[length(estado) + 1] <- 2
      dias[length(dias) + 1] <- sum(dias.baja[i])
    }
  }
  estado[length(estado) + 1] <- 3
  dias[length(dias) + 1] <- as.numeric(fecha2 - fechas.alta[i], unit = 'days')
  matriz <- data.frame(Estado = estado, Dias.baja = dias)
```

```
  return(matriz)
}

#Creamos una matriz de saltos con tres filas y tres columnas.
saltos <- matrix(0, nrow = 3, ncol = 3)
f <- c('ec','el','s')
c <- c('ec','el','s')
dimnames(saltos) <- list(f, c)

#Definimos una serie de vectores vacíos.
long.bajas <- c()
N.Bajas <- c()
Dias.Baja <- c()
Dias.Total <- c()

for(p in pacientes){
  historia <- extraer.historia(episodios.e1ye2[episodios.e1ye2$paciente ==
                                                 p,])
  #Acumulamos las longitudes de los distintos procesos de baja.
  long.bajas <- append(long.bajas, historia$Dias.baja[historia$Estado !=
                                                        3])
  #Guardamos el número de bajas de cada paciente.
  N.Bajas <- append(N.Bajas, sum(historia$Estado != 3))
  #Sumamos los días de bajas de cada paciente.
  Dias.Baja <- append(Dias.Baja, sum(historia$Dias.baja[historia$Estado !=
                                                          3]))
  #Contamos los días que cada paciente lleva en el sistema.
  Dias.Total <- append(Dias.Total, sum(historia$Dias.baja))
  for(i in 1:length(historia[, 2])){
    for(j in 1:3){
      if(historia[i, 1] == j){
        saltos[j, j] <- saltos[j, j] + (historia[i, 2] %/% 5)
        if((historia[i, 2] %% 5) == 0){
          saltos[j, j] <- saltos[j, j] - 1
        }
      }
    }
  }
  Estado <- historia$Estado
  for(i in 1:(length(historia[, 2])-1)){
    saltos[Estado[i], Estado[i+1]] <- saltos[Estado[i], Estado[i+1]] + 1
  }
}
```

```
#La  matriz  M  es  una  matriz de probabilidades obtenida a partir de los
#saltos.
M <- matrix(0, nrow = 3, ncol = 3)
filas <- c('ec', 'el', 's')
columnas <- c('ec', 'el', 's')
dimnames(M) <- list(filas, columnas)

M[1,] <- saltos[1,] / sum(saltos[1, ])
M[2,] <- saltos[2,] / sum(saltos[2, ])
M[3,] <- saltos[3,] / sum(saltos[3, ])

#Creamos un data.frame con los vectores definidos anteriormente.
tabla <- data.frame(N.Bajas = N.Bajas, Dias.Baja = Dias.Baja, Dias.Total =
                    Dias.Total)
```

A.7. Modelo 3

```
#Fichero: Modelo3.R

#Con este Scriptse obtenemos las probabilidades de la matriz de transición
#en  un proceso de cadenas de Markov con tres estados transitorios (sano -
#enfermo e1 - enfermo e2).

#Cogemos  todos  los pacientes que hayan tenido algún episodio de baja por
#Trastorno interno de rodilla o Esguince de rodilla y pierna.
episodios.e1ye2 <- datos.trabajo[(datos.trabajo$descripcion.diag ==
                                  'ESGUINCE RODILLA Y PIERNA') |
                                   (datos.trabajo$descripcion.diag ==
                                    'TRASTORNO INTERNO DE RODILLA'),]
pacientes <- datos.trabajo$paciente[(datos.trabajo$descripcion.diag ==
                                     'ESGUINCE RODILLA Y PIERNA') |
                                     (datos.trabajo$descripcion.diag ==
                                     'TRASTORNO INTERNO DE RODILLA')]

#[fecha1, fecha2]  constituye  en  intervalo de timepo en el que se mueven
#nuestros datos
fecha1 <- as.Date('2006/01/01')
fecha2 <- as.Date('2013/12/31')

#Definimos las siguientes variables:
e1 <- 'ESGUINCE RODILLA Y PIERNA'
e2 <- 'TRASTORNO INTERNO DE RODILLA'
```

```
#Con esta función se pretende calcular una tabla formada por dos columnas,
#la  primera  de  ellas con los estados (sano, enfermo e1 o enfermo e2) de
#cada  uno  de los pacientes y la segunda con el nº de días de cada uno de
#los estados.
extraer.historia = function(episodiospaciente){
  dias.baja <- episodiospaciente$dias.baja
  diagnostico <- episodiospaciente$descripcion.diag
  fechas.baja <- as.Date(episodiospaciente$fecha.baja, format = '%d-%m-%y')
  fechas.alta <- as.Date(episodiospaciente$fecha.alta, format = '%d-%m-%y')
  antig <- round(episodiospaciente$anios.antig * 365)
  #Tiempo que el trabajador lleva en el sistema
  alta <- max((fechas.baja[1] - antig[1]), fecha1)
  dias.enfermo <- sum(dias.baja)
  dias.sano <- as.numeric(fecha2 - alta - dias.enfermo, unit='days')
  estado <- 3
  dias <- as.numeric(fechas.baja[1] - alta, unit = 'days')
  for(i in 1:length(diagnostico)){
    if(i > 1){
      dias.sano[i] <- as.numeric(fechas.baja[i] - fechas.baja[i-1], unit =
                                   'days')
      if(dias.sano[i] > 30){#Con esto consideramos saltos  de  enfermo  a
                            #sano sólo cuando entre dos procesos de  baja,
                            #el trabajador está sano más de 30 días
        estado[length(estado) + 1] <- 3
        dias[length(dias) + 1] <- dias.sano[i] - sum(dias.baja[i-1])
      }
    }
    if(diagnostico[i] == e1){
      estado[length(estado) + 1] <- 1
      dias[length(dias) + 1] <- sum(dias.baja[i])
    }
    else{
      estado[length(estado) + 1] <- 2
      dias[length(dias) + 1] <- sum(dias.baja[i])
    }
  }
  estado[length(estado) + 1] <- 3
  dias[length(dias) + 1] <- as.numeric(fecha2 - fechas.alta[i], unit =
                                         'days')
  matriz <- data.frame(Estado = estado, Dias.Baja = dias)
  return(matriz)
}
```

```
#Creamos una matriz de saltos formada por tres filas y tres columnas.
saltos <- matrix(0, nrow = 3, ncol = 3)
fila <- c('e1','e2','s')
columna <- c('e1','e2','s')
dimnames(saltos) <- list(fila, columna)

#Definimos los siguientes vectores vacíos.
long.bajas1 <- c()
N.Bajas1 <- c()
Dias.Baja1 <- c()
long.bajas2 <- c()
N.Bajas2 <- c()
Dias.Baja2 <- c()
Dias.Total <- c()

for(p in pacientes){
  historia <- extraer.historia(episodios.e1ye2[episodios.e1ye2$paciente ==
                                                 p,])
  #Acumulamos las longitudes de los distintos procesos de baja teniendo en
  #cuenta el tipo de enfermedad.
  long.bajas1 <- append(long.bajas1, historia$Dias.Baja[historia$Estado ==
                                                          1])
  long.bajas2 <- append(long.bajas2, historia$Dias.Baja[historia$Estado ==
                                                          2])
  #Guardamos el número de bajas de cada paciente teniendo en cuenta el tipo
  #de enfernedad.
  N.Bajas1 <- append(N.Bajas1, sum(historia$Estado == 1))
  N.Bajas2 <- append(N.Bajas2, sum(historia$Estado == 2))
  #Sumamos los días de bajas de cada paciente teniendo en cuenta el tipo de
  #enfermedad.
  Dias.Baja1 <- append(Dias.Baja1, sum(historia$Dias.Baja[historia$Estado ==
                                                            1]))
  Dias.Baja2 <- append(Dias.Baja2, sum(historia$Dias.Baja[historia$Estado ==
                                                            2]))
  #Contamos los días que cada paciente lleva en el sistema.
  Dias.Total <- append(Dias.Total, sum(historia$Dias.Baja))
  for(i in 1:length(historia[, 2])){
    for(j in 1:3){
      if(historia[i, 1] == j){
        saltos[j, j] <- saltos[j, j] + (historia[i, 2] %/% 5)
        if((historia[i, 2] %% 5) == 0){
          saltos[j, j] <- saltos[j, j] - 1
        }
      }
```

```
    }
  }
  Estado <- historia$Estado
  for(i in 1:(length(historia.enfermoi[, 2])-1)){
    saltos[Estado[i], Estado[i+1]] <- saltos[Estado[i], Estado[i+1]] + 1
  }
}

#La matriz M es una matriz de probabilidades obtenida a partir de los saltos.
M <- matrix(0, nrow = 3, ncol = 3)
filas <- c('e1', 'e2', 's')
columnas <- c('e1', 'e2', 's')
dimnames(M) <- list(filas, columnas)

M[1,] <- saltos[1,] / sum(saltos[1, ])
M[2,] <- saltos[2,] / sum(saltos[2, ])
M[3,] <- saltos[3,] / sum(saltos[3, ])

#Nos  apoyamos  en los histogramas para representar la distribución  de las
#longitudes de las bajas de cada una de las enfermedades.
histograma.e1 <- hist(long.bajas1, breaks = c(seq(0, 250, by = 5), 750),
                    xlim = c(-1,150), plot = 'true',
                    main = 'Longitudes de las bajas por esguince de rodilla
                    y pierna')

#En rojo representamos la verdadera distribución de la geométrica
lines(0:150, dgeom(0:150, prob = M[1,3]), type = "s", col = "red")

#Repetimos el mismo proceso pero en este caso para e2
histograma.e2 <- hist(long.bajas2, breaks = c(seq(0, 250, by = 5), 750),
                    xlim = c(0,200), plot = 'true',
                    main = 'Longitudes de las bajas por trastorno interno
                    de rodilla')
lines(0:200, dgeom(0:200, prob = M[2,3]), type = "s", col = "red")
```

Bibliografía

[1] Michael J. Crawley. *The R Book*. John Wiley & Sons Ltd, England, 2007.

[2] Charles M. Gristead and J. Laurie Snell. *Intrduction to Probability*. American Mathematical Society, 2006.

[3] Kenneth Lange. Discrete-Time Markov Chains. *American Scientist*, 78(8):151–158, 1990.

[4] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[5] M. Mercedes Suárez Rancel. *Análisis de Datos Avanzados*. 2005. ISBN 84-609-3840-9.

BEHAVIOUR OF THE PROCESSES OF SICK LEAVES

Probabilistics models

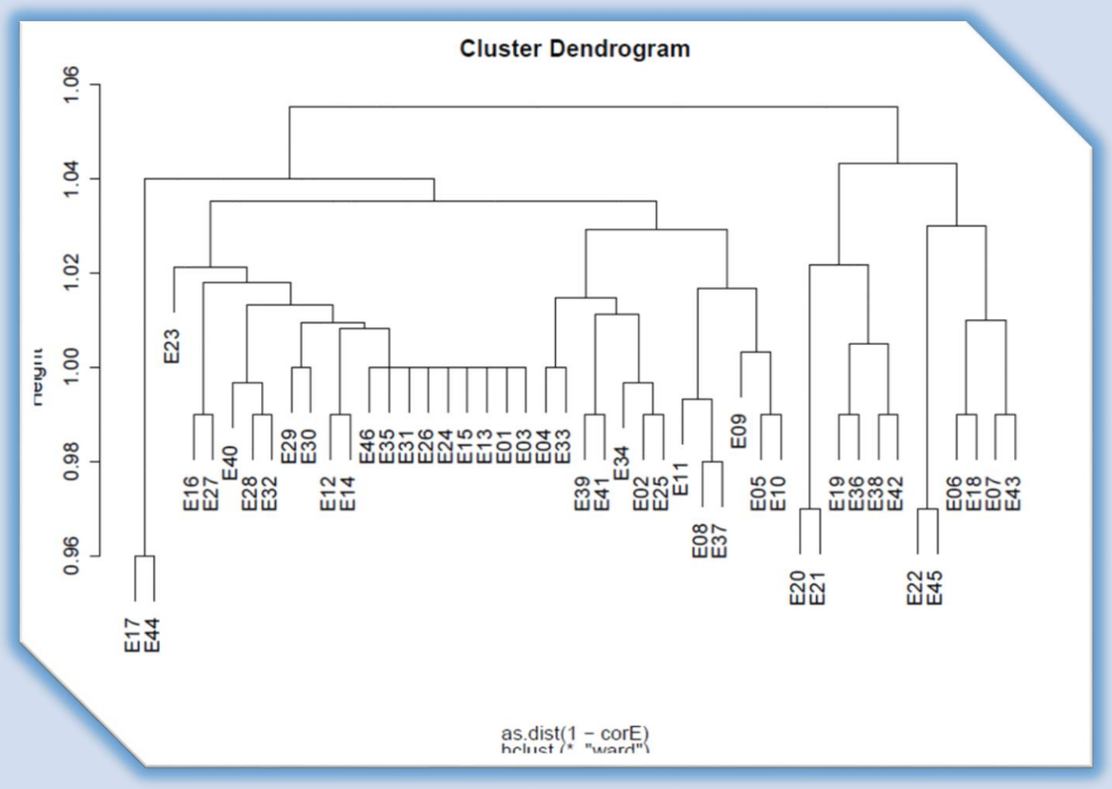
Alexandra García Lorenzo

Objective

The principal objective of this work is develop a series of models that help us to understand better the behavior of processes of sick leaves.

Relationship beteween illnesses

Before develop the models, we have find which are the illnesses that appear with more frequency and simultaneously in individuals. For this, we support to cluster analysis, thus obtaining the next dendrogram.

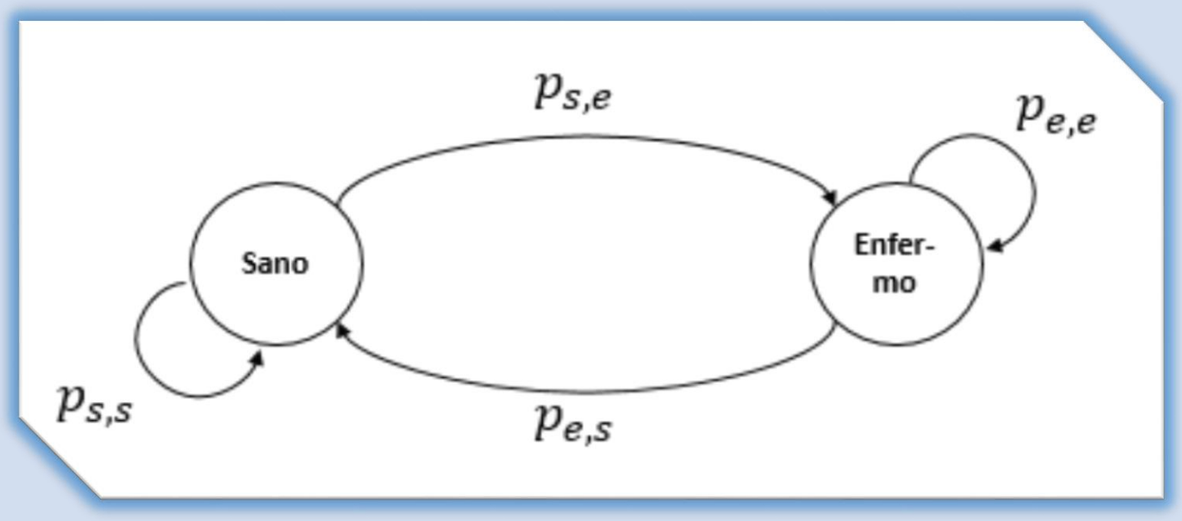


Markov's chains

Support us in the processes of Markovs chains ,we decide develop three models for study the behavior of sick leaves. These models are:

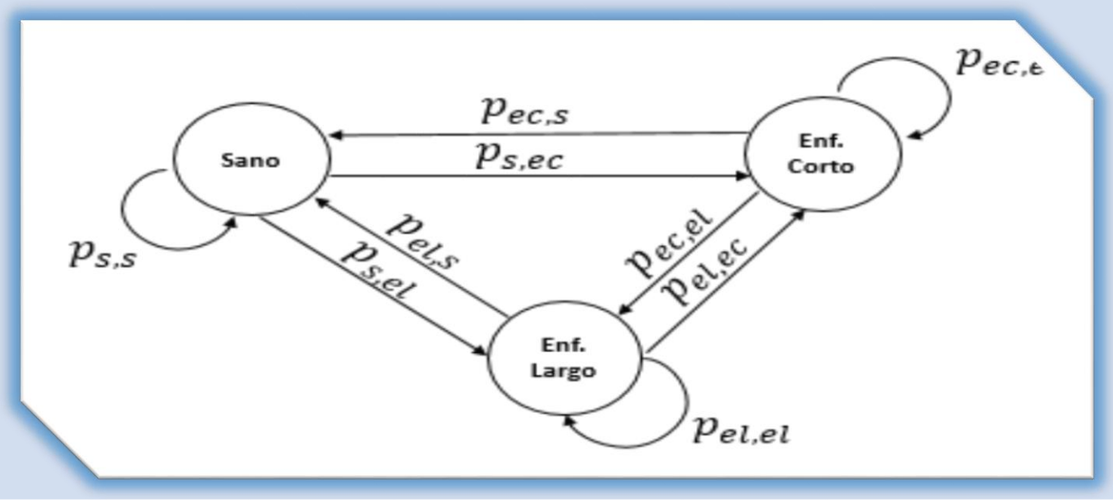
Model 1

This model is the easiest of the three. It consist in a Markov chain whith three transient states.



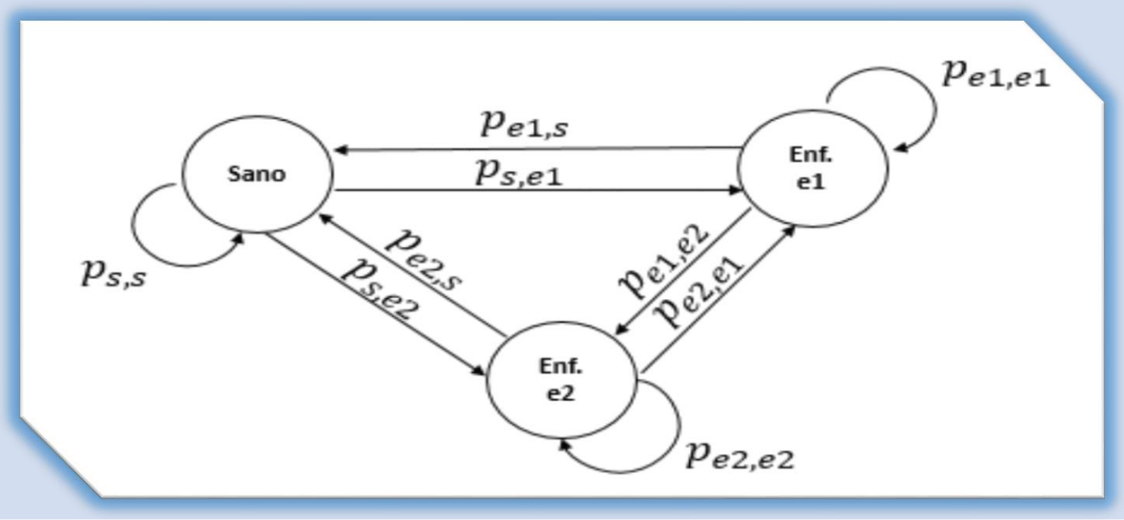
Model 2

This model try to improve the previous model, adding a new state with the purpose to difference the illnesses in long or short duration.



Model 3

This last model regard the pairs of illnesses most frequent. We obtain a Markov chain with three states.



Conclusion

For finally this memory, we will obtain conclusions about processes of the behivior fo sicke leaves.