Sección de Matemáticas
Universidad de La Laguna

Andrea Marcos Vargas

Evolución del alumnado de la Facultad de Ciencias mediante cadenas de Markov

Evolution of the students of the Faculty of Sciences by means of Markov chains

Trabajo Fin de Grado
Grado en Matemáticas
La Laguna, Julio de 2022

Dirigido por
Carlos González Alcón

Carlos González Alcón
Departamento de Matemáticas, Estadística e Investigación Operativa
Universidad de La Laguna
38204 La Laguna, Tenerife

Agradecimientos

A mi familia y amigos por apoyarme siempre. A mi tutor, Carlos, por ayudarme a avanzar cada semana. A mis compañeros, son lo mejor que me llevo de todos estos años. Muchísimas gracias.

Andrea Marcos Vargas
La Laguna, 6 de julio de 2022

Resumen · Abstract

Resumen

El abandono universitario es una problemática que ocupa la atención de organismos públicos de numerosos países debido a las consabidas repercusiones economicas y sociales que tiene. A pesar de que la tasa de abandono universitario se ha reducido respecto a años anteriores, situada en torno al 30 % *en 2009, todavía se mantiene en un porcentaje superior a la media europea (*16 %*), ya que se sitúa en torno al* 19 %*. En este trabajo se presentarán dos modelos probabilísticos obtenidos a partir de los datos ofrecidos por el Gabinete de Análisis y Planificación de La Universidad de La Laguna, y cuyo objetivo es asignar a cada grado de la facultad de Ciencias una cadena de Markov para conocer el rendimiento académico de sus alumnos. Las cadenas de Markov se analizan con el entorno y lenguaje de programación estadístico R.*

Palabras clave: *Cadena de Markov – Matriz de transición – Probabilidad de transición – Programación en R.*

Abstract

University dropout is a problem that occupies the attention of public bodies in many countries due to the well-known economic and social repercussions it has. Although the university dropout rate has decreased with respect to previous years, standing at around 30 % *in 2009, it is still higher than the European average (*16 %*), which is around* 19 %*. This paper presents two probabilistic models obtained from the data provided by the Analysis and Planning Office of the University of La Laguna, the aim of which is to assign a Markov chain to each degree of the Faculty of Science in order to find out the academic performance of its students. The Markov chains are analysed with the R statistical programming language and environment.*

Keywords: *Markov Chain – Transition Matrix – Transition Probability – R programming.*

Contenido

Agradecimientos III

Resumen/Abstract V

Introducción IX

1. Marco Teórico 1
1.1. Procesos Estocásticos 1
1.2. Cadenas de Markov 2
1.2.1. Cadenas de Markov absorbentes 8
1.2.2. Cadenas de Markov ergódicas 12

2. Análisis de los datos 15
2.1. Modelo 1: por créditos o asignaturas 16
2.1.1. Presentación del modelo 16
2.1.2. Preparación de los datos 17
2.1.3. Análisis realizado 18
2.1.4. Resultados obtenidos 20
2.2. Modelo 2: por cursos 28
2.2.1. Presentación del modelo 28
2.2.2. Análisis realizado 28
2.2.3. Resultados obtenidos 29

A. Apéndice 37
A.1. Código en el lenguaje R 37

Bibliografía 43

Lista de símbolos y abreviaciones 45

Poster 47

Introducción

El abandono universitario puede ser un problema relevante tanto a nivel institucional como social. Desde el punto de vista institucional, se produce un evidente coste de recursos públicos que no aumentan el capital humano del país, así como problemas organizativos y de reputación para las universidades que lo sufren. En el plano social, los estudiantes que abandonan se frustran al no alcanzar sus objetivos, considerando la univerdad como una pérdida de tiempo y recursos, sin retorno económico evidente.

La mala fama que acompaña a algunas titulaciones o su dificultad real hacen que el número de abandonos sea alarmante. Podría ser este el caso de los grados de la facultad de Ciencias, ya que dichas carreras son consideradas complejas porque más allá de memorizar resultados, se necesita llegar a comprenderlos. Según un estudio del *BBVA* [13], la tasa de abandono en las titulaciones de Ciencias puras es del 31.1 %.

Este trabajo se centra en la construcción de modelos probabilísticos basados en cadenas de Markov utilizando los datos que nos aporta el Gabinete de Análisis de la Universidad de La Laguna sobre la evolución del alumnado en los grados de la Facultad de Ciencias. Además, se comentan los resultados obtenidos de los modelos diseñados.

En el capítulo 1 se recoge toda la teoría que posteriormente se necesita para el desarollo del trabajo. En primer lugar nos centramos en la definición de proceso estocástico, que es una colección de variables aleatorias ordenadas X_t que dependen, en general, del tiempo. De este modo, podemos introducir las cadenas de Markov, que son los procesos estocásticos que verifican la propiedad de Markov, cuya comprensión es crucial para entender el capítulo que le sigue. Además, conoceremos las características de las matrices estocásticas, que han sido herramientas fundamentales para el análisis de las cadenas de Markov. Finalmente, se comentan los tipos de cadenas que existen y se presentan una serie de teoremas y proposiciones que aplicaremos a los datos reales para la obtención de resultados.

El capítulo 2 comienza explicando detalladamente las variables de los datos que se van a analizar mediante los modelos probabilísticos. En los datos se refleja si un alumno ha superado o no las asignaturas de las que estaba matriculado cada año. Se propone un primer modelo en el que se considera una cadena de Markov con 40 estados transitorios que representan las asignaturas y dos estados absorbentes que simbolizan el abandono y la finalización del grado. A continuación, se plantea otro modelo en el que se expone una cadena con cuatro estados transitorios, representando los cuatro cursos de cada uno de los grados, y dos estados absorbentes que vuelven a coincidir con el abandono y finalización en cada uno de los grados. Además, se muestran los resultados obtenidos del análisis realizado.

1

Marco Teórico

El objetivo de este primer capítulo es proporcionar la teoría necesaria que nos permita estudiar procesos estocásticos, concretamente las cadenas de Markov, de las cuales comentaremos sus principales características, sus representaciones, la clasificación de las mismas y de sus estados y demostraremos algunos resultados importantes principalmente sobre cadenas de Markov, que son las que utilizaremos en este trabajo.

1.1. Procesos Estocásticos

La teoría de los procesos estocásticos se centra en el estudio y modelización de sistemas que evolucionan a lo largo del tiempo, o del espacio, de acuerdo a unas leyes no determinísticas, esto es, de carácter aleatorio. La forma habitual de describir la evolución del sistema es mediante sucesiones o colecciones de variables aleatorias. De esta manera, se puede estudiar cómo evoluciona el proceso a lo largo del tiempo.

Definición 1.1 *Un proceso estocástico es una colección o familia de variables aleatorias $\{X_t\}$, ordenadas según el subíndice t que, en general, se suele identificar con el tiempo.*

Por tanto, para cada instante t tendremos una variable aleatoria distinta representada por X_t, con lo que un proceso estocástico puede interpretarse como una sucesión de variables aleatorias cuyas características pueden variar a lo largo del tiempo. A los posibles valores que puede tomar la variable aleatoria se le denominarán *estados*, y al conjunto de todos los posibles valores, *espacio de estados*, que denotaremos S. Tanto el espacio de estados S, como la variable *tiempo*, pueden ser de tipo discreto o continuo. Dependiendo de cómo sea el conjunto de subíndices, que denominaremos T, y el tipo de variable aleatoria dado por X_t se puede establecer la siguiente clasificación de procesos estocásticos:

	Tiempo (T)	
Espacio (S)	**Discreto**	**Continuo**
Discreto finito	Cadena	Proceso de ramificación
Discreto infinito	Camino aleatorio	Proceso puntual
Continuo	Sucesión de vv.aa.	Proceso continuo

Tabla 1.1. Clasificación de procesos estocásticos

Ejemplo (Procesos estocásticos.).

Cadena. Un sistema hidraúlico dispone de una válvula de seguridad que se revisa diariamente. Esta válvula presenta tres posibles estados: correcto, incorrecto o deteriorado. De este modo, se puede definir un proceso mediante las variables $X_n \equiv$ "estado en el que se encuentra la válvula en el día n al revisarse".

Camino aleatorio. Un jugador comienza una partida en la ruleta con cero puntos y va anotando su puntuación según una regla establecida: $+1$ punto si cae en rojo y -1 punto si cae en negro. Las variables que definen nuestro proceso serían $X_n \equiv$ "puntuación acumulada que obtiene el jugador en n partidas".

Proceso de ramificación. Una población se reproduce generación tras generación con descendencia aleatoriamente masculina o femenina. En los casos en que no exista descendencia masculina, el apellido familiar se acaba perdiendo. Definimos $X_t \equiv$ "número de descendientes varones en el instante t".

Proceso puntual. Nos encontramos en una cola a la van llegando personas en momentos aleatorios y se van atendiendo de manera ordenada. Se define, entonces, $X_t \equiv$ "número de personas esperando en la cola en un instante de tiempo t".

Sucesión de variables aleatorias. Una empresa petrolífera debe decidir cuánto petróleo extrae al mes para maximizar los beneficios. Se define, entonces, $X_n \equiv$ "cantidad de petróleo extraída en el mes n".

Proceso continuo. En la ventanilla de un banco se atiende a los clientes que hacen cola al llegar. Se puede definir $X_t \equiv$ "tiempo de espera de un cliente que llega en el instante t".

1.2. Cadenas de Markov

Las cadenas de Markov, como cualquier otro tipo de cadena, son procesos estocásticos en tiempo discreto y espacio discreto finito, que fueron introducidas por el matemático ruso Andrei Markov alrededor de 1905, con el objetivo de crear un modelo probabilístico para analizar la frecuencia con la que aparecen las vocales en poemas y textos literarios. A día de hoy, las cadenas de Markov pueden aplicarse en numerosos fenómenos científicos y sociales.

Propiedad de Markov

Para describir una cadena de Markov tomamos un conjunto S formado por r estados que sin pérdida de generalidad lo identificaremos con $\{1, 2, 3, \ldots, r\}$. El proceso comienza en uno de estos estados y se mueve de un estado a otro. A cada uno de estos movimientos se le llama *paso*. La principal característica de una cadena de Markov es verificar la *propiedad de Markov*, la cual refleja que en nuestro proceso, la probabilidad de que una cadena que en un periodo t está en un estado pase a otro estado en el siguiente paso, dependerá únicamente del estado actual en el que el proceso se encuentra, esto es:

$$P(X_{t+1} = s_{t+1} | X_0 = s_0, X_1 = s_1, \ldots, X_{t-1} = s_{t-1}, X_t = s_t) = \\ P(X_{t+1} = s_{t+1} | X_t = s_t)$$

con $s_i \in S$ para todo $i = 0, \ldots, t+1$. Esta propiedad es equivalente a establecer que la probabilidad condicionada de cualquier evento futuro dado cualquier evento pasado y el estado actual $X_t = s_t$, es independiente de los eventos pasados y solo depende del estado actual del proceso. Las probabilidades condicionadas

$$P(X_{t+1} = s_{t+1} | X_t = s_t)$$

se llaman *probabilidades de transición del paso* t.

Cuando las probabilidades no dependen del paso en que se encuentran sino que son las mismas para todo t hablaremos de probabilidades de transición estacionarias. En este caso se verifica que para todo t

$$P(X_{t+1} = j | X_t = i) = P(X_1 = j | X_0 = i),$$

y se denotan p_{ij}. La existencia de probabilidades de transición estacionarias también implica que, para todo par de estados $i, j \in S$ y $\forall m \in \mathbb{N}$, se verifica que

$$P(X_{t+m} = j | X_t = i) = P(X_m = j | X_0 = i).$$

Estas probabilidades condicionadas se denotan $p_{ij}^{(m)}$ y se les denomina probabilidades de transición en m pasos. Éstas deben ser no negativas y, deben satisfacer las siguientes propiedades:

- $p_{ij}^{(m)} \geq 0$, para todo $i, j \in S$;
- $\sum_{j \in S} p_{ij}^{(m)} = 1$, para todo $i \in S$.

Si un proceso estocástico cambia de estado verificando estas últimas propiedades, se denomina cadena de Markov estacionaria.

Definición 1.2 *Una* cadena de Markov *es un modelo estocástico en tiempo discreto que verifica la propiedad de Markov, lo cual quiere decir que la evolución del proceso en el futuro depende solo del estado presente y no de todo lo que haya ocurrido en el pasado.*

Matriz de transición

Dada una cadena de Markov que se encuentra en el estado i, la probabilidad de que en el siguiente paso se encuentre en el estado j la hemos denotado p_{ij}. A partir de estas probabilidades de transición obtenemos la *matriz de transición* que denotaremos $P = (p_{ij})$ y verifica:

- es una matriz cuadrada, porque X_{t+1} y X_t toman valores en el mismo espacio de estados S,
- la suma de sus filas es igual a 1:

$$\sum_{j=1}^{r} p_{ij} = \sum_{j=1}^{r} P(X_{t+1} = j | X_t = i) = 1.$$

Ejemplo. Supongamos que existe un lugar en la Tierra donde nunca se tiene buen tiempo atmosférico dos días consecutivos. Si un día hace calor, al día siguiente nevará o lloverá; y, en cambio, si hoy nieva o llueve, se tiene la misma probabilidad de que mañana haga el mismo tiempo y solamente la mitad de las veces hará sol. Distinguimos los siguientes estados: B, N y L, que representan el buen tiempo, la nieve y la lluvia respectivamente. Construimos una matriz cuadrada con las probabilidades de transición, que quedaría de la siguiente manera:

$$P = \begin{array}{c} \\ L \\ B \\ N \end{array}\!\!\begin{array}{c} \begin{array}{ccc} L & B & N \end{array} \\ \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix} \end{array}.$$

Las entradas de la primera fila de la matriz P representan las probabilidades de cambio o estabilidad climática tras un día de lluvia y, del mismo modo, las entradas de la segunda y tercera filas representan las probabilidades de que el tiempo cambie o se mantenga igual después de los días de buen tiempo y nieve, respectivamente. Esta matriz cuadrada sería nuestra matriz de transición y podríamos representarla mediante el grafo:

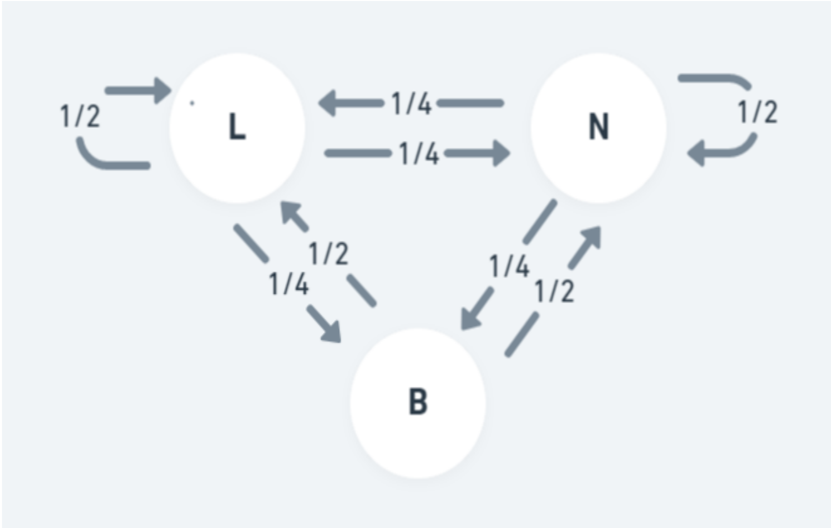


Figura 1.1. Representación de la cadena del ejemplo.

Supongamos ahora que queremos determinar la probabilidad de que, estando la cadena en el estado i hoy, dentro de dos días estuviera en el estado j y denotamos dicha probabilidad como $p_{ij}^{(2)}$. El hecho de que nieve dentro de dos días si hoy llueve sería la unión disjunta de los tres sucesos siguientes:

1. mañana llueve y dentro de dos días nieva;
2. mañana hace buen tiempo y dentro de dos días nieva;
3. mañana nieva y dentro de dos días nieva.

Considerando este primer suceso, la probabilidad de que llueva mañana habiendo llovido hoy se corresponde con la entrada $p_{11} = P(L|L)$ de la matriz P, y la probabilidad de que nieve al tercer día sería $P(N|L)$. Entonces, la probabilidad de que nieve después de que llueva dos días seguidos viene dada por $P(N|L) \cdot P(L|L)$. Podemos deducir:

$$\text{P(nieve pasado mañana | llueva hoy)} =$$
$$P(N|L) \cdot P(L|L) + P(N|B) \cdot P(B|L) + P(N|N) \cdot P(N|L),$$

es decir,

$$p_{13}^{(2)} = p_{13} \cdot p_{11} + p_{23} \cdot p_{12} + p_{33} \cdot p_{13}.$$

Observación.

$$p_{ij}^{(2)} = \sum_{k=1}^{r} p_{ik} \cdot p_{kj}.$$

Nota.
$P^(n) = p_{ij}^{(n)}$ es la matriz de probabilidades de transición dados n pasos en una cadena de Markov.

Teorema 1.3 *Sea $P^{(n)}$ la matriz de probabilidades de transición en n pasos de una cadena de Markov, se verifica que $p_{ij}^{(m+n)} = \sum_k p_{ik}^{(m)} \cdot p_{kj}^{(n)}$.*

Demostración. Sean i, $j \in S$. Utilizando la propiedad markoviana y la igualdad $P(A \cap B|C) = P(A|B \cap C) \cdot P(B|C)$, obtenemos:

$$p_{ij}^{(m+n)} = P(X_{m+n} = j|X_0 = i)$$

$$= \sum_k P(X_{m+n} = j, X_m = k|X_0 = i)$$

$$= \sum_k P(X_{m+n} = j|X_m = k, X_0 = i) \cdot P(X_m = k|X_0 = i)$$

$$= \sum_k P(X_{m+n} = j|X_m = k) \cdot P(X_m = k|X_0 = i)$$

$$= \sum_k p_{kj}^{(n)} \cdot p_{ik}^{(m)} = \sum_k p_{ik}^{(m)} \cdot p_{kj}^{(n)}.$$

□

Nota.

$$P(A \cap B|C) = \frac{P(A \cap B \cap C)}{P(C)} = \frac{P(A|B \cap C) \cdot P(B \cap C)}{P(C)}$$

$$= \frac{P(A|B \cap C) \cdot P(C) \cdot P(B|C)}{P(C)} = P(A|B \cap C) \cdot P(B|C).$$

Lema 1.4 *Sea P la matriz de transición de una cadena de Markov, y sea u^t la distribución de estados en el paso t. Entonces se verifica $u^{(m+n)} = u^{(m)} \cdot P^{(n)}$ y, por tanto, $u^{(n)} = u \cdot P^n$.*

Demostración. Sean i, $j \in S$. Tenemos

$$u_j^{(m+n)} = P(X_{m+n} = j) = \sum_i P(X_{m+n} = j|X_m = i) \cdot P(X_m = i)$$

$$= \sum_i p_{ij}^{(n)} \cdot u_i^{(m)} = \sum_i u_i^{(m)} \cdot p_{ij}^{(n)} = (u^{(m)} \cdot P^n)_j.$$

Por el teorema anterior, concluimos que $u^{(m+n)} = u^{(m)} \cdot P^{(n)}$ y $u^{(n)} = u \cdot P^n$.

□

Clasificación de estados de una cadena de Markov

Definición 1.5 *Un estado i de una cadena de Markov se denomina* recurrente *si comenzando la cadena en él, se tiene la certeza de volver en algún momento a él.*

Para tener una visión más clara sobre esta clase de estados, recordamos el ejemplo sobre el tiempo atmosférico en el que una vez que salimos de los estados N, L y B, estamos completamente seguros de que en algún momento volveremos a ellos y, por tanto, dichos estados son recurrentes.

Definición 1.6 *Un estado i de una cadena de Markov se denomina* absorbente *si es imposible salir de él, es decir, $p_{ii} = 1$.*

Definición 1.7 *Un estado i de una cadena de Markov se denomina* transitorio *si existe la posibilidad de abandonar el estado y no volver nunca a él.*

Definición 1.8 *Un estado j se dirá* accesible *o* alcanzable *desde el estado i, y lo denotaremos $i \to j$, si se verifica que $p_{ij}^{(n)} > 0$ para algún $n \geq 0$.*

Ejemplo. Un hombre camina por cuatro manzanas de tal manera que si está en la esquina 1, 2 o 3, camina hacia la izquierda o hacia la derecha con igual probabilidad y luego continúa hasta llegar a la esquina 4, que es un bar; o a la esquina 0, que es su casa. Si llega a casa o al bar, se queda allí. Dicho proceso se trata de una cadena de Markov la cual podemos representar mediante la matriz:

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

y el grafo

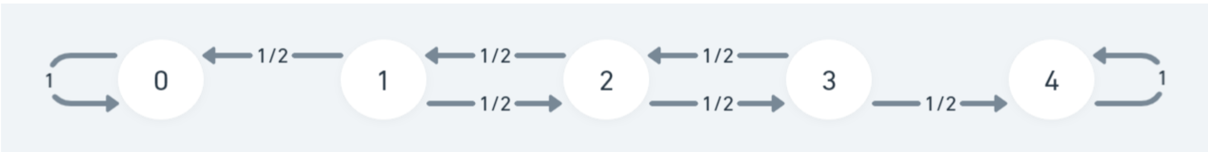


Figura 1.2. Representación de la cadena del ejemplo.

Concluimos que los estados 0 y 4 son *estados absorbentes* y el resto son *estados transitorios*. En esta cadena, todos los estados son alcanzables desde el estado 1, 2 y 3.

1.2.1. Cadenas de Markov absorbentes

Definición 1.9 *Una cadena de Markov es* absorbente *si tiene al menos un estado absorbente, y si desde cada estado no absorbente es posible ir a un estado absorbente, no necesariamente en un solo paso.*

Consideramos una cadena de Markov absorbente arbitraria con p estados transitorios y q estados absorbentes, reordenando los estados de forma que los p primeros sean los transitorios. Podemos dividir su matriz en cuatro submatrices de la siguiente manera:

$$P = \begin{pmatrix} Q & R \\ 0 & I \end{pmatrix},$$

donde

- Q es la matriz de dimensión $p \times p$ que contiene las probabilidades de paso de un estado transitorio a otro estado transitorio;
- R es la matriz de dimensión $p \times q$ que contiene las probabilidades de paso de un estado transitorio a otro absorbente;
- 0 es la matriz nula de dimensión $q \times p$;
- I es la matriz identidad de dimensión $q \times q$.

Anteriormente, hemos visto que la entrada $p_{ij}^{(n)}$ de la matriz P^n es la probabilidad de estar en el estado j comenzando en el estado i, después de n pasos. Podemos concluir ahora que P^n sería de la forma:

$$P^n = \begin{pmatrix} Q^n & * \\ 0 & I \end{pmatrix},$$

donde Q^n presenta las probabilidades de estar en cada uno de los estados transitorios después de n pasos para cada estado inicial, y * representa la matriz de transición de estados transitorios a absorbentes en n pasos, cuya expresión nos resulta irrelevante en este trabajo.

En nuestro primer teorema demostraremos que la probabilidad de encontrarse en un estado transitorio después de n pasos se aproxima a cero según n va creciendo. Así, cada entrada de Q^n debe acercarse a cero tanto como queramos a medida que n tiende a infinito, es decir, $Q^n \to 0$ a medida que $n \to \infty$.

Probabilidad de absorción

Teorema 1.10 *En una cadena de Markov absorbente, la probabilidad de que el proceso sea absorbido es 1.*

Demostración. Desde cada estado no absorbente j es posible alcanzar un estado absorbente. Sea m_j el número mínimo de pasos necesarios para alcanzar un estado absorbente partiendo de j. Sea p_j la probabilidad de que, partiendo de j, el proceso no alcance un estado absorbente en m_j pasos. Entonces, $p_j < 1$. Considerando todos los posibles estados j, sea m el mayor de los m_j y sea p el mayor de p_j. La probabilidad de no ser absorbido en m pasos es menor o igual que p, en $2m$ pasos menor o igual que p^2, etc. Como $p < 1$ estas probabilidades tienden a 0. Además, como la probabilidad de no ser absorbido en n es monótona decreciente, se verifica que $\lim_{n\to\infty}(Q^n) = 0$. $\square$

Teorema 1.11 *Para una cadena de Markov absorbente la matriz $I - Q$ tiene inversa N y $N = I + Q + Q^2 + \cdots$ La entrada n_{ij} de la matriz N es el número medio de veces que la cadena se encuentra en el estado j, comenzando en el estado i.*

Demostración. Sea $(I-Q)x = 0$; es decir, $x = Qx$. Entonces, mediante iteraciones obtenemos: $x = Q^n x$. Como $Q^n \to 0$, tenemos que $Q^n x \to 0$ por lo que $x = 0$. Así, $(I - Q)^{-1} = N$ existe. Obsérvese:

$$(I - Q)(I + Q + Q^2 + \cdots + Q^n) = I - Q^{n+1},$$

multiplicando ambos lados por N,

$$I + Q + Q^2 + \cdots + Q^n = N(I - Q^{n+1}),$$

haciendo $n \to \infty$, tenemos que

$$N = I + Q + Q^2 + \cdots$$

Sean i y j dos estados transitorios y sea $X(k)$ una variable aleatoria que es igual a 1 si la cadena está en el estado j después de k pasos, e igual a 0 en caso contrario. Para cada k, esta variable aleatoria depende tanto de i como de j. Tendríamos

$$P(X^{(k)} = 1) = q_{ij}^{(k)},$$

y

$$P(X^{(k)} = 0) = 1 - q_{ij}^{(k)},$$

donde $q_{ij}^{(k)}$ es la ij-ésima entrada de la matriz Q^k.

Entonces, el número esperado de veces que la cadena está en el estado j en los primeros n pasos, dado que comienza en el estado i, es claramente:

$$E(X^{(0)} + X^{(1)} + \cdots + X^{(n)}) = q_{ij}^{(0)} + q_{ij}^{(1)} + \cdots + q_{ij}^{(n)},$$

y haciendo tender n a infinito tenemos que,

$$E(X^{(0)} + X^{(1)} + \cdots) = q_{ij}^{(0)} + q_{ij}^{(1)} + \cdots = n_{ij}.$$

□

Definición 1.12 *Para una cadena de Markov absorbente representada por la matriz P, a la matriz $N = (I - Q)^{-1}$ se le denomina* matriz fundamental de P. *La entrada n_{ij} de N es el número esperado de veces en que el proceso está en el estado transitorio j si se inicia en el estado transitorio i.*

Ejemplo. Retomemos el ejemplo 1.2 del hombre que se mueve entre su casa y el bar, con matriz de transición

$$P = \left(\begin{array}{ccc|cc} 0 & 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right).$$

La matriz Q se corresponde con el primer bloque

$$Q = \begin{pmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix}$$

siendo

$$I - Q = \begin{pmatrix} 1 & -1/2 & 0 \\ -1/2 & 1 & -1/2 \\ 0 & -1/2 & 1 \end{pmatrix}$$

y hallando la inversa obtenemos

$$N = (I - Q)^{-1} = \begin{array}{c c} & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} & \begin{pmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{pmatrix} \end{array}.$$

Observando la fila central de N, concluimos que si empezamos en el estado 2, entonces el número esperado de veces de que el hombre se encuentre en las manzanas 1, 2 y 3 antes de descansar en su casa o en el bar, es respectivamente 1, 2 y 1.

Tiempo de absorción

Dado que la cadena comienza en un estado i, ¿cuál es el número esperado de pasos antes de que la cadena sea absorbida?

Teorema 1.13 *Sea t_i el número esperado de pasos antes de ser absorbido suponiendo que la cadena empieza en el estado i y sea t el vector columna cuya i-ésima entrada es t_i. Entonces, $t = Nc$, donde c es el vector columna cuyas entradas son todas 1.*

Demostración. Si sumamos las entradas de la fila i-ésima de N, obtendremos el número esperado de veces que nos encontramos en cualquiera de los estados transitorios dado un estado i inicial; es decir, el tiempo esperado antes de ser absorbido. Así, t_i es la suma de las entradas de la i-ésima fila de N. Al escribir matricialmente esta afirmación, se verificaría que $t = Nc$.

$\square$

Ejemplo. Tomando

$$N = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{pmatrix} \end{array},$$

obtenida en el ejemplo 1.2.1, tendríamos

$$t = Nc = \begin{pmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ 3 \end{pmatrix},$$

y por lo tanto, comenzando en las manzanas 1, 2 y 3, los tiempos esperados de absorción son 3, 4 y 3, respectivamente.

Probabilidad de absorción

Cuando existen varios estados absorbertes puede ser interesante conocer los siguientes resultados.

Teorema 1.14 *Sea b_{ij} la probabilidad de que una cadena absorbente sea absorbida en el estado j si comienza en el estado transitorio i. Sea B la matriz con entradas b_{ij}. Entonces B es una matriz que verifica que $B = NR$, donde N es la matriz fundamental y R es la de la forma canónica.*

Demostración. Tenemos:

$$B_{ij} = \sum_n \sum_k q_{ik}^{(n)} r_{kj} = \sum_k \sum_n q_{ik}^{(n)} r_{kj} = \sum_k n_{ik} r_{kj} = (NR)_{ij}.$$

$\square$

Ejemplo. Retomando el ejemplo 1.2, la R correspondiente es

$$R=\begin{array}{cc} & \begin{array}{cc} 0 & 4 \end{array} \\ \begin{array}{c}1\\2\\3\end{array} & \begin{pmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 1/2 \end{pmatrix} \end{array},$$

tenemos

$$B = NR = \begin{pmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{pmatrix} \cdot \begin{pmatrix} 1/2 & 0 \\ 0 & 0 \\ 0 & 1/2 \end{pmatrix} = \begin{array}{cc} & \begin{array}{cc} 0 & 4 \end{array} \\ \begin{array}{c}1\\2\\3\end{array} & \begin{pmatrix} 3/4 & 1/4 \\ 1/2 & 1/2 \\ 1/4 & 3/4 \end{pmatrix} \end{array}.$$

Partiendo del estado 1, hay probabilidad 3/4 de que el proceso sea absorbido por el estado 0 y probabilidad 1/4 de que el proceso sea absorbido por el estado 4; partiendo del estado 2, hay probabilidad 1/2 de absorción en el estado 0 y 1/2 en el estado 4 y, si partirmos del estado 3, tendremos probabilidad 1/4 de absorción por el estado 0 y 3/4 por el estado 4.

1.2.2. Cadenas de Markov ergódicas

Trataremos ahora con cadenas de Markov sin estados absorbentes.

Definición 1.15 *Una cadena de Markov se denomina* cadena ergódica *si es posible llegar desde cada estado al resto de ellos, no necesariamente en un solo paso. Las cadenas de Markov ergódicas también se denominan irreducibles.*

Definición 1.16 *Una cadena de Markov será* regular *si alguna potencia de la matriz de transición solo tiene elementos positivos. En otras palabras, si para algún n, es posible llegar desde cualquier estado hasta otro en exactamente en n pasos.*

De esta definición podemos concluir que toda cadena regular es ergódica. El recíproco no es cierto en general.

Ejemplo. Sea la matriz de transición de una cadena de Markov definida por:

$$P = \begin{array}{cc} & \begin{array}{cc} 1 & 2 \end{array} \\ \begin{array}{c}1\\2\end{array} & \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array},$$

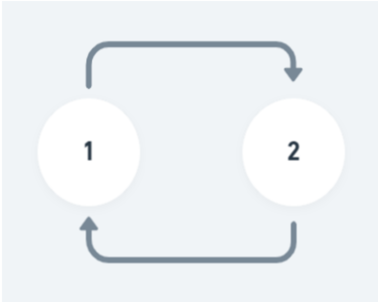


Figura 1.3. Representación de la cadena del ejemplo.

y por lo tanto, es claro que es posible pasar de un estado a cualquier otro, por lo que la cadena es ergódica.
Observamos que la cadena no es regular. En los pasos pares siempre nos encontraremos en el estado del que partimos, y en los impares, en el otro. $P^n = P$, por tanto siempre hay ceros.

Teorema 1.17 *Sea P la matriz de transición de una cadena regular. Entonces, a medida que $n \to \infty$, las potencias P^n se aproximan a una matriz W cuyas filas serán todas el vector w. Se trata de un vector de probabilidad, es decir, sus componentes son todas positivas y su suma es 1.*

Demostración. Queremos demostrar que las potencias P^n de una matriz de transición de una cadena regular tienden a una matriz con todas las filas iguales. Esto sería lo mismo que ver que P^n converge a una matriz con columnas constantes. Sabemos que la j-ésima columna de P^n es $P^n \cdot y$, donde y es un vector columna con un 1 en la j-ésima entrada y 0 en las demás. Por lo tanto, solo tendríamos que demostrar que para cualquier vector columna y, $P^n \cdot y$ se aproxima a un vector constante cuando $n \to \infty$. Como cada fila de P es un vector de probabilidad, $P \cdot y$ sustituye a y por las medias ponderadas de sus componentes según el vector de pasos de la fila de P y el resultado de este proceso será que las componentes de $P \cdot y$ sean más similares que los de y. A medida que hacemos más iteraciones de este proceso para obtener $P^n \cdot y$, la diferencia entre la componente máxima y mínima tenderá a 0 a medida que $n \to \infty$. Esto significa que $P^n \cdot y$ tiende a un vector constante. □

Definición 1.18 *Un vector fila w que verifique $w \cdot P = w$ se llama* vector de fila fijo *de P. Del mismo modo, un vector columna x tal que $P \cdot x = x$ se llama* vector de columna fijo *de P.*

Teorema 1.19 *Sea P una matriz de transición regular que verifica $W = \lim_{n\to\infty} P^n$, sea w la fila común de W (es el único vector que es a la vez un vector de fila fijo* *de P y vector de probabilidad), y c es el vector columna cuyas componentes son todas 1. Entonces se verifica:*

1. *$w \cdot P = w$, y cualquier vector fila v tal que $v \cdot P = v$ es un múltiplo constante de w.*
2. *$P \cdot c = c$, y cualquier vector columna x tal que $P \cdot x = x$ es un múltiplo de c.*

Demostración. Para demostrar 1 utilizaremos el teorema 1.17. Como $P^n \to W$, entonces $P^{n+1} = P^n \cdot P \to W \cdot P$, pero $P^{n+1} \to W$ con $W = W \cdot P$ y $w = w \cdot P$. Sea v un vector cualquiera con $v \cdot P = v$. Entonces $v = v \cdot P^n$ y pasando al límite $v = v \cdot W$. Si denotamos r a la suma de las componentes de v, entonces se comprueba fácilmente que $v \cdot W = r \cdot w$ y por lo tanto, $v = r \cdot w$.

Para demostrar 2 suponemos que $x = P \cdot x$. Entonces $x = P^n \cdot x$, y de nuevo aplicando que $P^n \to W$, $x = W \cdot x$. Como todas las filas de W son iguales, las componentes de $W \cdot x$ son todas iguales, por lo que x es un múltiplo de c.

$\square$

Teorema 1.20 *Sea P la matriz de transición de una cadena regular y v un vector de probabilidad arbitrario. Entonces $\lim_{n\to\infty} v \cdot P^n = w$, donde w es el único vector de probabilidad fijo para P.*

Demostración. Por el teorema 1.17 sabemos que

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

y por lo tanto,

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

pero las entradas de v suman 1, y cada fila de W es igual a w. A partir de estas afirmaciones se concluye que $v \cdot W = w$.

$\square$

2

Análisis de los datos

Este capítulo se centra en la construcción de dos modelos probabilísticos usando cadenas de Markov para modelizar el rendimiento académico, ayudar a entender la deserción y el éxito académico como el tiempo que el alumnado permanece en el sistema universitario en los grados de la Facultad de Ciencias. Nos basamos en los datos que nos aporta el Gabinete de Análisis de la Universidad de La Laguna. Dichos datos recogen información sobre los expedientes académicos de los alumnos de la Facultad de Ciencias entre los cursos académicos 2009-2010 y 2020-2021, ambos inclusive. Cada una de las entradas de la tabla de datos es una asignatura de un alumno en un curso académico. De entre toda la información proporcionada, propuestos haremos uso de las siguientes variables:

- **Curso académico**
  Variable de tipo caracter (`char`) que representa el año en el que cada alumno realizó sus estudios.
- **Identificador del alumno**
  Variable de tipo numérico (`num`) de 6 cifras asociada a cada alumno, que no varía al pasar de curso y que, por tanto, nos facilita el seguimiento académico de cada uno de ellos.
- **Titulación**
  Variable de tipo `char` que en nuestro caso representa los grados en Química, Física, Biología, Matemáticas y Ciencias Ambientales, cada uno de ellos representado con los identificadores **G017**, **G019**, **G023**, **G034** y **G055** respectivamente.
- **Asignatura y su tipología**
  Cada una de las materias se representa con un número de 9 cifras, por lo que es de tipo `num`. Existen asignaturas de formación básica **T**, obligatorias **B**, optativas **O**, prácticas externas **E** y trabajo fin de grado **P**, por tanto, la tipología es una variable de tipo `char`.

- **Crédito**
  Variable de tipo `num` que mide el trabajo académico de cada estudiante. Las asignaturas tienen una valoración que viene determinada por el número de créditos.
- **Presenta**
  Variable de tipo factor (`fact`) que hemos codificado con el valor **0** si el alumno no se ha presentado a la asignatura y con el valor **1** si el alumno si se ha presentado a la asignatura.
- **Supera**
  Variable de tipo `fact` que hemos codificado con el valor **0** si el estudiante no ha superado la asignatura y con el valor **1** si la ha superado.

2.1. Modelo 1: por créditos o asignaturas

2.1.1. Presentación del modelo

Este primer modelo que presentamos consiste en un sistema de cadenas de Markov en tiempo discreto con 39 estados transitorios que se corresponden con los créditos que aprueba un alumno en cada año académico. Vamos a realizar el estudio en cinco grados diferentes: Biología, Matemáticas, Física, Química y Ciencias Ambientales, cuyas asignaturas son, por lo general, de seis créditos, lo que quiere decir que si un alumno ha superado n asignaturas, tendrá $6n$ créditos. Para el caso particular de las Prácticas Externas y TFG, hemos dividido sus créditos totales entre seis para así considerarlos como más de una asignatura, abarcando varios estados de la cadena. Por ello en la cadena utilizamos múltiplos de seis. Además, el modelo consta de otros dos estados que representan el abandono del grado, -1, y la finalización del mismo, 40. Los alumnos comienzan en el estado 0 y avanzan conforme los créditos que vayan superando, hasta que consigan llegar al estado 40 finalizando el grado, o en su defecto, alcancen el estado -1 abandonando sus estudios. A continuación, podemos observar un esquema en el que se ilustra el modelo comentado:

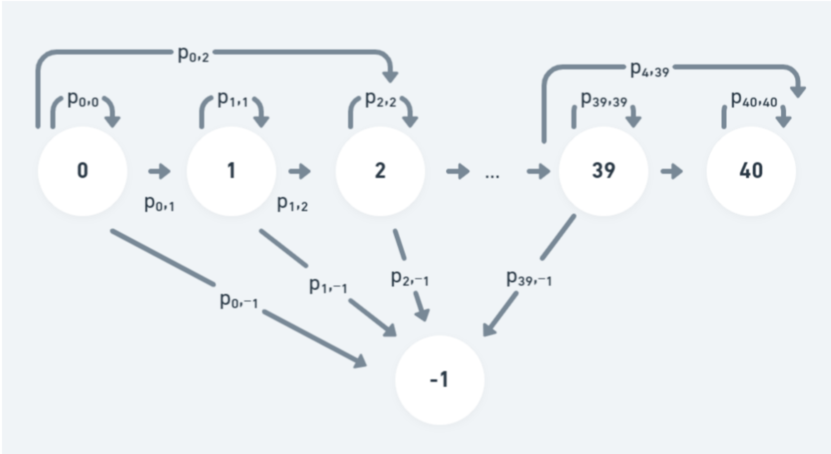


Figura 2.1. Representación gráfica del primer modelo.

donde $p_{i,j}$ con i y $j \in \{$ -1, 0, 1, 2, ..., 41$\}$ son las probabilidades de transición que conformarán la matriz de transición, una distinta para cada uno de los cinco grados.

El GAP nos ha informado de que en la Universidad de La Laguna se considera que un alumno ha abandonado sus estudios si en las tablas de datos no existen registros suyos dos años académicos después de su última aparición.

2.1.2. Preparación de los datos

A través de la matriz de transición conoceremos la probabilidad que existe de que los alumnos matriculados pasen de tener i a tener j asignaturas aprobadas. Dichas matrices presentarán la siguiente estructura:

$$P = \begin{pmatrix} p_{0,0} & p_{0,1} & \cdots & p_{0,41} & p_{0,-1} \\ p_{1,0} & p_{1,1} & \cdots & p_{1,41} & p_{1,-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{41,0} & p_{41,1} & \cdots & p_{41,41} & p_{41,-1} \\ p_{-1,0} & p_{-1,1} & \cdots & p_{-1,41} & p_{-1,-1} \end{pmatrix}. \qquad (2.1)$$

Los cálculos se han realizado con el lenguaje R. En primer lugar, conectamos Google Colaboratory con Google Drive mediante el código

```
from google.colab import drive
drive.mount('/content/gdrive')
```

para poder trabajar los datos que nos facilitaron en un archivo de hoja de cálculo (.xls) previamente subidas al Google Drive. Posteriormente, hemos realizado una serie de cambios en los nombres de las variables de los datos originales, además de dividirlos en cinco tablas, una por cada grado. Seguidamente, comenzamos a trabajar con los datos.

2.1.3. Análisis realizado

Necesitamos conocer el total de asignaturas diferentes aprobadas por cada alumno en cada curso a lo largo de su carrera para obtener los coeficientes de la matriz 2.1. Para ello, creamos una lista de vectores para cada uno de los grados, cuyas posiciones representan las asignaturas totales superadas por cada alumno cada año que está matriculado. Si existen asignaturas de 12 o 18 créditos contarán el doble o triple, respectivamente. Se procede a definir una función para generar unas matrices de ocurrencias que recogerán, tras recorrer los vectores, el número de alumnos que van del estado i al j en cada paso. Finalmente, definimos una función para generar las matrices de transición de la forma de 2.1, es decir, dividimos cada fila de la matriz de ocurrencias entre la suma total de los elementos de su fila. Al implementar el script en el que trabajamos con los datos reales obtenemos la siguiente matriz:

$$B = \begin{array}{c} \\ \textit{0} \\ \textit{1} \\ \textit{2} \\ \vdots \\ \textit{39} \\ \textit{40} \\ \textit{-1} \end{array} \begin{array}{c} \begin{array}{ccccccc} \textit{0} & \textit{1} & \textit{2} & \cdots & \textit{39} & \textit{40} & \textit{-1} \end{array} \\ \begin{pmatrix} 0{,}036 & 0{,}022 & 0{,}029 & \cdots & 0 & 0 & 0{,}020 \\ 0 & 0{,}200 & 0{,}100 & \cdots & 0 & 0 & 0{,}450 \\ 0 & 0 & 0{,}160 & & 0 & 0 & 0{,}540 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0{,}250 & 0{,}750 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix} \end{array};$$

para el grado en Biología, en el que vamos a centrarnos en el informe del trabajo.

A continuación, necesitamos considerar una función complementaria debido a que tenemos pocos datos para la gran cantidad de (p_{ij}) que hay que estimar. La denominamos

`suavizar_matriz`

y la aplicamos a las matrices de ocurrencias nos ha ayudado a suavizar tanto las probabilidades como las irregularidades (picos y valles) de los gráficos y, de este modo, poder reconocer fácilmente las tendencias de los alumnos. Para este proceso, tomamos cada entrada de la matriz de ocurrencias y dejamos un % de su valor en su misma posición, mientras que otro % se suma a las posiciones de arriba, abajo, izquierda y derecha.

De este modo, obtuvimos la matriz de transición

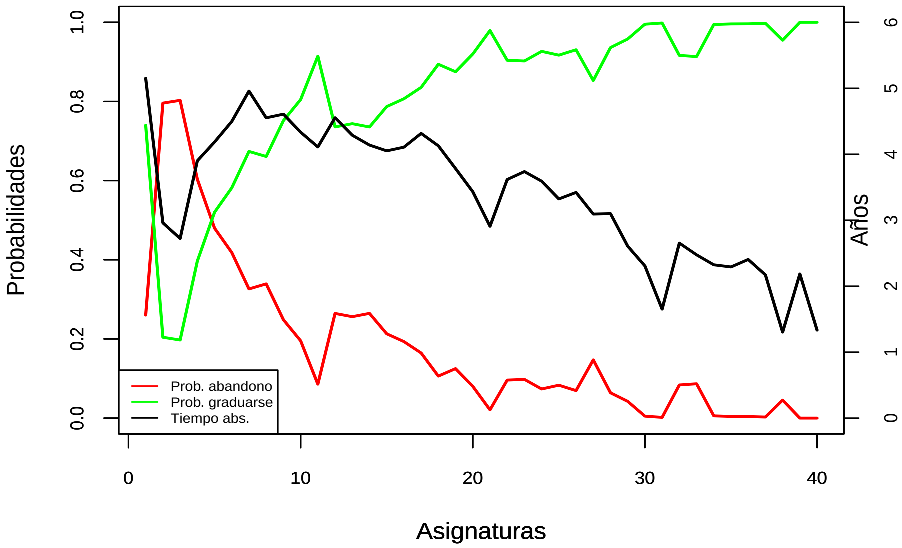


Figura 2.2. Representación gráfica de las probabilidades para abandonar y finalizar el grado en Biología y del tiempo que van a estar matriculados los alumnos según las asignaturas que tengan superadas a partir de la matriz B.

$$
B' = \begin{array}{c} \\ \mathit{0} \\ \mathit{1} \\ \mathit{2} \\ \vdots \\ \mathit{39} \\ \mathit{40} \\ \mathit{-1} \end{array}
\begin{array}{c}
\begin{array}{ccccccc} \mathit{0} & \mathit{1} & \mathit{2} & \cdots & \mathit{39} & \mathit{40} & \mathit{-1} \end{array} \\
\begin{pmatrix}
0{,}020 & 0{,}031 & 0{,}030 & \cdots & 0 & 0{,}007 & 0{,}010 \\
0{,}038 & 0{,}027 & 0{,}040 & \cdots & 0 & 0{,}010 & 0{,}040 \\
0 & 0{,}070 & 0{,}070 & & 0 & 0{,}110 & 0{,}290 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0{,}540 & 0{,}270 & 0{,}140 \\
0 & 0 & 0 & \cdots & 0{,}250 & 0{,}500 & 0{,}250 \\
0 & 0 & 0 & \cdots & 0 & 0{,}660 & 0{,}330
\end{pmatrix}
\end{array};
$$

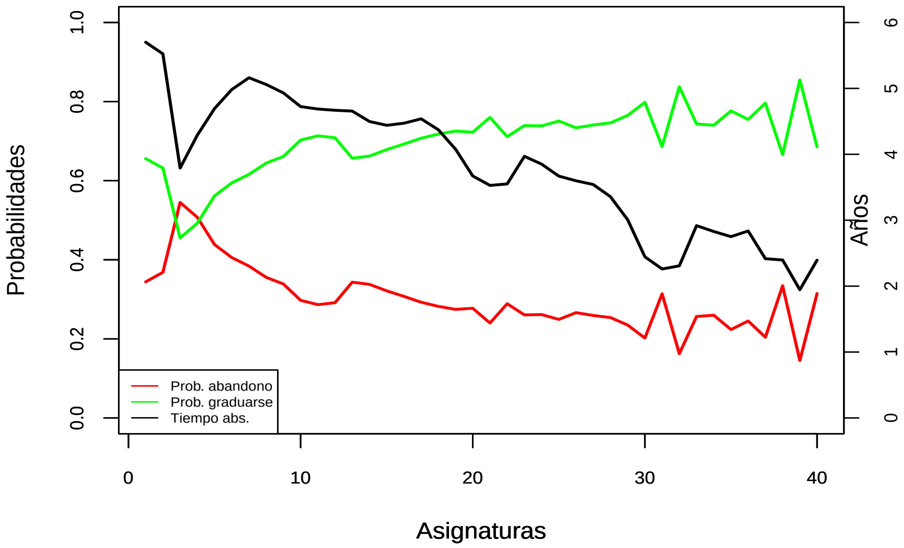


Figura 2.3. Representación gráfica de las probabilidades para abandonar y finalizar el grado en Biología y del tiempo que van a estar matriculados los alumnos según las asignaturas que tengan superadas, a partir de la matriz de transición suavizada B'.

Mostramos el código en A.1.

A partir de la matriz de transición, se obtienen los tiempos de absorción y las probabilidades tanto para graduarse como para abandonar cada uno de los grados. En el caso de Biología, podemos apreciarlos en las figuras 2.2 y 2.3.

2.1.4. Resultados obtenidos

De nuestro análisis de datos hemos obtenido los siguientes resultados

Asig. aprobadas	**Prob. abandonar**	**Prob. graduarse**	**Tiempo de absorción**
0	0.34	0.66	5.70
1	0.37	0.63	5.52
2	0.54	0.46	3.79
3	0.51	0.49	4.29
4	0.44	0.56	4.69
5	0.41	0.59	4.98
6	0.38	0.62	5.16
7	0.36	0.64	5.06
8	0.34	0.66	4.93
9	0.30	0.70	4.72
10	0.29	0.71	4.69
11	0.29	0.71	4.67
12	0.34	0.66	4.66
13	0.34	0.66	4.50
14	0.32	0.68	4.44
15	0.31	0.69	4.47
16	0.29	0.71	4.54
17	0.28	0.71	4.37
18	0.27	0.73	4.08
19	0.28	0.72	3.67
20	0.24	0.76	3.53
21	0.29	0.71	3.55
22	0.26	0.74	3.97
23	0.26	0.74	3.85
24	0.25	0.75	3.67
25	0.27	0.73	3.60
26	0.26	0.74	3.54
27	0.25	0.75	3.36
28	0.23	0.77	3.01
29	0.20	0.80	2.45
30	0.31	0.69	2.26
31	0.16	0.84	2.31
32	0.26	0.74	2.92
33	0.26	0.74	2.83
34	0.22	0.78	2.75
35	0.25	0.75	2.84
36	0.20	0.80	2.42
37	0.33	0.67	2.40
38	0.15	0.85	1.95
39	0.31	0.69	2.39

Tabla 2.1. Tabla que recoge las probabilidades de graduarse o abandonar de un alumno con un determinado número de asignaturas aprobadas, así como los tiempos medios de permanecer en el grado.

Observando la figura 2.3, se puede deducir que cuando el alumnado comienza sus estudios, el tiempo de absorción es, en general, más alto. Esto se debe a que en un principio los alumnos deben permanecer más tiempo en el grado para lograr finalizar sus estudios. Por tanto, la función decrece conforme más asignaturas se superan. Aún así, se aprecia que cuando un alumno ha superado las primeras asignaturas, el tiempo de absorción decrece repentinamente, coincidiendo con el resaltable crecimiento de la probabilidad de abandono. Esto significa que muchos de los alumnos abandonan sus estudios con tan solo unas pocas asignaturas superadas. Con respecto a la probabilidad de graduarse, existe un gran decrecimiento de la función en el momento en que el alumnado cuenta con escasas asignaturas superadas en el grado. Esto ocurre porque la probabilidad de que los alumnos se gradúen con tan pocas asignaturas superadas es muy baja. Por este motivo, la función es generalmente creciente.

Hemos realizado el mismo análisis para todos los grados de la Facultad de Ciencias. Se adjuntan tres gráficos donde se observarán los resultados obtenidos para el resto de grados.

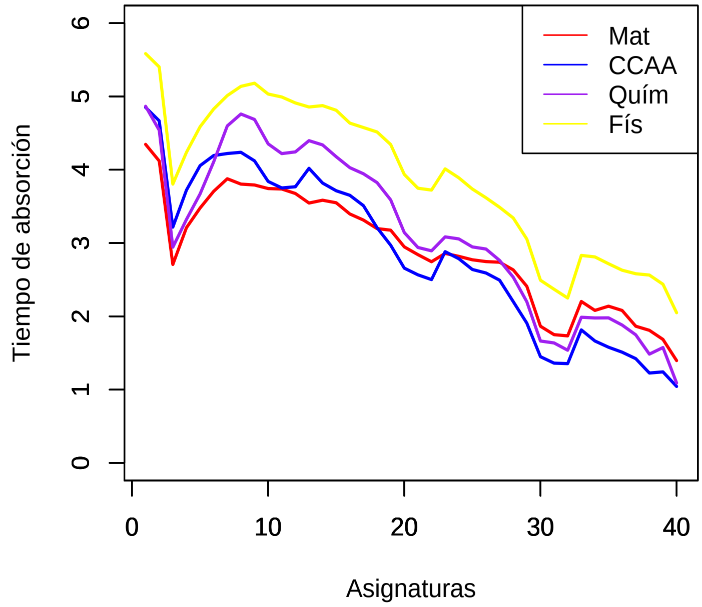


Figura 2.4. Representación gráfica del tiempo de absorción del resto de grados.

Observamos que el tiempo de absorción decrece de manera similar en todos los grados, existiendo una coincidencia en todos ellos: decrece de manera pronunciada en las primeras asignaturas superadas. Esto significa que el tiempo medio que permanecerán los alumnos en el grado tras haber superado unas pocas asignaturas es bajo, es decir, que existe una elevada probabilidad de abandono en ese número concreto de asignaturas, lo que se ve en el gráfico superior de la figura 2.5. El alumnado del grado en Física obtiene los tiempos medios más altos a lo largo de toda la carrera, mientras que los jóvenes matemáticos comienzan con los tiempos de absorción más bajos, aunque a mitad del grado, Ciencias Ambientales le arrebate el puesto.

Veámos ahora las probabilidades de abandono y graduarse en el resto de grados:

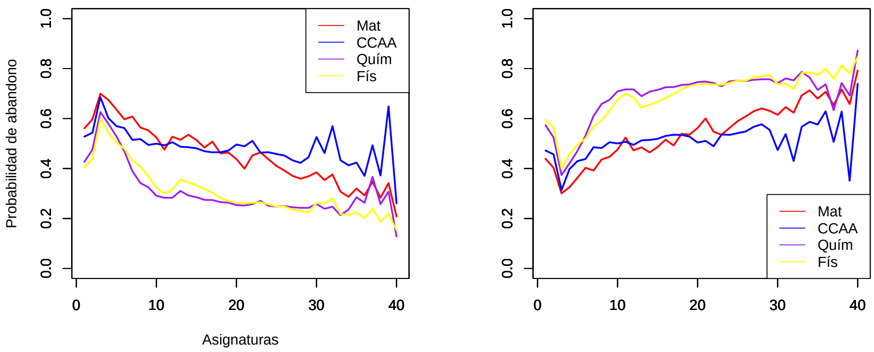


Figura 2.5. Representación gráfica de la probabilidad de abandono y graduarse del resto de grados.

La probabilidad de abandono en el grado en Matemáticas es considerablemente mayor que en el resto de grados, excepto en las últimas asignaturas, donde Ciencias ambientales tiene mayor probabilidad de abandono. El alumnado de Física y Química posee menor posibilidad de renunciar a sus estudios. Por otro lado, la probabilidad de graduarse va creciendo conforme más asignaturas ha superado el alumno. Notamos que en Matemáticas siempre es menor con respecto a los otros grados, a excepción de Ciencias Ambientales, cuyos alumnos tienen menor probabilidad de graduarse en las últimas asignaturas de la titulación.

A continuación calculamos los tiempos de absorción condicionados a cada uno de los estados absorbentes: abandonar y graduarse. Para ello, debemos eliminar de nuestra matriz de ocurrencias la fila y columna que corresponda al estado absorbente que no sea el asociado en cada caso. Por tanto, nos quedaría una matriz de una dimensión menos, con la cual generaríamos una nueva matriz de transición, forzando a que sus filas sumen uno.

Asignaturas	Biología	Matemáticas	CCAA	Química	Física
0	9.15	5.56	5.28	5.96	9.74
1	8.85	5.22	5.11	5.55	9.30
2	6.20	3.48	3.64	3.79	6.44
3	6.95	4.12	4.25	4.28	7.33
4	7.74	4.50	4.53	4.69	8.03
5	8.18	4.86	4.60	5.13	8.45
6	8.43	4.92	4.72	5.66	9.00
7	8.49	4.99	4.61	5.86	9.26
8	8.41	4.97	4.49	5.76	9.57
9	8.37	4.90	4.18	5.46	9.67
10	8.35	5.03	4.10	5.30	9.78
11	8.34	4.79	4.08	5.34	9.58
12	8.10	4.82	4.36	5.48	9.32
13	7.97	4.72	4.16	5.43	9.40
14	7.99	4.72	4.06	5.26	9.41
15	8.09	4.67	4.02	5.13	9.34
16	8.22	4.47	3.89	5.04	9.37
17	8.10	4.50	3.59	4.95	9.43
18	7.84	4.42	3.34	4.70	9.35
19	7.32	4.20	2.98	4.23	8.92
20	7.45	4.20	2.89	4.03	8.74
21	7.08	3.96	2.79	3.98	8.68
22	7.81	4.11	3.26	4.19	9.05
23	7.68	4.12	3.15	4.16	8.98
24	7.57	4.13	3.00	4.04	8.89
25	7.39	4.18	2.97	3.99	8.77
26	7.24	4.13	2.91	3.92	8.77
27	7.04	3.86	2.69	3.74	8.68
28	6.64	3.19	2.34	3.45	8.46
29	5.51	3.09	1.75	2.80	7.45
30	6.82	3.02	1.66	2.64	7.28
31	6.77	3.79	1.59	2.61	6.90
32	6.62	3.75	2.25	3.12	8.29
33	6.79	3.68	2.10	3.14	8.30
34	6.69	3.71	1.96	2.93	8.10
35	6.55	3.25	1.98	3.03	8.20
36	6.53	3.39	1.67	2.50	7.76
37	6.52	3.40	1.54	2.53	8.21
38	6.40	3.01	1.34	2.35	7.75
39	5.60	2.91	1.17	1.80	7.82

Tabla 2.2. Tabla que recoge los tiempos medios de los alumnos que abandonan el grado.

En el primero de los casos, el tiempo medio de absorción asociado al estado de abandonar decrece ligeramente en todos los grados. Ciencias ambientales posee los tiempos de absorción más bajos mientras que Física los mayores. Podemos apreciarlo en la figura 2.6.

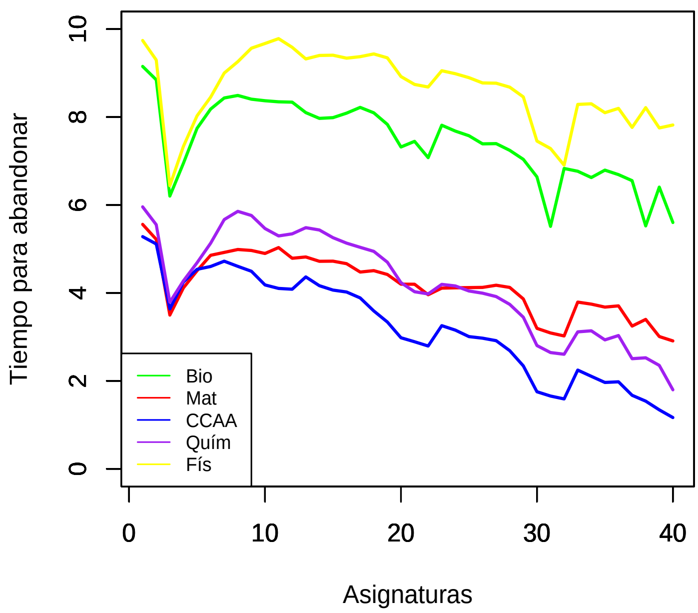


Figura 2.6. Representación gráfica del tiempo medio de absorción asociado al estado de abandonar.

Asignaturas	Biología	Matemáticas	CCAA	Química	Física
0	6.81	6.19	5.66	6.23	7.19
1	6.80	6.25	5.63	6.25	7.27
2	6.37	5.69	5.52	5.58	6.96
3	6.62	6.06	5.31	5.69	6.97
4	6.50	5.98	5.28	5.70	6.98
5	6.53	5.88	5.27	5.65	7.00
6	6.53	6.00	4.99	5.50	6.79
7	6.22	5.58	4.91	5.30	6.69
8	5.96	5.42	4.60	4.77	6.41
9	5.49	5.06	4.28	4.58	5.92
10	5.36	4.78	4.16	4.61	5.73
11	5.40	4.96	4.19	4.91	5.97
12	5.69	4.89	4.43	4.76	5.90
13	5.48	4.96	4.22	4.55	5.76
14	5.32	4.76	4.11	4.35	5.48
15	5.28	4.46	4.03	4.26	5.33
16	5.27	4.44	3.87	4.11	5.15
17	5.04	4.12	3.56	3.86	4.91
18	4.70	4.06	3.33	3.36	4.43
19	4.25	3.66	3.01	3.15	4.23
20	4.01	3.48	2.85	3.12	4.20
21	4.14	3.72	2.23	3.37	4.55
22	4.55	3.56	3.13	3.28	4.40
23	4.43	3.41	2.97	3.12	4.21
24	4.20	3.31	2.80	3.37	4.09
25	4.18	3.25	2.91	3.28	3.92
26	4.11	3.13	2.81	3.15	3.78
27	3.90	2.91	2.54	3.13	3.48
28	3.50	2.32	2.28	2.97	2.97
29	2.84	2.11	1.85	2.76	2.81
30	2.86	2.14	1.62	2.43	2.72
31	2.63	2.55	1.76	1.88	3.23
32	3.46	2.40	2.15	2.16	3.21
33	3.37	2.40	1.96	2.18	3.13
34	3.20	2.51	1.86	2.25	2.98
35	3.34	2.40	1.73	2.12	3.01
36	2.81	2.23	1.73	2.09	2.86
37	3.08	2.08	1.37	1.66	2.80
38	2.21	2.00	1.62	1.78	2.79
39	3.03	1.50	1.05	1.11	2.25

Tabla 2.3. Tabla que recoge los tiempos medios de graduación: años que emplea un estudiante en finalizar sus estudios.

Nos percatamos de que el tiempo medio de absorción asociado al estado de graduarse también decrece ligeramente en todos los grados. Esto se debe a que el tiempo medio que un alumno tarda en graduarse con un bajo número de asignaturas aprobadas es mayor que en el resto de los casos. Ciencias Ambientales posee los tiempos para graduarse más bajos, mientras que Física comienza con los mayores, hasta que a mitad de asignaturas superadas en el grado, Biología le hace relevo.

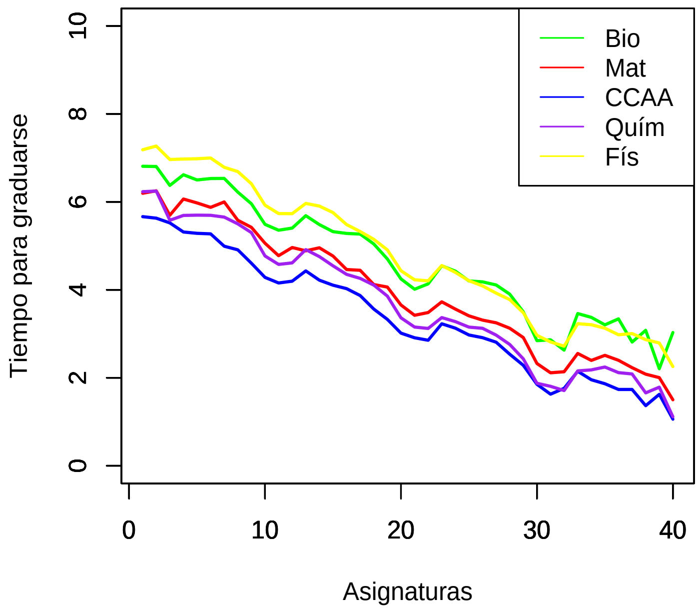


Figura 2.7. Representación gráfica del tiempo medio de absorción asociado al estado de graduarse.

2.2. Modelo 2: por cursos

2.2.1. Presentación del modelo

El segundo modelo que proponemos se trata de una cadena de Markov en la que se analizan cada uno de los grados de la facultad de Ciencias curso a curso. La cadena consta de cuatro estados transitorios que se corresponden con los cuatro cursos que tiene cada grado, y dos estados absorbentes, abandonar y graduarse, que representaremos por -1 y -2 respectivamente. Podemos observar un esquema que ilustra el modelo comentado:

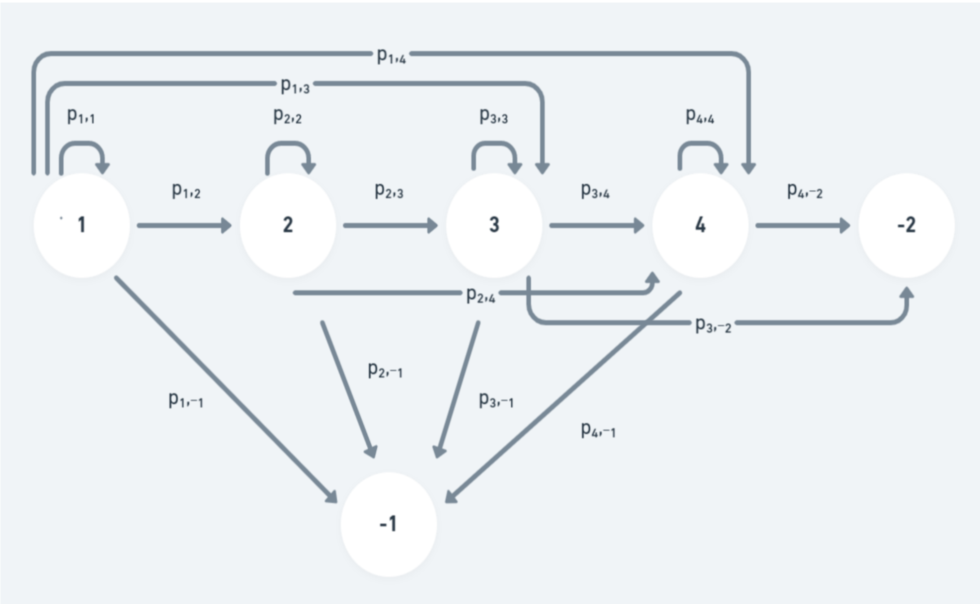


Figura 2.8. Representación gráfica del segundo modelo.

donde $p_{i,j}$ con i y $j \in \{$ -2, -1, 1, 2, 3, 4 $\}$ son las probabilidades de transición entre cada par de estados.

2.2.2. Análisis realizado

En este modelo las matrices serán de 6×6 y presentarán la siguiente estructura:

$$P' = \begin{pmatrix} p_{1,1} & p_{1,2} & p_{1,3} & p_{1,4} & p_{1,-2} & p_{1,-1} \\ p_{2,1} & p_{2,2} & p_{2,3} & p_{2,4} & p_{1,-2} & p_{2,-1} \\ p_{3,1} & p_{3,2} & p_{3,3} & p_{3,4} & p_{3,-2} & p_{3,-1} \\ p_{4,1} & p_{4,2} & p_{4,3} & p_{4,4} & p_{4,-2} & p_{4,-1} \\ p_{-2,1} & p_{-2,2} & p_{-2,3} & p_{-2,4} & p_{-2,-2} & p_{-2,-1} \\ p_{-1,1} & p_{-1,2} & p_{-1,3} & p_{-1,4} & p_{-1,-2} & p_{-1,-1} \end{pmatrix}.$$

Utilizamos el mismo enlace para visualizar el código mediante el que hemos obtenido las probabilidades de transición y también nos basamos en los mismos datos que para el modelo 1. Además, tomamos la matriz de ocurrencias del modelo anterior y sumamos las filas y columnas correspondientes a las asignaturas de cada curso para así obtener una matriz de menor dimensión, como explicamos en A.1. Una vez aplicamos la función

```
suavizar_matriz
```

construimos la nueva matriz de transición para cada uno de los grados. Para el caso de Biología, la matriz es la siguiente:

$$B' = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ -2 \\ -1 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & -2 & -1 \end{array} \\ \begin{pmatrix} 0{,}390 & 0{,}460 & 0{,}080 & 0 & 0{,}020 & 0{,}050 \\ 0{,}010 & 0{,}320 & 0{,}550 & 0{,}090 & 0{,}010 & 0{,}030 \\ 0 & 0{,}003 & 0{,}300 & 0{,}620 & 0{,}060 & 0{,}020 \\ 0 & 0 & 0{,}003 & 0{,}600 & 0{,}310 & 0{,}090 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array};$$

Por como se obtiene B', podemos decir que pasa de curso el alumno que tiene todas las asignaturas de ese curso aprobadas. A partir de la matriz de transición, se obtienen los tiempos de absorción y las probabilidades tanto para graduarse como para abandonar cada uno de los grados.

2.2.3. Resultados obtenidos

Del análisis realizado hemos obtenido los siguientes resultados

Curso	Prob. abandonar	Prob. graduarse	Tiempo de absorción
1	0.30	0.69	5.79
2	0.26	0.74	4.78
3	0.23	0.77	3.67
4	0.22	0.78	2.50

Tabla 2.4. Tabla que recoge los resultados del análisis del grado en Biología.

En la figura 2.9, observamos que en un principio el tiempo de absorción es mayor y va decreciendo conforme los alumnos pasan de curso en el grado. Esto se debe a que en un principio a los alumnos les queda más tiempo para finalizar sus estudios en el grado. Obtenemos una función casi constante que representa la evolución de la probabilidad de graduarse. Podemos apreciar un ligero crecimiento a partir del segundo curso, lo cual quiere decir que a partir de ese momento, la probabilidad de que los alumnos se gradúen es mayor que en los cursos previos.

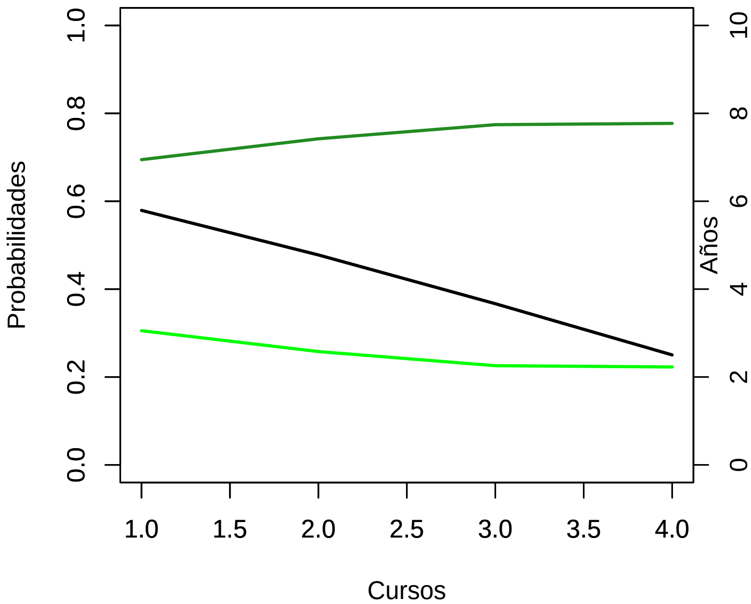


Figura 2.9. Representación gráfica de las probabilidades para abandonar y finalizar el grado en Biología y del tiempo que van a estar matriculados los alumnos conforme avancen de curso.

Nota.

Hemos realizado el mismo análisis para todos los grados de la facultad de Ciencias. En las figuras 2.10 y 2.11 se pueden observar los resultados obtenidos para el resto de grados.

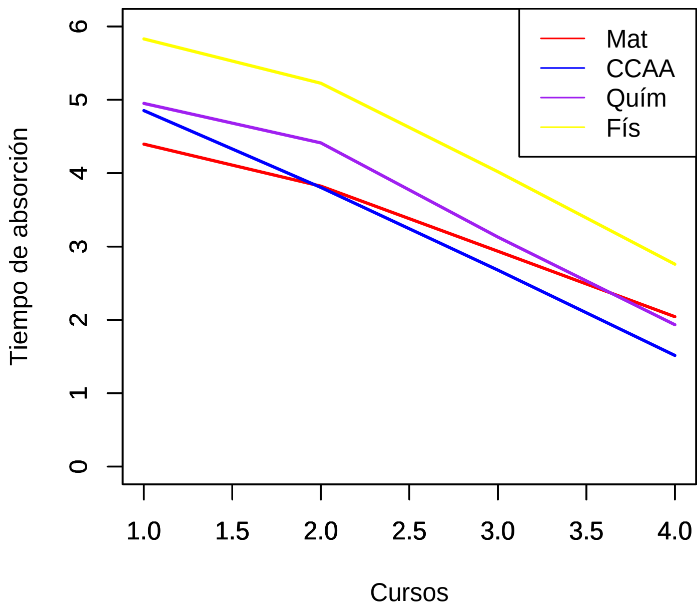
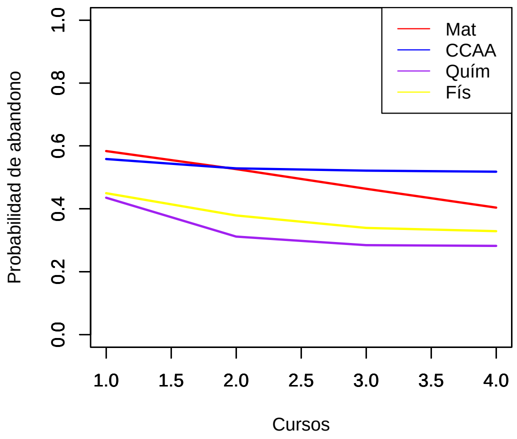


Figura 2.10. Representación gráfica del tiempo de absorción del resto de grados.

El tiempo de absorción es mayor en el grado en Física que en el resto de titulaciones, pero podemos apreciar que la función sigue una misma trayectoria decreciente en todos los grados. Es lógico puesto que cuanto más avancen los alumnos en el grado, menos tiempo les quedará matriculados. Ciencias Ambientales posee los menores tiempos de absorción una vez se ha llegado al segundo curso, mientras que Matemáticas comienza con los más bajos.



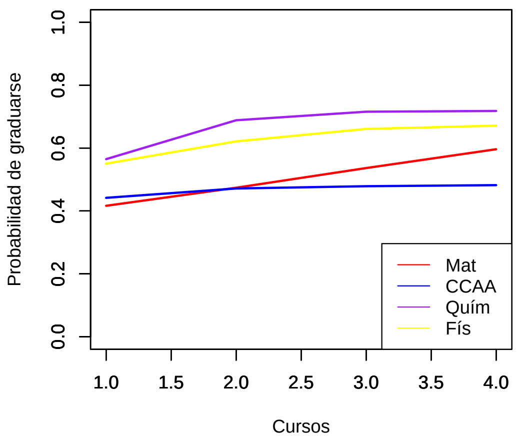


Figura 2.11. Representación gráfica de la probabilidad de abandono y graduarse del resto de grados.

En este caso, como podemos apreciar en la figura 2.11, la probabilidad de abandono comienza siendo más alta en el grado en Matemáticas, aunque a medida que pasan los cursos pasa a ser mayor en Ciencias Ambientales. Observamos que la menor probabilidad de abandono se corresponde con el grado en Química, seguido de Física.

Por otro lado, la probabilidad de graduarse va creciendo conforme los alumnos van avanzando de curso a lo largo de todas las titulaciones. Esto se debe a que cuánto más avancen los alumnos en el grado, mayor es la probabilidad de que consigan finalizarlo. Notamos que la probabilidad de graduarse en Matemáticas comienza siendo menor con respecto al resto de grados. Esto también sucedía en el modelo 1.

A continuación, calculamos los tiempos de absorción asociados a cada uno de los estados absorbentes: abandonar y graduarse.

Curso	Biología	Matemáticas	CCAA	Química	Física
1	10.66	5.72	5.45	7.13	9.01
2	9.90	5.19	4.33	6.83	8.67
3	9.00	4.39	3.21	5.62	7.63
4	7.74	3.58	1.99	4.31	6.35

Tabla 2.5. Tabla que recoge los tiempos medios para abandonar de cada grado.

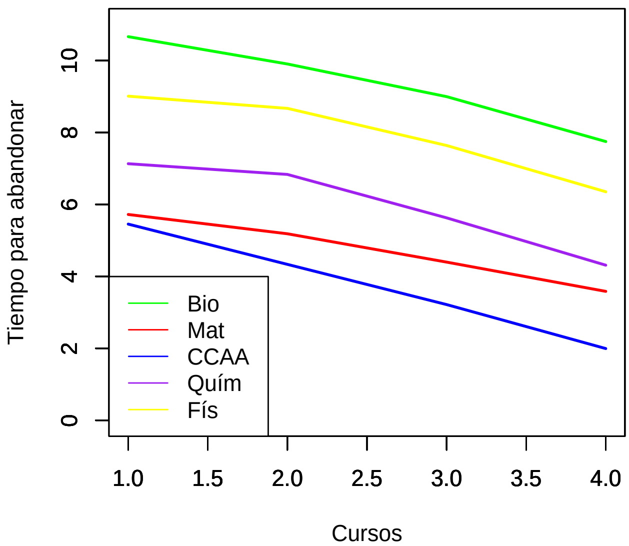


Figura 2.12. Representación gráfica del tiempo de absorción asociado al estado de abandonar.

El tiempo medio de absorción asociado al estado de abandonar es mayor cuando el alumnado comienza sus estudios, y decrece conforme se avanza en el grado.

Curso	Biología	Matemáticas	CCAA	Química	Física
1	6.80	6.62	5.93	6.54	7.84
2	5.42	5.32	4.48	4.98	6.48
3	4.14	3.94	3.31	3.55	4.97
4	2.93	2.75	2.07	2.30	3.62

Tabla 2.6. Tabla que recoge los tiempos de absorción asociados al estado de graduarse de cada grado.

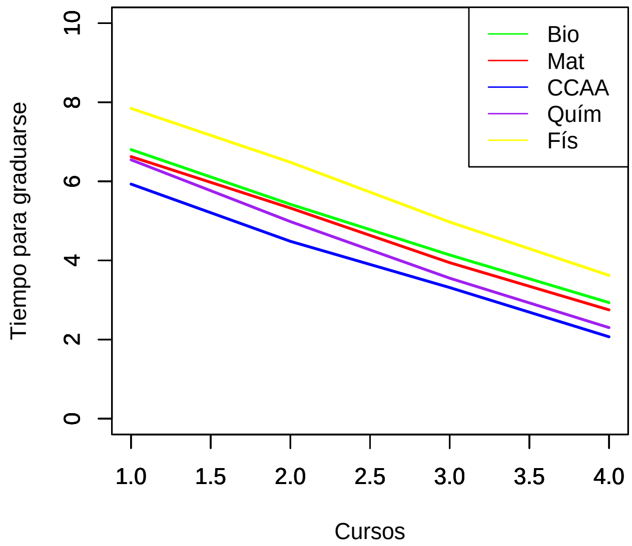


Figura 2.13. Representación gráfica del tiempo medio para graduarse cada curso.

El tiempo medio de absorción asociado al estado de graduarse decrece conforme el alumnado va avanzando de curso en el grado. Es lógico que cuanto más avanza un alumno en sus estudios, menos tiempo necesita para finalizarlos.

Conclusiones

En este trabajo se ha modelizado la evolución académica del alumnado de la Facultad de Ciencias como una cadena de Markov: un proceso estocástico sin memoria.

En el trabajo se exponen los fundamentos teóricos necesarios para conocer las cadenas de Markov, así como la herramienta fundamental para trabajar con ellas: las matrices estocásticas. Se presentan una serie de resultados para la obtención de las probabilidades y de los tiempos de absorción.

La parte más interesante de este trabajo ha sido el análisis de los datos cedidos por el GAP, que se presenta en el segundo capítulo. Para ello, aplicamos todos los conceptos teóricos expuestos en el capítulo uno.

Se hicieron dos modelos. En el primero, se presenta una cadena de Markov de 39 estados transitorios que se corresponden con los créditos que aprueba un alumno en cada año académico y 2 estados absorbentes: abandonar y graduarse. En el segundo, la cadena de Markov propuesta consta de 4 estados transitorios que se corresponden con los cuatro cursos de una titulación, y 2 estados absorbentes: abandonar y graduarse.

Podemos destacar que ambos modelos, aunque mostrando gráficos diferentes, concluyen resultados similares. Los tiempos de absorción decrecen en todos los grados en ambos modelos; la probabilidad de abandonar primero decrece de manera considerable y luego continúa decreciendo ligeramente; y, sin embargo, la probabilidad de graduarse, en ambos modelos, crece poco a poco a lo largo del grado. Además, consideramos que en los grados de nuestra facultad existe una alta probabilidad de abandono, sobre todo en el primer curso.

El lenguaje de programación utilizado, fue *R*. Resultó de gran utilidad para realizar los cálculos gracias a su amplia variedad de paquetes y su sencillez para visualizar datos.

Para ampliar este trabajo se podría realizar un análisis de los mismos datos mediante un cadena de Markov de tres estados: permanecer en la carrera, cambiar de carrera dentro de la misma facultad y retirarse de la universidad.

A

Apéndice

A.1. Código en el lenguaje R

Se podrá encontrar todo el código utilizado para analizar los datos mediante los modelos propuestos en el siguiente enlace: https://colab.research.google.com/drive/1tUdBglm_kCYvcXSp23OzJoRjJEHaVHBJ?usp=sharing, aunque, a continuación se explica brevemente basándonos en el primero de los modelos probabilísticos propuestos.

Antes de comenzar, conectamos Google Colaboratory con Google Drive mediante el código

```
from google.colab import drive
drive.mount('/content/gdrive')
```

para poder trabajar con las hojas de datos previamente subidas al Google Drive. Previamente, deben cargarse las librerías

`readr, data.table, rlist, readxl, dplyr, sqldf y matrixcalc.`

Se observa si un alumno aprueba o suspende el TFG mediante la función:

```
alumno_suspende_tfg <- function(id, tabla){

supera <- sqldf(sprintf("SELECT Supera FROM tabla WHERE
Id=%d AND Asignatura='TRABAJO FIN DE GRADO' AND Supera==1", id))

supera <- unlist(supera)

if(length(supera)==0){
    return(TRUE)
}
else{
    return(FALSE)
```

```
    }
  }
```

Seguidamente, se crea una lista de vectores que contiene la suma acumulativa de las asignaturas que aprueba un alumno cada año que esté matriculado, utilizando la función

```
generar_vectores_rendimientos
```

Un ejemplo de este vector podría ser $[0, 10, 20, 30, 40]$. Esto indicaría que el alumno ha superado 10 asignaturas en cada curso.

La siguiente función que necesitamos es

```
generar_matriz <- function(a,b,lista_vectores_rendimientos)
```

y se encarga de crear la matriz de ocurrencias de un curso a otro, es decir, una matriz que representa, por ejemplo, cuántos alumnos pasaron de tener 10 asignaturas aprobadas a tener 20 en el siguiente año. Los parámetros a y b, son números que representan el paso de un curso a otro en la carrera.
Nota.

A continuación mostraremos una función que necesitamos para suavizar las irregularidades (picos y valles) de los gráficos y, de este modo, reconocer fácilmente las tendencias de los alumnos. Dicha función es:

```
suavizar_matriz <- function(sumatorio_matrices,
suavizado_principal, suavizado_secundario){

matriz_auxiliar <- matrix(nrow=42, ncol=42)

matriz_auxiliar[is.na(matriz_auxiliar)] <- 0

matriz_auxiliar <- sumatorio_matrices*suavizado_principal +
(shift.up(sumatorio_matrices, rows = 1, fill =
0)*suavizado_secundario) + (shift.down(sumatorio_matrices,
rows = 1, fill = 0)*suavizado_secundario) +
(shift.left(sumatorio_matrices, cols = 1, fill =
0)*suavizado_secundario) + (shift.right(sumatorio_matrices_
biologia, cols = 1, fill = 0)*suavizado_secundario)

return(matriz_auxiliar)
}.
```

Una vez se obtiene la matriz de ocurrencias aplicándole el suavizado, se procede a hallar la matriz de transición en cada uno de los grados, mediante la función

```
generar_matriz_estocastica,
```

y una vez conseguida, se puede aplicar la teoría expuesta en el capítulo 1. Vamos a mostrar el código para el grado en Biología. Para el resto de grados se procede de forma análoga. Primero, se hallan las matrices R, Q y N:

```
matriz_Q_biologia =  matriz_estocastica_biologia[(1:40),(1:40)]

matriz_R_biologia = matriz_estocastica_biologia[(1:40),(41:42)]

matriz_I_identidad = diag(40)

matriz_N_previa_biologia = matriz_I_identidad - matriz_Q_biologia

matriz_N_biologia <- solve(matriz_N_previa_biologia)
```

A continuación, generamos el vector t y así se obtienen los tiempos de absorción y las matrices de probabilidad B:

```
vector_t = matrix(data=1, nrow=40, ncol=1)

tiempos_absorcion_biologia <- matriz_N_biologia%*%vector_t

matriz_B_biologia = matriz_N_biologia%*%matriz_R_biologia
```

Por último, representamos gráficamente los resultados del análisis anterior. Dichas representaciones pueden encontrarse en el capítulo 2. El gráfico ??, se consigue tras compilar:

```
plot(tiempos_absorcion_biologia, type="l", col="green",
main="Grado en Biologa", lwd=2, xlab="Asignaturas",
ylab="Prob. abandono-tiempo de absorcion")

par(new=TRUE)

plot(matriz_B_biologia[,2], type="l", col="green",
main="Grado en Biologia", lwd=2, xlab="Asignaturas",
ylab="Prob. abandono-tiempo de absorcion")
```

Luego, se decide buscar los tiempos de absorción asociados a cada uno de los estados absorbentes. Para ello se necesita modificar la función de obtención de matrices de transición a:

```
generar_matriz_estocastica_1,
```

ya que se debe eliminar la columna y fila que representa al estado absorbente del que no vamos a hallar el tiempo de absorción, es decir, se tiene una matriz cuadrada de una dimensión menos que antes.

Por tanto, se tiene:

```
matriz_tiempoab_biologia <- sumatorio_matrices_biologia[-41, -41]

matriz_tiempoab_biologia <- generar_matriz_estocastica_1(matriz_
tiempoab_biologia)
```

para el caso de abandonar el grado en Biología. La matriz

```
sumatorio_matrices_biologia
```

se trata de la matriz de ocurrencias.

La matriz final sobre la que luego se aplica la teoría del capítulo 1 sería:

```
matriz_tiempoab_biologia <- matriz_tiempoab_biologia[1:41,
1:41]*(1/(1-matriz_estocastica_biologia[(1:42)[-41], 41]))
```

Para el segundo modelo probabilístico propuesto, se necesita modificar nuevamente la función

```
generar_matriz_estocastica_2,
```

por el cambio de la dimensión de la matriz de ocurrencias.

Luego, se suman las filas y columnas que correspodían en el modelo 1 con cada uno de los cursos, o sea, de 10 en 10, y se obtiene así una matriz de $6x6$, teniendo en cuenta que se siguen teniendo dos estados absorbentes. Utilizamos el código:

```
convertir_matriz <- function(sumatorio_matrices_grado){
  mat_aux = matrix(nrow = 6, ncol = 6)
  mat_aux[is.na(mat_aux)] <- 0
  aux1 <- 1
  aux2 <- 10
  aux3 <- 1
  aux4 <- 10
  for(i in 1:4){
    for(j in 1:4){
      mat_aux[i,j] = sum(sumatorio_matrices_biologia[c(aux1:aux2),
      c(aux3:aux4)])

      if(aux3 == 1){
        aux3 = aux3+9
```

```
    }
    else{
      aux3 = aux3+10
    }
    aux4 = aux4+10
  }

  aux3 <- 1
  aux4 <- 10

  if(aux1 == 1){
    aux1 = aux1+9
  }
  else{
    aux1 = aux1+10
  }

  aux2 = aux2+10
}

aux1 <- 1
aux2 <- 10
aux3 <- 41
for(i in 1:4){
  for(j in 5:6){

    mat_aux[i,j] = sum(sumatorio_matrices_biologia[c(aux1:aux2),aux3])
    aux3 <- 42

  }
  if(aux1 == 1){
    aux1 = aux1 + 9
  }
  else{
    aux1 = aux1 + 10
  }
  aux2 = aux2 + 10
  aux3 <- 41
}
```

```
    mat_aux[5,5] <- 1
    mat_aux[6,6] <- 1

    return(mat_aux)
}
```

y la matriz de transición del segundo modelo para el grado en Biología, se consigue mediante la línea de código:

```
matriz_modelo2_bio <- convertir_matriz(sumatorio_matrices_biologia)
```

Bibliografía

[1] William Feller. *An Introduction to Probability Theory and Its Applications.* 1950, vol 1, third edition. Princeton University.

[2] G.Casella, S.Fienberg and I.Olkin. *Applied Probability.* Publisher: Springer, 2010.

[3] Charles M. Grinstead and J. Laurie Snell. *Introduction to Probability.* Version dated 4 July 2006, pp. 405–470.

[4] Henry Maltby, Worranat Pakornrat, Jeremy Jackson. *Markov Chains* [en línea]. Disponible en: https://brilliant.org/wiki/markov-chains/ [Consulta: 17-10-2021].

[5] Markov Chains. *Probability Theory* [en línea]. Disponible en: https://en.wikipedia.org/wiki/Markov_chain [Consulta: 05-11-2021].

[6] Saul Spatz. *Expected time till absorption in specific state of a Markov chain* [en línea]. Disponible en: https://math.stackexchange.com/questions/2852105/expected-time-till-absorption-in-specific-state-of-a-markov-chain [Consulta: 29-05-2022].

[7] R Core Team. *R: A language and environment for statistical computing*, 2019 [en línea] R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ [Consulta: 15-01-2022]

[8] Freddy Hernández y Olga Usuga. *Manual de R* [en línea] [2021-07-26] Disponible en: https://fhernanb.github.io/Manual-de-R/ [Consulta: 20-01-2022]

[9] Asael Alonzo. *Cadenas de Markov en R* [en línea] Disponible en: https://rpubs.com/asael_am/409332 [Consulta: 03-03-2022]

[10] Curso: *Introduction to the Tidyverse* [en línea] Disponible en: https://www.coursera.org/ [Consulta: 10-04-2022]

[11] Curso: *Importing Data in the Tidyverse* [en línea] Disponible en: https://www.coursera.org/ [Consulta: 10-04-2022]

[12] Mará Fernández-Mellizo. *Análisis del abandono de los estudiantes de grado en las universidades presenciales en España* [en línea] Disponible en: https://www.universidades.gob.es/stfls/universidades/ministerio/ficheros/Informe_Abandono_Universitario_completo_MFMS.pdf [Consulta: 23-06-2022]

[13] Fundación BBVA. *Abandono universitario* [en línea] Disponible en: https://www.fbbva.es/noticias/un-33-de-los-alumnos-no-finaliza-el-grado-que-inicio-y-un-21-\abandona-sin-terminar-estudios-universitarios/ [Consulta: 01-07-2022]

Lista de símbolos y abreviaciones

$\boldsymbol{X_t}$	Familia de variables aleatorias
$\boldsymbol{S}$	Espacio de estados
$\boldsymbol{p_{ij}}$	Probabilidad de transición del estado $\boldsymbol{i}$ al estado $\boldsymbol{j}$
$\boldsymbol{E(X)}$	Esperanza matemática
$\boldsymbol{I}$	Matriz identidad
$\boldsymbol{t_i}$	Número esperado de pasos antes de ser absorbido empezando en $\boldsymbol{i}$
$\boldsymbol{b_{ij}}$	Probabilidad de que una cadena absorbente sea absorbida en el estado $\boldsymbol{j}$ si parte del estado $\boldsymbol{i}$

Study of the evolution of the student body in the Faculty of Science

Sección de Matemáticas
Universidad de La Laguna

Andrea Marcos Vargas
Facultad de Ciencias · Sección de Matemáticas
Universidad de La Laguna
alu0101105345@ull.edu.es

Abstract

IN THIS PAPER, *we will present two probabilistic models whose main tool is the Markov chain. The objective of these two proposals is to assign to each grade of the Faculty of Science a Markov chain for the academic performance of its students. In the first proposal we analyze such performance through the credits passed by each student in each year. In the alternative model we represent the performance as a function of the students who pass and the repeaters. Both models are analyzed and compared using data from the Analysis Office of the University of La Laguna. In this report, examples with data from 2009 to 2020, including 2020, will be included.*

1. Chapter 1

A stochastic process is a collection of random variables X_t ordered according to the subindex t which, in general, is usually identified with time.

To describe a Markov chain we take a set S consisting of r states and without loss of generality we will identify it with $\{1, 2, 3, \ldots, r\}$. The process starts in one of these states and moves from one state to another. Each of these movements is called a "step". The main characteristic of a Markov chain is to verify the Markov property, which reflects that in our process, the probability that a chain that in a period t is in a state passes to another state in the next step, will depend only on the current state in which the process is, that is:

$$P(X_{t+1} = s_{t+1} | X_0 = s_0, X_1 = s_1, \ldots, X_{t-1} = s_{t-1}, X_t = s_t) = P(X_{t+1} = s_{t+1} | X_t = s_t)$$

for any set $\{s_0, \ldots, s_{t+1}\} \in S$.
A *Markov chain* is a discrete-time stochastic model that verifies Markov's property. Markov chains can be represented by means of transition matrixes.

A Markov chain is *absorbing* if it has at least one absorbing state (if it is impossible to exit from it, i.e., $p_{ii} = 1$), and if from each non-absorbing state it is possible to go to an absorbing state, not necessarily in a single step.
After the computational programming and execution in Google Colaboratory analyzed with R software, we obtained a set of results.

2. Chapter 2

2.1 Model 1

This first model that we present consists of a discrete-time Markov chain system with 40 transient states corresponding to each of the subjects of a grade. In addition, the model consists of 2 other absorbing states that represent the abandonment of the degree, -1, and the completion of the degree, 41. Next, we can observe a scheme in which the commented model is illustrated:

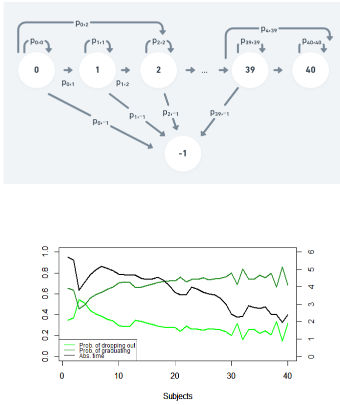


Figure 1: *Graphical representation of the probabilities of dropping out and completing the degree in Biology and of the time students will be enrolled according to the subjects they have passed.*

2.2 Model 2

In this second model, we analyze each of the degrees of the Faculty of Science on a course-by-course basis. We will do this by means of a Markov chain with 4 transient states that correspond to the 4 courses that each degree has, and 2 absorbing states, drop out and graduate, which we will represent by -1 and 100 respectively. We can observe a scheme that illustrates the commented model:

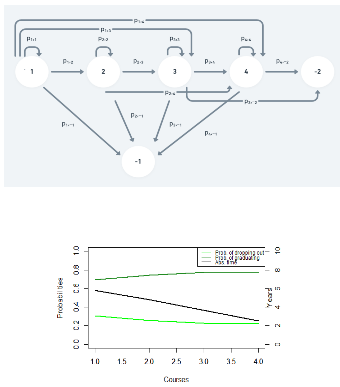


Figure 2: *Graphical representation of the probabilities of dropping out and completing the degree in Biology and of the time students will be enrolled as they advance in the course.*

References

[1] Grinstead and Snell's. *Introduction to Probability*. Version dated 4 July 2006, pp. 405–470.

[2] R Core Team. *R: A language and environment for statistical computing*. 2019. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.