Universidad de Las Palmas de Gran Canaria
Departamento de Matemáticas

Tesis Doctoral

Contribuciones al reemplazamiento óptimo de sistemas reparables y al dimensionamiento óptimo de sistemas de colas

Juan José González Henríquez

Las Palmas de Gran Canaria, Marzo de 2003

Universidad de Las Palmas de Gran Canaria
Departamento de Matemáticas

Tesis Doctoral

Contribuciones al reemplazamiento óptimo de sistemas reparables y al dimensionamiento óptimo de sistemas de colas

Autor: Juan José González Henríquez

Directores: Pedro Saavedra Santana y
Angelo Santana del Pino

DON PEDRO SAAVEDRA SANTANA, Catedrático de Universidad del Área de Conocimiento de Estadística e Investigación Operativa del Departamento de Matemáticas de la Universidad de Las Palmas de Gran Canaria y

DON ÁNGELO SANTANA DEL PINO, Titular de Universidad del Área de Conocimiento de Estadística e Investigación Operativa del Departamento de Matemáticas de la Universidad de Las Palmas de Gran Canaria,

CERTIFICAN que la presente memoria titulada,

CONTRIBUCIONES AL REEMPLAZAMIENTO ÓPTIMO DE SISTEMAS REPARABLES Y AL DIMENSIONAMIENTO ÓPTIMO DE SISTEMAS DE COLAS,

ha sido realizada bajo la dirección de ambos por el Licenciado en Matemáticas D. JUAN JOSÉ GONZÁLEZ HENRÍQUEZ, y constituye su Tesis para optar al grado de Doctor en Matemáticas.

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos a que haya lugar, firmamos la presente en Las Palmas de Gran Canaria, a 25 de marzo de dos mil tres.

fdo.: D. Pedro Saavedra Santana

fdo : D Angelo Santana del Pino

Índice General

Lista de Símbolos v

Agradecimientos vii

Prefacio ix

1 Generalidades sobre procesos puntuales 11

1.1 Introducción 11

1.2 Parámetros principales de un proceso puntual en $\mathbb{R}^+$ 16

1.3 Estacionariedad 19

1.4 La función de intensidad condicional 21

1.5 Modelos Básicos 24

1.5.1 Proceso de Poisson 24

1.5.2 Proceso de renovación 27

1.6 Superposición de procesos de renovación 30

1.7 El problema general de regla de parada 32

1.7.1 Caso monótono 34

1.8 El problema de parada y el reemplazamiento de sistemas reparables . 35

2 Modelos de sistemas reparables 39

2.1 Introducción 39

2.2 Reparaciones mínimas y perfectas 41

2.3 Reparaciones imperfectas 43
2.4 Otros modelos de reparaciones imperfectas 44
2.4.1 Modelos lineales generalizados y modelos loglineales 45
2.4.2 Proceso de renovación modulados 45
2.4.3 Procesos de renovación con tendencia 46
2.4.4 Proceso gamma no homogéneo 47
2.5 Modelos de sistemas reparables con costes 47
2.5.1 Políticas básicas de reemplazamiento 49

3 Un modelo de reparaciones mínimas para un sistema reparable compuesto. 53

3.1 Introducción 53
3.2 Determinación de la Estrategia Óptima 57
3.3 Análisis de un caso particular 60

4 Teoría de paradas óptimas y fiabilidad 63

4.1 Introducción 63
4.2 Modelo de Boland-Proschan y procesos de decisión semi-Markovianos 65
4.3 Modelo de Boland-Proschan y teoría de paradas óptimas 66

5 Dimensionamiento óptimo de un sistema de colas 69

5.1 Motivación y presentación del modelo 69
5.2 Descripción del estado del sistema 76
5.3 Cálculo de las probabilidades estacionarias del sistema 79
5.4 Probabilidades de pérdida. 84
5.5 Tiempos de espera. 85
5.6 Optimización del rendimiento del sistema. 90
5.7 Análisis de los resultados 91
5.7.1 Caso 1: $\lambda_R >> \lambda_N$, $\rho = 0.95$ 92
5.7.2 Caso 2: $\lambda_R << \lambda_N$, $\rho = 0.95$ 99

5.7.3 Caso 3: $\lambda_R \cong \lambda_N$, $\rho = 0.95$ 102
5.7.4 Caso 4: $\rho << 0.95$ 108
5.8 Conclusión. 109

Bibliografía **111**

Lista de símbolos

Concepto	**Definición**
$\mathbb{N}$	Conjunto de los números naturales incluido el 0
$\mathbb{R}$	Conjunto de los números reales
$\mathbb{R}^+$	Conjunto de los números reales positivos
$\mathcal{B}$	σ-álgebra de Borel sobre $\mathbb{R}$
$o(x)$	Una función que dividida entre x se aproxima a cero a media que x tiende a cero
$r(t)$	Función de razón de fallos.
$\lambda(t\|\mathcal{H}_t)$	Función de intensidad condicional.
$I_R(t)$	Función indicatriz del conjunto E.
$\lfloor x \rfloor$	Parte entera de x.

Agradecimientos

En primer lugar, quiero expresar mi agradecimiento a los profesores Dr. D. Pedro Saavedra Santana y Dr. D. Angelo Santana del Pino por su orientación, estímulo y eficaz ayuda en la dirección de este trabajo. También quiero expresar mi gratitud a muchos compañeros del departamento de Matemáticas, en especial a aquellos que se siente atraídos por la Estadística. Todos ellos, en en algún momento, me escucharon y me aconsejaron tanto profesionalmente como anímicamente. También agradezco a D. Eduardo Rodríguez Barrera, a D^a. Nancy Sanín y al Dr. D. Gustavo Montero sus sugerencias y aportaciones LaTeX en la redacción de este documento.

Por último agradezco a mis amigos y familiares su apoyo y entusiasmo a lo largo de esta empresa. Todavía recuerdo emotivamente aquellas simpáticas palabras de mi madre: "anímate y ponte a trabajar con tu gemelo". Un especial agradecimeinto a mi mujer por su santa paciencia (aunque no siempre; y con razón) en mis horas delante del ordenador.

Prefacio

En esta memoria analizamos tres problemas de optimización, dos en el campo de la fiabilidad de sistemas y uno en el campo de la teoría de colas. Los tres problemas comparten, desde un punto de vista teórico, la misma función objetivo. Se trata de la función del beneficio o coste por unidad de tiempo a largo plazo (lo que proceda) que aparece en la teoría de la renovación con recompensa.

En los problemas de fiabilidad, el concepto de renovación adquiere un significado particular: reemplazamiento de un sistema. El acelerado rítmo económico, junto con el gran desarrollo tecnológico, que se experimenta en nuestro tiempo hace que todo usuario de un sistema deba en algún momento estudiar la conveniencia de continuar con el sistema disponible o reemplazarlo. Con el tiempo y debido al desgate, la propensión al fallo de un sistema aumenta. De esta manera, el coste de reemplazamiento del sistema puede compensar económicamente el coste de las reparaciones que un sistema de edad considerada generaría en un futuro. Por tanto el objetivo en un problema de éstas características es determinar el instante óptimo para reemplazar un sistema por otro (de características similares) de tal manera que el beneficio por unidad de tiempo a largo plazo para la empresa sea el mejor posible.

El problema relacionado con la teoría de colas es un problema que surge en el campo de la comunicación de sistemas informáticos. En concreto, se trata de un problema de modelo de elecciones *(polling* model), es decir, un problema donde es necesario planificar la manera (cíclica o aleatoria) en que un único servidor atiende a múltiples colas. El caso que nos ocupa se trata de un sistema formada por dos colas con *buffers* finitos de tamaño N y R, respectivamente, que son atendidas por un

servidor mediante un esquema de prioridades controlado por otro buffer de tamaño M. Estos parámetros de control, M, N y R, deben elegirse cuidadosamente, de forma que se cumplan los requisitos exigidos por el tráfico de la red. Estos requisitos se concretan en minimizar el retardo para el tráfico que llega a la cola I y minimizar las pérdidas para el tráfico que llega a la cola II, sin que ello se consiga a costa de incrementar excesivamente las pérdidas para el tráfico que llega a la cola I y el retardo para el tráfico que llega a la cola II. Al final y de acuerdo con una estructura de costes adecuada, los parámetros M, N y R son aquellos que minimizan un determinada función de coste por unidad de tiempo, y a largo plazo.

Esta memoria se ha dividido en cinco capítulos. En el primero de ellos se exponen los conceptos matemáticos necesarios para resolver los problemas que nos hemos planteado. Estos problemas proceden de realidades diferentes pero puede ser modelados a través de un mismo modelo probabilístico : los procesos puntuales. Por tanto, y más concretamente, en el primer capítulo vamos a revisar los conceptos de la teoría de procesos puntuales que vamos necesitar. Además se expondrán resultados básicos de la teoría de paradas óptimas que serán determinantes a la hora de resolver los problemas de reemplazamiento de sistemas. Asimismo será el capítulo donde se fijará la notación que se utilizará a lo largo de esta memoria.

En el segundo capítulo se realizará una revisión de los problemas que comparten cierta similutud con los problemas que vamos a resolver. En primer lugar, revisaremos los problemas relacionados con los problemas de fiabilidad: los problemas de reemplazamiento de sistema, y posteriomente daremos un pequeña revisión de los modelos de elecciones (*polling* model).

En el tercer y cuarto capítulo se tratarán los problemas de reemplazamientos de sistemas. En el qinto capítulo se tratará el problema de teoría de colas.

Capítulo 1

GENERALIDADES SOBRE PROCESOS PUNTUALES

Como se ha mencionado en el prefacio de esta memoria, en este capítulo expondremos los conceptos teóricos necesarios para el desarrollo de los capítulos posteriores. Para evitar una exposición sobrecargada de conceptos propios de la generalidad y el rigor matemático de los temas que vamos a tratar, le daremos importancia a los aspectos intuitivos de ciertos conceptos que han sido ya establecidos con rigor por otros autores. Una buena parte de este capítulo esta dedicada a la teoría de procesos puntuales en la recta real y al estudio de ciertos modelos especiales. Finalizaremos el capítulo con dos secciones dedicadas a la relación entre la teoría de paradas óptimas y el problema de reemplazamiento de sistemas.

1.1 Introducción

Un proceso puntual estocástico es la abstracción matemática que surge, por ejemplo, al modelar poblaciones de objetos u organismos que se distribuyen aletoriamente en algún territorio o, eventos que se distribuyen a lo largo del tiempo. Por lo general hay un espacio de estados T y un conjunto de puntos $\{T_i\}$ del espacio T que representan las localizaciones de los diferentes miembros de la población, o los tiempos en que ocurren los eventos. Si T es el plano, los puntos $\{T_i\}$ pueden representar las coordenadas donde se situan los hormigueros o la maleza en una determinada parcela de terreno. Si T es la recta real, los puntos pueden ser los instantes en el que un sistema falla o el momento en que los clientes llegan a una

cola. Para terminar, si T es la superficie de una esfera los puntos pueden representar los epicentros de grandes terremotos o la ubicación de volcanes activos.

La mayor parte de la teoría formal de los procesos puntuales puede llevarse a cabo con cierta generalidad en un espacio T que sea métrico, completo, separable y σ-compacto. Sin embargo y teniendo en cuenta que los problemas que queremos resolver son problemas aplicados, hemos preferido desarrollorar la teoría de procesos puntuales y algunos resultados en el espacio donde se dan la mayoría de las aplicaciones: la recta real.

Un proceso puntual en $\mathbb{R}$ puede considerarse como un modelo para las ocurrencias de cierto fenómeno de carácter instantáneo y recurrente que se manifiesta según un patrón aleatorio sobre una línea temporal o espacial. Así, después de haber observado dicho fenómeno y haber fijado un punto de referencia en dicha línea, disponemos de una sucesión ordenada $\{T_i\}$ de tiempos o distancias (con i perteneciente a un adecuado conjunto de índices I) que determina la ubicación de los eventos recurrentes obervados. Como en la práctica raramente tendremos ocasión de tratar con situaciones donde existan más de un número finito de puntos $\{T_i\}$ dentro de un conjunto acotado de la recta, haremos dicha suposición a lo largo de esta memoria. Este hecho junto con la propiedad de σ-compacidad de la recta real implica que la población que consideraremos será a lo sumo numerable. También nos gustaría asumir desde este momento que $T_m \neq T_n$ si $m \neq n$, es decir, que no podrán producirse dos o más ocurrencias simútaneas (proceso puntual simple); sin embargo, esta cuestión la estudiaremos posteriormente. De entrada, vamos a permitir ocurrencias simultáneas.

A partir de las consideraciones anteriores veremos que las trayectorias de un proceso puntual sobre la recta real pueden describirse de cuatro formas equivalentes, a saber:

(i) Como medidas de recuento;

(ii) Como funciones en escalera no decrecientes con recorrido en $\mathbb{N}$;

(iii) Como una sucesión de puntos $\{T_i\}$, y

(iv) Como una sucesión de periodos entre saltos.

Para ello, sea $\mathcal{B}$ la σ-álgebra de los subconjuntos Borel de la recta real y sea $N(A)$ el número de ocurrencias del proceso en el conjunto A con $A \in \mathcal{B}$; es decir,

$$N(A) = \sharp\{i : T_i \in A\} = \textit{número de índices } i \text{ para los cuales } T_i \in A. \tag{1.1}$$

Esta función de conjunto verifica, naturalmente, las siguientes propiedades:

1. $N\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} N(A_i)$ siendo $A_i \cap A_j = \emptyset$ para $i \neq j$ y $A_i \in \mathcal{B}$, $\forall\, i \in \mathbb{N}$.
2. $N(A)$ es un valor entero no negativo que puede ser infinito.
3. $N(A) < \infty$ para todo conjunto Borel A acotado de la recta real.

Toda función de conjunto $N(A)$ que satisfaga las tres primeras propiedades anteriormente mencionadas recibe el nombre de medida de recuento sobre la σ-álgebra $\mathcal{B}$ de los subconjuntos Borel de la recta real.

Para ser consistentes con el concepto de función de conjunto, deberíamos escribir, por ejemplo, $N((a,b])$ cuando A es el intervalo simiabierto $(a,b]$; sin embargo, por brevedad, preferimos escribir $N(a,b]$. A partir de aquí, podemos definir la función real $N(t)$ con $t > 0$ como

$$N(t) = N(0,t] = N((0,t]). \tag{1.2}$$

Esta función $N(t)$ es no decreciente, continua a la derecha y con recorrido en $\mathbb{N}$; en resumen, una función en escalera. Para procesos puntuales sobre la simirrecta real positiva, el conocimiento de $N(t)$ para todo $t > 0$ es suficiente para determinar $N(A)$ para todo Boreliano $A \subset (0,\infty)$ de la misma manera que una función de distribución determina una medida de probabilidad sobre los conjuntos Borel. Cuando el proceso puntual esta definido sobre la recta real, la definición (1.2) puede extenderse a,

$$N(t) = N((0,t]) \cdot I_{(0,\infty)}(t) + 0 \cdot I_{\{0\}}(t) - N((t,0]) \cdot I_{(-\infty,0)}(t). \tag{1.3}$$

De esta manera, la función $N(t)$ conserva la propiedad de ser en escalera sobre toda la recta real. También, esta útima función $N(t)$ determina $N(A)$ para todos los conjuntos Borel A y por lo tanto describe un proceso puntual a través de las funciones en escalera. Por tanto, en vez de haber comenzado con las medidas de recuento $N(A)$ podríamos haber descrito las trayectorias como funciones no decrecientes continuas a la derecha $N(t)$ $(-\infty < t < \infty)$ que son nonegativas y entero valoradas para todo $t > 0$, y no positivas y entero valoradas para todo $t < 0$, con $N(0) = 0$.

La tercera forma de describir las trayectorias de un proceso puntual en la recta adquiere notable sencillez cuando el proceso se define sobre la recta real positiva. Sea,

$$T_i = \inf\{t > 0 : N(t) \geq i\} \qquad (i = 1, 2, \ldots). \tag{1.4}$$

De esta definición se deduce fácilmente una una importante consecuencia,

$$T_i \leq t \text{ si y sólo si } N(t) \geq i. \tag{1.5}$$

Esta relación pone de manifiesto que especificar la sucesión creciente de puntos $\{T_i\}$ es equivalente a especificar las funciones $N(t)$ en el caso en que $N(-\infty, 0] = 0$. Si el proceso tiene puntos en toda la recta real, la extensión más simple consistente con (1.4) se obtiene definiendo,

$$\begin{aligned} T_i &= \inf\{t : N(t) \geq i\} \qquad (1.6) \\ &= \begin{cases} \inf\{t > 0 : N(0, t] \geq i\} & (i = 1, 2, \ldots), \\ -\inf\{t > 0 : N(-t, 0] \geq -i + 1\} & (i = 0, -1, \ldots). \end{cases} \end{aligned}$$

Esta doble sucesión infinita de puntos tiene la propiedad,

$$T_i \leq T_{i+1} \ (\forall i) \ \text{ y } \ T_0 \leq 0 < T_1 \tag{1.7}$$

Finalmente, si definimos,

$$X_i = T_i - T_{i-1} \quad \text{con} \quad \{T_i\} \text{ como en (1.6)}, \tag{1.8a}$$

el proceso queda completamente determinado también por la sucesión de periodos entre saltos $\{X_i\}$ (en adelante a las variables X_i las denominaremos variables intervalos o simplemente intervalos) y uno de los puntos $\{T_i\}$, por lo general T_0. La definición (1.8a) para la recta real positiva, sería igual, salvo que los puntos $\{T_i\}$ serían como en (1.4) y con el convenio de que $T_0 = 0$. Obsérvese que $X_i \geq 0$, y que si $N(t) \to \infty$ $(t \to \infty)$ entonces $\sum_{i=1}^{n} X_i \to \infty$ $(n \to \infty)$, mientras que si $N(t) \nrightarrow \infty$ $(t \to \infty)$ entonces X_i no está definido para $i > \lim_{t\to\infty} N(t)$.

Sea $\mathcal{N}$ el espacio formado por las medidas de recuento (con las propiedades anteriores), $\mathcal{N}_t$ el espacio de la funciones en escalera, $\mathcal{T}$ el espacio de las secuencias de puntos y $\mathcal{X}$ el espacio de la sucesión de intervalos. Si dotamos a cada uno de estos espacios con la σ-álgebra adecuada, puede asumirse, de acuerdo de los párrafos anteriores, que existen aplicaciones medibles biyectivas entre dos cualesquiera de los espacios $\mathcal{N}$, $\mathcal{N}_t$, $\mathcal{T}$ y $\mathcal{X}$. De esta manera, existe una correspondencia biyectiva entre las distribuciones de probabilidad definidas sobre uno de los espacios y las distribuciones de probabilidad definidas sobre otro de los restantes espacios. Así, al fijar las distribuciones finito-dimensionales en uno cualquiera de estos espacios, con las correspondientes condiciones de consistencia de Kolmogorov, tenemos definido un proceso puntual en la recta real. Sin embargo, como ocurre en la teoría general de procesos estocásticos, las distribuciones finito-dimensionales, salvo que el proceso sea separable, no nos dan información de las propiedades topológicas (continuidad, derivabilidad, etc) de las trayectorias del proceso. Por tanto, es necesario comprobar ciertas propiedades analíticas del proceso que veremos posteriormente, para estar seguros de que nuestras trayectorias son como queremos: funciones no decrecientes continuas a la derecha $N(t)$ $(-\infty < t < \infty)$ con recorrido en los enteros potivos para todo $t > 0$; y con recorrido en los enteros negativos para todo $t < 0$; además, $N(0) = 0$. Todo lo anteriormente expuesto puede encontrarse bien formalizado en [1] (capítulos 6 y 7). Resumimos la discusión anterior en el siguiente teorema.

Teorema 1 *Cualquiera de las siguientes condiciones determina la distribución de probabilidad de un proceso puntual sobre* $\mathbb{R}$*:*

(i) Las distribuciones finito-dimensionales $P\{N(A_i) = n_i \; ; i = 1, 2, \ldots, n\}$ *siendo* A_i *conjutos Borel acotados*.*

(ii) Las distribuciones finito-dimensionales de la sucesión de puntos $\{T_i\}_{i=-\infty}^{\infty}$.

(iii) Las distribuciones finito-dimensionales de la sucesión de intervalos $\{T_0, X_i : i = 0, \pm 1, \ldots\}$.

(iv) Las distribuciones finito-dimensionales del proceso $\{N(t) : t \in \mathbb{R}\}$.

1.2 Parámetros principales de un proceso puntual en $\mathbb{R}^+$

Una forma natural de medir el número de ocurrencias en un intervalo determinado en un proceso puntual es a través de la función $M(t) = E[N(t)]$. A lo largo de esta sección estudiaremos la importancia de esta función y su relación con aspectos interesantes de los procesos puntuales en $\mathbb{R}^+$. Se puede comprobar que $M(t)$ es una función no decreciente continua a la derecha y tal que el límite por la izquierda $M(t-) = \lim_{s \uparrow t} M(t)$ existe para todo t. Respecto a $M(t)$ podemos decir lo siguiente:

- A partir de la equivalencia (1.5) se deduce que $P(N(t) \geq i) = P(T_i \leq t)$. Si denotamos por $G_i(t)$ a la función de distribución de la variable T_i tenemos que,

$$M(t) = E[N(t)] = \sum_{i=1}^{\infty} P(N(t) \geq i) = \sum_{i=1}^{\infty} G_i(t). \tag{1.8b}$$

  Por lo tanto si las funciones de distribución $G_i(t)$ son absolutamente continuas, $M(t)$ es absolutamente continua y su derivada $m(t)$ (salvo en un conjunto de

*Un teorema sobre medidas aleatorias afirma que la distribución de probabilidad de una medida aleatoria queda determinada al conocer las distribuciones finito-dimensionales de toda familia $A_1, \ldots, A_k$ de conjuntos disjuntos de un semianillo de conjuntos acotados que genera la σ-álgebra de Borel correspondiente. Por tanto los conjuntos A_i puede ser intervalos disjuntos de la forma $(a, b]$.

medida de Lebesgue nula) es de la forma,

$$m(t) = \sum_{i=1}^{\infty} g_i(t) \tag{1.8c}$$

siendo $g_i(t)$ una densidad de $G_i(t)$. Denominaremos a $m(t)$ función de razón de ocurrencias de puntos del proceso; representa la razón instantánea de variación del número esperado de puntos con respecto al tiempo. La integral de $m(t)$ en un intervalo nos da el número medio de puntos en ese intervalo.

- Si $M(t)$ existe en un entorno de t_0 y $M(t_0-) \neq M(t_0)$ entonces, hay una probabilidad positiva de que el suceso de interés ocurra al menos un vez en t_0. El recíproco también es cierto.

Además de la función razón de ocurrencia de puntos, en la teoría de procesos puntuales es de interés la llamada función de intensidad definida por,

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(N(t, t+\Delta t] \geq 1)}{\Delta t}. \tag{1.8d}$$

Si $\lambda(t)$ existe, $\lambda(t)\Delta t$ es aproximadamente la probabilidad de que haya al menos un punto en el intervalo $(t, t+\Delta t]$. Se puede probar que si $M(t) < \infty$ para todo t entonces la función $\lambda(t)$ existe.

En la primera sección de este capítulo hablábamos del interés especial que tienen para nosostros los procesos puntuales simples, aquellos procesos sin múltiples ocurrencias. En términos probabilísticos, aquellos procesos puntuales para los que $P\{N(\{x\} = 0 \text{ o } 1; \forall x \in \mathbb{R}\ \} = 1$. También decíamos que cuando se define un proceso puntual a través de las distribuciones finito-dimensionales con las correspondientes condiciones de consistencias de las que habla el teorema de extensión de Kolmogorov, no queda claro como son las trayectorias del proceso (continuas, derivables , etc). En el caso que nos ocupa sabemos que las trayectorias $N(t)$ $(t \geq 0)$ son funciones con recorrido en $\mathbb{N}$, no decrecientes, continuas a la derecha y con $N(0) = 0$. Nos falta saber si estas funciones en escalera son o no de saltos unitarios. Desde luego,

si los saltos son unitarios el proceso puntual es simple. A continuación daremos dos condiciones suficientes para garantizar que el proceso puntual es simple.

Definición 1

$$\Lambda(t) = \sup \sum_{i=1}^{n} P\{N(t_{i-1}, t_i] \geq 1\}, \tag{1.8e}$$

cuando el supremo recorre todas las particiones $0 = t_0 < t_1 \ldots < t_n = t$ *del intervalo* $(0, t]$. $\Lambda(t)$ *es la integral superior de la función intervalo* $P[N(a, b] \geq 1], a \geq b$.

Teorema 2 *Si* $M(t) < \infty$ *para cada* $t \geq 0$,

$$M(t) = \Lambda(t) \Longleftrightarrow N \text{ es simple.}$$

Por lo tanto, una condición suficiente para que un proceso puntual integrable $(M(t) < \infty)$ sea simple es que $M(t) = \Lambda(t)$. También son de interés los siguientes teoremas.

Teorema 3 *Si* $M(t) < \infty$ *para cada* $t \geq 0$, *entonces:*

a)$E[\xi(t)] = \Lambda(t)$, *siendo* $\xi(t)$ *el número de discontinuidades de* $N(\cdot)$ *sobre el intervalo* $(0, t]$.

b)$\lambda(t)$ *existe en casi todo punto y* $\Lambda(t) = \int_0^t \lambda(x)\, dx$.

Corolario 1 *Si un proceso puntual integrable es simple entonces*

$$M(t) = \Lambda(t) = \int_0^t \lambda(x)\, dx \quad y \quad m(t) = \lambda(t) \quad \text{en casi todo punto.}$$

Del corolario anterior se deduce que, para un proceso puntual con esperanza finita y sin ocorrencias simultáneas, puede hablarse sin ambigüedad de la función de razón de ocurrencia de puntos $m(t)$, la cual coincide con la funcion $\lambda(t)$. En términos de fiabilidad de sistemas, la igualdad $m(t) = \lambda(t)$ indica que la razón de ocurrencia de fallos es igual a la razón con la que se producen los instantes de fallo, dado que sólo se produce un fallo en cada instante de fallo. Esta igualdad no se da cuando

el sistema está sujeto a circunstancias que pueden causar muchos fallos simultáneos. Así ocurre, por ejemplo, con ciertos fenómenos (una tormenta de granizo, una lluvia de meteoritos) que pueden causar varias averías simultáneas en un sistema. En este caso, el conocimiento de la función $m(t)$ es para la empresa de mayor utilidad que el conocimiento de $\lambda(t)$ pues, a partir de ella podrá decidir el número de operarios de mantenimiento necesarios para hacer frente a las múltiples averías que se producirán en un instante incierto t.

Hemos dicho que si $M(t) < \infty$ y $M(t) = \Lambda(t)$ para todo t entonces el proceso puntual es simple. Aparte de esta condición, existen otras condiciones, que tratan aspectos infinitesimales del proceso, para garantizar que un proceso puntual es simple. En [1], (capítulo 7) se habla con rigor de todas estas condiciones infinitesimales. De todas ellas nos merece especial atención la siguiente: un proceso puntual en la recta real es *orderly* si,

$$\lim_{\delta \to 0+} \frac{P(N(t, t+\delta] > 1)}{\delta} = 0. \tag{1.8f}$$

Puede probarse que, un proceso puntual que verifica (1.8f) es simple. Para finalizar diremos que en un proceso puntual simple se verifican las siguientes aproximaciones,

$$M(x+\delta) - M(x) = E\left[N(x, x+\delta]\right] \simeq P\left\{N(x, x+\delta] = 1\right\} \simeq P\left\{N(x, x+\delta] \geq 1\right\} \tag{1.8g}$$

1.3 Estacionariedad

En esta sección vamos a exponer algunos conceptos sobre procesos puntuales estacionarios que emplearemos en esta memoria.

Definición 2 *Un proceso puntual sobre* $\mathbb{R}$ *es estacionario si para todo* $r = 1, 2, \ldots,$ *y para todo subconjunto Borel acotado* $A_1, \ldots, A_r$ *de* $\mathbb{R}$ *la distribución conjunta del vector,*

$$\{N(A_1 + t), \ldots N(A_r + t)\}$$

no depende de t con $-\infty < t < \infty$.

En el caso de que el proceso puntual se defina sobre la semirrecta real positiva, los conjuntos A_i serán subconjuntos Borel de $(0, \infty)$ y t tomará valores reales positivos. Esta forma de estacionariedad puede debilitarse de diversas maneras; por ejemplo restringiendo los conjuntos $A_1, \ldots, A_r$ a ser intervalos y/o acotando por un valor b los valores de r. Entre estos casos se encuentra la denominada estacionariedad simple. Un proceso puntual es estacionario simple o simplemente estacionario si la distribución del número de puntos en un intervalo depende de su longitud pero no de su localización[†]; esto es, si

$$p_k(x) \equiv P\{N(t, t+x] = k\} \qquad (x > 0, k = 0, 1, \ldots), \tag{1.8h}$$

depende de la longitud de x pero no de la localización de t. Se puede comprobar ([1], pág 43) que para un proceso puntual con esta propiedad $M(t) = m \cdot t$. De acuerdo con (1.8c) m nos da la razón de ocurrencia de puntos del proceso. ¿Cuándo existirá la función de intensidad (1.8d) de un proceso simplemente estacionario? Si existe, ¿Será constante e igual a m? Y, ¿Cuándo es simple un proceso simplemente estacionario? El siguiente teorema da respuesta a estas preguntas.

Teorema 4 *(a) (Khintchine) La función de intensidad de un proceso simplemente estacionario existe y es una constante* λ *con* $0 \leq \lambda \leq \infty$

(b) (Korolyuk) Si un proceso puntual es simple y simplemente estacionario entonces $\lambda = m$.

(c) (Dobrushin) Si un proceso puntual es simplemente estacionario y $m < \infty$ *entonces el proceso es simple (y* ordely (1.8f)*)*

Definición 3 *Un proceso puntual tiene intervalos estacionarios cuando para todo* $r = 1, 2, \ldots,$ *y para todo los enteros* $i_1, \ldots, i_r$ *la distribución conjunta de* $\{X_{i_1+k}, \ldots, X_{i_r+k}\}$ *no depende de* k $(k = 0, \pm 1, \ldots)$.

[†]A esta propiedad también se le conoce como propiedad de incrementos estacionarios.

De acuerdo con, (1.8a) para que las trayectorias de un proceso con esta propiedad esté bien definido es preciso elegir un punto arbitrario t_0 a partir del cual se ubican los puntos del proceso. Una posibilidad, quizá la más natural, es tomar $t_0 = 0$. Por lo general, el proceso de la definición anterior no será estacionario en el sentido de la definición 2.

Uno de los resultados más interesantes de esta sección es la correspondencia biyectiva que existe entre las distribuciones de probabilidad $\mathcal{P}$ de procesos estacionarios simples sobre $\mathbb{R}$ con razón de ocurrencias m finita y las distribuciones de probabilidad $\mathcal{P}_o$ de las dobles sucesiones estacionarias de variables aleatorias positivas con media m^{-1}([1], pág 475). Sin ánimo de extendernos en cuestiones que pertenecen a la llamada teoría de Palm, únicamente decir que de esta correspondencia se deduce una de las relaciones más conocidas en teoría de colas, las famosas ecuaciones de Palm-Kinchin:

$$\begin{aligned} p_k(x) &= -m\int_0^x \{\pi_k(u) - \pi_{k-1}(u)\}du \quad (k = 1, 2, \dots), \\ p_0(x) &= 1 - m\int_0^x \pi_0(u)du \end{aligned} \tag{1.8i}$$

siendo $\pi_k(x) = \lim_{\delta \to 0^+} P\{N(0,x] = k \mid N(-\delta, 0] > 0\}$ con x fijo y $k = 0, 1, 2, \dots$ Es claro que para cada x las $\pi_k(x)$ constituyen una distribución discreta de probabilidad, llamada distribución de Palm. Las ecuaciones (1.8i) adquieren relevancia práctica cuando observamos un proceso estacionario simple en un punto (un evento) arbitrario del proceso (muestreo síncrono).

1.4 La función de intensidad condicional

A partir de las observaciones reales de tiempos de vida, tanto en fiabilidad como en el análisis de supervivencia, es difícil elegir un modelo entre las diversas funciones no simétricas de probabilidad. Por ejemplo, las diferencias entre la función de distribución gamma, Weibull y lognormal son significativas tan sólo en la cola de la distribución. Por otra parte, y debido a la limitación del tamaño muestral, los tiempos

de vida suelen estar diseminados en la cola de la distribución por lo que resulta difícil discriminar entre estas funciones de distribución. Es necesario, por tanto, definir un concepto que nos permita distiguir entre dos posibles funciones de distribución. Tal concepto es la función de razón de riesgo o función de razón de fallos definida como:

$$r(t) = \lim_{\Delta \to 0} \frac{P(t \leq T < t + \Delta \mid t \geq T)}{\Delta}. \tag{1.8j}$$

Si la distribución de fallos F tiene función de densidad f, la función de razón de fallos $r(t)$, para aquellos valores de t tales que $F(t) < 1$, adquiere la forma:

$$r(t) = \frac{f(t)}{\overline{F}(t)} \tag{1.8k}$$

donde $\overline{F}(t) = 1 - F(t)$. Esta función tiene una interpretación probabilística muy útil; $r(t)\Delta$ representa la probabilidad de que un objeto de edad t falle en el intervalo $[t, t + \Delta)$. Atendiendo a la siguiente igualdad,

$$\overline{F}(t) = \exp\{-\int_0^t r(s)ds\} \tag{1.8l}$$

es claro que F queda determinada unívocamente por su función de riesgo. Es por esta razón y por su interpretación probabilística por lo que en fiabilidad y en supervivencia se suele utilizar la función de riesgo como guía para elegir el modelo que mejor se ajusta a los tiempos de vida observados. Precisamente algo parecido ocurre cuando se observan datos de un fenómeno que se puede modelar a través de un proceso puntual simple. En este caso la función recibe el nombre de función de intensidad condicional y se define como:

$$\lambda(t|\mathcal{H}_t) = \lim_{\Delta \to 0} \frac{P(N(t + \Delta) - N(t) \geq 1 \mid \mathcal{H}_t)}{\Delta}; t > 0 \tag{1.8m}$$

donder $\mathcal{H}_t$ denota la historia del proceso de fallos hasta el instante t e incluyendo t. Por lo general, se entenderá que $\mathcal{H}_t$ está generada por $\{N(s) : 0 \leq s \leq t\}$ y que $P(N(t + \Delta) - N(t) > 1) = o(\Delta)$. Esto último es una condición necesaria para que el proceso sea simple. En fiabilidad, por ejemplo, $\lambda(t|\mathcal{H}_t)\Delta$ es aproximadamente

la probabilidad de un fallo en el intervalo $[t, t+\Delta t)$, condicionado a la historia del proceso de fallos hasta el instante de tiempo t.

Desde el punto de vista teórico la función de intensidad condicional es sumamente importante, pues a partir de ella queda completamente determinada la estructura probabilística de un proceso puntual simple en la recta real positiva. Para una demostración rigurosa de este enunciado puede consultarse [1] (cap. 13) o [2]. Desde el punto de vista práctico, la elección de un modelo para unas observaciones reales se realiza más fácilmente a través de la función de intensidad condicional que a través de otras formas equivalentes que sirven para definir un proceso puntual simple. De hecho, en la modelización de fallos de sistemas reparables (sistemas que pueden ser puestos en funcionamiento tras un fallo) es costumbre desde hace algún tiempo dar la función de intensidad condicional, cuando se quiere especificar un modelo de fallos para un sistema reparable. Haremos una revisión de estos modelos en el siguiente capítulo de esta memoria.

A continuación trataremos una situación interesante donde la función de intensidad condicional adquiere una expresión concreta. Para ello, sea $\{T_i\}$ la sucesión de puntos de un proceso puntual simple con $0 = T_0 < T_1 < T_2 < \ldots$ y sea $F_n(x|T_0, T_1, \ldots, T_{n-1})$ la función de distribución condicional de la variable $X_n = T_n - T_{n-1}$ condicionada a $\mathcal{H}_{T_{n-1}} = \sigma(T_0, \ldots, T_{n-1})$. Supongamos además que para todo $n \geq 1$, $F_n(x|T_0, T_1, \ldots, T_{n-1})$ admite una función de densidad $f_n(x)$, es decir,

$$F_n(x|\mathcal{H}_{T_{n-1}}) = \int_0^x f_n(u)du. \tag{1.8n}$$

Bajo esta suposiciones se puede demostrar que la función de intensidad condicional adquiere la forma concreta,

$$\lambda(t|\mathcal{H}_t) = \frac{f_n(t-T_n)}{1-F_n(t-T_n|\mathcal{H}_{T_{n-1}})} \quad \text{donde } \{T_n \leq t < T_{n+1}\}. \tag{1.8o}$$

Una demostración heurística de este resultado puede encontrarse en la páginas 59−61 de [2]. Si por ejemplo, la sucesión de intervalos $\{X_i\}$ está formada por variables

aleatorias independientes con distribución común $F(t)$ con densidad $f(t)$ y con función de riesgo $r(t)$ (proceso de renovación), la función de intensidad condicional con respecto a la historia interna $\mathcal{H}_t$ viene dada por,

$$\lambda(t|\mathcal{H}_t) = r(t - T_{n-1}) \quad \text{donde } \{T_{n-1} \leq t < T_n\} \tag{1.8p}$$

Nota 1 *Respecto a la definición de función de intensidad condicional hemos de decir que algunos autores condicionan respecto de* $\mathcal{H}_{t-} = \sigma\{N(s) : 0 \leq s < t\}$. *Esto se hace para garantizar que* $\lambda(t)$ *sea predecible y por tanto única. La definición dada por nosotros coincide con Cox [3] y Bremaud [2]. Este último afirma en su libro (página 31) que si existe un función de intensidad puede siempre encontrarse una versión predecible.*

1.5 Modelos Básicos

1.5.1 Proceso de Poisson

De acuerdo con el teorema 1, un proceso puntual N en $\mathbb{R}$ queda determinado al conocer $P\{N(A_i) = n_i \ ; i = 1, 2, \ldots n\}$ para toda familia finita $A_1, \ldots, A_n$ de conjuntos disjuntos de un semianillo de conjuntos acotados que genera a la σ-álgebra de Borel $\mathcal{B}$. Como la clase de conjuntos formada por los intervalos de la forma $(a, b]$ es un semianillo de conjuntos acotados que genera a la σ-álgebra de Borel $\mathcal{B}$ podemos decir que, un proceso puntual N en la recta queda determinado al conocer para todo n,

$$P\{N(a_i, b_i] = n_i; i = 1, 2, \ldots, n\} \tag{1.8q}$$

donde $a_i < b_i \leq a_{i+1}$ con $i = 1, \ldots, n$.

Un proceso de Poisson en $\mathbb{R}$ es un proceso puntual para el cual

$$P\{N(a_i, b_i] = n_i; i = 1, 2, \ldots, n\} = \prod_{i=i}^{n} \frac{(H(b_i) - H(a_i))^{n_i}}{n_i!} \exp\{-(H(b_i) - H(a_i))\} \tag{1.8r}$$

siendo $H(t)$ una función no negativa, no decreciente y continua a la derecha. A partir de la definición se deduce fácilmente que $M(t) = E[N(t)] = H(t)$. Por esta razón, la función $H(t)$ recibe el nombre de función de medias. La función $H(t)$ determina las propiedades del proceso de Poisson :

1. Si $H(t)$ está definida en un intervalo de la recta real entonces el proceso está definido en dicho intervalo.

2. Si $H(t)$ tiene asíntotas verticales entonces el proceso es explosivo, es decir, el proceso tiene un número infinito de puntos en tiempo finito.

3. Si $H(t)$ es tal que $\lim_{t\to\infty} H(t) < \infty$ entonces se puede demostar que el proceso es finito, es decir, el número de puntos del proceso es finito con probabilidad 1.

En este trabajo nos vamos a centrar en el proceso de Poisson definido en $\mathbb{R}^+$, no explosivo y no finito. Por esta razón, la función $H(t)$ va a estar definida en $[0, \infty)$ con $\lim_{t\to+\infty} H(t) = +\infty$.

Desde luego la función $H(t)$ puede ser continua, discontinua, derivable, etc. Estas propiedades afectan a las propiedades del proceso de Poisson:

1. El proceso de Poisson es simple si y sólo $H(t)$ es una función continua.

2. Si $H(t)$ es discontinua en t_0 entonces la variable aleatoria $N\{t_0\}$ tiene una distribución de Poisson de parámetro $J = H(t_0) - H(t_0-)$.

3. Si $H(t)$ es absolutamente continua entonces el proceso es simple y existe una función $H'(t) = m(t)$ la cual representa la razón instantánea de variación del número esperado de puntos con respecto al tiempo. Esta función $m(t)$ recibe el nombre de función de intensidad del proceso.

Aunque la función $H(t)$ pueda ser continua singular, en este trabajo no vamos a considerar estos casos. Por lo general, trataremos con funciones $H(t)$ de la forma

$$H(t) = \int_{-\infty}^{t} m(u)du. \tag{1.8s}$$

Si $m(u) = m$ se puede demostrar que el proceso de Poisson es simplemente estacionario (propiedad de incrementos estacionarios). En este caso el proceso recibe el nombre de proceso de Poisson homogéneo. Si $m(u)$ no es constante el proceso de Poisson es no estacionario y el proceso recibe el nombre de proceso de Poisson no homogéneo.

En una sección anterior hablábamos de la importancia de la función de intensidad condicional en la modelación de fenómenos puntuales. Al respecto se puede demostrar (ver Snyder y Miller [4] pág 51), que si un proceso puntual $N(t)$ definido en la semirrecta real positiva verifica las siguientes propiedades:

$$\begin{aligned} &P\{N(t+\delta) - N(t) = 1 \mid \mathcal{H}_t\} = m(t)\delta + o(\delta) \\ &P\{N(t+\delta) - N(t) > 1 \mid \mathcal{H}_t\} = o(\delta) \\ &N(0) = 0 \end{aligned} \tag{1.8t}$$

siendo $\mathcal{H}_t = \sigma\{N(s) : 0 \leq s \leq t\}$, el proceso puntual resultante es un proceso de Poisson con función de medias absolutamente continua, es decir, con función de medias de la forma (1.8s). El recíproco también es cierto. Es decir, la función de intensidad condicional de un proceso de Poisson con función de medias de la forma (1.8s) es $\lambda(t|\mathcal{H}_t) = m(t)$. Un aspecto importante de las probabilidades (1.8t) es que éstas no dependen de $\mathcal{H}_t$, es decir, la probabilidad de encontrar un punto en $(t, t+\delta]$ no depende de la abundancia o escasez de puntos justo antes del instante t, e incluso no depende de si ha habido un punto justo en t. Obsérvese que la segunda condición es suficiente para garantizar que el proceso es simple. A continuación relacionamos algunas propiedades importantes de proceso de Poisso:.

1. Sean $T_1, T_2, \ldots$ las sucesivas localizaciones de un proceso puntual $N(t)$, entonces, $N(t)$ es un proceso de Poisson homogéneo con función de intensidad constante $m(u) = \rho$ si y sólo si los tiempos entre localizaciones $T_1, T_2 - T_1, \ldots$ son variables aleatorias independientes e idénticamente distribuidas con distribución exponencial de parámetro ρ.

2. Sea M una función continua y no decreciente. $T_1, T_2, \ldots$ son las localizaciones de un proceso de Poisson no homogéneo $N(t)$ con $E[N(t)] = M(t)$ si y sólo si $M(T_1), M(T_2), \ldots$ son las localizaciones de un proceso de Poisson homogéneo con función de intensidad $m(u) = 1$.

3. Si para un proceso de Poisson simple sabemos que hay n localizaciones del proceso en el intervalo $[0, t_0]$ entonces, las localizaciones $T_1, T_2, \ldots, T_n$ son los estadísticos ordenados de una variable aleatoria con función de distribución $M(t)/M(t_0)$ con $t < t_0$.

Basándonos en esta última propiedad se puede realizar (entre otras cosas) un test estadístico, llamado en la literatura test de Laplace (ver [5] pág 86), para comprobar si las localizaciones $T_1, T_2, \ldots, T_n$ de un proceso puntual no son las localizaciones de un proceso de Poisson homogéneo. También debemos decir que las propiedades 2 y 3 nos proporcionan un instrumento para simular un proceso de Poisson simple tanto homogéneo como no homogéneo.

1.5.2 Proceso de renovación

La teoría de la renovación tuvo su origen en el estudio de problemas que podían plantearse como grupos autorrenovables (traducción española de "self-renewing aggregates"), esto es, en el estudio de problemas de poblaciones de organismos vivos u objetos de vida finita que son capaces de autorregenerarse de tal manera que puedan estabilizarse. Entre estos problemas se encuentran problemas de reemplazamiento industrial, problemas actuariales y problemas de poblaciones biológicas, cuya resolución final, es la solución de una determinada ecuación integral llamada ecuación de renovación. Una revisión de estos problemas y su solución puede encontrarse en Lotka [6]. En este apartado resumiremos las ideas y los resultados más importantes de la teoría de la renovación, empezando, como no, por su definición. Un estudio bastante completo de estos procesos pueden encontrarse en [7] y [8].

A partir del apartado (iii) del teorema (1) se puede definir, de forma muy

sencilla, un proceso puntual en la recta real positiva. Para ello basta considerar la sucesión de intervalos de la que allí se hablaba como una sucesión $\{X_i\}_{i=1}^{\infty}$ de variables aleatorias no negativas independientes con distribución común F no degenerada en el origen. El proceso puntual así definido recibe el nombre de proceso de renovación ordinario. A cada uno de las localizaciones T_n del proceso con $T_n = X_1 + \cdots + X_n$ $(n = 1, 2, \ldots)$ y $T_0 = 0$ las denominaremos instantes de renovación o simplemente renovaciones. Para este proceso son de interés, por lo general, las siguientes funciones o variables:

(i) Tiempo hasta la n-ésima renovación. Está claro que esta variable aleatoria es precisamente T_n.

(ii) Número de renovaciones $N(t)$ hasta un instante determinado t. Por supuesto, $N(0) = 0$, esto es, la renovación del instante T_0 no será contada como tal.

(iii) La función de renovación, es decir, la función $M(t) = E[N(t)]$.

(iv) La densidad de renovación, es decir $m(t) = M'(t)$. De acuerdo con (1.8b) y (1.8c) esta función existe cuando la distribución F es absolutamente continua. A igual que en el caso general que comentábamos en la sección 2, esta función nos da la razón instantánea de variación del número esperado de renovaciones con respecto al tiempo.

(v) Tiempo trancurrido desde la última renovación hasta un instante fijo t. Denotaremos a esta cantidad por U_t.

(vi) Tiempo trancurrido hasta la siguiente renovación desde un instante fijo t. Denotaremos a esta cantidad por V_t

En la práctica las variables aleatorias $X_1, X_2, \ldots$ puede interpretarse como los tiempos de vida de una máquina que tras fallar es reemplazada instantáneamente por otra exactamente igual. Además, por lo general, se empieza con una máquina

nueva en el instante $t = 0$, esto es, en el instante $t = 0$ comienza la observación del proceso con una localización en ese preciso instante. Sin embargo esto no siempre es así. Por ejemplo, en el instante $t = 0$ puede empezarse con una máquina ya usada. Para permitir otras condiciones iniciales, una variable aleatoria X_0 independiente de la sucesión $\{X_i\}_{i=1}^{\infty}$ y distribuida con una distribución F_1 se antepone al proceso de renovación $\{X_i\}_{i=1}^{\infty}$. A esta nueva sucesión, $X_0, X_1, X_2, \ldots$ se le denomina proceso de renovación general o proceso de renovación con retardo. Cuando la distribución de X_0 tiene distribución $F_1(x) = 1/E[X_1] \int_0^x (1 - F(t))dt$, el proceso de renovación con retardo es estacionario y además la variable aleatoria V_t (tiempo trancurrido hasta la siguiente renovación desde un instante fijo t) tiene la misma distribución F_1 independientemente de la posición del instante fijo t ([1], pág 72). Por otro lado, el proceso de renovación ordinario no es estacionario pero tiene, evidentemente, la propiedad de intervalos estacionarios.

Algunas propiedades:

1. El proceso de renovación ordinario es simple si y sólo si la distribución F es tal que $F(0) = 0$.

2. Para cualquier proceso de renovación (tanto ordinario como con retardo) $M(t) \to t/\mu$, $m(t) \to 1/\mu$, y $P(V_t < x) \to 1/\mu \int_0^x (1 - F(t))dt$ a medida que t tiende a ∞. Por tanto está claro que asintóticamente todo proceso de renovación es estacionario.

3. Si F tiene distribución exponencial de parámetro ρ, obtenemos un proceso de Poisson homogéneo.

También es de interés en esta memoria los procesos de renovación compuestos. Un proceso de renovación compuesto es un proceso acumulativo $\Pi(t) = \sum_{n=1}^{N(t)} Y_n$ donde $N(t)$ cuenta el número de renovaciones de un proceso de renovación hasta un instante determinado t. Además, la sucesión de vectores (T_n, Y_n) es independiente,

siendo T_n el instante de la n-ésima renovación del proceso de renovación. Las cantidades Y_n pueden representar cantidades (de masa, energía, costes, etc.) que se acumulan en cada instante de salto del proceso.

En las aplicaciones suele ser de interés el comportamiento asintótico del proceso $\Pi(t)$, el cual viene expresado en el siguiente teorema:

Teorema 5 *Si* $E[Y_1] < \infty$ *y* $E[T_1] < \infty$ *entonces se tiene:*

$$\frac{\Pi(t)}{t} \underset{t\to\infty}{\to} \frac{E[Y_1]}{E[T_1]} \ c.s. \qquad y \qquad \frac{E[\Pi(t)]}{t} \underset{t\to\infty}{\to} \frac{E[Y_1]}{E[T_1]}$$

Si cada vez que ocurre una renovación decimos que se completa un ciclo, este teorema establece que la cantidad esperada acumulada a lo largo de la historia del proceso es igual a la cantidad media acumulada durante un ciclo dividida por la duración media de cada ciclo.

1.6 Superposición de procesos de renovación

La frecuencia con la que aparece la distribución normal en estadística se explica por la multitud de fenómenos aleatorios donde subyace el efecto del teorema central del límite. En procesos puntuales el papel de la distribución normal lo ocupa el proceso de Poisson (homogéneo y no homogéneo) ya que este aparece a menudo como la superposición de múltiples fenómenos puntuales independientes. En esta sección revisamos los conceptos teóricos que justifican la presencia del proceso de Poisson (y por tanto de la distribución exponencial) tanto en teoría de colas como en fiabilidad.

La superposición de n procesos de renovación independientes es un proceso de renovación, únicamente cuando los n procesos son de Poisson. En este caso, la superposición es también un proceso de Poisson [9]. Sin embargo, se ha probado que la superposición de un número infinito de procesos de renovación independientes estacionarios, es un proceso de Poisson homogéneo [10].

Desde el punto de vista de la fiabilidad la hipótesis de que n es muy grande es razonable dado que muchos sistemas están compuestos de un número elevado de

componentes. Sin embargo es muy cuestionable que cada componente pueda estar en equilibrio, dado que muchos sistemas conservan componentes originales de fábrica durante toda su vida. En otras palabras es poco realista asumir que t es bastante grande. Por lo tanto, los resultados de Gligelionis [11] para tiempos finitos son bastante interesantes.

Sea $\{N_{nr}(t), 0 \leq t < \infty\}$ con $1 \leq r \leq n$ una familia de n procesos de renovación independientes tal que para cualquier valor t fijo, satisface las siguientes condiciones:

(a) $\lim_{n\to\infty} \sum_{r=1}^{n} P\{N_{nr}(t) \geq 2\} = 0$

(b) $\lim_{n\to\infty} \left(\max_{1\leq r\leq n} F_{nr}(t) \right) = 0$ donde $F_{nr}(t)$ es la distribución del primer fallo del dispositivo r-ésimo.

Sea $S_n(t)$ el proceso formado por la superposición de éstos n procesos, es decir, $S_n(t) = \sum_{r=1}^{n} N_{nr}(t)$. En este contexto se tiene el siguiente teorema.

Teorema 6 *$\{S_n(t), 0 \leq t < \infty\}$ converge en probabilidad a un proceso de Poisson no homogéneo con función de medias $M(t)$ si y sólo si para todo valor fijo $t > 0$, $\lim_{n\to\infty} \sum_{r=1}^{n} F_{nr}(t) = M(t)$.*

Nota 2 *Además si cada $F_{nr}(t)$ satisface la condición de que $0 < F'_{nr}(0) \equiv f_{nr}(0) < \infty$ entonces $\{S_n(t), 0 \leq t < \infty\}$ converge en probabilidad a un proceso de Poisson homogéneo para todo $t > 0$. Las distribuciones utilizadas con frecuencia en fiabilidad, esto es, la distribución de Weibull, la distribución log-normal y la distribución gamma no satisfacen esta condición. No obstante, en los sistemas industriales es de esperar que algunas partes fallen en cuanto el sistema se ponga en marcha, por lo que $f_X(0) \neq 0$.*

La condición necesaria (a) significa que no hay posibilidad de que ocurra dos o más fallos en cada componente. Es decir, queda excluida la posibilidad de que un segundo fallo se produzca en cada componente o dispositivo. En este caso la

superposición de los procesos de renovación es más bien un proceso de estadísticos ordenados. Si embargo, en este caso los fallos en cada dispositivo son independientes pero no tienen que ser indénticamente distribuidos. Sin duda, esta condición no es plausible para sistemas reales, dado que para valores de t suficientemente grandes es posible que ciertos dispositivos sean reemplazados dos o más veces. Esta condición parece más apropiada en fiabilidad del software donde un fallo adecuadamente reparado no vuelve a aparecer nunca más.

La función de razón de fallos de la distribución exponencial es constante. Esto quiere decir que si un equipo falla de acuerdo a esta distribución la propensión al fallo no cambia con el uso. Por esta razón, la distribución exponencial fue raramente considerada como distribución de fallos de un sistema. Sin embargo, por las razones expuestas anterioremente hay situaciones donde la distribución exponencial juega un papel prominente.

1.7 El problema general de regla de parada

Supongamos que vamos observando gradualmente una sucesión de variables aleatorias $y_1, y_2, \ldots$ y que debemos parar este proceso de observación en alguna etapa n. Si paramos en dicha etapa n, recibiremos una "recompensa" Z_n, que es función de los valores $y_1, \ldots, y_n$. El problema de regla de parada consiste en encontrar reglas de parada que maximice o minimicen nuestra "recompensa" esperada. Para definir formalmente este problema son necesarios los siguientes elementos:

(i) Un espacio de probabilidad $(\Omega, \mathcal{F}, P)$.

(ii) Una sucesión creciente $\mathcal{F}_n$ de sub-σ-álgebras de $\mathcal{F}$.

(iii) Una sucesión de variableas aleatorias $Z_1, Z_2, \ldots$, tal que Z_n medible respecto a $\mathcal{F}_n$, $n = 1, 2, \ldots$.

A partir de estos objetos hacemos las siguientes definiciones:

Definición 4 *Una regla de parada es una variable aleatoria* ξ *con valores* $1, 2, 3, \ldots, +\infty$, *tal que* $P(\xi < \infty) = 1$ *y el suceso* $\{\xi = n\} \in \mathcal{F}_n$.

Definición 5 *El par de sucesiones* $\{Z_n, \mathcal{F}_n\}_{n=1}^{\infty}$ *recibe el nombre de sucesión estocástica. Una sucesión estocástica se dice integrable si* $E[|Z_n|] < \infty$ $(n \geq 1)$.

A partir de estas definiciones, el problema de regla de parada puede plantearse de la siguiente manera: dada una sucesión estocástica integrable $\{Z_n, \mathcal{F}_n\}_{n=1}^{\infty}$ encontrar dentro del conjunto de reglas de parada Γ aquella regla ξ^* tal que $E[Z_{\xi^*}] = \sup E[Z_\xi] = V$ donde el supremo recorre el conjunto de regla de parada tal que $E[Z_\xi]$ exista, siendo

$$Z_\xi = \sum_{n=1}^{\infty} Z_n I_{\{\xi=n\}} = \begin{cases} Z_n, \{\xi = n\} \\ 0, \{\xi = \infty\} \end{cases}$$

Para este problema el interés se centra en responder a las siguientes preguntas:

- ¿Existe la regla de parada óptima ξ^* tal que $E[Z_{\xi^*}] = V$?¿Cómo es?
- ¿Cómo podemos calcular el valor V?

En algunas ocasiones, el problema de parada óptima se estudiará en un subconjunto propio D de Γ. Precisamente en el campo donde vamos aplicar estos conceptos, la fiabilidad, existen numerosos problemas de reemplazamiento que han sido resueltos para un subconjunto D concreto de reglas, las cuales, dependen por lo general de algún número finito de parámetros. A este subconjunto se le suele conocer en fiabilidad , y en otros contextos, con el nombre de política. Ahora vamos a comentar como debe entenderse cada uno de los elementos de un problema general de parada para los problemas de fiabilidad que quieren resolverse, a saber; los problemas de reemplazamiento de sistemas. En primer lugar, debe observarse, que el problema general de parada se ha planteado en tiempo discreto. Esto se debe precisamente a que en los problemas de reemplazamiento que estudiaremos, y por razones que

posteriomente se justificarán, el reemplazamiento del sistema se realizará en uno de los instantes de fallos a los que está sujeto el sistema. Es decir, después de haber observado n instantes de fallos $T_1, T_2, \ldots T_n$ habrá que decidir si se continúa con la producción o si por el contrario se realiza el reemplazamiento del sistema. Así, la sub-σ-álgebra $\mathcal{F}_n$ debe entederse como la σ–álgebra generada por las variables aleatorias $T_1, T_2, \ldots T_n$.

A partir del problema general de problema de regla de parada descrito anteriormente, surge otro problema muy interesante al considerar finita la sucesión estocástica $\{Z_n, \mathcal{F}_n\}_{n=1}^{m}$. En este caso, de no haber parado antes de observar T_m hay que parar obligatoriamente en la etapa m. A este problema se le conoce como problema de regla de parada con horizonte finito. Este problema puede verse también como un caso especial del problema de horizonte infinito, donde $Z_{m+1} = \cdots = Z_\infty = -\infty$. Este tipo de problemas pueden resolverse de forma óptima por el llamado método de inducción regresiva.

En general, los problemas de horizonte infinito no tienen una solución que pueda expresarse mediante una expresión matemática concreta y por tanto es necesario el estudio de métodos o reglas que permitan acercarse a la solución óptima. En principio, es posible encontrar soluciones aproximadas del problema al considerar una versión truncada (eligiendo un valor de m bastante grande) del problema original. Aparte de este método, llamado método de truncamiento, existen otros caminos que nos permiten encontrar otras soluciones aproximadas del problema. Tal es el caso de la regla de parada de las k etapas futuras; una regla simple que mejora el método de truncamiento. En la siguiente sección describimos esta regla y sus propiedades, para finalmete estudiar el llamado caso monótono de un problema de regla de parada.

1.7.1 Caso monótono

En un problema de regla de parada, la regla de las k etapas futuras (k-ef) es aquella regla que estando en la etapa n para o continúa si la regla óptima entre las etapas $n+1$ y $n+k$ (ambas inclusivas) para o continúa. Una propiedad importante de

estas reglas es que si en una determinada etapa n la regla k-*ef* recomienda continuar entonces la regla $(k-1)$-*ef* también recomienda continuar. Además, si la regla óptima existe, y si la k-*ef* recomienda continuar entoces la regla óptima también recomienda continuar. Esta propiedad sugiere una simplificación de la k-*ef* : aplicar la regla 1-*ef* hasta que pare, después aplicar la regla 2-*ef* hasta que pare, y así sucesivamente hasta llegar a la regla k-*ef*.

En este contexto resulta interesante saber cuanto de próximo está la regla óptima de la más simple de las reglas k-*ef*, la 1-*ef* (llamada tanbién regla miope), la cual puede expresarse de esta otra manera,

$$N_1 = \{n \geq 0 : Z_n \geq E(Z_{n+1}|Z_1, \ldots, Z_n)\} \tag{1.9}$$

Así, la regla miope es aquella que para en el primer n donde la ganancia obtenida por parar es al menos tan grande como la ganancia esperada al parar en la siguiente etapa.

Definición 6 *Denotamos por* $A_n = \{Z_n \geq E(Z_{n+1}|\mathcal{F}_n)\}$. *Diremos que un problema de regla de parada es monótono si,*

$$A_0 \subset A_1 \subset A_2 \subset \ldots \quad c.s.$$

Teorema 7 *Si* $E(\sup_n Z_n) < \infty$, $\lim_{n\to\infty} Z_n = Z_\infty$ *c.s. y las variables* $H_n = \sup_{j\geq n}\{Z_j - Z_n\}$ *son uniformemente integrables entoces el problema de regla de parada monótono tiene por regla óptima a la regla miope.*

Para un estudio completo en teoría de paradas óptimas puede consultarse [12].

1.8 El problema de parada y el reemplazamiento de sistemas reparables

En la industria es habitual que una máquina deba ser reemplazada por otra de características similares. Debido a que esta operación de reemplazo se realiza reiteradamente a lo largo del tiempo, es conveniente reemplazar la máquina de tal

manera que consigamos el mayor beneficio posible. Como objetivo, parece apropiado maximizar el beneficio promedio por unidad de tiempo. Esto es, al repetirse el problema de reemplazamiento un número n de veces de forma independiente y de acuerdo con la misma regla de parada ξ, conseguiríamos $\xi_1, ..., \xi_n$ reglas de parada independientes e idénticamente distribuidas y $Z_{\xi_1}, ..., Z_{\xi_n}$ beneficios, independientes e idénticamente distribuidos. De esta manera el beneficio total conseguido en el tiempo total $T_{\xi_1} + ... + T_{\xi_n}$ es $Z_{\xi_1} + ... + Z_{\xi_n}$, por lo que el beneficio promedio por unidad de tiempo es el cociente $(Z_{\xi_1} + ... + Z_{\xi_n})/(T_{\xi_1} + ... + T_{\xi_n})$. Si tanto numerador como denominador se divide entre n entonces, por la ley de los grandes números, el cociente anterior converge a $E[Z_\xi]/E[T_\xi]$, siempre y cuando las esperanzas correspondientes existan. Por lo tanto, se trata de elegir una regla de parada ξ que maximice el cociente $E[Z_\xi]/E[T_\xi]$, siendo Z_ξ el beneficio conseguido hasta el momento T_ξ (momento en el cual reemplazamos la máquina actual por la nueva). Con esta función objetivo, el problema de parada puede replantearse de la siguiente manera: dada una sucesión estocástica integrable $\{Z_n, \mathcal{F}_n\}_{n=1}^{\infty}$ encontrar dentro del conjunto de reglas de parada Γ aquella regla ξ^* tal que $E[Z_{\xi^*}]/E[T_{\xi^*}] = \sup E[Z_\xi]/E[T_\xi] = V$ donde el supremo ahora recorre el conjunto de regla de parada tal que $E[Z_\xi] < \infty$ y $\xi \geq 1$ (al menos observaremos un fallo). A primera vista parece que se trata de otro problema diferente al problema clásico de parada óptima antes definido, pero en realidad, es una generalización de este problema ya que permite que los tiempos entre etapas sean diferentes y no, unitarios y fijos como el problema clásico. No obstante, en los problemas de fiabilidad que resolveremos es preferible trabajar con una formulación clásica del problema de parada. En resumidas cuentas lo que queremos es eliminar el denominador en nuestra función objetivo $E[Z_\xi]/E[T_\xi]$. Para ello haremos lo siguiente:

Como $E(Z_\xi) \leq VE(T_\xi)$, para toda $\xi \in \Gamma$, podemos decir que ξ^* es óptima si

$$E(Z_{\xi^*} - VT_{\xi^*}) = \sup_{\xi \in \Gamma} E(Z_\xi - VT_\xi) \tag{1.11}$$

Con otras palabras, la sucesión de variables aleatorias $Z_1, Z_2, \ldots,$ que aparecen en el problema clásico de parada es ahora de la forma $Z_1' = Z_1 - VT_1, Z_2' =$

$Z_2 - VT_2, \ldots,$. La equivalencia entre estas dos formas de ver el problema queda confirmada con el siguiente teorema.

Teorema 8 *a) Si para algún* $\gamma, \sup_{\xi \in \Gamma} E(Z_\xi - \gamma T_\xi) = 0$, *entonces* $\sup_{\xi \in \Gamma} E[Z_\xi]/E[T_\xi] = \gamma$. *Además, si* $\sup_{\xi \in \Gamma} E(Z_\xi - \gamma T_\xi) = 0$ *se alcanza en* $\xi^* \in \Gamma$, *entonces* ξ^* *es un regla óptima para maximizar* $E[Z_\xi]/E[T_\xi]$.

(b) Recíprocamente, si $\sup_{\xi \in \Gamma} E[Z_\xi]/E[T_\xi] = \gamma$ *y el supremo se alcanza en* $\xi^* \in \Gamma$, *entonces* $\sup_{\xi \in \Gamma} E(Z_\xi - \gamma T_\xi) = 0$ *y el supremo se alcanza con la regla* ξ^*.

En caso de encontrar la regla óptima ξ^*, la expresión matemática de la regla óptima va a depender del valor óptimo V, es decir, $\xi^* = \xi^*(V)$. De esta manera para resolver completamente el problema es necesario calcular dicho valor V. Ahora bien, si consideramos el valor V no como un valor fijo sino como un valor variable, llamémoslo γ, tendremos una familia de reglas de paradas $\xi(\gamma)$ con $\gamma \in \mathbb{R}$. Así, para cada valor de γ tendremos una regla $\xi(\gamma)$ cuyo beneficio promedio a largo plazo es $V(\gamma) = E(Z_{\xi(\gamma)} - \gamma T_{\xi(\gamma)})$. En general, $\lim_{\gamma \to -\infty} V(\gamma) = \infty$ y $\lim_{\gamma \to \infty} V(\gamma) = -\infty$.y de acuerdo con el teorema anterior hay que buscar el valor γ tal que $V(\gamma) = 0$. Además, según resultados de teoría de paradas óptimas se tiene el siguiente teorema.

Teorema 9 $V(\gamma)$ *es decreciente y convexa*

A partir de este resultado podemos describir un simple método iterativo para calcular el valor γ tal que $V(\gamma) = 0$ y la regla de parada óptima. Se trata del método de Newton, el cual, converge cuadráticamente. Sea γ_0 un valor inicial para el valor óptimo. En el punto γ_0, la recta $y = V(\gamma_0) - E[T_{\xi(\gamma_0)}](\gamma - \gamma_0)$ es un hiperplano soporte, ya que $V(\gamma_0) - E[T_{\xi(\gamma_0)}](\gamma - \gamma_0) = E[T_{\xi(\gamma_0)}] - \lambda V'(\gamma_n) \leq V(\gamma)$. De acuerdo con el método de Newton, los sucesivos valores de γ_n , $n = 0, 1,...$ se calculan mediante la recurrencia,

$$\gamma_{n+1} = \gamma_n + V(\gamma_n)/V'(\gamma_n). \tag{1.12}$$

Y como $V'(\gamma_n) = -E[T_{\xi(\gamma_n)}]$ se tiene que,

$$\gamma_{n+1} = \gamma_n - \frac{V(\gamma_n)}{E[T_{\xi(\gamma_n)}]} = \frac{E[Z_{\xi(\gamma_n)}]}{E[T_{\xi(\gamma_n)}]}. \tag{1.13}$$

Por tanto, el valor V es el punto fijo de la función,

$$U(\gamma) = \frac{E[Z_{\xi(\gamma_n)}]}{E[T_{\xi(\gamma_n)}]}. \tag{1.14}$$

Capítulo 2

MODELOS DE SISTEMAS REPARABLES

¿Qué es un sistema reparable? ¿Qué es una reparación mínima? ¿ Cuáles son los modelos existentes en la actualidad para sistemas reparables ? En el caso de sistema reparables reparados mínimamente y teniendo en cuenta costes y beneficios, ¿Cuál es el momento idóneo de reemplazamiento? Estas y otras cuestiones tendrán respuesta a lo largo de este capítulo.

2.1 Introducción

En la actualidad, el rendimiento de la industria y de cualquier persona en su actividad diaria depende, por lo general, de la disponibilidad de numerosos aparatos y/o máquinas*. Desde luego, la disponibilidad de la mayoría de estos sistemas depende, en buena medida, de la acciones de mantenimiento preventivo y/o correctivo que se lleven a cabo. El mantenimiento correctivo, también llamado reparación, tiene por objeto restaurar las funciones propias del sistema cuando estas han cesado por alguna rotura. El mantenimiento preventivo, como su nombre indica, se realiza para prevenir posibles fallos y para reducir al máximo los procesos de desgaste propios del funcionamiento del sistema. A aquellos sistemas a los que únicamente se mantiene correctivamente se les denomina sistemas reparables.

De acuerdo con Ascher y Feingold [13], un sistema reparable puede ser definido

*En adelante, para referirnos genéricamente a todas esas herramientas utilizaremos el término, sistema industrial, o simplemente, sistema

como una colección de dos o más dispositivos que después de perder una o varias de sus funciones puede recuperarlas satisfactoriamente a través de cualquier forma distinta al reemplazamiento total del sistema . En esta memoria nos dedicaremos al estudio de sistemas reparables que recuperan inmediatamente la operatividad tras perderla. Esta suposición teórica se debe a que en la práctica el tiempo empleado en realizar una reparación al sistema es insignificante en comparación al tiempo en funcionamiento. Los modelos probabilísticos más apropiados para estudiar la disposición de los fallos de un sistema reparable a lo largo del tiempo son los procesos puntuales.

Después de una reparación es normal que el propietario de un sistema quiera conocer su estado. Entre otras cosas desearía saber si la propensión al fallo del sistema va a ser la propia de un sistema de su edad que nunca ha fallado o como la de un sistema nuevo. A las reparaciones que dejan al sistema en una de esas situaciones extremas se les conoce como reparaciones mínimas y reparaciones máximas, respectivamente. En la práctica, por lo general, la reparaciones del sistema serán algo intermedio entre una reparación mínima y una reparación máxima. Fijar el tipo o los tipos de reparaciones posibles que se van a realizar en un sistema es fijar un modelo para el sistema reparable en cuestión. En las tres secciones siguientes vamos a revisar los modelos existentes hasta la fecha en la literatura para modelos reparables.

Otra cuestión importante para el propietario de un sistema es saber, de acuedo con algún criterio económico, cuando es el momento idóneo para reemplazarlo por uno igual o de mejor calidad. En este caso es importante conocer los parámetros económicos que influyen en el sistema: costes de las distintas reparaciones, coste de reemplazamiento, los beneficios de la producción, etc. A un modelo de sistema reparable al cual se le añade una estructura de costes y beneficios al objeto de hallar el instante óptimo de reemplazamineto del sistema por otro exactamente igual se le denomina modelo de sistema reparable con costes. En la última sección de este capítulo haremos una revisión de los distintos modelos de sistemas reparables con

costes.

2.2 Reparaciones mínimas y perfectas

El concepto de reparación mínima fue introducido por Barlow y Proschan [14] para mimetizar el comportamiento de un sistema complejo cuando una de sus múltiples componentes es sustituida o reparada tras un fallo. Un ejemplo ilustrativo es la reparación del pinchazo de un neumático de un coche. En términos matemáticos Barlow y Proschan definen el concepto de la siguiente manera: si un sistema (con función de distribución de fallos $F(x)$ absolutamente continua y función de razón de fallos $r(x)$) falla en el instante t entonces, después de ser reparado mínimamente, la probabilidad de que falle después de $s+t$ unidades de tiempo es $1-F(t+s)/(1-F(t)$ y su razón de fallos es $r(s+t)$. De acuerdo con esta definición el proceso de fallos de un sistema reparable sujeto a reparaciones mínimas tiene función de intensidad condicional $\lambda(t|\mathcal{H}_t) = r(t)$; y es por tanto un proceso de Poisson no homogéneo con función de intensidad $r(t)$.

Sin embargo, si el sistema es muy simple (como por ejemplo un sistema en paralelo con dos componentes c_1 y c_2, ambas con distribución de vida exponencial de parámetro 1) la reparación mínima física de un fallo del sistema (en el sentido de Barlow y Proschan de abrir el sistema y sustituir o reparar la pieza que causó el fallo del mismo) está lejos de corresponderse con el significado matemático de reparación mínima. Si el sistema en paralelo anterior falla en un instante t es porque las dos componetes han fallado y como ambas no pueden fallar simultáneamente, la función de razón de fallos después de t es constante e igual a la unidad en cualquiera de los casos. Sin embargo, la función de razón de fallos después de t, cuando se realiza una reparación mínima en el sentido matemático es $r(t) = 2(1-e^{-t})/(2-e^{-t})$. Por esta razón Bergman [15] denominó a la reparación mínima en el sentido matemático reparación mínima estadística y a la reparación mínima en el sentido físico de Barlow y Proschan reparación mínima física. Desde un punto de vista estríctamente

práctico la reparación mínima estadística se realiza sustituyendo el sistema por otro de características similares con la misma edad y que no haya fallado desde su puesta en funcionamiento. Para mayor información acerca del concepto de reparación mínima y sus matices puede consultarse [16]. En esta memoria cuando nos refiramos a reparación mínima entenderemos que se trata de la reparación mínima estadística y por tanto el proceso puntual suyacente será un proceso de Poisson no homogéneo.

El concepto opuesto a repación mínima es el concepto de reparación máxima o perfecta. Después de cada fallo, la propensión al fallo de un sistema reparado máximamente es igual al momento en que este fue puesto en funcionamiento. El proceso de fallos para un sistema reparado de esta manera es un proceso de renovación. De acuerdo (1.8p) la función de intensidad condicional de este proceso es $\lambda(t|\mathcal{H}_t) = r(t - T_{N(t)})$ donde $r(t)$ es la correspondiente función de razón de fallos correspondiente a la distribución entre fallos F, y $t - T_{N(t)}$ es el tiempo transcurrido desde el último fallo anterior a t.

A la hora de analizar un conjunto de datos de sistemas reparables, es muy habitual realizar realizar los siguientes pasos. En primer lugar se aplica un test de tendencia (saber si los fallos son más frecuentes o mas escasos con el tiempo) a los tiempos entre fallos X_i. Si no hay significación estadística para justificar la existencia de tendencia, el proceso de fallos puede ser modelado como un proceso de renovación y de esta manera los datos puede ser analizados utilizando los procedimientos estadísticos para analizar observaciones independientes e idénticamente distribuidas. En caso contrario, el proceso de fallos se puede modelar como un proceso de Poisson no homogéneo, fijando la función de intensidad $\lambda(t)$ ($\lambda(t) \neq \lambda \forall t$) que mejor se ajusta a la tendencia que presenta los datos. Por ejemplo, a un sistema que deteriora (mejora) con el tiempo le corresponde una función de intensidad $\lambda(t)$ creciente (decreciente).

2.3 Reparaciones imperfectas

En la práctica, el estado en el que queda un sistema después de ser reparado es, por lo general, un estado intermedio entre los extremos: reparación perfecta o máxima y reparación mínima. Un modelo de sistema reparable cuya calidad de reparación oscile de acuerdo con cierto criterio entre estas dos reparaciones extremas, ambas incluidas, se le denomina un modelo de reparaciones imperfectas. Uno de los modelos pioneros en este sentido fue el modelo de reparaciones imperfectas de Brow y Proschan [17]. En este modelo, la reparación es mínima con probabilidad p y es una reparación perfecta con probabilidad $1-p$. Posteriormente, Block y otros [18], extendieron el modelo de Brown y Proschan, permitiendo que la probabilidad p anterior dependiese del tiempo. En un trabajo reciente, Dorado y otros [19], sugirieron un modelo general de reparaciones el cual contiene un buen número de modelos de reparaciones imperfectas. La función de intensidad completa del modelo viene dada por,

$$\lambda(t|\mathcal{H}_t) = \theta_{N(t)+1} \cdot r\left(a_{N(t)+1} + \theta_{N(t)+1}(t - T_{N(t)})\right) \tag{2.1}$$

donde $\{\theta_i\}_{i\geq 1}$ y $\{A_i\}_{i\geq 1}$ $(\theta_1 = 1, A_1 = 0)$ son dos sucesiones de variables aleatorias con dominio en $[0,1]$ y $[0,\infty)$ respectivamente que reciben de vida suplementaria y edad efectiva. Además, $A_i \leq A_{i-1} + \theta_{i-1}X_{i-1}$ con $i > 1$. La función $r(t)$ es la función de razón de fallos de la variable primer tiempo de fallo. A partir de (2.1) se deduce que la distribución condicional de los tiempos entre fallos es,

$$F_{X_n}(t|A_i, \theta_i, X_1, ..., X_n) = 1 - \overline{F}(\theta_i t + A_i)/\overline{F}(A_i) \tag{2.2}$$

donde $\overline{F}$ denota la función de supervivencia de la variable primer tiempo de fallo. Es fácil comprobar que para si para todo i, $\theta_i = 1$ y $A_i = 0$ tenemos el modelo de reparaciones perfectas y que para si para todo $i, \theta_i = 1$ y $A_i = T_{N(t)}$ tenemos el modelo de reparaciones mínimas. También son casos particulares del modelo de Dorado los modelos de Kijima. Para conseguir el modelo tipo I de Kijima para todo i, $\theta_i = 1$ y

$A_i = \sum_{j=1}^{i-1} D_j X_j$ donde las D_j son variables aleatorias independientes e idénticamente distribuidas. En el caso de que las variables D_j sean determinísticamente igual a $1-\rho$, el modelo tipo I de Kijima coincide con el modelo de Kijima y otros cuya función de intensidad condicional es,

$$\lambda(t|\mathcal{H}_t) = r(t - \rho T_{N(t)}). \tag{2.3}$$

Para conseguir el modelo tipo II de Kijima para todo i, $\theta_i = 1$ y $A_i = \sum_{j=1}^{i-1} \prod_{k=j}^{i-1} D_k X_j$. A su vez, si en este último caso las variables aleatorias D_j tienen distribución Bernoulli entonces, el modelo tipo II de Kijima resulta ser el modelo de Brown y Proschan.

Otro modelo interesante de reparaciones imperfectas es el modelo de reducción de la edad con memoria m propuesto recientemente por Doyen y Gaudoin [20] el cual tiene por función de intensidad condicional,

$$\lambda(t|\mathcal{H}_t) = r\left(t - \rho \sum_{j=0}^{a} (1-\rho)^j T_{N(t)-j}\right) \tag{2.4}$$

donde $a = \min(m-1, N(t))$. Si las variables D_j de los dos modelos de Kijima son determinísticamente igual a $1-\rho$ resulta que el modelo de Doyen y Gaudoin incluyen a los modelos de Kijima.

Finalmente si en el modelo de Dorado la sucesión $\{\theta_i\}_{i\geq 1}$ es independiente de sucesión de instantes de fallos $\{T_i\}_{i\geq 1}$ y con $A_i = 0$ el proceso de fallos resulta ser un proceso de cuasi-renovación: tiempos entre fallos independientes pero no idénticamente distribuidos. Si particularmente, $\theta_i = 1/\alpha^{i-1}$ tenemos el modelo propuesto por Wang y Pham [21] en 1996.

2.4 Otros modelos de reparaciones imperfectas

Aparte de los modelos de reparaciones imperfectas anteriormente estudiados existen una diversidad de procesos puntuales estudiados en su mayoría por investigadores de escandinavia que pueden ajustarse, por supuesto, al patrón de

fallos del algún sistema reparable. De forma somera vamos a revisar los modelos más importantes hasta la fecha.

2.4.1 Modelos lineales generalizados y modelos loglineales

En el año 1992, Berman y Turner [22] consideraron la estimación del modelo paramétrico general,

$$\lambda(t|\mathcal{H}_t) = g\{\sum_{k=1}^{p} \alpha_k Q_k(t|\mathcal{H}_t)\} \tag{2.5}$$

donde g es una conocida función continua monótona y $Q_k(t|\mathcal{H}_t)$ son funciones conocidas dependientes de t y de la historia $\mathcal{H}_t$. Los autores analizaron datos completos de tiempos de fallo y mostraron como resolver el problema utilizando el software existente en la actualidad para resolver los modelos lineales generalizados. El caso especial, $g(y) = e^y$ fue aplicado por Lawless y Thiagarajah [23] a los sistemas reparables. En particular, consideraron el modelo,

$$\lambda(t|\mathcal{H}_t) = e^{\alpha+\beta g_1(t)+\gamma g_2(t-T_{N(t)})} \tag{2.6}$$

donde α, β y γ son parámetros desconocidas y g_1, g_2 son funciones conocidas. Este modelo, es una caso particular de proceso de renovación modulado.

2.4.2 Proceso de renovación modulados

En el año 1972 Cox [24] introdujo una clase de procesos puntuales, los procesos de renovación modulados, al perturbar la función de intensidad condicional de un proceso de renovación por un factor dependiente de cantidades que se creen que influyen en las probabilidades de ocurrencia del fenómeno en cuestión, estando cada cantidad combinada con un desconocido regresor. En términos matemáticos,

$$\lambda(t|\mathcal{H}_t) = e^{\sum_{i=1}^{p} \beta_i z_i(t)} r(t - T_{N(t)}). \tag{2.7}$$

Cox sugirió que la estimación de los parámetros del modelo debe hacerse a través de la función de verosimilitud parcial, de igual manera que en el modelo de regresión de Cox del análisis de supervivencia.

2.4.3 Procesos de renovación con tendencia

Un aspecto interesante de los procesos de revovación modulados y del modelo presentado por Lawless y Thiagarajah [23] es que contienen como caso particular a los procesos de renovación y a los procesos de Poisson no homogéneos. Así, realizando los oportunos test estadísticos de razón de verosimilitudes a los parámetros del modelo es posible decidir si los datos del sistema reparable que analizamos se ajustan a un modelo de reparaciones mínimas o a un modelo de reparaciones perfectas. A continuación presentaremos un modelo alternativo el cual a primera vista parece similar al modelo (2.6) pero que en realidad no está considerado en la fórmula general (2.5).

El proceso de renovación con tendencia introducido por Lindqvist [25] generaliza la siguiente propiedad del proceso de Poisson no homogéneo: sea $M(t)$ una función continua y no decreciente. $T_1, T_2, \ldots$ son las localizaciones de un proceso de Poisson no homogéneo $N(t)$ con $E[N(t)] = M(t)$ si y sólo si $M(T_1), M(T_2), \ldots$ son las localizaciones de un proceso de Poisson homogéneo con función de intensidad $m(u) = 1$. A continuación damos una definición formal de proceso de renovación con tendencia.

Definición 7 *Sea $m(t)$ una función no negativa definida para todo $t \geq 0$, $M(t) = \int_0^t m(u)du$ y F una distribución de probabilidad positiva con esperanza unitaria. Un proceso puntual simple $T_1, T_2, \ldots$ es un proceso de renovación con tendencia $PRT(F, \lambda(t))$ si el proceso transformado $M(T_1), M(T_2), \ldots$ es un proceso de renovación ordinario cuya secuencias entre localizaciones tiene distribución F.*

A partir de la definición se deduce que la función de intensidad condicional del proceso de renovación con tendencia es,

$$\lambda(t|\mathcal{H}_t) = r(M(t) - M(T_{N(t)}))m(t) \tag{2.8}$$

donde $r(t)$ es la función de riesgos de la función de distribución F. El proceso de renovación ordinario con función de distribución de tiempos entre llegadas $G(t) =$

$F(\lambda t)$ (siendo λ el inverso del valor esperado de los tiempos entre llegadas) coincide con un $PRT(F, \lambda)$.

2.4.4 Proceso gamma no homogéneo

El proceso gamma no homogéneo desarrollado por Berman [26] es una caso especial de proceso de renovación con tendencia $PRT(G, \lambda(\cdot))$, con G con distribución gamma con parámetro de escala 1. En su artículo Berman propuso el siguiente modelo con función de intensidad condicional,

$$\lambda(t) = \rho \exp\{\sum_{i=1}^{p} \beta_i z_i(t)\} \tag{2.9}$$

donde ρ y los β_i son parámetros desconocidos y las $z_i(\cdot)$ son funciones conocidas. De particular interés es el modelo de tendencias cíclicas y frecuencia fija, para el cual,

$$\lambda(t) = \rho \exp\{\beta_1 \cos(\omega t) + \beta_2 \sin(\omega t)\} \tag{2.10}$$

donde w es fijo. El contraste de hipótesis nula $\beta_1 = \beta_2 = 0$ equivale a un constraste de hipótesis nula "no hay tendencia cíclica" y también equivale al contraste de hipótesis nula "proceso de renovación con tiempos entre fallos con distribución gamma".

2.5 Modelos de sistemas reparables con costes

En esta sección nos vamos a centrar en el modelo de reparaciones mínimas. En la práctica del mantenimiento industrial, una buena parte de los sistemas complejos con varias componentes son considerados como una simple unidad. Sin embargo, la operatividad de un sistema complejo depende de las componentes individuales. Así, cuando una componente de un sistema complejo falla, éste habitualmente se refleja en el sistema. Ante un fallo del sistema es conveniente determinar si reemplazar el sistema es más económico que reparar o reemplazar la componente dañada. Como la razón de fallos, por lo general, incrementa con la edad en la mayor parte de los sistemas, el gasto en mantener operativo al sistema mediante reparaciones mínimas se vuelve muy costoso. En estas circunstancias, es interesante saber cuándo es óptimo

reemplazar el sistema en vez de seguir realizando reparaciones mínimas. En otras palabras, el reemplazamiento de un sistema reparable encargado de la producción initerrumpida de un determinado bien debe realizarse en algún momento del tiempo, bien por sus costes de mantenimiento o bien por un mal desempeño de sus funciones. Si asumimos que el sistema se reemplaza por otro de naturaleza idéntica, que los fallos del sistema son reparados mínimamente, y que esta situación se repite indefinidamente en el tiempo podemos plantearnos el siguiente problema: encontrar la regla de parada ξ que minimice la función objetivo

$$C(\xi) = E[Z_\xi]/E[T_\xi] \tag{2.11}$$

siendo T_ξ la variable aleatoria que mide el tiempo entre dos reemplazamientos consecutivos (un ciclo) del sistema y Z_ξ la variable aleatoria que mide el coste de mantemiento de un ciclo. Esta última variable puede ser a veces una recompensa más que un coste. En ese caso habría que maximizar en (2.11), en vez de minimizar. La función objetivo (2.11) tiene su fundamento en el teorema 5 del primer capítulo (véase también el primer párrafo de la última sección del mismo capítulo). El conjunto de reglas de parada de una determinada manera, y dependiente de un parámetro (o varios parámetros) recibe el nombre de política. Muchas veces y debido a la dificultad en encontrar la regla óptima el problema anterior se restringe a un subconjunto de reglas posibles (una política, por ejemplo).

Los modelos de reparaciones mínimas con costes consideran generalmente las siguientes hipótesis:

1. La función razón de fallos $r(t)$ del sistema subyacente es creciente.

2. El coste de las reparaciones mínimas c_m es menor que el coste de reemplazamiento del sistema c_r .

3. Los fallos del sistemas son detectados y reparados mínimamente de forma inmediata

2.5.1 Políticas básicas de reemplazamiento

En los últimos años, los investigadores han prestado especial atención al problema de la edad óptima de reemplazamiento de sistemas complejos sujeto a reparaciones mínimas. El primer trabajo en este sentido se remonta al año 1960 cuando Barlow y Hunter [27] trataron por primera vez el problema, usando un modelo de reemplazamiento periódico con reparaciones mínimas. El objetivo de este modelo de reparaciones mínimas era encontrar un edad de reemplazamiento t^* que minimizara el coste esperado de reemplazamientos y reparaciones mínimas a largo plazo. La política estudiada fue:

POLÍTICA 1: Los fallos que ocurran antes de t (fijo) son reparados mínimamente. Cuando el sistema alcance la edad t se realiza el reemplazamiento.

El coste de mantenimiento esperado por unidad de tiempo a largo plazo es:

$$C_{p_1}(t) = \frac{c_r + c_m E[N(t)]}{t} \tag{2.12}$$

donde $N(t)$ representa el número de fallos (reparaciones mínimas) en el periodo $(0, t]$.

La edad óptima de reemplazamiento t^* es el valor t que satisface la ecuación

$$t \cdot r(t) - \int_0^t r(u)du = \frac{c_r}{c_m} \tag{2.13}$$

Un solución única τ^* existe si $r(\infty) = \infty$. En este caso el coste de mantenimiento esperado por unidad de tiempo a largo plazo es $C_{p_1}(\tau^*) = c_m r(\tau^*)$

Este modelo básico de reparaciones mínimas ha sido generalizado y modificado por muchos autores con el objeto de ajustarse a situaciones más reales. Tilquin y Cléroux [28] añadieron a los costes del sistema, un coste de ajuste $c_a(ik)$ en la edad ik, $i = 1, 2, 3, \ldots$ y $k > 0$, e investigaron una política óptima de reemplazamiento. Respecto a los valores de $c_a(ik)$, ellos suponen que $c_a(0) = 0$ y que $c_a(s) \geq 0, \forall s = ik$.

Tilquin y Cléroux argumenta que su modelo es más verosímil que el modelo básico, pues los costes de ajuste $c_a(ik)$ pueden usarse para reflejar coste de depreciación, intereses bancarios, costes de ajuste, etc. Al igual que en el modelo básico, el problema es encontrar la edad óptima t^* de reemplazamiento del sistema que minimice el coste esperado por unidad de tiempo a largo plazo, que este caso viene dado por,

$$C_{p_1}(t) = \frac{c_m E[N(t)] + c_r + c_a^*(v(t))}{t}, \tag{2.14}$$

donde $c_a^*(v(t)) = \sum_{i=0}^{v(t)} c_a(ik)$ y $v(t)$ representa el número de ajustes realizados en el periodo $(0, t]$. Tilquin y Cléroux probaron que el mínimo global de la ecuación (2.14) existe en el intervalo $[0, \infty)$.

En otro modelo de reparaciones mínimas Muth [29] estudió una política que presentaba un menor coste esperado a largo plazo que el modelo estándar.

POLÍTICA 2: Todos los fallos son reparados mínimamente. Sin embargo, el sistema es reemplazado en el primer fallo después de una edad t.

A diferencia del modelo estándar el tiempo de reemplazamiento ya no es fijo. En este caso es un tiempo aleatorio superior a un valor fijo t. La función de costes a minimizar en el modelo de Muth para obtener t es:

$$C_{p_2}(t) = \frac{c_m N(t) + c_r}{t + q(t)} \tag{2.15}$$

donde $q(t) = E[\tau - t \mid \tau > t]$ es la función de vida media residual del sistema en el instante t. En la ecuación (2.15) el valor óptimo t^* puede hallarse con simples procedimientos de cálculo infinitesimal. Sin embargo, en los casos en que dichas técnicas no son aplicables Muth no dió algotitmo alguno para hallar t^*.

POLÍTICA 3: Seguir la política 2. Sin embargo, si no ha habido reemplazamiento en el intervalo $[\tau, T]$ con $\tau < T$, entonces se realiza un reemplazamiento preventivo en el instante T.

Makabe y Morimura [30] modificaron por completo el concepto de tiempo de reemplazamiento bajo reparaciones mínimas introducido por Barlow and Hunter [27]. Posteriomente, el modelo fue estudiado por Park [31] para el caso Weibull. En este modelo se estudió la siguiente política:

POLÍTICA 4 : Reparar mínimamente los primeros $n-1$ fallos y reemplazar el sistema en el siguiente fallo.

El coste de mantenimiento esperado por unidad de tiempo a largo plazo es,

$$C_{p_4}(n) = \frac{c_r + (n-1)c_m}{E(T_n)} \tag{2.16}$$

donde el valor T_n es el instante en que ocurre el n-ésimo fallo. El valor óptimo n^* es el entero más pequeño que satisface la desigualdad:

$$E(T_n) - (n - 1 + \frac{c_r}{c_m})E(T_{n+1} - T_n) \geq 0 \tag{2.17}$$

Para una de las distribuciones de fallo más habituales en fiabilidad, la distribución de Weibull (función de razón de fallos $r(t) = \rho t^k, t > 0, k > 1$), Park resolvió el modelo y encontró una expresión explícita para el cálculo del valor óptimo n^*, a saber,

$$n^* = \left\lfloor \frac{1}{k-1}\left(\frac{c_r}{c_m} - 1\right)\right\rfloor + 1 \tag{2.18}$$

Park comparó su nueva solución con la solución tradicional basada en edades óptimas propuesta por Barlow y Hunter. Su política da a largo plazo mejores rendimientos que la política de Barlow y Hunter. Además, el número de fallos antes del reemplazamiento es menor en la política de conteo de Park que en la política de tiempos de Barlow y Hunter. Sin embargo, todos estos resultados fueron expuestos numéricamente para la distribución de Weibull y no se justificó matemáticamente.

Phelps [32] comparó los modelos de reparaciones mínimas estudiados por Barlow y Hunter [27], Muth [29], y Park [31] suponiendo que la razón de fallos

es creciente. Phelps demuestra que la política de Muth (reemplazar en el primer fallo después de un tiempo t^*) es la mejor de las tres políticas y además pone en evidencia que la política de Muth es mejor que la política de Barlow y Hunter. En un artículo posterior, Phelps [33] generalizó su trabajo usando los resultados de procesos de decisión semi-Markovianos para probar que la política de Muth es óptima en el conjunto de todas las políticas de reemplazamientos posibles para modelos de reparaciones mínimas con razón de fallos creciente.

Capítulo 3

UN MODELO DE REPARACIONES MÍNIMAS PARA UN SISTEMA REPARABLE COMPUESTO.

Las estrategias de reemplazamiento para diversos modelos de sistemas reparables con reparaciones mínimas están ampliamente estudiadas en la literatura. Sin embargo, cuando un sistema está formado por varios subsistemas, surgen nuevos problemas en relación con la determinación del instante óptimo de reemplazamiento. El propósito de este capítulo es estudiar un modelo para este nuevo problema y encontrar dicho instante óptimo. El contenido de este capítulo ha sido publicado en la revista, *Revista Canaria de La Ciencia*.

3.1 Introducción

Gran parte de los trabajos publicados sobre sistemas reparables asumen que cuando tras una avería se repara un equipo, éste queda en el mismo estado que si fuese nuevo. Tal hipótesis permite elaborar modelos tratables desde el punto de vista matemático, pero resultan poco verosímiles en la práctica. Algo más realista son los modeos considerados por Ascher y Feingold [13] que tratan sistemas cuyo estado, después de una reparación, es idéntico al estado inmediatamente anterior a la ocurrencia del fallo. Los modelos en los que se asume esta hipótesis reciben el nombre de modelos de reparaciones mínimas. Brown y Proschan [17] consideran también modelos de reparaciones mínimas en sistemas de fiabilidad. Block, Borges y Savits [18] estudian un sistema en el que no siempre es posible la reparación mínima, siendo

la probabilidad de que tal reparación sea posible dependiente de la edad del equipo. Stadje y Zuckerman [34] consideran un sistema para el que los sucesivos periodos durante los que está operativo son estocásticamente decrecientes, mientras que los tiempos que duran las reparaciones crecen también en sentido estocástico. Dagpunar y Jack [35] estudian también sistemas de reparaciones mínimas con periodos de reparación no nulos.

Una característica común a todos estos trabajos es que las distintas estrategias de mantenimiento se analizan exclusivamente en sistemas consistentes en una única estructura operativa. Sin embargo, cuando se consideran sistemas compuestos por varios subsistemas con distintas funciones, surgen nuevos problemas en la determinación de la estrategia óptima de mantenimiento. En este capítulo se considera un equipo formado por una estructura principal y un subsistema que se encarga de la producción de ciertos ítems. La duración de la estructura principal del equipo hasta que ocurra una avería, sigue una distribución de probabilidad con razón de fallo creciente, siendo siempre posible la realización de reparaciones mínimas. No obstante, cada fallo de la estructura principal puede deteriorar el subsistema de producción, de tal forma que la tasa residual de producción de ítems disminuye aleatoriamente después de cada fallo. El subsistema, por su parte, no es reparable, y por tanto la tasa de producción sólo puede restaurarse a su valor inicial mediante el reemplazamiento del equipo.

De acuerdo con esta descripción, cualquier estrategia admisible de reemplazamiento deberá tener en cuenta no sólo la edad del equipo sino también su tasa residual de producción de ítems, que supondremos observable después de cada fallo. No obstante, si este valor no fuese directamente observable, podría sustituirse por alguna estimación adecuada del mismo. En los epígrafes siguientes determinaremos la estrategia óptima de reemplazamiento de este sistema para una cierta estructura de costes y beneficios, y probaremos que, bajo condiciones específicas, el reemplazamiento del equipo se lleva a cabo casi seguramente en tiempo finito. Por último

analizaremos un caso particular de sistema con estas características, obteniendo una aproximación de la regla óptima de reemplazamiento.

Denotamos por X a la variable aleatoria que mide el tiempo que tarda en producirse el primer fallo en el equipo principal. Por hipótesis, X es de razón de fallo creciente, siendo $\overline{F}(t) = P(X > t)$ su función de supervivencia. Suponemos que tras cada fallo es siempre posible efectuar una reparación mínima, que se realiza de forma instantánea. Llamaremos $\{X_n\}$ a las duraciones de los sucesivos periodos entre reparaciones mínimas consecutivas, $T_n = \sum_{i=1}^{n} X_i$ el tiempo transcurrido hasta el n-ésimo fallo ($T_0 = 0$), y $N(t) = \sum_{n=1}^{\infty} I_{[0,t]}(T_n)$ al proceso de recuento de fallos, siendo I_E la función indicatriz del conjunto E. Bajo estas condiciones $\{N(t), t \geq 0\}$ es un proceso de Poisson con función de medias $M(t) = E[N(t)] = -\ln\left(\overline{F}(t)\right)$ [18]. Asimismo, suponemos que durante el periodo comprendido entre los fallos $i-1$ e i-ésimo, el equipo produce ítems a una tasa α_i , tomando la tasa inicial siempre un valor fijo α_1 , y siendo $\{\alpha_i\}$ una sucesión aleatoria no creciente casi seguramente con $E[\alpha_i] \leq \alpha_1 \beta^{i-1}$ y $0 < \beta < 1$. Por último, consideramos que cada ítem producido reporta un beneficio B, cada reparación mínima tiene un coste C y la sustitución por un equipo nuevo un coste D.

Al igual que en los modelos de reparaciones mínimas con costes del capítulo anterior, el problema aquí es encontrar la estrategia de mantenimiento ξ que maximiza el cociente

$$\psi(\xi) = \frac{E[Z_\xi]}{E[T_\xi]} \tag{3.1}$$

siendo T_ξ la variable aleatoria que mide el tiempo entre dos reemplazamientos consecutivos (un ciclo) del sistema y Z_ξ la variable aleatoria que mide el beneficio en un ciclo. Las distribuciones de probabilidad tanto de la duración del correspondiente periodo de renovación T_ξ como del rendimiento Z_ξ obtenido durante este periodo, quedarán determinadas por la estrategia de reemplazamiento ξ elegida. Deduciremos a continuación la expresión de esta función de utilidad en el modelo considerado.

Para ello, y como es habitual en los modelos de reparaciones mínimas,

asumiremos que después de cada reparación la supervivencia del equipo es independiente del número de fallos y subsiguientes reparaciones realizadas hasta ese momento. Por tal motivo, la estrategia óptima de reemplazamiento deberá depender del tiempo de supervivencia hasta el siguiente fallo y no de los fallos anteriores. Como además cada fallo del sistema afecta a la tasa de producción de ítems $\{\alpha_n\}$, la estrategia óptima dependerá también de la tasa de productividad residual en el instante inmediatamente posterior a la avería. Resulta claro, además, que la sustitución del equipo debe realizarse en algún instante de fallo T_n .

El estado del sistema en el ν-ésimo instante de fallo queda completamente especificado por el vector $(T_\nu, \alpha_{\nu+1})$. El proceso $\{(T_\nu, \alpha_{\nu+1})\}_{\nu \geq 0}$ es un proceso markoviano con valores en $[0,\infty) \times [0,\infty)$. De acuerdo con lo indicado en el párrafo anterior, la estrategia de mantenimiento óptima deberá basarse en este proceso. Como ya se ha señalado, no procede realizar la sustitución entre dos fallos consecutivos, por lo que las reglas de sustitución admisibles consisten en efectuar el reemplazamiento en algún instante de fallo ν, cuando el sistema se encuentra en el estado $(T_\nu, \alpha_{\nu+1})$. Denotaremos por Γ al conjunto de dichas reglas. Si una regla $\xi \in \Gamma$ dispone la sustitución del equipo cuando el sistema está en el estado ν , el valor de la función de utilidad (3.1) es:

$$\psi(\xi) = \frac{B \cdot E\left[\sum_{i=1}^{\nu} \alpha_i \cdot (T_i - T_{i-1})\right] - C \cdot (E[\nu] - 1) - D}{E[T_\nu]} \tag{3.2}$$

Es inmediato observar que ψ está acotada superiormente dado que $\alpha_n \leq \alpha_1$, $\forall n$ casi seguramente y por tanto $\psi(\xi) \leq B \cdot \alpha_1$ para cualquier regla $\xi \in \Gamma$. Sea entonces:

$$\psi^* = \sup_{\xi \in \Gamma} \psi(\xi) \tag{3.3}$$

Una estrategia ξ^* es entonces óptima si:

$$\psi^* = \psi(\xi^*) \tag{3.4}$$

Deduciremos en la siguiente sección la forma de la estrategia óptima ξ^*.

3.2 Determinación de la Estrategia Óptima

De acuerdo con la sección 1.9 del capítulo 1, determinar la estrategia óptima de nuestro problema es equivalente a determinar la regla óptima de parada de la sucesión estocástica $\{Z_\nu\}_{\nu=1}^{\infty}$ donde,

$$Z_\nu = B \cdot \sum_{i=1}^{\nu} \alpha_i \cdot (T_i - T_{i-1}) - C \cdot (\nu - 1) - \psi^* T_\nu. \tag{3.5}$$

A continuación probaremos que el problema es monótono y que la regla miope es óptima. En primer lugar verificaremos las tres hipótesis del teorema 16 del capítulo inicial de esta memoria.

- $E[\sup_\nu Z_\nu] < \infty$.

  $Z_\nu = B \cdot \sum_{i=1}^{\nu} \alpha_i \cdot (T_i - T_{i-1}) - C \cdot (\nu - 1) - \psi^* T_\nu \leq B \cdot \sum_{i=1}^{\nu} \alpha_i \cdot (T_i - T_{i-1}) \leq B \cdot \sum_{i=1}^{\infty} \alpha_i \cdot T_1$. Por lo tanto, $E[\sup_\nu Z_\nu] \leq B \cdot \sum_{i=1}^{\infty} E[\alpha_i \cdot T_1] = B \cdot E[T_1] \sum_{i=1}^{\infty} E[\alpha_i] \leq B \cdot E[T_1] \cdot \alpha_1 \sum_{i=1}^{\infty} \beta^{i-1} < \infty$.

- $\lim_{\nu \to \infty} Z_\nu = -\infty$.

  Dado que $Z_\nu = B \cdot \sum_{i=1}^{\nu} (\alpha_i - \psi^*) \cdot (T_i - T_{i-1}) - C \cdot (\nu - 1)$, $\psi^* > 0$ y la sucesión α_n es decreciente a cero casi seguramente se tiene que, $P((\alpha_i - \psi^*) \geq 0, \forall i) = 0$. Así, la probabilidad de que un número numerable de términos $\alpha_i - \psi^* < 0$ es 1. Y por lo tanto $\lim_{\nu \to \infty} Z_\nu = -\infty$, casi seguramente.

- La sucesión de variables aleatorias $Y_\nu = \sup_{j \geq \nu}(Z_j - Z_\nu)$ es uniformemente integrable.

  En este caso, al ser la sucesión α_n decreciente a cero casi seguramente se puede comprobar con facilidad que $E[|Y_\nu|] \longrightarrow 0$ a medida que $\nu \to \infty$ y por consiguiente es uniformemente integrable la sucesión de variables aleatorias Y_ν.

Tan sólo resta por comprobar la monotonía del problema. Para ello vamos a demostrar la implicación $Z_\nu \geq E(Z_{\nu+1} | Z_1, \cdots, Z_\nu) \Longrightarrow Z_{\nu+1} \geq$

$E(Z_{\nu+2}|Z_1,\cdots,Z_{\nu+1})$. Como $Z_\nu = Z_{\nu-1} + (B\alpha_\nu - \psi^*)(T_\nu - T_{\nu-1}) - C$, la veracidad de esta implicación se reduce a la veracidad de esta otra,

$$(B\alpha_{\nu+1} - \psi^*)E[(T_{\nu+1} - T_\nu) \mid T_\nu] - C \leq 0 \Longrightarrow (B\alpha_{\nu+2} - \psi^*)E[(T_{\nu+2} - T_{\nu+1}) \mid T_{\nu+1}] - C \leq 0$$

Ahora bien, como sucesión α_ν es decreciente y $E[(T_{\nu+1} - T_\nu) \mid T_\nu] \geq E[(T_{\nu+2} - T_{\nu+1}) \mid T_{\nu+1}]$ se tiene que $(B\alpha_{\nu+2} - \psi^*)E[(T_{\nu+2} - T_{\nu+1}) \mid T_{\nu+1}] - C \leq (B\alpha_{\nu+1} - \psi^*)E[(T_{\nu+2} - T_{\nu+1}) \mid T_{\nu+1}] - C \leq (B\alpha_{\nu+1} - \psi^*)E[(T_{\nu+1} - T_\nu) \mid T_\nu] - C \leq 0$ y por tanto el problema es monótono. Así, de acuerdo con el teorema 16 la estrategia óptima de reemplazamiento es sustituir cuando el sistema alcance el estado $(T_{\nu^*}, \alpha_{\nu^*+1})$, siendo:

$$\nu* = \min\{\nu;\ (B \cdot \alpha_{\nu+1} - \psi*) \cdot g(T_\nu) \leq C\}, \tag{3.6}$$

donde $g(T_\nu) = E[(T_{\nu+1} - T_\nu) \mid T_\nu]$, esto es, la vida media residual del equipo principal en el instante T_ν.

Supondremos a continuación que la distribución del tiempo de supervivencia del equipo (periodo transcurrido hasta el primer fallo) es absolutamente continua, siendo $f(t)$ su función de densidad. En tal caso existe la correspondiente función de razón de fallos, que puede expresarse como $r(t) = \frac{f(t)}{\overline{F}(t)}$ .Obsérvese que $r(t)$ es además la función de intensidad del proceso $N(t)$ correspondiente al número de reparaciones mínimas realizadas dentro de un periodo de renovación. La hipótesis de razón de fallo creciente significa que esta función es no decreciente. Cabe preguntarse en este punto cuál es la probabilidad de que el óptimo de la función de utilidad empleada no se alcance en tiempo finito, lo que significaría que la máquina no se reemplaza nunca, solventándose todos los fallos siempre mediante reparaciones mínimas. El siguiente teorema prueba bajo determinadas condiciones que con probabilidad 1, ν^* es finito, lo que supone que el reemplazamiento de la máquina se realiza en tiempo finito casi seguramente.

Teorema 10 *Supongamos que la distribución de* T_1 *es tal que* $E[T_1] < \infty$, $r(t)$ *es no decreciente y* $\lim_{t\to\infty} r(t) = \infty$. *Supongamos además que* $P(\alpha_n \leq K,\ \forall n) = 1$ *para algún* K. *Bajo estas condiciones* $P(\nu* < \infty) = 1$

Demostración. $\{\nu* = \infty\} = \{(B \cdot \alpha_{n+1} - \psi*) \cdot g(T_n) > C;\ \forall n\}$. Ahora bien, dado que $\{T_n\}$ son los instantes de salto de un proceso de Poisson con función de intensidad $r(t)$, $T_n \underset{n\to\infty}{\to} \infty$ c.s. , lo que supone que $\overline{F}(T_n) \underset{n\to\infty}{\to} 0$ c.s. Además, $E[T_1] = \int_0^\infty \overline{F}(t)\,dt < \infty$ por lo que $\int_t^\infty \overline{F}(z) \cdot dz \underset{t\to\infty}{\to} 0$. Sea

$$\phi(t) = \frac{\overline{F}(t)}{\int_t^\infty \overline{F}(z) \cdot dz} - r(t).$$

Dado que $\phi(t) \underset{t\to\infty}{\to} 0$, entonces $\phi(T_n) \underset{n\to\infty}{\to} 0$ c.s.Tenemos por tanto: $\frac{\overline{F}(T_n)}{\int_{T_n}^\infty \overline{F}(z)\cdot dz} = \phi(T_n) + r(T_n) \underset{n\to\infty}{\to} \infty \Longrightarrow g(T_n) = \frac{1}{\overline{F}(T_n)} \cdot \int_{T_n}^\infty \overline{F}(z) \cdot dz \underset{n\to\infty}{\to} 0$. Dado que $P(\alpha_n < K,\ \forall n) = 1$ para algún K se tiene que $(B \cdot \alpha_{n+1} - \psi*) \cdot g(T_n) \underset{n\to\infty}{\to} 0$. Así pues, $\limsup P((B \cdot \alpha_{n+1} - \psi*) \cdot g(T_n) > C) = 0 \Rightarrow P(\nu* < \infty) = 1$ ■

Nota 3 *El teorema sólo exige a la sucesión* $\{\alpha_n\}$ *que esté acotada con probabilidad uno. Naturalmente, esto ocurre para una sucesión que verifique las hipótesis de nuestro modelo.*

Analizaremos a continuación algunos casos en los que la función de riesgo $r(t)$ no converge a infinito.

El caso de $r(t) = r_0\ \forall t$ corresponde a una duración hasta el primer fallo exponencialmente distribuida y por tanto en este caso el proceso de reparaciones mínimas $\{N(t)\}_{t\geq 0}$ es un proceso de Poisson homogéneo. Si la sucesión de tasas de producción $\{\alpha_n\}$ es casi seguramente no creciente y $\alpha_n \underset{n\to\infty}{\to} \alpha_\infty$, se tiene entonces que,

$$(B \cdot \alpha_{n+1} - \psi*) \cdot g(T_n) \underset{n\to\infty}{\to} \frac{(B \cdot \alpha_\infty - \psi*)}{r_0}$$

Si $\alpha_\infty = 0$ (de hecho lo es por hipótesis) con probabilidad uno, obviamente $P(\nu* = \infty) = 0$. Sin embargo, si r_∞ es tal que $P\left(\frac{(B\cdot\alpha_\infty - \psi*)}{r_0} > C\right) > 0$, con esta misma probabilidad el instante de reemplazamiento no se alcanza en tiempo finito.

Para el caso en el que $r(t)$ es IFR, pero $r(t) \underset{t\to\infty}{\to} r_0 < \infty$, puede hacerse un análisis análogo al caso exponencial, que sintetizamos en el siguiente corolario.

Corolario 2 *Supongamos que la distribución de* T_1 *es tal que* $E[T_1] < \infty$, $r(t)$ *es no decreciente y* $\lim_{t\to\infty} r(t) = r_0$. *Supongamos además que* $\alpha_n \underset{n\to\infty}{\to} \alpha_\infty$. *Entonces:* $P(\nu = \infty) = P\left(\frac{(B\cdot\alpha_\infty - \psi*)}{r_0} > C\right)$.

3.3 Análisis de un caso particular

Presentamos a continuación, a modo de ejemplo, el análisis mediante simulación de un caso particular del modelo propuesto. Hemos considerado para la duración del equipo hasta el primer fallo una distribución de Weibull con parámetros $\rho = 0.01$ y $k = 1.4$ $(r(t) = \rho \cdot k \cdot t^{k-1})$ lo que significa que el proceso de fallos es un proceso de Poisson no homogéneo con función de intensidad $r(t)$. Las sucesivas cantidades α_i pueden interpretarse como el número de elementos operativos del sistema en el periodo comprendido entre T_{i-1} y T_i. Cada uno de estos elementos produce un ítem por unidad de tiempo, por lo que α_i puede identificarse con la tasa de producción del sistema. Supondremos que $\alpha_1 = 100$. En cada instante de fallo se ha supuesto que la probabilidad de que uno de estos elementos quede improductivo es $p = 0.025$. Por tanto, si la tasa de producción en el periodo comprendido entre T_{i-1} y T_i es α_i, la tasa de producción en la etapa siguiente, α_{i+1}, es una variable aleatoria con distribución binomial de parámetros $n = \alpha_i$ y $p = 0.975$. Finalmente hemos supuesto la estructura de costes $B = 1$, $C = 5$ y $D = 900$.

La trayectoria $\{(T_i, \alpha_{i+1}),\ i = 1, \ldots, \nu\}$ es observable en cada instante de fallo v. A partir de estos datos calculamos el rendimiento empírico del sistema por unidad

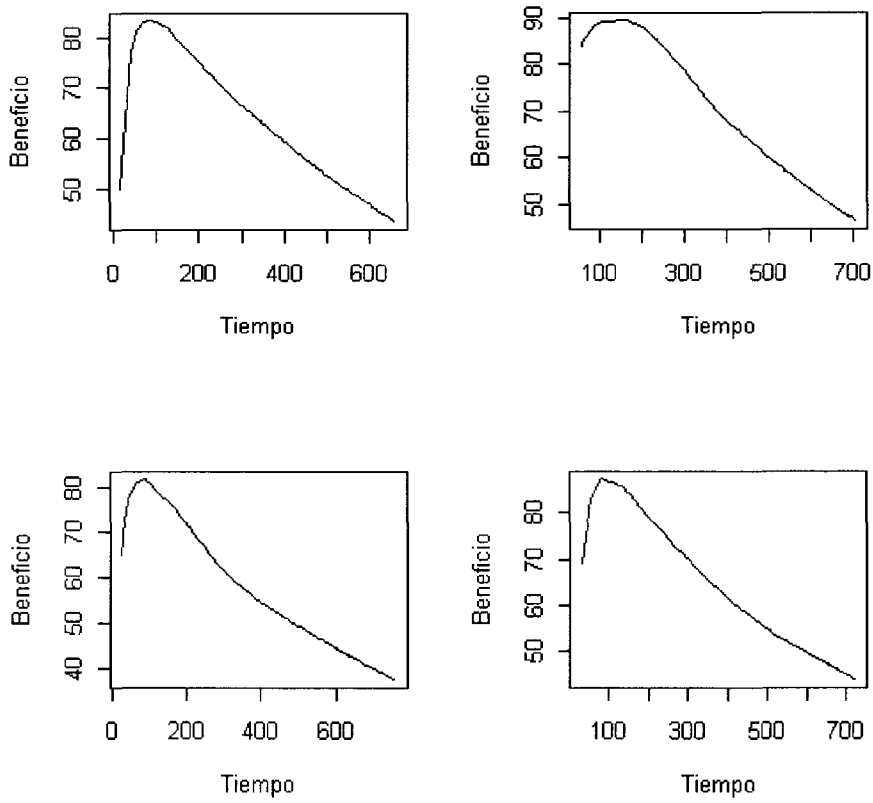

Figura 1 Cuatro posibles trayectorias de $\eta(\nu)$

de tiempo como:

$$\eta(\nu) = \frac{B \cdot \sum_{i=1}^{\nu} \alpha_i \cdot (T_i - T_{i-1}) - C \cdot (\nu - 1) - D}{T_\nu} \tag{3.7}$$

En la figura 1 se muestra la gráfica de una de las trayectorias simuladas del proceso estocástico $\eta(\nu)$.

De acuerdo con los resultados teóricos la estrategia óptima de reemplazamiento para este problema es:

$$\nu* = \min\{\nu;\ (\alpha_{\nu+1} - \psi*) \cdot g(T_\nu) \leq 5\}. \tag{3.8}$$

Para que esta estrategia de reemplazamiento pueda ser aplicada en este caso particular es necesario calcular el valor óptimo $\psi*$. Para ello, seguiremos las ideas

expuestas justo al final del primer capítulo. Esto es, consideramos el valor $\psi*$ como un valor variable. Llamémoslo γ. De esta manera tenemos la familia de reglas de paradas $\xi(\gamma) = \min\{\nu;\ (\alpha_{\nu+1} - \gamma) \cdot g(T_\nu) \leq 5\}$. Al fijar el valor γ disponemos de una regla de parada que tiene como beneficio por unidad de tiempo a largo plazo el valor $V(\gamma)$. Teóricamente, de acuerdo con la sección 1.9 del primer capítulo, la función $V(\gamma)$ posee un punto fijo. Precisamente, ese valor es el valor óptimo $\psi*$. Para calcular dicho valor hemos implementado un programa con MATHEMATICA. Empezando con el valor con el valor $\gamma = 0$ hemos obtenido los siguientes resultados,

Iteración	1	2	3	4	5	6	7
γ	31.233	63.815	79.8912	84.6504	85.5415	**85.549**	**85.549**

Concretamente, en 6 iteraciones se alcanza el valor deseado, siendo en este caso ψ^*= **85.549**. Por lo tanto la estrategia óptima para este problema es

$$\nu* = \min\left\{\nu;\ \alpha_{\nu+1} \leq \frac{5}{g(T_\nu)} + \mathbf{85.549}\right\} \tag{3.9}$$

En la siguiente tabla damos una estimación Monte Carlo (20000 simulaciones) de la función de probabilidad de la regla de parada $\nu*$.

$\nu*$	2	3	4	5	6	7	8	9
$P(\nu* = n)$	$5 \cdot 10^{-5}$	$6 \cdot 10^{-3}$	$5 \cdot 10^{-2}$	0.17	0.26	0.242	0.15	$7 \cdot 10^{-2}$

10	11	12	13	14	15	16
$3 \cdot 10^{-2}$	$1 \cdot 10^{-3}$	$3 \cdot 10^{-3}$	$7 \cdot 10^{-4}$	$4 \cdot 10^{-4}$	$5 \cdot 10^{-5}$	$5 \cdot 10^{-5}$

Capítulo 4

TEORÍA DE PARADAS ÓPTIMAS Y FIABILIDAD

En este capítulo mostramos como la teoría de paradas óptimas presenta ciertas ventajas con respecto a la teoría de procesos de decisión semi-Markovianos en cuanto a la determinación de estrategias óptimas de reemplazamiento. El contenido de este capítulo ha sido publicado en la revista *Jounal of Operational Research Society.*

4.1 Introducción

En el primer trabajo de reemplazamiento de sistemas reparables Barlow y Hunter [27] asumieron que los fallos que ocurren antes de un tiempo t pueden ser reparados mínimamente y que cuando el sistema alcance la edad t^* se realiza el reemplazamiento. Con el tiempo otros autores (Muth [29], y Park [31]) estudiaron otras políticas de reemplazamientos más acertadas que la propuesta por Barlow y Hunter. En todos estos trabajos se consideraron fijos tantos los costes de reemplazamiento como los costes de las reparaciones mínimas. Desde luego podía haberse seguido con el estudio nuevas políticas, pero lo que había que buscar era la política óptima de reemplazamiento. Fue Phelps [32] en 1981 quien comparó los modelos de reparaciones mínimas con costes fijos de reparaciones y reemplazamientos; demostró, suponiendo que la razón de fallos es creciente, que la política de Muth (reemplazar en el primer fallo despues de un tiempo t^*) es la mejor de las planteadas hasta entonces. En un artículo posterior, Phelps [33] generalizó su trabajo usando los resultados de procesos de decisión semi-Markovianos para probar que la política de

Muth es óptima en el conjunto de todas las políticas de reemplazamientos posibles para modelos de reparaciones mínimas con razón de fallos creciente y costes fijos de reparaciones y reemplazamiento.

En otro artículo Boland junto con Proschan [36] estudiaron el problema de reemplazamiento de sistemas reparables a través de un modelo con reparaciones mínimas y costes de reparaciones crecientes que depende del número de reparaciones en el ciclo*. Para este modelo consideraron una política de reemplazamiento al estilo de la estudiada por Barlow y Hunter. Sin embargo, esta política no era óptima. Hubo que esperar a 1992 para que Makis and Jardine [37] encontraran la política óptima de reemplazamiento para este modelo. En su artículo Makis and Jardine consideraron un modelo general con reparaciones imperfectas que incluía como caso especial al modelo de Boland y Proschan. Al igual que Phelps, Makis y Jardine resolvieron el problema de encontrar la política óptima utilizando resultados de la teoría de los procesos de decisión semi-Markovianos. No obstante, las condiciones que deben verificarse para poder aplicar los teoremas de la teoría de procesos de decisión semi-Markovianos que permiten encontar la política óptima de reemplazamiento son numerosas y no tan elementales de verificar.

En este capítulo se obtiene un nuevo procedimiento, dentro del marco de la teoría paradas óptimas, para encontrar la política óptima de reemplazamiento del modelo de Boland y Proschan. En este caso las condiciones que permiten la obtención de la política óptima son más elemetales que las propuestas por Makis y Jardine. La teoría de paradas óptimas parece, como lo demuestran las publicaciones recientes [38], una teoría próspera en el estudio de políticas optimas de reemplazamientos de sistemas reparables.

*Un ciclo es el tiempo entre dos reemplazamientos vecinos.

4.2 Modelo de Boland-Proschan y procesos de decisión semi-Markovianos

En esta sección vamos a explicar con un poco más de detalle los resultados del artículo de Makis y Jardine en lo que se refiere al hallazgo de la política óptima del modelo de Boland y Proschan. Debe tenerse en cuenta que el modelo estudiado por Makis y Jardine es un modelo más general que incluye al modelo de Boland y Proschan como caso particular. De acuerdo con nuestro objetivo usaremos a lo largo de esta sección la siguiente notación:

* c_r = Coste de reemplazamiento.

* c_n = Coste de reparación en el n-ésimo fallo, siendo t la edad del sistema.

* $F_x(t) = P(T \leq x+t|T > x)$ =Distribución condicional de vida residual cuando el sistema tiene edad x.

* $m(x)$ = Vida media residual.

* T_n = n-ésimo fallo dentro de un ciclo de reemplazamiento.

Las hipótesis del modelo son las siguientes:

(i) $c_r > 0$. $c_n \leq k \ \forall n$.

(ii) Existe un valor $\sigma > 0$ tal que para todo $x \geq 0$,$m(x) \geq \sigma$, $m(0) < \infty$.

(iii) c_n es un sucesión no decreciente.

(iv) La razón de fallos del sistema es creciente.

A partir de la estructura de costes se deduce que sólo deben tenerse en cuenta políticas que reemplazan en un instante de fallo. Por tanto en cada fallo n las decisiones a tomar son reemplazar con coste c_r ó reparar con coste c_n. A partir de las hipótesis (i) y (ii) anteriores podemos aplicar un resultado de la teoría de procesos de

decisión semi-Markovianos [39] que garantiza la existencia de una función acotada w y una cosntante $g*$ tal que,

$$w(n,x) = \min\left\{c_n + \int_0^{+\infty} w(n+1, x+t)F_x(dt) - g^*\tau(x), c_r + \int_0^{+\infty} w(1,y)F(dy) - g^*\tau(0)\right\}.$$

De acuerdo con un teorema (ver, [39]), la política que minimiza el término derecho de la función anterior es óptima y g^* es el coste promedio esperado por unidad de tiempo. A partir de ese resultado Makis y Jardine obtienen el siguiente teorema (que resuelve el problema).

Teorema 11 *Bajo las hipótesis (i)-(iv), la política óptima de reemplazamiento del modelo de Boland y Proschan es la siguiente :*

Reemplazar en el n-ésimo fallo si $T_n \geq t_n^*$ *donde* T_n *es el instante del n-ésimo fallo y* $t_n^* = \inf\{t \geq 0 : c_n \geq g^*\tau(t)\}$

La sucesión $\{t_n^*\}$ no es creciente, $t_1^* \geq t_2^* \geq \ldots$

4.3 Modelo de Boland-Proschan y teoría de paradas óptimas

En esta sección vamos a obtener la política óptima de reemplazamiento del modelo de Boland y Proschan utilizando algunos resultados de la teoría de paradas óptimas. Recuédese que es preciso tener en cuenta que deben considerarse tan solo las políticas que remplacen en instante de fallo. Como criterio de optimalidad para encontar la política óptima, adoptaremos el criterio usual del costo promedio a largo plazo. Por argumentos clásicos de la teoría de la renovación, el coste promedio a largo plazo del sistema gobernado por la política de reemplazamiento T viene dado por,

$$V_T = E(R_T)/E(T) \tag{4.1}$$

donde R_T es el coste total necesario para mantener el sistema en funcionamiento hasta el instante T. Sea $V^* = \inf_{T\in\Gamma} V_T$, donde Γ es el conjunto de todas las políticas que

reemplazan en un instante de fallo. Una política de reemplazamiento T^* es óptima si,

$$V_{T^*} = \inf_{T\in\Gamma} R_T \tag{4.2}$$

Obviamente, $E(R_T) \geq V^*E(T)$, para todo $T \in \Gamma$, and T^* es óptima si y solo sí

$$E(R_{T^*} - V^*T^*) = \inf_{T\in\Gamma} E(R_T - V^*T) \tag{4.3}$$

Atendiendo a la anterior ecuación y no olvidando que las únicas políticas que deben tenerse en cuenta son aquellas que paran en un fallo, el problema de encontrar la política óptima para el modelo de Boland y Proschan puede plantearse como: hallar la regla de parada óptima para la sucesión estocástica,

$$Z_n = c_r + \sum_{i=1}^{n-1} c_i - V^*T_n \ , \quad n \geq 1 \tag{4.4}$$

Proposición 1 *El problema de regla de parada anterior es un problema de caso monótono.*

Demostración. Tenemos que probar la siguiente implicación,

$$Z_n \leq E(Z_{n+1}|Z_1, \cdots, Z_n) \Longrightarrow Z_{n+1} \leq E(Z_{n+2}|Z_1, \cdots, Z_{n+1})$$

Es claro que, $Z_{n+1} = Z_n + c_n - V^*(T_{n+1} - T_n)$. Entoces teniendo en cuenta las hipótesis anteriores $Z_n \leq E(Z_{n+1}|Z_1, \cdots, Z_n)$ es equivalente a $c_n \geq V^*m(T_n)$. Además, como $c_{n+1} \geq c_n$ y $m(T_n) \geq m(T_{n+1})$, se tiene que, $Z_{n+1} \leq E(Z_{n+2}|Z_1, \cdots, Z_{n+1})$. ■

Por tanto y de acuerdo con los teoremas de la teoría de paradas óptimas (ver [12], pág 113) la regla de parada,

$$N = \inf\{n \geq 1 \mid Z_n \leq E(Z_{n+1}|Z_1, \cdots, Z_n)\}$$

es óptima. Es decir , la política óptima de reemplazamiento es

$$N = \inf\{n \geq 1 \mid c_n \geq V^*m(T_n)\} \tag{4.5}$$

Desde luego esta política no estará completamente determinada hasta que no se conozca el valor V^* que aparece en (4.5). Para calcular dicho valor podemos utilizar el algoritmo de punto fijo que mencionamos en la última sección de primer capítulo. Otra alternativa sería aplicar el algoritmo propuesto por Makis y Jardine, el cual se basa en un algoritmo mucho más general dado por Aven y Berman [40]. Para ver comparaciones entre la política óptima de reemplazamiento y la política de reemplazamiento periódico propuesta por Boland para el caso particular de distribuciones Weibull y exponencial puede consultarse de nuevo el trabajo de Makis y Jardine.

Capítulo 5

DIMENSIONAMIENTO ÓPTIMO DE UN SISTEMA DE COLAS

Hemos visto cómo la teoría de la renovación proporciona una potente herramienta para resolver problemas de optimización asociados a procesos estocásticos que se caracterizan por la presencia de instantes de renovación, a partir de los cuales el proceso repite su comportamiento (en sus aspectos probabilistas). En particular ello nos ha servido para obtener instantes óptimos de reemplazamiento en el contexto del problema de reemplazamiento de sistemas reparables. En este capítulo veremos una nueva aplicación de la teoría de la renovación; esta vez en el contexto de un problema de colas.

5.1 Motivación y presentación del modelo

Las actuales redes de comunicaciones de banda ancha están diseñadas para soportar y acomodar, de manera flexible y eficiente, una amplia variedad de servicios tales como voz, vídeo, datos y sus combinaciones multimedia. Para cumplir con estos objetivos se han diseñado diversos estándares de red, entre los que cabe destacar ATM como uno de los más importantes. En este estándar la información que viaja por la red se organiza en pequeños paquetes de longitud fija llamados células. Los diferentes tipos de aplicaciones que hacen uso de la red pueden variar en sus requerimientos de servicio. Así, por ejemplo, las aplicaciones en tiempo real, tales como la videoconferencia, requieren prestaciones extremas en términos de volumen de tráfico (*throughput*), de retardo, de la variación del retardo (*delay jitter*, *jitter* o varianza en el retardo) y de la tasa de pérdidas. La cada vez mayor generalización

de estas aplicaciones ha convertido en urgente la necesidad de proporcionar servicios de red con prestaciones garantizadas y desarrollar los algoritmos que soporten estos servicios.

Uno de los mecanismos fundamentales para poder proporcionar servicios con prestaciones garantizadas es la elección de la disciplina de servicio de paquetes en los conmutadores, que son los dispositivos físicos encargados de combinar el tráfico procedente de distintas fuentes. En una red de conmutación de paquetes, los paquetes de distintas conexiones interactúan unos con otros en cada conmutador, y sin el control apropiado estas interacciones pueden llegar a afectar a las prestaciones de la red. La disciplina de servicio del nodo de conmutación controla el orden en el que los paquetes son servidos y determina cómo interactúan los paquetes de las distintas conexiones. Una forma de conseguir este objetivo es mediante la implementación, en los buffers del conmutador, de mecanismos de prioridad capaces de controlar el tiempo o espacio dedicado a cada célula. Debido a su simplicidad, rapidez y más bajo coste de implementación, en los conmutadores de alta velocidad se prefieren los mecanismos de prioridad que controlan el espacio (capacidad) disponible en el buffer, a los que deben tener en cuenta el instante de llegada de cada célula y controlar su tiempo de residencia en el conmutador.

En este trabajo se presenta y analiza un esquema de dimensionado y gestión de buffers que tiene en cuenta el patrón de llegada de dos clases de tráfico. En particular, hemos considerado que estas clases de tráfico corresponden, respectivamente, a tráfico en tiempo real (voz, video) y a tráfico en tiempo no real (datos). Por simplicidad llamaremos TR al tráfico de la primera clase y TNR al de la segunda. En los últimos años se han desarrollado numerosas políticas de servicio para el tráfico resultante de la mixtura de estas dos clases, con el objetivo fundamental de que la red pueda ofrecer calidad de servicio al tráfico en tiempo real, a la vez que se garantiza un determinado nivel de prestaciones para el tráfico en tiempo no real. En particular la calidad de servicio para tráfico TR exige un bajo nivel de retardo en el tránsito de los paquetes

(no es posible mantener una conversación telefónica si los paquetes de voz digitalizada llegan de manera muy asincrónica), mientras que para el tráfico TNR se requiere una muy baja tasa de pérdidas (este es el tráfico habitual en internet correspondiente a la descarga de ficheros; lo importante es no perder el contenido del fichero aunque tarde algo más en llegar a destino). Ello no quiere decir que para tráfico en tiempo real sea admisible una tasa de pérdidas excesivamente elevada, ni que el tráfico en tiempo no real pueda llegar a demorarse cualquier cantidad de tiempo, sino más bien que las pérdidas admisibles en tráfico TR son varios órdenes de magnitud superiores a las del tráfico TNR; y a la inversa, el nivel de retardo admisible en tráfico TNR puede llegar a ser también varios órdenes de magnitud superior al admisible para tráfico TR.

En [41] puede encontrarse un análisis comparativo de diversas políticas para el multiplexado de tráfico en tiempo real con tráfico en tiempo no real, incluyendo una revisión de trabajos de investigación previos. La más simple de estas políticas es la clásica política FIFO, que es la que ofrece peores resultados para ambos tipos de tráfico. Los resultados mejoran cuando se asigna prioridad más alta a los paquetes del tráfico TR, siendo atendidos antes que los paquetes TNR, que resultan de esta forma penalizados con unos retardos excesivos. Otras políticas alternativas y mejores son MLT (Minimum laxity threshold) que asigna prioridades en función del tiempo que ya hayan permanecido en cola los paquetes de las distintas clases y del tiempo máximo en cola admisible para los paquetes TR, y la QLT (Queue Length Threshold) en que se da prioridad a los paquetes NRT sólo cuando el número de los mismos en cola supera un cierto umbral. El rendimiento de estas dos políticas es similar, si bien la última es más fácil de implementar físicamente, toda vez que no requiere contabilizar el tiempo de permanencia en el buffer de cada uno de los paquetes presentes en el mismo, y resultando con ello, como ya hemos dicho más arriba, la política preferida.

Otra clase de políticas, muy habitual para gestionar la multiplexación de estas dos clases de tráfico, es la formada por los mecanismos basados en rondas, consistentes básicamente en alternar el servicio entre los distintos canales de entrada

al conmutador. La más simple de estas políticas consiste en atender una célula de cada canal, desde el primero al último, y vuelta a empezar desde el primer canal (este es el mecanismo round-robin clásico [42], [43]). Una variación de esta política consiste en asignar prioridades a los canales en función de la clase de tráfico (weighted round robin), atendiendo en cada ronda más células en los canales de mayor prioridad que en los canales de prioridad más baja; por ejemplo, se podrían atender tres células de un canal de video por cada célula de un canal de datos. Se han presentado numerosas propuestas sobre la forma en que debería distribuirse la atención del servidor entre los diversos canales. Algunas de estas políticas fijan las prioridades de cada clase de tráfico de modo estático (de una vez para siempre) [44], [45] mientras que otros asignan las prioridades dinámicamente, en función de las condiciones de la red y de la carga del sistema [46], [47]. Muchas de estas políticas para la asignación dinámica de prioridades especifican umbrales: mientras en el canal A haya menos de n células esperando, se atiende al canal B; una vez que en A haya n células, el servicio se dedica exclusivamente a A hasta que en A queden menos de m células. Estos umbrales pueden también ser fijos, determinados en el diseño del hardware que constituye el conmutador [48], o flexibles/adaptativos, controlados por software en función de las condiciones de tráfico en cada momento [49], [50], [45].

El mecanismo que presentamos en esta tesis, está basado en un sistema de rondas que incorpora a los buffers de entrada al conmutador propios de cada canal, un buffer adicional que es compartido por las dos clases de tráfico. Este sistema es gestionado por una política que, en cierta medida, prioriza al tráfico TR, para evitarle retardos excesivos, a la vez que permite cursar el tráfico TNR con un mínimo nivel de pérdidas. Un correcto dimensionado de los buffers que componen este mecanismo permite que esta política pueda proporcionar una reducción en los retardos de las células de alta prioridad (tiempo real), sin que por ello se vean afectadas considerablemente las prestaciones dadas al tráfico de baja prioridad (tiempo no real).

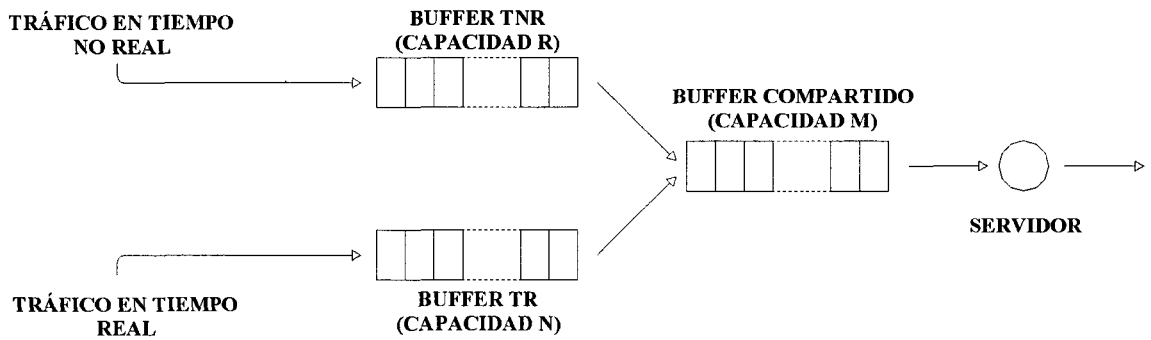

Figura 1 Diseño del sistema de colas

Tal como se observa en la figura 1, el tráfico en tiempo real (TR) accede a su propio buffer, que tiene una capacidad limitada para R células. Asimismo, el tráfico en tiempo no real (TNR) accede a un segundo buffer con capacidad N. Las células que salen de ambos buffers se mezclan en un tercer buffer compartido, con capacidad para M células. La salida de este tercer buffer es el canal de transmisión, sobre el que se multiplexan ambos tráficos. La gestión del sistema se produce del siguiente modo:

1. En cada instante, en el buffer compartido puede haber sólo una célula TNR, y como máximo $M - 1$ células TR.

2. Si no hay ninguna célula TNR en el sistema, la primera célula TNR que llegue al mismo pasa directa e instantáneamente al buffer compartido. Si hay ya una célula TNR en el buffer compartido, cualquier nueva célula TNR que llegue al sistema se incorpora al buffer TNR.

3. Cada vez que llega una célula TR, accede directamente al buffer compartido si en éste hay menos de $M-1$ células TR; en otro caso se incorpora el buffer TR.

4. Cada vez que una célula TR (TNR) es transmitida, la primera célula del buffer TR (TNR), si es que hay alguna, pasa instantáneamente a la cola del buffer compartido.

5. Cuando un buffer de entrada está lleno, las células que lleguen al mismo son rechazadas.

Como puede apreciarse en esta descripción, el mecanismo de gestión de estas colas es muy sencillo y, por tanto, su funcionamiento en la práctica es muy rápido, lo que constituye una de sus principales ventajas desde el punto de vista técnico. Los parámetros de control del sistema son los tamaños de los buffers, N, R y M. Estos valores deben elegirse cuidadosamente, de forma que se cumplan los requisitos exigidos por el tráfico de la red. Estos requisitos se concretan en minimizar el retardo para el tráfico TR y minimizar las pérdidas para el tráfico TNR, sin que ello se consiga a costa de incrementar en demasía las pérdidas para el tráfico TR y el retardo para el tráfico TNR. Si valoramos en α el coste del retardo por unidad de tiempo y por célula para el tráfico TR y en β el coste de la pérdida de una célula en el tráfico TNR, nuestro objetivo es encontrar los valores de N, R y M que minimizan el coste medio por unidad de tiempo a largo plazo para el sistema:

$$\phi(M,N,R) = \lim_{t\to\infty} \frac{\alpha E\left[Y_{RT}(t)\right] + \beta E\left[L_{TNR}(t)\right]}{t} \tag{5.1}$$

donde $E[Y_{TR}(t)]$ es el retardo medio total acumulado por los clientes TR hasta t, y $E[L_{TNR}(t)]$ es el número medio de células TNR perdidas hasta t. Este mínimo debe hallarse con la restricción de que el tiempo medio de espera de los clientes TNR no supere un umbral W_1 y la tasa media de pérdida para los clientes TR no rebase tampoco el umbral g_1.

La asignación de costes (tales como el a y el b que se acaban de citar) a la gestión de sistemas orientados a ofrecer calidad de servicio con multiplexación de

tráfico de diversas clases se utiliza cada vez más como herramienta de regulación del uso de las redes [51]. Aunque en este trabajo hemos considerado que los costes a y b son constantes para todos los paquetes, en la actualidad empiezan ya a implementarse sistemas en los que los usuarios pagan más o menos dependiendo del nivel de calidad que quieran alcanzar. En la práctica, esto se traduce en que si dos paquetes TR pertenecientes a dos usuarios distintos se encuentran simultáneamente en un conmutador, se asignará prioridad al paquete del usuario que pague la tarifa más alta.

Hemos llevado a cabo diversas simulaciones comparando esta política de gestión con otras dos políticas basadas en rondas. Para la comparación, hemos utilizado una política WRR (Weighted Round Robin), consistente en atender k células de clase TR por cada célula de clase TNR, y una política HOL (Head of the Line), consistente en vaciar la cola TR cada vez que atiende a un número predeterminado j de células TNR. El objetivo de las simulaciones ha sido comprobar el comportamiento de las tres políticas bajo distintas condiciones de carga y relaciones de servicio para cada clase de cliente. En todos los casos se han dispuesto los parámetros de control de los distintos mecanismos de gestión descritos de forma que las tasas de servicio dedicadas a cada tipo de tráfico sean lo más parecidas posible, para que tenga sentido la comparación.

Como resultado de estas simulaciones puede apreciarse que la política presentada es, en general, mejor que la HOL (produce casi siempre retardos menores o iguales que ésta para las dos clases de tráfico en todas las condiciones de carga), y reduce los tiempos de espera para el tráfico TR en comparación con la política RR, aunque a costa de incrementar ligeramente los tiempos de espera del tráfico TNR. Debe destacarse también que en las simulaciones se observa que la nueva política disminuye los retardos máximos en ambos tipos de tráfico, causando por tanto una menor distorsión en el patrón del tráfico a la salida del sistema. Esta comprobación empírica de las buenas propiedades de esta política de gestión invita a realizar un

estudio analítico de la misma, que nos permita disponer en cada caso de los valores óptimos de los parámetros de control del sistema.

En lo que sigue supondremos que ambas clases de tráfico llegan al sistema según sendos procesos de Poisson, de parámetros lR y lN respectivamente. Asimismo, supondremos que el tiempo de servicio sigue una distribución de probabilidad general, idéntica para ambas clases de tráfico, aunque posteriormente particularizaremos los resultados para servicio determinista. Ello es razonable para una red de comunicaciones, donde la velocidad del canal (que es el servidor del sistema) es constante e idéntica cualquiera que sea el tipo de información que se transmita. Más discutible es la hipótesis del proceso de Poisson para las llegadas, que se utiliza básicamente porque permite la suficiente simplicidad en el tratamiento analítico del problema. No obstante, la técnica empleada es generalizable a modelos con llegadas de Erlang o con distribuciones tipo fase, que permiten abarcar tipos de tráfico más generales. En cualquier caso, el mecanismo de gestión y control presentado ha demostrado su eficiencia en simulaciones con tráficos de entrada que no son de Poisson. En particular, se han realizado diversas pruebas cuando las células de ambas clases siguen patrones de tráfico a ráfagas, generados mediante la mezcla de varias fuentes ON-OFF. En cada caso, durante los periodos OFF, de duraciones exponenciales con medias diferentes según la fuente de procedencia, no se generan células; durante los periodos ON, de duraciones también exponenciales se generan células espaciadas regularmente. En las simulaciones puede observarse que, con estos patrones de tráfico, el comportamiento a grandes rasgos de los retardos producidos en el conmutador siguiendo las tres políticas señaladas no difiere del observado para el tráfico de Poisson, aunque los valores son, desde luego, distintos.

5.2 Descripción del estado del sistema

El estado del sistema de colas anterior en el instante t puede describirse mediante la terna $(N_{TR}(t), N_{TNR}(t), S_{TNR}(t))$ donde:

$N_{TR}(t)$: es el número total de clientes TR en el sistema en el instante t.

$N_{TNR}(t)$: es el número total de clientes TNR en el buffer TNR en el instante t.

$S_{TR}(t)$: es la posición que ocupa el (único) cliente TNR en la cola compartida, en caso de haber alguno de estos clientes en dicha cola; en caso contrario esta variable vale 0.

Dado que los búffers son finitos, este sistema alcanza necesariamente el equilibrio. Llamaremos entonces:

$$(N_{TR}, N_{TNR}, S_{TNR}) = \lim_{t\to\infty} (N_{TR}(t), N_{TNR}(t), S_{TNR}(t)) \tag{5.2}$$

De acuerdo con la política de gestión del sistema, sólo son posibles los estados de la forma:

$$\begin{aligned} &(k,0,0), \quad 0 \leq k \leq R+M-1 \\ &(k,n,m), \quad m-1 \leq k \leq R+M-1; 0 \leq n \leq N; 1 \leq m \leq M \end{aligned} \tag{5.3}$$

que constituyen un total de $M + R + M(N+1)(R + \frac{M+1}{2})$ estados posibles.

Nuestro primer objetivo será determinar las probabilidades en el equilibrio, π_{ijk}, de los distintos estados del sistema. Para ello, de igual modo que en el tratamiento clásico de las colas **M/G/1**, podríamos calcular estas probabilidades a partir de la cadena de Markov encajada en los instantes de salida. Para ello basta calcular las probabilidades de transición entre dos instantes de salida del sistema sucesivos. Esta es una tarea sencilla pero trabajosa. Así, llamando a_k a la probabilidad de que durante un tiempo de servicio lleguen k clientes TR, y c_j a la probabilidad de que lleguen j TNR tendríamos, por ejemplo:

$$p_{000,k00} = a_k c_0 \; ; \qquad 0 \leq K \leq R+M-3 \tag{5.4}$$

El cálculo es más complicado cuando entre dos salidas se producen llegadas de las dos clases de clientes; si en la primera salida no quedó ningún cliente TNR en la

cola compartida, hay que determinar el orden de llegada para saber en qué posición de esta cola queda el primer cliente TNR que accede a ella. Si llamamos $\varphi_{rs}(j) =$ Prob(si entre dos salidas sucesivas llegan r clientes TR y s clientes TNR, el primero de los clientes TNR llegue en la j-ésima posición, $1 \leq j \leq r+s$), puede probarse que:

$$\varphi_{rs}(j) = \frac{\binom{r+s-j}{s-1}}{\binom{r+s}{s}} \tag{5.5}$$

y tendríamos, por ejemplo,

$$p_{000,knm} = \frac{\lambda_R}{\lambda_R + \lambda_N} a_k c_{n+1} \varphi_{k(n+1)}(m), 1 \leq k \leq R+M-3, 1 \leq m \leq M-2, 0 \leq n \leq N-1$$

Si denotamos por E_m al conjunto de estados para los que $S_{TNR} = m$ (ordenando dentro de este conjunto los estados en orden creciente de N_{TR} y N_{TNR}), puede demostrarse que la matriz de transiciones es de la forma:

$$P = \begin{pmatrix} A_{00} & A_{01} & A_{02} & \cdots & A_{0M-1} & A_{0M} \\ A_{10} & A_{11} & A_{12} & \cdots & A_{1M-1} & A_{1M} \\ 0 & A_{21} & 0 & \cdots & 0 & 0 \\ 0 & 0 & A_{32} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A_{MM-1} & 0 \end{pmatrix} \tag{5.6}$$

donde las A_{ij} son matrices que se pueden descomponer en subcajas de matrices triangulares o muy huecas. Si bien es posible obtener las probabilidades estacionarias π_{ijk} a partir de esta matriz mediante la resolución del sistema:

$$\begin{aligned} \pi P &= \pi \\ \|\pi\|_1 &= 1 \end{aligned}, \tag{5.7}$$

resulta un problema numérico complejo. Por ello, procederemos de otro modo para encontrar las π_{ijk}

5.3 Cálculo de las probabilidades estacionarias del sistema

El proceso $\{(N_{TR}(t), N_{TNR}(t), S_{TNR}(t)), t \geq 0\}$ es regenerativo. Este proceso se regenera a sí mismo en aquellos instantes en que se produce una llegada que encuentra el sistema vacío. Si definimos un ciclo como el tiempo transcurrido entre dos llegadas que encuentran el sistema vacío, y llamamos:

T = Longitud de un ciclo.

T_{ijk} = Tiempo total, durante un ciclo, en que el sistema se encuentra en el estado (i, j, k).

Entonces, de acuerdo con la teoría de procesos regenerativos:

$$\pi_{ijk} = \frac{E[T_{ijk}]}{E[T]} \tag{5.8}$$

En particular, debido a que las llegadas son de Poisson, se tiene $E[T_{000}] = 1/(\lambda_R + \lambda_N)$, y por tanto:

$$\pi_{000} = \frac{1}{(\lambda_R + \lambda_N) E[T]} \tag{5.9}$$

Siguiendo la línea argumental de Tijms [52] es posible construir ecuaciones recursivas para calcular los π_{ijk} utilizando las expresiones anteriores. Para ello definimos:

N_{ijk} = Número de clientes atendidos durante un ciclo que a su salida dejan el sistema en el estado (i, j, k).

$A_{(i,j,k),(a,b,c)}$ = Tiempo medio, durante un servicio que comenzó con el sistema en el estado (i, j, k), en que el sistema permanece en el estado (a, b, c).

Observando ahora que:

- el primer servicio de un ciclo comienza con un cliente presente (que puede ser TR o TNR),

- durante un servicio el sistema puede alcanzar el estado $(a,0,0)$ solamente si comenzó en un estado $(i,0,0)$ con $i \leq a$,
- durante un servicio, el sistema puede alcanzar el estado (a,b,c) con $1 \leq c \leq M$ sólo si el servicio comenzó con el sistema en algún estado (i,j,c), con $i \leq a$, $j \leq b$ (ya que la posición c del cliente TNR en la cola compartida no se altera durante un tiempo de servicio), o bien si el sistema empezó en algún estado $(i,0,0)$ con $i \leq a$, $i < c$, podemos escribir las siguientes ecuaciones:

Si $c-1 \leq a \leq R+M-1$, $1 < c \leq M$, $0 \leq b \leq N$:

$$E\left[T_{abc}\right] = \frac{\lambda_R}{\lambda_R+\lambda_N} A_{(1,0,0)(a,b,c)} + \sum_{i=c-1}^{a} \sum_{j=0}^{b} E\left[N_{ijc}\right] A_{(i,j,c)(a,b,c)} + \sum_{i=1}^{a} E\left[N_{i00}\right] A_{(i,0,0)(a,b,c)} \tag{5.10}$$

Si $0 \leq a \leq R+M-1$, $0 \leq b \leq N$

$$E\left[T_{ab1}\right] = \frac{\lambda_N}{\lambda_R+\lambda_N} A_{(0,0,1)(a,b,1)} + \sum_{i=0}^{a} \sum_{j=0}^{b} E\left[N_{ij1}\right] A_{(i,j,1)(a,b,1)} \tag{5.11}$$

Si $0 \leq a \leq R+M-1$

$$E\left[T_{a00}\right] = \frac{\lambda_R}{\lambda_R+\lambda_N} A_{(1,0,0)(a,0,0)} + \sum_{i=0}^{a} E\left[N_{i00}\right] A_{(i,0,0)(a,0,0)} \tag{5.12}$$

Dado que las llegadas y salidas en este sistema se producen de una en una, cuando el sistema está en equilibrio el número medio $E[N_{abc}]$ de salidas que dejan el sistema en el estado (a,b,c) es igual al número medio de llegadas que encuentran el sistema en este estado. Ahora bien, puesto que las llegadas se producen según un proceso de Poisson, (siempre que el sistema disponga de espacio disponible para alojar a cada nueva llegada), este último número medio es $\lambda^*_{abc} E\left[T_{abc}\right]$, donde:

$$\lambda^*_{abc} = \begin{cases} \lambda_R+\lambda_N & si\ a < R+M-1\ y\ b+c < N+1 \\ \lambda_R & si\ a < R+M-1\ y\ b+c \geq N+1 \\ \lambda_N & si\ a \geq R+M-1\ y\ b+c < N+1 \\ 0 & si\ a \geq R+M-1\ y\ b+c \geq N+1 \end{cases}$$

Por tanto:

$$E\left[N_{abc}\right] = \lambda^*_{abc} E\left[T_{abc}\right]$$

Sustituyendo estos valores en el sistema de ecuaciones (5.10),(5.11)y(5.12), llamando $p_R = \frac{\lambda_R}{\lambda_N+\lambda_R}$, $p_N = \frac{\lambda_N}{\lambda_N+\lambda_R}$ y dividiendo todos los términos por la duración media del ciclo, $E[T]$, y teniendo en cuenta (5.8) y (5.9), se llega a:

Si $c-1 \leq a \leq R+M-1$, $1 < c \leq M$, $0 \leq b \leq N$:

$$\pi_{abc} = p_R \lambda^*_{abc} \pi_{000} A_{(1,0,0)(a,b,c)} + \sum_{i=c-1}^{a} \sum_{j=0}^{b} \lambda^*_{abc} \pi_{abc} A_{(i,j,c)(a,b,c)} + \sum_{i=1}^{a} \lambda^*_{abc} \pi_{i00} A_{(i,0,0)(a,b,c)} \tag{5.13}$$

Si $0 \leq a \leq R+M-1$, $0 \leq b \leq N$

$$\pi_{ab1} = p_N \lambda^*_{abc} \pi_{000} A_{(0,0,1)(a,b,1)} + \sum_{i=0}^{a} \sum_{j=0}^{b} \lambda^*_{abc} \pi_{ij1} A_{(i,j,1)(a,b,1)} \tag{5.14}$$

Si $0 \leq a \leq R+M-1$

$$\pi_{a00} = p_R \lambda^*_{abc} \pi_{000} A_{(1,0,0)(a,0,0)} + \sum_{i=0}^{a} \lambda^*_{abc} \pi_{i00} A_{(i,0,0)(a,0,0)} \tag{5.15}$$

Si en este sistema se dividen todas las ecuaciones por π_{000}, llamando $\theta_{abc} = \pi_{abc}/\pi_{000}$, pueden obtenerse con facilidad de modo recursivo las θ_{abc}. Las π_{abc} se obtienen luego fácilmente sin más que observar que:

$$\pi_{000} + \sum\nolimits_{(a,b,c)\neq(0,0,0)} \pi_{abc} = 1 \Rightarrow 1 + \sum\nolimits_{(a,b,c)\neq(0,0,0)} \theta_{abc} = \frac{1}{\pi_{000}}$$

de donde:

$$\pi_{000} = \frac{1}{1 + \sum_{(a,b,c)\neq(0,0,0)} \theta_{abc}} \quad y \quad \pi_{abc} = \frac{\theta_{abc}}{1 + \sum_{(a,b,c)\neq(0,0,0)} \theta_{abc}} \tag{5.16}$$

Obviamente, aún falta determinar los términos $A_{(i,j,k)(a,b,c)}$.

Comencemos por calcular $A_{(i,j,c)(a,b,c)}$ para $1 \leq c \leq M$, $0 \leq i \leq a$, $0 \leq j \leq b$. Para ello definimos la variable aleatoria:

$$\chi_{(i,j,k)(a,b,c)} = \begin{cases} 1 & \begin{cases} \text{si en el instante } t \text{ el sistema está en el estado } (a,b,c), \text{ y aún} \\ \text{no ha terminado un servicio que comenzó en el instante 0} \\ \text{con el sistema en el estado } (i,j,c) \end{cases} \\ 0 & \text{en otro caso} \end{cases}$$

Es evidente entonces que:

$$A_{(i,j,c)(a,b,c)} = E\left[\int_0^\infty \chi_{(i,j,c)(a,b,c)}(t)dt\right] = \int_0^\infty E\left[\chi_{(i,j,c)(a,b,c)}(t)\right] dt$$

Ahora bien:

$$E[\chi_t] = P(\chi_t = 1) = (1 - B(t))\, e^{-\lambda_R t}\frac{(\lambda_R t)^{a-i}}{(a-i)!}e^{-\lambda_N t}\frac{(\lambda_N t)^{b-j}}{(b-j)!}$$

donde $B(t)$ es la función de distribución del tiempo de servicio. En el caso particular de servicio determinista de duración D se tiene:

$$B(t) = \begin{cases} 0 & t < D \\ 1 & t \geq D \end{cases}$$

Luego, en este caso:

$$A_{(i,j,c),(a,b,c)} = \int_0^D e^{-\lambda_R t}\frac{(\lambda_R t)^{a-i}}{(a-i)!}e^{-\lambda_N t}\frac{(\lambda_N t)^{b-j}}{(b-j)!}dt$$

Si llamamos:

$$G(m,n) = \int_0^D e^{-\lambda_R t}\frac{(\lambda_R t)^{a-i}}{(a-i)!}e^{-\lambda_N t}\frac{(\lambda_N t)^{b-j}}{(b-j)!}dt$$

$$F(m,n) = e^{-(\lambda_R+\lambda_N)D}\frac{\lambda_R^m \lambda_N^n}{m!n!}\frac{D^{m+n}}{\lambda_R+\lambda_M}$$

puede comprobarse fácilmente la relación recursiva:

$$G(m,n) = \frac{\lambda_R}{\lambda_R+\lambda_M}G(m-1,n) + \frac{\lambda_N}{\lambda_R+\lambda_M}G(m,n-1) - F(m,n) \qquad (5.17)$$

con valores iniciales:

$$G(m,0) = \frac{1}{\lambda_R+\lambda_M}\left[\left(\frac{\lambda_R}{\lambda_R+\lambda_M}\right)^m - e^{-(\lambda_R+\lambda_n)D}\sum_{h=0}^{m}\left(\frac{\lambda_R}{\lambda_R+\lambda_M}\right)^h \frac{(\lambda_R D)^{m-h}}{(m-h)!}\right] \qquad (5.18)$$

$$G(0,n) = \frac{1}{\lambda_R + \lambda_M}\left[\left(\frac{\lambda_R}{\lambda_R+\lambda_M}\right)^n - e^{-(\lambda_R+\lambda_n)D}\sum_{h=0}^{n}\left(\frac{\lambda_R}{\lambda_R+\lambda_M}\right)^h \frac{(\lambda_R D)^{n-h}}{(n-h)!}\right] \tag{5.19}$$

De esta forma,

$$A_{(i,j,c)(a,b,c)} = G(a-i, b-j), 1 \leq c \leq M, 0 \leq i \leq a, 0 \leq j \leq b \tag{5.20}$$

y su valor puede calcularse utilizando recursivamente (5.17) con los valores iniciales (5.18) y (5.19).

Para calcular ahora $A_{(i,0,0)(a,b,c)}$, con $1 \leq c < M$ debemos notar que para que durante un tiempo de servicio se pase del estado $(i,0,0)$ al (a,b,c), deben haber llegado $a-i$ clientes TR y $b+1$ clientes TNR, con la condición de que el primer cliente TNR haya llegado cuando hay exactamente $c-1$ clientes TR en el sistema. Utilizando la propiedad del proceso de Poisson de que si en un periodo $(0,t)$ se producen n ocurrencias del mismo, éstas se distribuyen uniformemente el el intervalo, la probabilidad de que la primera llegada de un cliente TNR sea la que ocupe la posición c en la cola compartida habiendo llegado $a-i$ clientes TR y $b+1$ clientes TNR en $(0,t)$, viene dada por:

$$\frac{(a-i)!}{(a-c+1)!}\frac{(a+b+1-c)!}{(a+b+1-i)!}(b+1)! \tag{5.21}$$

Procediendo ahora como para el cálculo de $A_{(i,j,c)(a,b,c)}$, llegamos a:

$$A_{(i,0,0),(a,b,c)} = \frac{(a-i)!}{(a-c+1)!}\frac{(a+b+1-c)!}{(a+b+1-i)!}(b+1)!G(a-i,b+1) \tag{5.22}$$

Un razonamiento similar, teniendo en cuenta ahora que para que un cliente TNR ocupe la posición M debe haber llegado cuando en el sistema hay $M-1$ ó más clientes TR nos permite obtener:

$$A_{(i,0,0),(a,b,M)} = \frac{(a-i)!}{(a-M+1)!}\frac{(a+b+2-M)!}{(a+b+1-i)!}G(a-i,b+1) \tag{5.23}$$

Por último, también de modo muy sencillo se deduce que:

$$A_{(i,0,0)(a,0,0)} = G(a-i,0) \tag{5.24}$$

De esta forma, las ecuaciones (5.20), (5.22), (5.23) y (5.24) junto con (5.17) nos proporcionan el esquema recursivo preciso para obtener finalmente las probabilidades en el equilibrio π_{ijk} a partir de (5.13),(5.14),(5.15) y (5.16).

5.4 Probabilidades de pérdida.

Un cliente TR es rechazado por el sistema cuando a su llegada no hay espacio en buffer para alojarlo; ésto ocurre cuando en total en el sistema hay $R+M-1$ cientes TR. Por tanto, la probabilidad de pérdida en el estado estacionario para los clientes TR viene dada por:

$$\gamma_{TR} = \pi_{R+M-1,0,0} + \sum_{c=1}^{M}\sum_{b=0}^{N} \pi_{R+M-1,b,c}$$

Dado que durante un ciclo tratan de acceder al sistema por término medio $\lambda_R E[T]$ clientes TR, si llamamos $L_{TR}(T)$ al número de clientes TR rechazados por ciclo, se tiene, haciendo uso de (5.9):

$$E\left[L_{TR}(T)\right] = \gamma_{TR}\lambda_R E\left[T\right] = \frac{\gamma_{TR}\lambda_R}{(\lambda_R+\lambda_N)\,\pi_{000}}$$

Asimismo, la probabilidad de pérdida en es estado estacionario para los clientes TNR viene dada por:

$$\gamma_{TNR} = \sum_{a=0}^{R+M-1}\sum_{c=1}^{M} \pi_{aNc}$$

y el número medio de clientes TNR rechazados por ciclo es:

$$E\left[L_{TR}(T)\right] = \gamma_{TR}\lambda_N E\left[T\right] = \frac{\gamma_{TR}\lambda_N}{(\lambda_R+\lambda_N)\,\pi_{000}} \tag{5.25}$$

5.5 Tiempos de espera.

Situémonos en nuestro sistema en el preciso instante en que se acaba de completar un servicio (se transmite una célula), que ha dejado el sistema en estado (a, b, c), y sea , con $i \leq a$, el tiempo que aún debe esperar en cola el cliente TR que tiene delante $i-1$ clientes TR. Obviamente, si $i \leq M-1$:

$$\omega_{ibc} = \begin{cases} \sum_{j=1}^{i-1} X_j & si \quad i < c \text{ ó } \quad c = b = 0 \\ \sum_{j=1}^{i} X_j & si \quad i \geq c \geq 1 \end{cases} \tag{5.26}$$

siendo X_j la duración de un tiempo de servicio. Ahora bien, si $i \geq M$, el tiempo de espera de este cliente no se verá afectado por los nuevos clientes TR que lleguen a partir de ahora, pero sí podría verse afectado por los clientes TNR que llegasen a partir de este momento. En efecto, existe la posibilidad de que durante el tiempo que tarde nuestro cliente TR en llegar al buffer compartido, en virtud de la política de gestión empleada puedan acceder a este buffer clientes TNR que llegaron después que él, pero que encontraron ninguna o poca cola en su buffer específico y pudieron entrar pronto en el buffer compartido. Podemos establecer entonces las siguientes relaciones de recurrencia para el tiempo de espera de nuestro cliente TR:

- Si $c = 0$:

$$\omega_{i00} = \begin{cases} X + \omega_{i-1,0,0} & \text{con prob. } \beta \\ X + \omega_{i-1,j,M-1} & \text{con prob. } \beta_{j+1},\ 0 \leq j < N \\ X + \omega_{i-1,N,M-1} & \text{con prob. } \beta^*_{N+1} \end{cases} \tag{5.27}$$

  donde X es la duración de un tiempo de servicio y β_k es la probabilidad de que durante un tiempo de servicio lleguen al sistema k clientes TNR. Por su parte, β^*_kes la probabilidad de que durante un servicio lleguen k ó más clientes TNR.

- Si $c = 1$:

$$\omega_{i01} = \begin{cases} X + \omega_{i,0,0} & \text{con prob. } \beta_0 \\ X + \omega_{i,j,M} & \text{con prob. } \beta_{j+1},\ 0 \leq j < N \\ X + \omega_{i,N-1,M} & \text{con prob. } \beta_N^* \end{cases} \tag{5.28}$$

$$\omega_{ib1} = \begin{cases} X + \omega_{i,b-1+j,M} & \text{con prob. } \beta_j, 0 \leq j < N-b, 0 < b < N \\ X + \omega_{i-1,j,M-1} & \text{con prob. } \beta_{N-b}^* \end{cases} \tag{5.29}$$

$$\omega_{K,N,1} = X + \omega_{K,N-1,M} \tag{5.30}$$

- Si $1 < c \leq M$:

$$\omega_{ibc} = \begin{cases} X + \omega_{i-1,b+j,c-1} & \text{con prob. } \beta_j, 0 \leq j < N-b, 0 \leq b < N \\ X + \omega_{i-1,N,c-1} & \text{con prob. } \beta_{N-b}^* \end{cases} \tag{5.31}$$

$$\omega_{i,N,c} = X + \omega_{i-1,N,c-1} \tag{5.32}$$

De estas ecuaciones de pueden obtenerse las siguientes relaciones de recurrencia para los tiempos medios de espera:

- Si $i \leq M-1$:

$$E\left[\omega_{ibc}\right] = \begin{cases} (i-1)E[X] & \text{si. } i < c \text{ ó } c = b = 0 \\ iE[X] & \text{si } i \geq c \geq 1 \end{cases} \tag{5.33}$$

- Si $i \geq M$:

$$
\begin{aligned}
E[\omega_{i00}] &= E[X] + \beta_0 E[\omega_{i-1,0,0}] + \sum_{j=0}^{N-1} \beta_{j+1} E[\omega_{i-1,j.M-1}] + \beta^*_{N+1} E[\omega_{i-1,N,M-1}] \\
E[\omega_{i01}] &= E[X] + \beta_0 E[\omega_{i,0,0}] + \sum_{j=0}^{N-2} \beta_{j+1} E[\omega_{i,j.M}] + \beta^*_N E[\omega_{i,N-1,M}] \\
E[\omega_{ib1}] &= E[X] + \sum_{j=0}^{N-b} \beta_j E[\omega_{i,b-1.M}] + \beta^*_{N-b} E[\omega_{i,N-1,M}] \text{ con } 0 < b < N \\
E[\omega_{iN1}] &= E[X] + E[\omega_{i,N-1,M}] \\
E[\omega_{ibc}] &= E[X] + \sum_{j=0}^{N-b} \beta_j E[\omega_{i-1,b+j.c-1}] + \beta^*_{N-b} E[\omega_{i-1,N,c-1}] \text{ con } 0 \leq b < N \\
E[\omega_{iN1}] &= E[X] + E[\omega_{i-1,N,c-1}]
\end{aligned}
\tag{5.34}
$$

Las ecuaciones (5.34) junto con los valores iniciales dados por (5.33) permiten obtener recursivamente las esperanzas $E[\omega_{ijk}]$ para todos los estados (i, j, k) posibles. La resolución de estas ecuaciones será más o menos difícil en función de cuál sea la función de distribución de probabilidad del tiempo de servicio. En el caso particular de tiempo de servicio determinista de duración D se tiene:

$$
E[X] = D \quad , \quad \beta_k = e^{-\lambda_N D} \frac{(\lambda_N D)^k}{k!} \quad , \quad \beta^*_k = 1 - \sum_{j=0}^{k-1} e^{-\lambda_N D} \frac{(\lambda_N D)^j}{j!} \tag{5.35}
$$

Ahora bien, las esperanzas obtenidas en (5.34) corresponden a los tiempos medios de espera en cola medidos a partir del momento en que termina un servicio. El tiempo de espera global de un cliente TR que a su llegada encuentra el sistema en estado (a, b, c) es:

- Si $a < M - 1$:

$$
W_{abc} = \begin{cases} X_{RES} + \sum_{i=1}^{a-1} X_i & \text{si.} c = b = 0 \\ X_{RES} + \sum_{i=1}^{a} X_i & \text{si } c > 1 \end{cases} \tag{5.36}
$$

donde X_{RES} es el tiempo de servicio residual que falta para que termine el servicio del cliente que ocupa la cabecera de la cola compartida a la llegada del cliente TR.

- Si $a \geq M-1$:

Si durante el tiempo de servicio residual del cliente que ocupa la cabecera de la cola llegan nuevos clientes TR, éstos no afectan al tiempo de espera del cliente TR que acaba de llegar. Sin embargo, por la misma razón señalada más arriba, sí que afectan los clientes TNR que lleguen durante este tiempo. Si llamamos ak a la probabilidad de que durante el tiempo de servicio residual posterior a la llegada del cliente TR lleguen k clientes TNR, y a la probabilidad de que lleguen k ó más, tenemos:

$$W_{abc} = \begin{cases} X_{RES} + \omega_{a-1,j,M} & \text{con prob. } \alpha_{j+1}, 0 \leq j \leq N & (b=c=0) \\ X_{RES} + \omega_{a-1,N,M} & \text{con prob. } \alpha^*_{N+1} & (b=c=0) \\ X_{RES} + \omega_{a-1,0,0} & \text{con prob. } \alpha_0 & (c=1, b=0) \\ X_{RES} + \omega_{a,j,M} & \text{con prob. } \alpha_{j+1}, 0 \leq j \leq N-2 & (c=1, b=0) \\ X_{RES} + \omega_{a,N-1,M} & \text{con prob. } \alpha^*_N & (c=1, b=0) \\ X_{RES} + \omega_{a,b+j-1,M} & \text{con prob. } \alpha_j, 0 \leq j \leq N-b & (c=1, b>0) \\ X_{RES} + \omega_{a,N,M} & \text{con prob. } \alpha^*_{N-b} & (c=1, b>0) \\ X_{RES} + \omega_{a-1,b+j,c-1} & \text{con prob. } \alpha_j, 0 \leq j \leq N & (c>1) \\ X_{RES} + \omega_{a-1,N,c-1} & \text{con prob. } \alpha^*_{N-b} & (c>1) \\ X_{RES} + \omega_{a-1,0,0} & \text{con prob. } \alpha_0 & (b=c=0) \end{cases} \tag{5.37}$$

De modo similar a como hicimos anteriormente para las ω_{ijk}, podemos ahora hallar a partir de (5.37) el tiempo medio de espera en cola para un cliente TR que a

su llegada encuentra el sistema en el estado (a, b, c):

$$\begin{aligned}
E[W_{a00}] &= E[X_{RES}] + \alpha_0 E[\omega_{a-1,0,0}] + \sum_{j=0}^{N} \alpha_{j+1} E[\omega_{a-1,j,M}] + \alpha^*_{N+1} E[\omega_{a-1,N,M}] \\
E[W_{a01}] &= E[X_{RES}] + \alpha_0 E[\omega_{a,0,0}] + \sum_{j=0}^{N} \alpha_{j+1} E[\omega_{a,j,M}] + \alpha^*_{N+1} E[\omega_{a,N,M}] \\
E[W_{aba}] &= E[X_{RES}] + \sum_{j=0}^{N-b} \alpha_j E[\omega_{a,b+j-1,M}] + \alpha^*_{N-b} E[\omega_{a,N,M}], \text{ con } 0 < b \leq N \\
E[W_{aba}] &= E[X_{RES}] + \sum_{j=0}^{N-b} \alpha_j E[\omega_{a-1,b+j-1,c-1}] + \alpha^*_{N-b} E[\omega_{a-1,N,c-1}], \text{ con } 0 < b \leq N, c \geq 1
\end{aligned} \tag{5.38}$$

Estas esperanzas pueden calcularse recursivamente utilizando como valores iniciales las esperanzas que se obtienen directamente de (5.36):

$$\begin{aligned}
E[W_{a00}] &= E[X_{RES}] + (a-1)E[X] \quad , 1 \leq a \leq M-1 \\
E[W_{abc}] &= E[X_{RES}] + aE[X] \qquad\qquad c > 1
\end{aligned} \tag{5.39}$$

y la condición obvia: $E[W_{000}] = 0$.

Como ya ocurrió con las $E[\omega_{ijk}]$, la obtención de las esperanzas en (5.38) depende de la dificultad de cálculo de las α_k. En el caso particular de servicio determinista de duración D puede probarse que:

$$\alpha_k = \int_0^D e^{-\lambda_N (D-t)} \frac{(\lambda_N (D-t))^k}{k!} \frac{1}{D} dt = \frac{1}{\lambda_N D} \left[1 - e^{-\lambda_N D} \sum_{j=0}^{k} \frac{(\lambda_N D)^j}{j!} \right] \tag{5.40}$$

En este caso, además $E[X_{RES}] = D/2$.

De esta forma, estamos ya en condiciones de poder calcular el tiempo medio de espera de un cliente TR arbitrario. Simplemente condicionando por el estado del sistema a la llegada de este cliente tenemos:

$$E[W^{TR}] = \sum_{(a,b,c)} E[W_{abc}] \pi_{abc} \tag{5.41}$$

El valor de esta esperanza se calcula haciendo uso de (5.38) y (5.39), con las probabilidades estacionarias halladas en (5.13,5.14,5.15) y (5.16).

Por último, el tiempo medio de espera para los clientes TNR puede hallarse utilizando la fórmula de Little. En este sistema concreto, la fórmula de Little adopta la forma:

$$E[N_{TR}] + E[N_{TNR}] = [\lambda_R(1-\gamma_{TR}) + \lambda_N(1-\gamma_{TNR})]\, E[W] \tag{5.42}$$

siendo $E[W] = p_{TR}E[W^{TR}] + p_{TNR}E[W^{TNR}]$ y donde p_{TR} y p_{TNR} denotan los porcentajes de clientes TR y TNR respectivamente, que han pasado por el sistema durante un ciclo. El número medio de clientes de TR que han pasado por el sistema durante un ciclo es $\lambda_R E[T]$, por lo que $p_{TR} = \lambda_R/(\lambda_R + \lambda_N)$. Los números medios $E[N_{TR}]$ y $E[N_{TNR}]$ pueden calcularse fácilmente a partir de las π_{ijk}:

$$\begin{aligned} E[N_{TR}] &= \sum_{a=0}^{R+M-1} a\pi_{a00} + \sum_{a=0}^{R+M-1} a \sum_{b=0}^{N}\sum_{c=1}^{M} \pi_{abc} \\ E[N_{TNR}] &= \sum_{a=0}^{R+M-1}\sum_{b=0}^{N}\sum_{c=1}^{M}(b+c)\pi_{abc} \end{aligned} \tag{5.43}$$

y de (5.41) y (5.42) se tiene que:

$$E[W^{TNR}] = \frac{(E[N_{TR}] + E[N_{NTR}])(\lambda_R + \lambda_N)}{[\lambda_R(1-\gamma_{TR}) + \lambda_N(1-\gamma_{TNR})]\,\lambda_N} - \frac{\lambda_R}{\lambda_N}E[W^{TR}] \tag{5.44}$$

5.6 Optimización del rendimiento del sistema.

Como se ha dicho en el epigrafe 1, los valores de M, N y R deben elegirse de modo que se minimice el coste medio por unidad de tiempo a largo plazo que suponen los retardos para los clientes TR y las pérdidas para los TNR:

$$\phi(M,N,R) = \lim_{t\to\infty} \frac{\alpha E\left[Y_{RT}(t)\right] + \beta E\left[L_{TNR}(t)\right]}{t} \tag{5.45}$$

Este mínimo debe hallarse con la restricción de que $E[W_{TNR}] < W_1$ y $\gamma_{TR} < g_1$. Dado que el proceso $(N_{TR}(t), N_{TNR}(t), S_{TNR}(t))$ es regenerativo, la teoría de la renovación nos permite calcular el límite anterior de la forma:

$$\phi(M,N,R) = \frac{\alpha E\left[Y_{RT}(T)\right] + \beta E\left[L_{TNR}(T)\right]}{T} \tag{5.46}$$

esto es, el coste medio para el sistema a largo plazo coincide con el coste medio durante un periodo de renovación. La ecuación (5.9) nos permite calcular $E[T]$, y la ecuación (5.25) nos proporciona $E[L_{TNR}(T)]$. Nos falta por determinar solamente $E[Y_{TR}(T)]$, el tiempo de espera acumulado por todos los clientes TR atendidos durante un periodo de renovación. Esta cantidad se halla fácilmente como:

$$E\left[Y_{TR}(T)\right] = \lambda_R(1-\gamma_{TR})E\left[T\right]E\left[W^{RT}\right] \tag{5.47}$$

esto es, el número medio de clientes TR que llegan durante un periodo de renovación, multiplicado por el tiempo medio de espera de cada uno, ya obtenido en (5.41).

Dado que todos los términos que aparecen en (5.46) se han calculado numéricamente, no podemos hallar una fórmula explícita para obtener los valores de M, N y R que minimizan (5.46), y éstos han de hallarse mediante un algoritmo de búsqueda que vaya incrementando y/o decrementando sucesivamente los valores de estos parámetros cumpliendo además las restricciones $E[W_{TNR}] < W_1$ y $\gamma_{TR} < g_1$. Debe señalarse que la forma recursiva de computar todos los elementos necesarios para el cálculo de las funciones involucradas en (5.46) simplifica mucho los cálculos necesarios en el algoritmo de búsqueda, ya que los términos obtenidos en una iteración pueden ser reutilizados en la siguiente sin necesidad de recalcularlos.

5.7 Análisis de los resultados

Hemos llevado a cabo el cálculo numérico de las expresiones obtenidas en los apartados anteriores para las probabilidades de pérdida y tiempos medios de espera de ambas clases de clientes (en tiempo real y en tiempo no real) bajo diversas condiciones de tráfico y diferentes tamaños de buffers y tasas de servicio. En esta sección mostramos un breve resumen del comportamiento observado en el sistema, y las conclusiones que de ello se derivan para el correcto dimensionamiento del mismo.

5.7.1 *Caso 1:* $\lambda_R >> \lambda_N$, $\rho = 0.95$

Este caso representa una situación en la que el tráfico en tiempo real es muy superior al otro. Este podría ser el caso de un sistema orientado básicamente a la transmisión de voz o vídeo, pero que reserva parte de su ancho de banda para transmisión de datos. El caso $\rho = 0.95$ supone además que la tasa de servicio del sistema es muy similar a la tasa de llegada de clientes, lo que significa que nos encontramos ante un sistema cargado donde pueden producirse colas de longitud importante.

En estas condiciones, la causa de que el sistema se encuentre cargado es precisamente el tráfico en tiempo real. El tiempo de espera de los paquetes de este tráfico se incrementa con el tamaño R de su buffer reservado, y también con el tamaño M del buffer compartido. Dada la baja tasa de llegadas relativa de los clientes TNR, éstos prácticamente no influyen en el retardo del tráfico TR, que se produce precisamente como consecuencias de la alta tasa de llegada relativa de los clientes TR. De esta forma, la reducción en el tiempo de espera TR sólo puede conseguirse a costa de reducir los tamaños de los buffers M y R, lo que obviamente significa un incremento en las pérdidas de TR. Debe observarse, además, que si el tamaño del buffer compartido se mantiene alto, éste se verá ocupado por colas relativamente largas de clientes TR, lo que incrementa notablemente el tiempo de espera de cada cliente TNR que accede a ese buffer.

En la figura 2 se han representado las distintas variables de interés (probabilidades de pérdida y tiempos medios de espera) frente a la suma de los tamaños de los buffers M y R. Los puntos de distintos colores representan el valor de N, el tamaño del buffer para los clientes TNR (Negro N =1, Rojo $N = 2$, Verde $N = 3$, Azul $N = 4$, Celeste $N = 5$, etc). Puede apreciarse claramente que la probabilidad de pérdida para los TNR disminuye a medida que aumenta N, y que, para valores de N pequeños depende de $M+R$, creciendo con este valor hasta alcanzar una tendencia asintótica; para valores de N grandes se aprecia que la probabilidad de

pérdida apenas depende de $M + R$, siendo pequeña y aproximadamente constante. El tiempo de espera de los TNR, como hemos señalado más arriba, se incrementa con el valor de N, y también con $M + R$, si bien, a partir de cierto valor de $M + R$ en adelante ya este tiempo medio se estabiliza (debido a que a partir de cierto tamaño de $M + R$ en adelante, la cola de clientes TR en estos buffers ya no se incrementa más, lo que indirectamente redunda en que no se incrementa el tiempo de espera para los TNR). En cuanto a las probabilidade de pérdida para los clientes TR se aprecia que disminuyen con $M + R$ sin que el valor de N les afecte. A su vez, el tiempo medio de espera para esta clase de clientes se incrementa con el valor de $M + R$, apareciendo en este caso una cierta dependencia con N que se vuelve más acusada a medida que aumenta $M + R$, observándose, para valores elevados de $M + R$ los menores tiempos de espera cuando más pequeño es N.

Para aislar el efecto por separado de los tamaños M y R podemos observar las figuras 3, 4, 5 y 6, correspondientes a los casos $N = 1$, $N = 2$, $N = 4$ y $N = 8$,respectivamente. Ahora se ha representado el valor de M en abcisas y los colores se han utilizado para los distintos valores de R (Negro $R = 1$, Rojo $R = 2$, Verde $R = 3$, Azul $R = 4$, Celeste $R = 5$, etc.). Puede apreciarse que, para valores de M y N pequeños, las menores pérdidas en clientes TNR se producen cuanto más pequeño sea R, si bien este efecto se va amortiguando a medida que crece N. De hecho vemos que para $N = 8$ las pérdidas TNR son siempre pequeñas cualesquiera que sean los valores de M y R. Algo similar ocurre con los tiempos medios de espera de los clientes TNR, coincidiendo los tiempos de espera más cortos en esta clase con las mayores probabilidades de pérdida en la misma. En cuanto a los clientes en tiempo real, observamos que para su tasa de pérdidas es decreciente tanto en R como en M. Si se deseara reducir esta tasa, cuando el valor de M es pequeño, debe elegirse un R grande, y viceversa, si M es grande puede elegirse un R pequeño. En lo que se refiere a los tiempos de espera de esta clase de clientes se observa también que son crecientes tanto con M como con R.

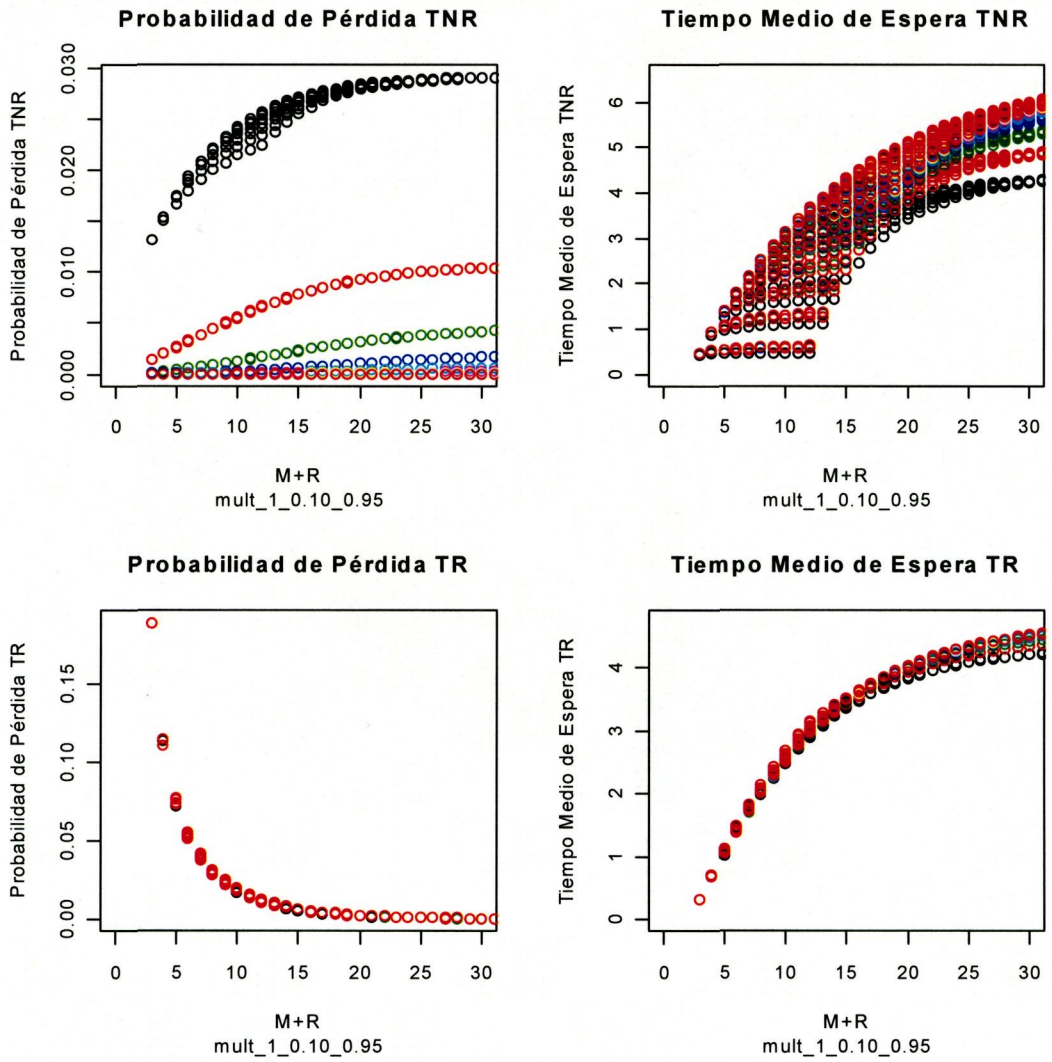

Figura 2 Probabilidades de pérdida y tiempos medios de espera frente a la suma de los tamaños de los buffers M y R. (Negro $N = 1$, Rojo $N = 2$, Verde $N = 3$, Azul $N = 4$, Celeste $N = 5$, etc.)

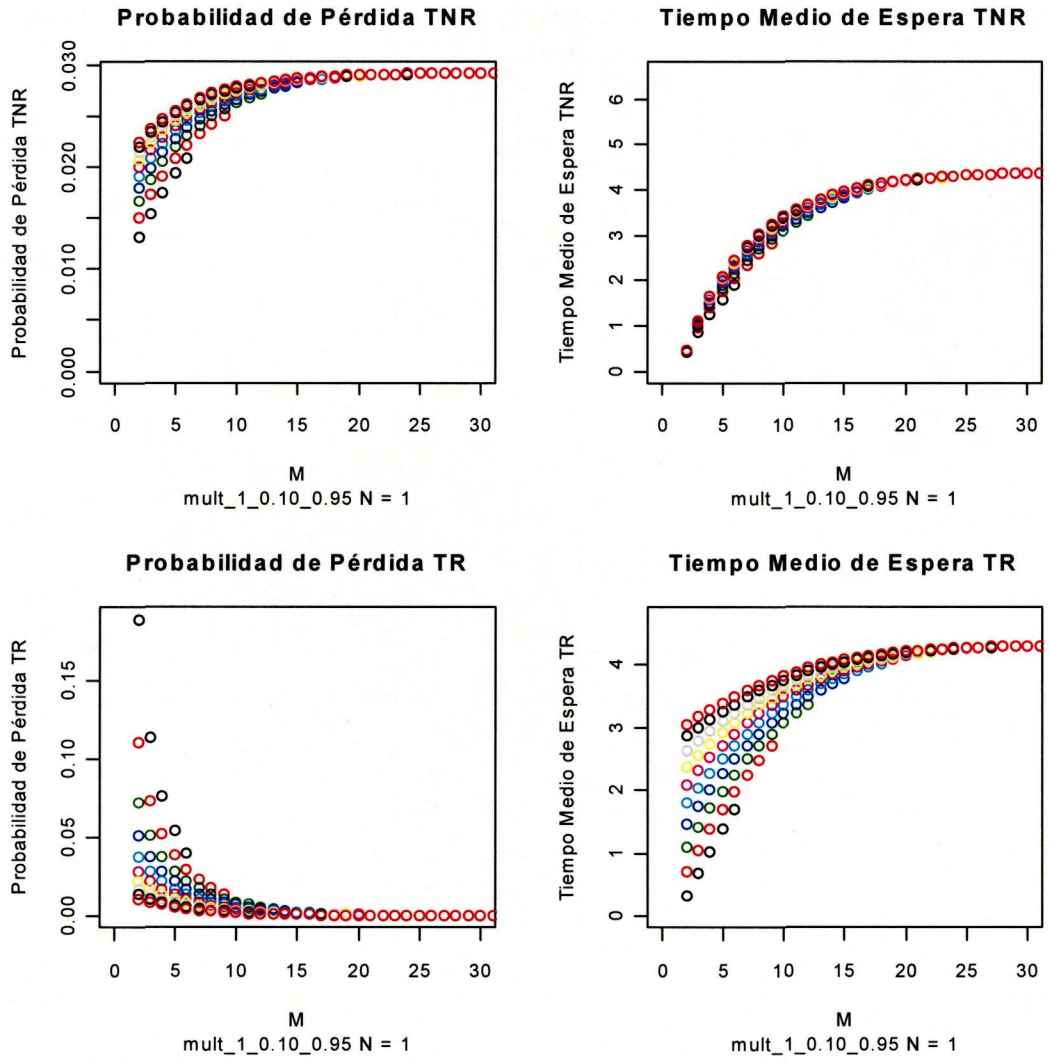

Figura 3 Caso $N = 1$. Probabilidades de pérdida y tiempos medios de espera frente a la suma de los tamaños de los buffers M y R. (Negro $N = 1$, Rojo $N = 2$, Verde $N = 3$, Azul $N = 4$, Celeste $N = 5$, etc.)

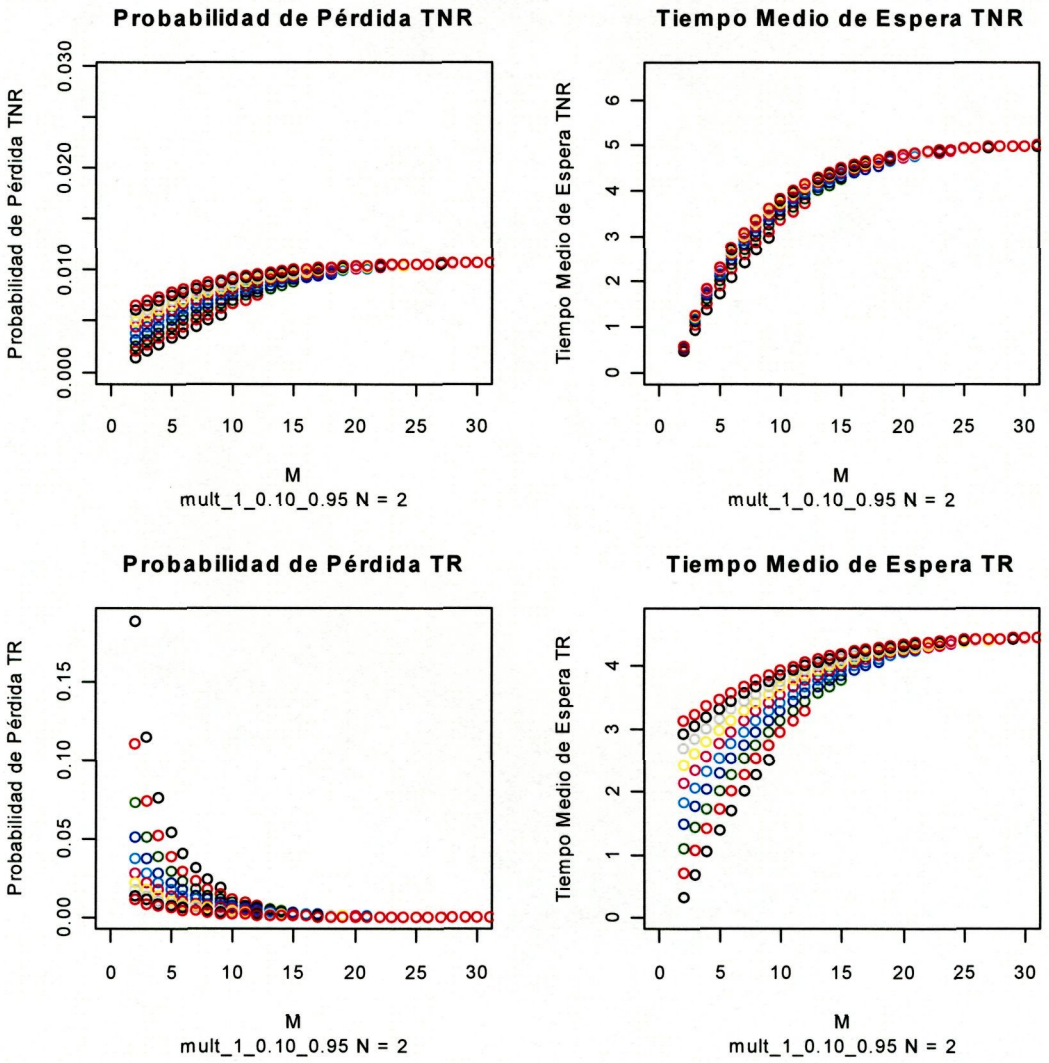

Figura 4 Caso $N = 2$. Probabilidades de pérdida y tiempos medios de espera frente a la suma de los tamaños de los buffers M y R. (Negro $N = 1$, Rojo $N = 2$, Verde $N = 3$, Azul $N = 4$, Celeste $N = 5$, etc.)

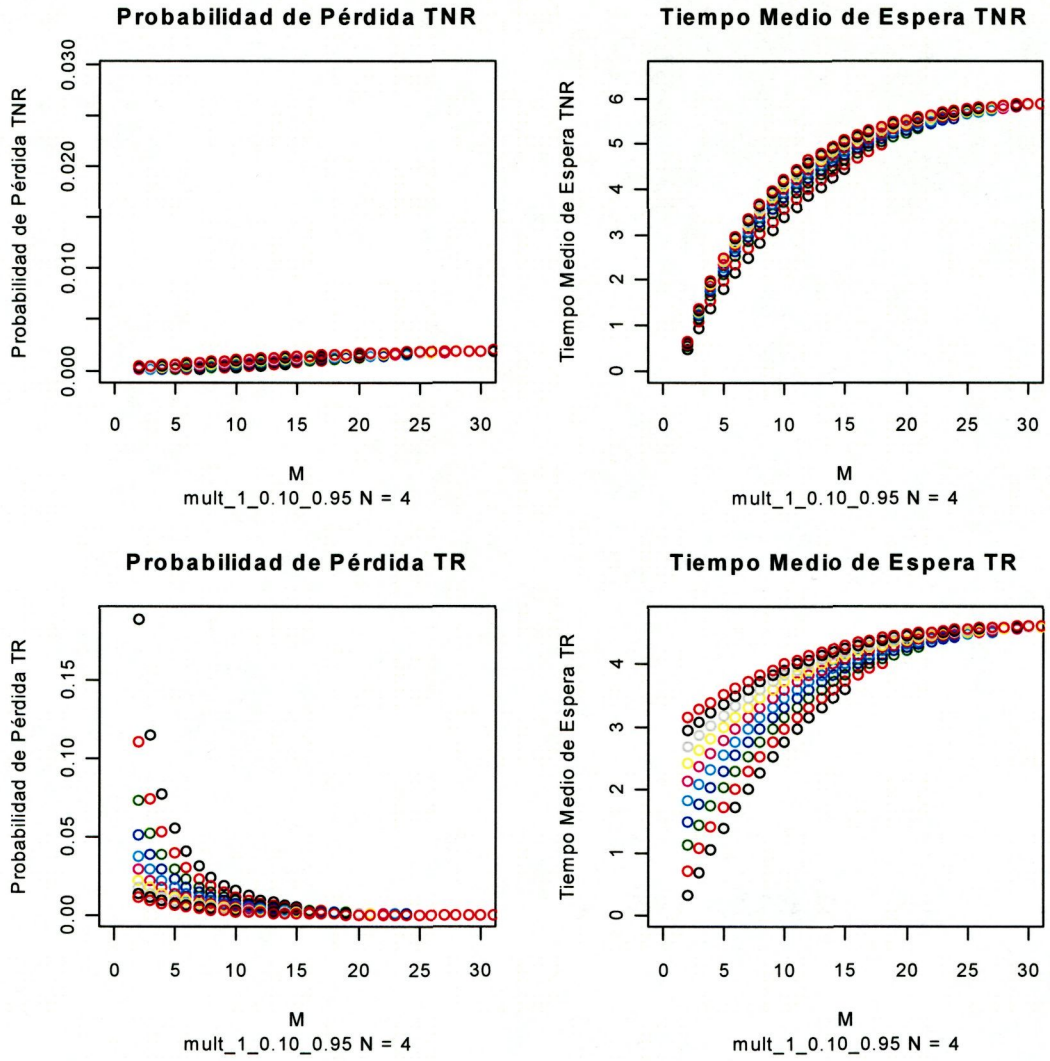

Figura 5 Caso $N = 4$. Probabilidades de pérdida y tiempos medios de espera frente a la suma de los tamaños de los buffers M y R. (Negro $N = 1$, Rojo $N = 2$, Verde $N = 3$, Azul $N = 4$, Celeste $N = 5$, etc.)

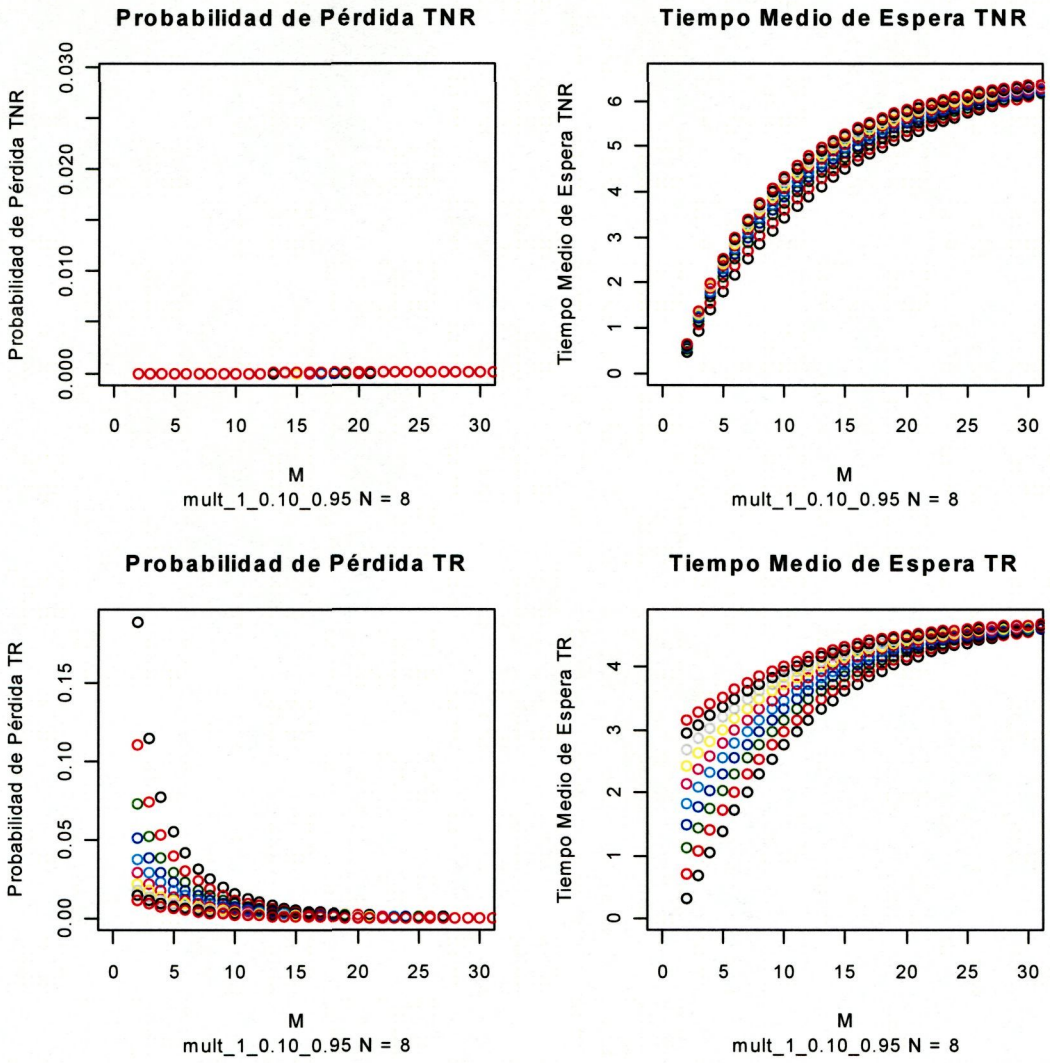

Figura 6 Caso $N = 8$. Probabilidades de pérdida y tiempos medios de espera frente a la suma de los tamaños de los buffers M y R. (Negro $N = 1$, Rojo $N = 2$, Verde $N = 3$, Azul $N = 4$, Celeste $N = 5$, etc.)

De esta forma, observamos que, en estas condiciones, el diseño óptimo del sistema consiste en:

- Un buffer compartido de tamaño reducido, que acortaría los tiempos de espera de los clientes TNR, a la vez que no causaría gran impacto en los tiempos de espera de los TR, ya que los TNR llegan con una tasa relativa muy baja.

- Un buffer específico para los clientes TR con un tamaño R ajustado de tal manera que el tiempo de espera de estos clientes no sea excesivo, a la vez que las pérdidas que necesariamente se producen no superen el umbral preespecificado.

- Un buffer específico para los clientes TNR que garantice que las pérdidas para esta categoría serán reducidas; dado que estamos en condiciones en las que λ_N es pequeño, dicho buffer podrá tener un tamaño N no demasiado grande.

Las expresiones obtenidas en este capítulo para probabilidades de pérdida y tiempos medios de espera permiten determinar exactamente los valores M, N, y R que cumplan los requisitos especificados.

5.7.2 Caso 2: $\lambda_R << \lambda_N$, $\rho = 0.95$

Ahora nos encontramos ante un sistema en el que la tasa de llegadas dominante corresponde al tráfico TNR, como ocurre en redes orientadas fundamentalmente a la transmisión de datos y que pueden emplearse de modo más o menos esporádico para la transmisión de voz o video. Asimismo hemos considerado una intensidad de tráfico elevada (0.95) lo que, al igual que en el caso anterior, nos indica que el sistema se encuentra cargado, con un servidor cuya velocidad de servicio se equipara prácticamente a la velocidad con que llegan los clientes, que en su mayor parte son de clase TNR. Por ello podemos esperar que se poduzcan colas largas de clientes TNR, lo que significa que para evitar pérdidas en esta categoría deberemos contar con un buffer específico para este tráfico con un tamaño N holgado, aún a costa de incrementar los tiempos de espera para esta clase de clientes.

De nuestro análisis de los resultados numéricos obtenidos con el modelo, que se muestran en la figura 7, podemos observar que la tasa de pérdidas en el tráfico TNR depende sólo de N y es bastante insensible al tamaño del buffer compartido M, y al tamaño del buffer R de los clientes TR. No obstante, observamos que el tiempo de espera de los clientes TR (que, recordemos, nos interesa que sea reducido), sí que se incrementa con N, sobre todo si el tamaño del buffer compartido es pequeño. Esto es fácil de entender, ya que al ser muy elevada la tasa de llegadas de clientes TNR, en el buffer compartido habrá casi siempre un cliente TNR esperando ser atendido. Si nos situamos en el caso extremo de que fuese $M = 2$, ello signficaría que la cola de clientes TR se formaría en su buffer específico y cada cliente TR que accediera al buffer compartido debería esperar casi siempre por un cliente TNR. Ello terminaría dando lugar a tiempos de espera inaceptables para el tráfico TR.

Por tanto, el diseño óptimo de este sistema sería el siguiente:

- Un buffer compartido con un tamaño M adecuado a la tasa de llegadas (que es baja) de los clientes TR y que garantice que sus tiempos de espera no sean elevados.

- Un buffer específico para los clientes TR que no debe ser demasiado grande ya que con su baja tasa de llegadas podrían caber casi siempre en el buffer compartido; esa baja tasa de llegadas, a su vez, garantiza que las pérdidas no serán elevadas.

- Un buffer específico para los clientes TNR que sí que debe tener el tamaño suficiente para que puedan esperar sin que se produzcan pérdidas significativas.

Al igual que en el caso anterior, las expresiones obtenidas en este capítulo para probabilidades de pérdida y tiempos medios de espera permiten determinar exactamente los valores M, N, y R que cumplan los requisitos especificados.

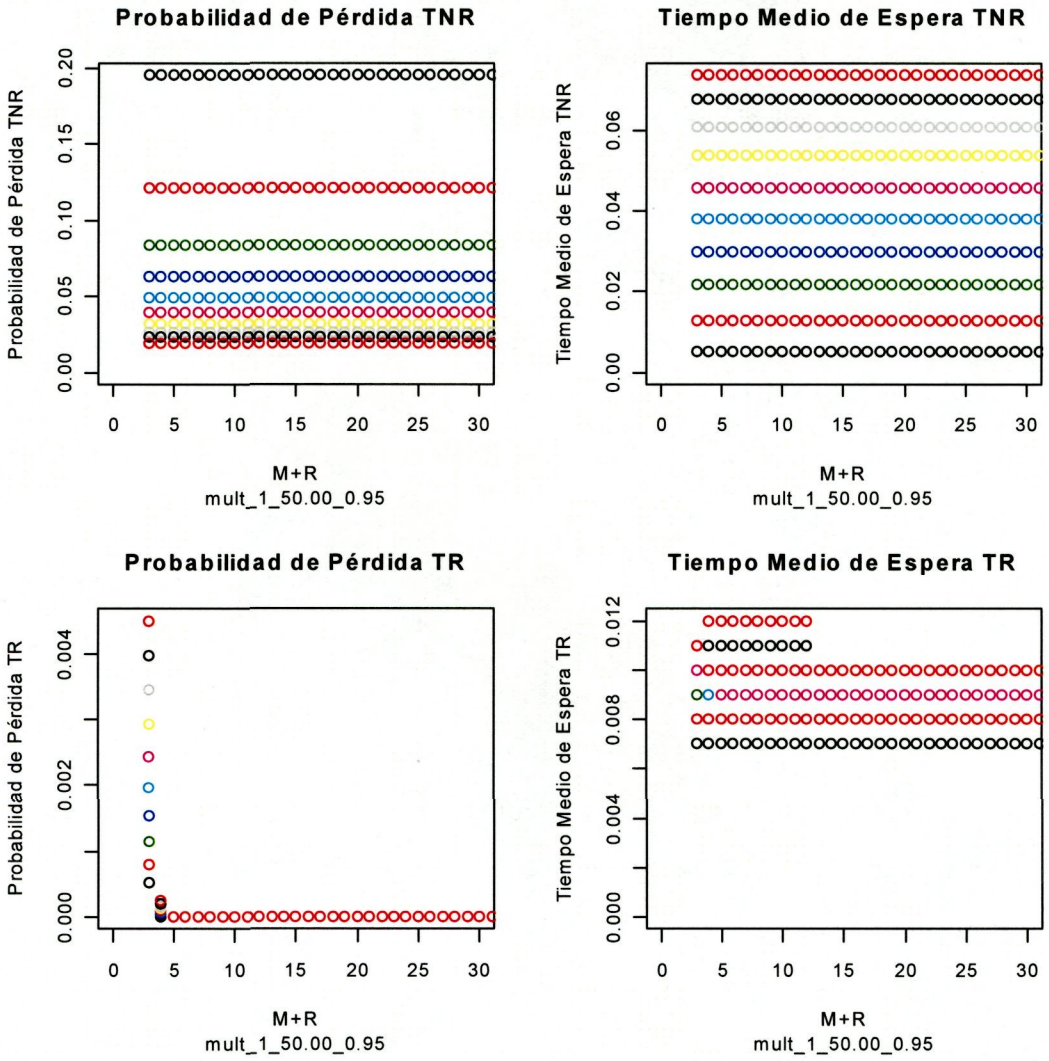

Figura 7 Probabilidades de pérdida y tiempos medios de espera frente a la suma de los tamaños de los buffers M y R.

5.7.3 *Caso 3:* $\lambda_R \cong \lambda_N$, $\rho = 0.95$

Analizamos ahora algunos caso en que las tasas λ_R y λ_M son comparables, siendo también ahora alta la intensidad de tráfico, del orden del 95%. Hemos considerado el caso $\lambda_N = 0.5\lambda_R$ (figura 8) y $\lambda_N = 2\lambda_R$ (figura 9). Puede apreciarse que en los dos casos el comportamiento del sistema es similar. La probabilidad de pérdida de los clientes TNR disminuye a medida que aumenta N. Asimismo, cualquiera que sea el valor de N, las tasas de pérdida más bajas para esta clase de clientes se alcanzan cuando el tamaño $M + R$ es pequeño, si bien a partir de un cierto valor de $M + R$ en adelante estas tasas se estabilizan. El tiempo medio de espera para los clientes TNR aumenta con N y con $M + R$, y al igual que en el caso anterior, a partir de cierto valor de $M + R$ en adelante este tiempo se estabiliza. Por tanto, el mejor rendimiento ofrecido por el sistema para los clientes TNR se alcanza para valores de N grandes y valores de $M + R$ pequeños.

En cuanto a los clientes TR, puede apreciarse también que los menores tiempos medios de espera corresponden a valores $M + R$ pequeños, si bien a dichos valores corresponden también tasas de pérdida altas. Incrementar ligeramente el valor de $M + R$ redunda en un rápido decrecimiento de la tasa de pérdidas de este tráfico a costa de cierta ralentización en sus tiempos de espera. Puede observarse que cuando se incrementa $M + R$, los tiempos de espera más cortos para el tráfico TR se obtienen para los valores de N más pequeños.

Para evaluar el efecto por separado de M y R, observemos las figuras 10, 11, 12; todos ellos obtenidos para $N = 10$, representando M en abcisas y utilizando un color distinto para cada valor de R:

En lo que se refiere al tráfico TNR, vemos que tanto las probabilidades de pérdida como los tiempos medios de espera son crecientes tanto con M como con R; para cada M fijo las pérdidas y tiempos de espera crecen con R; para cada R fijo dichas variables crecen con M. Como puede observarse se alcanza una tendencia asintótica relativamente pronto; a partir de ciertos valores M_0 y R_0 (que dependen

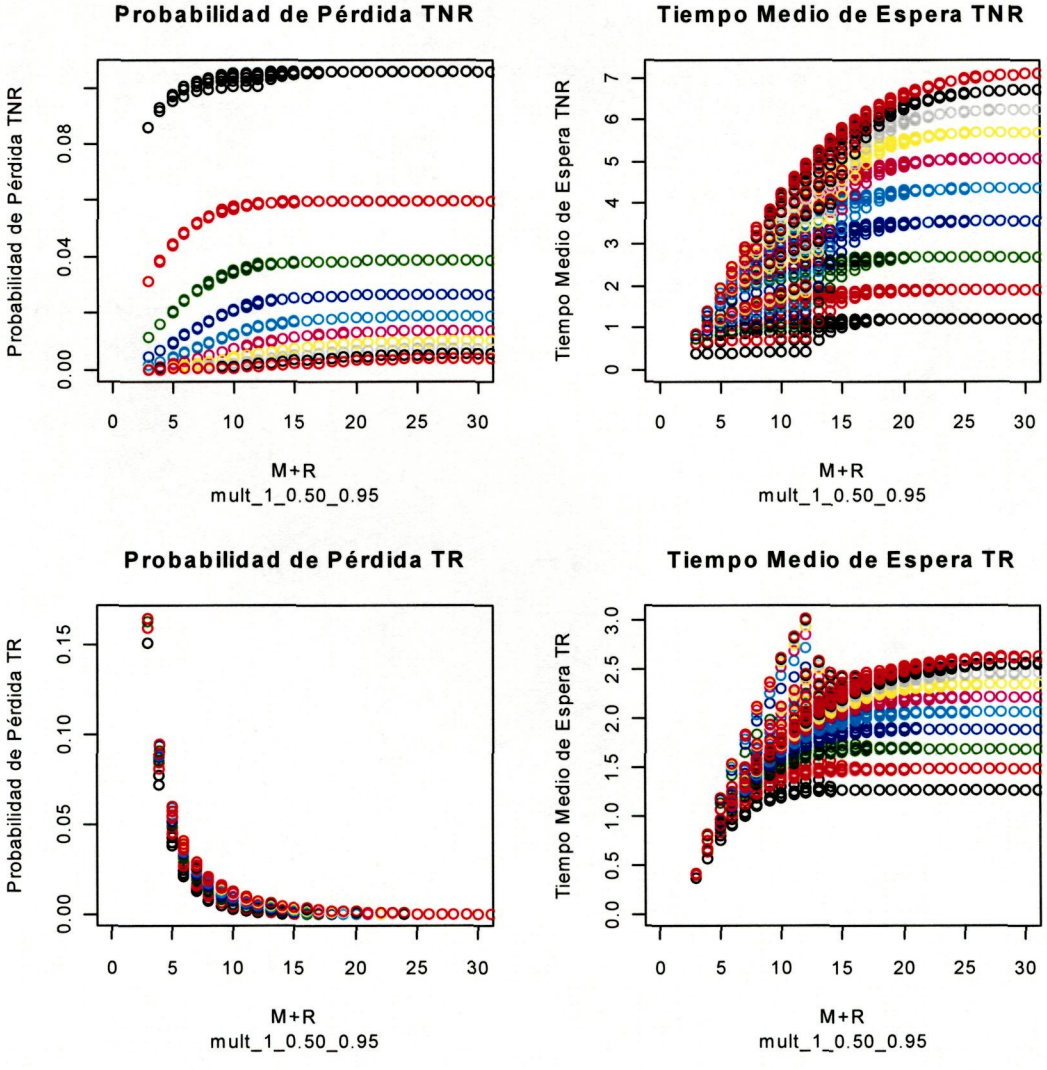

Figura 8 $\lambda_N = 0.5\lambda_R$. Probabilidades de pérdida y tiempos medios de espera frente a la suma de los tamaños de los buffers M y R.

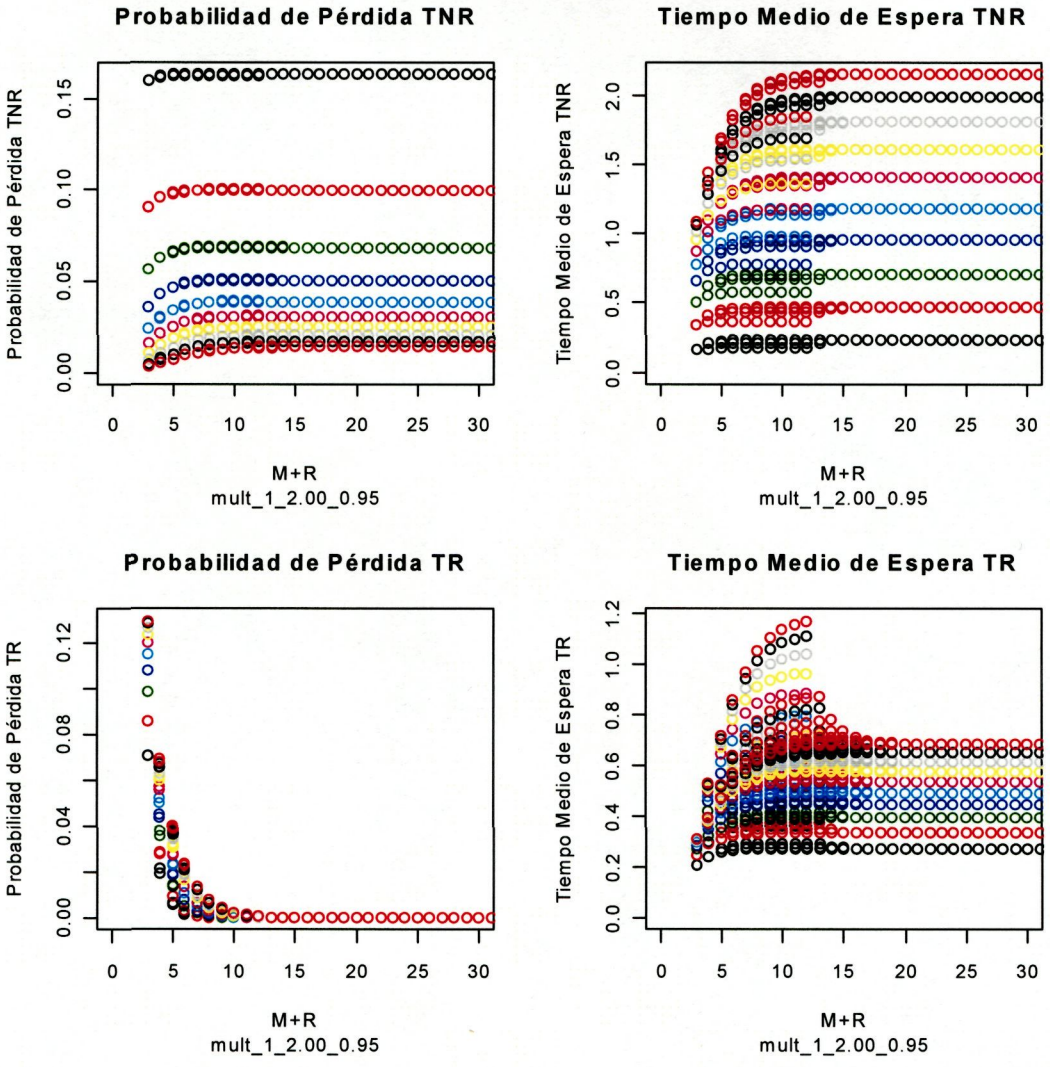

Figura 9 $\lambda_N = 2\lambda_R$. Probabilidades de pérdida y tiempos medios de espera frente a la suma de los tamaños de los buffers M y R.

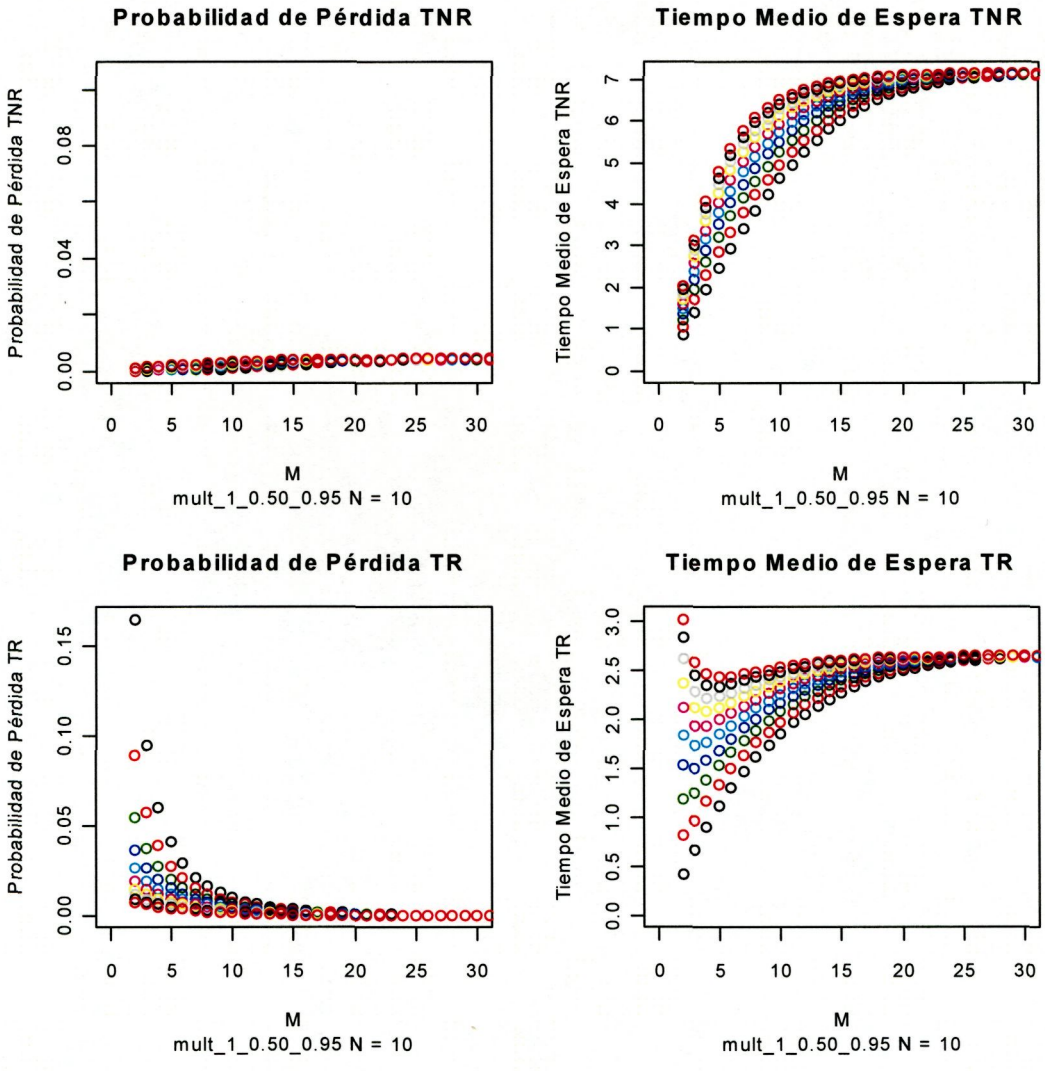

Figura 10 Probabilidades de pérdida y tiempos medios de espera frente M.

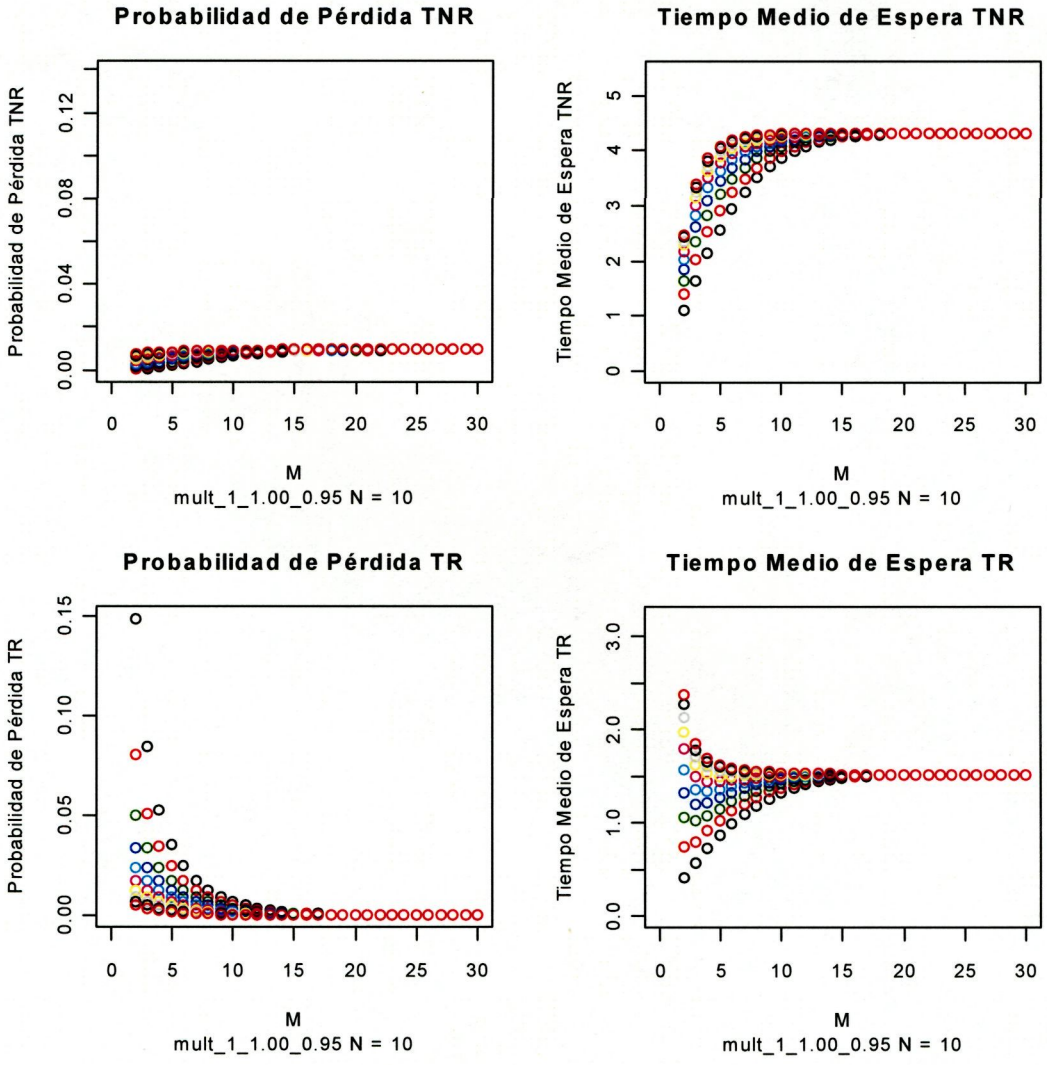

Figura 11 Probabilidades de pérdida y tiempos medios de espera frente M.

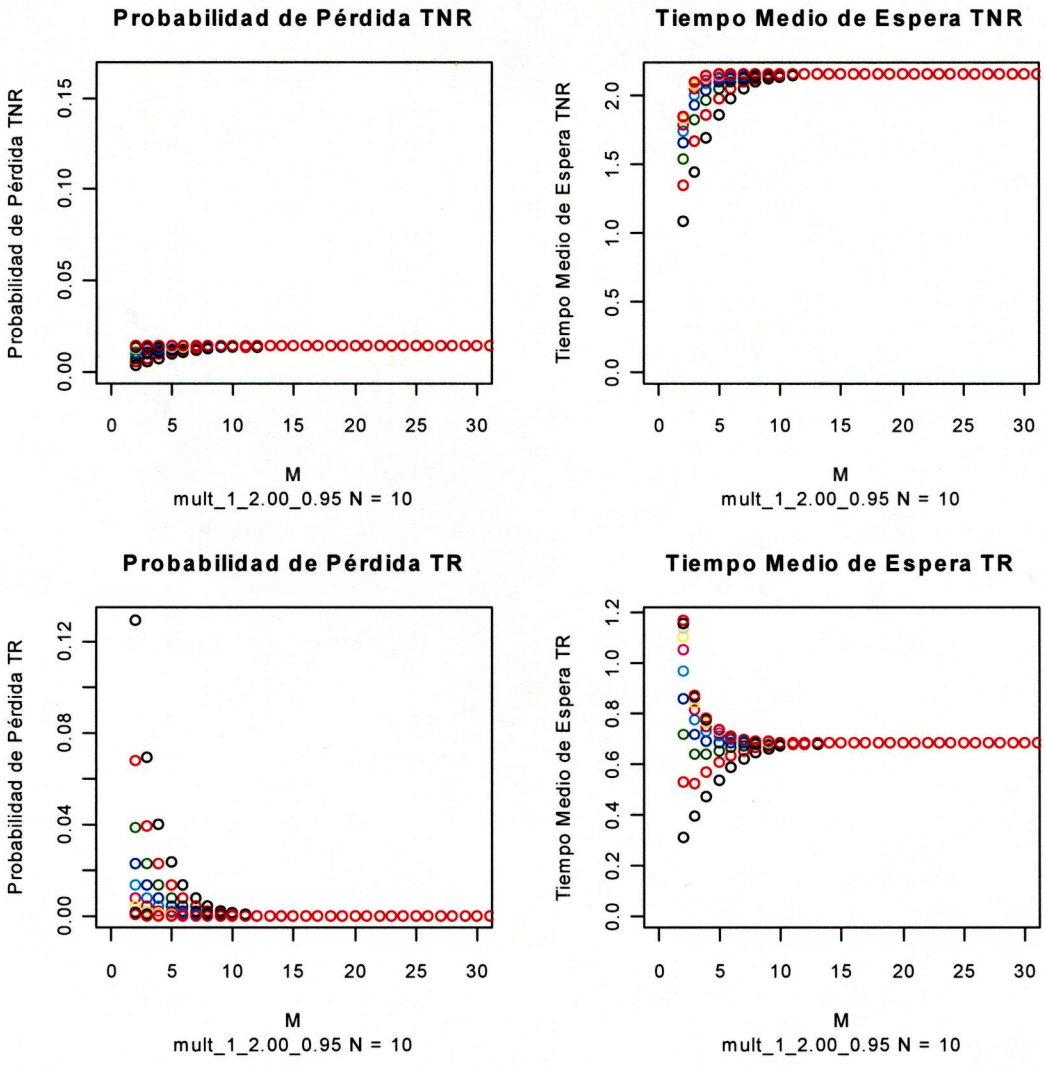

Figura 12 Probabilidades de pérdida y tiempos medios de espera frente M.

de λ_R, λ_N y μ) ni las pérdidas ni los tiempos medios de espera se incrementan por mucho que crezcan M y R. Ello se debe a que con esos valores M_0 y R_0 es posible acumular todo el tráfico TR entrante y mayores valores de M y R no dan lugar a que se produzcan colas mayores de este tipo de tráfico. Dada la forma de operar del sistema, la presencia de más clientes TR a la fuerza repercute en mayores tasas de pérdida y tiempos de espera para los clientes TNR (al menos si el tamaño N es pequeño. Si N es grande, las pérdidas en TNR son pequeñas aunque sean grandes sus tiempos de espera).

Asimismo, en lo que se refiere al tráfico TR se observa que los tiempos de espera más cortos se obtienen cuando M es pequeño y R es también pequeño: obviamente un M pequeño da lugar a que no se incrementen demasiado los tiempos de espera causados por la entrada de clientes TNR; y un R pequeño produce que no crezcan los tiempos de espera causados por el propio tráfico TR. Es evidente también que si se eligen M y R pequeños, se alcanzan las tasas de pérdida de tráfico TR más altas. Si estas pérdidas resultan excesivas, se observa que, manteniendo M fijo, disminuyen eligiendo valores de R más altos; no obstante, a partir de cierto valor de R en adelante es preferible incrementar el valor de M ya que ello da lugar a menores tiempos de espera. Este efecto es tanto más acusado cuanto mayor sea λ_R en relación con λ_N; en efecto, cuanto mayor sea la tasa λ_R respecto a λ_N ello quiere decir que el incremento en el tiempo de espera de los TR se debe precisamente a los clientes TR que se acumulan; bajar el tamaño de R redunda en que se limita el tamaño de las colas TR (incrementando obviamente sus pérdidas) y por tanto se mejoran los tiempos de espera.

5.7.4 Caso 4: $\rho << 0.95$

Los comportamientos que hemos observado hasta ahora con intensidades de tráfico elevadas se mantienen, aunque en menor medida, cuando la intensidad de tráfico ρ se reduce. En general la disminución de ρ lleva aparejada la reducción global de las tasas de pérdida y los tiempos medios de espera en las dos clases de clientes,

que llegan prácticamente a anularse para valores de ρ muy pequeños (inferiores a 0.10). La interpretación de este resultado es obvia: valores de ρ pequeños significan que el servidor opera muy rápido con respecto a la tasa de llegada de clientes, lo que significa que prácticamente nunca habrá cola y aunque los tamaños de buffer M, N y R sean pequeños rara vez llegan a desbordarse, con lo que apenas se producen pérdidas.

5.8 Conclusión.

En este capítulo hemos analizado el rendimiento de una política de gestión de colas en un conmutador de red tal como se muestra en la figura 1, que recibe dos clases de tráfico con distintos requisitos de calidad en lo que se refiere a retardos y pérdidas. Hemos visto como, de forma recursiva, es posible calcular numéricamente las probabilidades de estado de este sistema, y a partir de las mismas hemos mostrado como obtener los retardos medios de ambas clases de tráfico, así como sus tasas de pérdida. Todas estas cantidades intervienen en el cálculo del rendimiento del sistema a través de una función objetivo que representa una ponderación del coste causado por los retardos del tráfico prioritario (en tiempo real) y las pérdidas causadas al tráfico no prioritario (tiempo no real). Los valores de los parámetros óptimos de control del sistema se obtienen finalmente mediante la aplicación de un algoritmo adecuado que minimice la función objetivo. El diseño de un algoritmo de optimización cuyo objetivo sea minimizar el coste en que se incurre por pérdidas en tráfico TNR y por retardos en tráfico TR, a la vez que se respetan restricciones relativas a la tasa máxima admisible de pérdidas en tráfico TR y retardos en tráfico TNR, es una tarea compleja dada la dimensión del espacio de búsqueda de los valores óptimos de M, N y R. Las observaciones cualitativas que hemos realizado en el análisis de los resultados numéricos que hemos obtenido con la implementación de nuestro modelo sugieren estrategias que permitirían dirigir de manera adecuada la búsqueda de la solución óptima haciendo más eficientes los algoritmos de optimización. Debe

señalarse también que de nuestro análisis puede deducirse que para determinados valores λ_N, λ_R y μ fijos, pueden no existir valores de M, N y R que den lugar a un adecuado rendimiento de un multiplexor como el aquí descrito: así, por ejemplo, si el sistema está muy cargado (valor de ρ muy alto) es muy posible que ninguna combinación de M, N y R produzca, a la vez, tiempos de espera y tasas de pérdida admisibles para cada clase de clientes (los tiempos de espera reducidos se lograrían a costa de tasas de pérdida muy altas; y a la inversa, las tasas de pérdida bajas se conseguirían a costa de tiempos de espera excesivos). En tales condiciones, la única solución pasa por incrementar la velocidad del servidor (o el número de servidores).

Bibliografía

1. D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes.* Springer-Verlag, 1988.
2. P. Brémaud, *Point Processes and Queues.* Springer-Verlag, 1981.
3. D. R. Cox and V. Isham, *Point Processes.* Chapman Hall, 1980.
4. D. L. Snyder and M. I. Miller, *Random Point Processes in Time and Space.* Springer, segunda ed., 1991.
5. L. Wolstenholme, *Reliability Modelling : A Statistical Approach.* Chapman Hall / CRC, 1999.
6. A. J. Lotka, "A contribution to the theory of self-renewing aggregates, with special reference to industrial replacement," *Annal of Mathematics Statistics*, vol. 10, pp. 1–25, 1939.
7. S. M. Ross, *Probability Models.* Academic Press, 1993.
8. D. R. Cox, *Renewal Theory.* London: Methuen, 1961.
9. E. Çinlar, "Introduction to stochastic processes," 1975.
10. R. F. Drenik, "The failure law of complex equipment," *Journal of Society Industrial of Applied Mathematic*, vol. 8, pp. 680–690, 1960.
11. B. B. Griglelionis, "On the convergence of sums of random step processes to a poisson process," *Theory of Probability Applied*, vol. 8, pp. 177–182, 1963.
12. Y. S. Chow, H. Robbins, and D. Siegmund, *Great Expectations: The Theory of Optimal Stopping.* Houghton-Mifflin,New York, 1971.
13. H. Ascher and H. Feingold, *Repairable Systems Reliability*, vol. 7 of *Lecture Notes in Statistics.* Marcel Dekker, Inc., 1984.

14. R. E. Barlow and F. Proschan, *Mathematical Theory of Reliability.* John Wiley Sons, 1965.

15. B. Bergman, "On reliability theory and its applications," *Scandinavian Journal of Statistics*, vol. 12, pp. 1–41, 1985.

16. M. S. Filkelstein, "Some notes on two types of minimal repair," *Advances in Applied Probability*, vol. 24, pp. 226–228, 1992.

17. M. Brown and F. Proschan, "Imperfect repair," *Journal of Applied Probability*, vol. 20, pp. 851–859, 1983.

18. H. W. Block, W. S. Borges, and T. H. Savits, "Age-dependent minimal repair," *Journal of Applied Probability*, vol. 22, pp. 370–385, 1985.

19. C. Dorado, M. Hollander, and Y. Sethuraman, "Nonparametric estimation for a general repair model," *Annals of Statistics*, vol. 25, pp. 1140–1160, 1997.

20. L. Doyen and O. Gaudoin, "Modelling and assessment of maintenance efficiency for repairable systems," in *ESREL*, 2002.

21. H. Pham and H. Wang, "Imperfect maintenance," *European journal of Operations Research*, vol. 94, pp. 425–428, 1996.

22. M. Berman and T. R. Turner, "Approximate point process likelihoods with GLIM," *Applied Statistics*, vol. 41, pp. 31–38, 1992.

23. J. Lawless and K. Thiagarajah, "A point process model incorporating renewals and time trends, with application to repairable systems," *Technometrics*, vol. 38, pp. 131–138, 1996.

24. D. Cox, "The statistical analysis of dependencies in point processes," in *Stochastic Point Processes* (P. Lewis, ed.), pp. 55–66, New York: Wiley, 1972.

25. B. H. Lindqvist, "The thend-renewal process, a useful model for repairable systems," in *Society of Reliability Engineers, Scandinavian Chapter. Anual Conference*, 1993. Malmö, Sweden.

26. M. Berman, "Inhomogeneous and modulated gamma processes," *Biometrika*, vol. 68, pp. 143–152, 1981.

27. R. Barlow and L. Hunter, "Optimun preventive maintenance policies," *Operations Research*, vol. 8, pp. 90–100, 1960.

28. C. Tilquin and R. Cléroux, "Periodic replacement with minimal repair at failure and adjustment costs," *Naval Reserach Logistics Quarterly*, vol. 22, pp. 243–254, 1975.

29. E. J. Muth, "An optimal decision rule for repair vs replacement," *IEEE Transactions on Reliability*, vol. R-26, pp. 179–181, 1977.

30. H. Makabe and H. Morimura, "A new policy for preventive maintenance," *Journal of the Operational Research Society of Japan*, vol. 5, no. 2, pp. 17–47, 1963.

31. K. S. Park, "Optimal number of minimal repair before replacement," *IEEE Transactions on Reliability*, vol. R-28, pp. 137–140, 1979.

32. R. Phelps, "Replacement policies under minimal repair," *Journal of the Operational Research Society*, vol. 32, pp. 549–554, 1981.

33. R. I. Phelps, "Optimal policy for minimal repair," *Journal of the Operational Research Society*, vol. 34, pp. 425–427, 1983.

34. W. Stadje and D. Zuckerman, "Optimal maintenance strategies for repairables systems with general repair," *Journal of Applied Probability*, vol. 28, pp. 384–396, 1990.

35. J. S. Dagpunar and N. Jack, "Optimizing system availability under minimal repair with non-negligible and replacement times," *Journal of Operational Reseach Society*, vol. 11, pp. 1097–1103, 1993.

36. P. J. Boland and F. Proschan, "Periodic replacement with increasing minimal repair costs at failure," *Operations Research*, vol. 30, pp. 1183–1189, 1982.

37. V. Makis and A. Jardine, "Optimal replacement policy for a general model with imperfect repair," *Journal of the Operational Research Society*, vol. 43, no. 2, pp. 111–120, 1992.

38. V. Makis, X. Jiang, and K. Cheng, "Optimal preventive replacement under minimal repair and random repair cost," *Mathematics of Operations Research*, vol. 25, pp. 141–156, febrero 2000.

39. S. Ross, "Average cost semi-markov decision processes," *Journal of Applied Probability*, vol. 7, pp. 649–656, 1970.

40. T. Aven, "Optimal replacement under a minimal repair strategy-a general set-up," *Advances in Applied Probability*, vol. 15, pp. 198–211, 1983.

41. R. Chipalkatti, J. Kurose, and D. Towsley, "Scheduling policies for real-time and non-real-time traffic in a stastistical multiplexer," tech. rep., Universidad de Massachusets, Enero 1989.

42. L. Kleinrock, *Queueings Sytems. Vol. 1: Theory.* Wiley, 1975.

43. L. Kleinrock, *Queueings Systems. Vol. 2: Computer Application.* Wiley, 1976.

44. K. I. A. T. Takine, H. Sunahara, and Y. Oie, "Delay analysis for CBR traffic under static-priority scheduling," *IEEE/ACM transactions on networking*, vol. 9, no. 2, pp. 177–185, 2001.

45. J. Shim, W. Lee, J. Pyun, and S. Ko, "A new implementable sheduling algorithm supporting various traffics in ATM networks- AWRR/DT," *IEEE Transactions on Communications*, vol. 38, no. 10, pp. 974–977, 1999.

46. D. Gan and S. McKenzie, "Performance of an ATM networking with multimedia traffic - a simulation study," in *International Broadcasting Convention*, no. Coference Publication No. 413, pp. 263–268, IEE, 14-18 septiembre 1995.

47. F. Ishizaki, T. Takine, and Y. Oie, "Delay analysis for real-time and non real-time traffic streams under a priority cell scheduling," *IEEE Transactions on Communications*, vol. 23, no. 5, pp. 3007–3012, 1998.

48. D. Choi, B. Choi, and D. Sung, "Performance analysis of priority leaky bucket scheme with queue-length-threshold scheduling policy," *IEE Proc.-Commun.*, vol. 145, no. 6, pp. 395–401, 1998.

49. W. Zhu and S. T. Chanson, "Adaptive threshold-based sheduling for real-time and non-real traffic," *IEEE transactions on Computers*, vol. C-36, no. 8, pp. 125–132, 1992.

50. M.Ñakamura and I. S. A. S. Mori, "Two parallel ques with dynamic routing under a

threshold-type scheduling," *IEEE Trans.*, vol. COM-34, no. 12, pp. 1145–1449, 1989.

51. S. Faizullah and I. Marsic, "Pricing QoS : Simulation and analysis," *IEEE/ACM transactios on Networking*, vol. 1, no. 6, pp. 193–199, 2001.

52. H. C. Tijms, *Stochastic Modelling and Analysis: A Computational Approach.* Wiley, 1986.