Universidad de la República

Facultad de Ciencias Económicas y Administración

Pasantía para obtener el grado de Licenciado en Estadística

Optimización del número de operadores de un Call Center

por
Matías Barrenechea y Gustavo González

Tutores: Leonardo Moreno - Guillermo Zoppolo

Montevideo
2016

UNIVERSIDAD DE LA REPÚBLICA

FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

El tribunal docente integrado por los abajo firmantes aprueba el trabajo de Pasantía:

Optimización del número de operadores de un Call Center

Matías Barrenechea - Gustavo González

Tutores: Leonardo Moreno, Guillermo Zoppolo

Puntaje:

Tribunal
Profesor: ..... Álvarez, Ramón.
Profesor: ..... Roche, Hugo.
Profesor: ..... Zoppolo, Guillermo.

Fecha:

Resumen

El presente trabajo de pasantía consistió en la optimización de los recursos humanos de un call center, que permitiera brindar un servicio de calidad. Para este propósito se analizó el problema desde la perspectiva de teoría de colas, en la cual es necesario el estudio de tres procesos: arribos de llamadas, tiempo de servicio y tiempo de espera del cliente.

En la práctica, el número de operadores (servidores) de un call center se aproxima utilizando el modelo M/M/c (Erlang-C). Esto indica que el proceso de arribos es un proceso markoviano, es decir que el tiempo entre dos llamadas consecutivas en el proceso de arribos tiene una distribución exponencial, que el tiempo de servicio en el proceso de servicio también se distribuye exponencial y que el sistema cuenta con c servidores. En este modelo se asume que los clientes tienen paciencia "infinita", lo que hace que no exista el abandono. Con los datos de dicho call center se llegó a la conclusión que los arribos de llamadas siguen un proceso de Poisson no homogéneo y su intensidad se estimó de manera no paramétrica. Para los tiempos de servicio, se detecta una distribución bimodal, que, en logaritmo, se ajusta a una mezcla de distribuciones normales. El tiempo de espera del cliente se asumió como determinístico. Con la modelización de estos procesos se construyeron dos funciones, una que simula el funcionamiento del call center para un día de la semana y una segunda que optimiza dicho funcionamiento sujeto al cumplimiento de determinadas medidas de performance.

Es importante aclarar que para los datos observados no se conoce el número de operadores, aunque se sabe que estaba distribuido de manera fija en tres turnos. Los resultados que se presentan son para los días lunes que no son necesariamente repre-

sentativos del resto de los días de la semana. Con estas consideraciones y teniendo en cuenta las medidas de performance requeridas, se encontró el número óptimo de operadores que fue 33, distribuidos en los tres turnos de la siguiente manera: 14 en el primer turno, 5 en el segundo y 14 en el último.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

En el presente estudio se analizó la información de un centro de llamadas (call center) que brindaba servicios en Uruguay. El mismo era uno de los más grandes que operaba a nivel nacional, contaba con una plantilla de más de 500 empleados y recibía mensualmente un volumen superior a las 400.000 llamadas. Estas llamadas corresponden a distintas campañas que éste brindaba.

En particular el objeto de estudio fue una de estas campañas. La misma comenzó a funcionar en el año 2010 y lo realizó durante 4 años seguidos, operando de lunes a viernes durante 12 horas seguidas. Tenía como objetivo atender el 95 % de las llamadas y a su vez que el 80 % de los clientes no esperaran más de 20 segundos hasta ser atendidos.

1.1. Los call centers

Un call center o centro de llamadas constituye un conjunto de recursos (personas y tecnologías) que permiten brindar un servicio telefónico. Existen diversas modalidades de call centers dependiendo del tipo y la forma de servicio que brindan. Los más comunes son:

1. *Inbound Call Center:* también llamado centro de llamadas entrantes, estos manejan predominantemente o exclusivamente llamadas entrantes, iniciadas por el cliente.

2. *Outbound Call Center:* o centro de llamadas salientes, son aquellos donde los agentes del call center realizan las llamadas a sus clientes o potenciales clientes.

3. *Blended Call Center:* combina llamadas entrantes y salientes.

4. *Contact Center:* son centros que utilizan las empresas para gestionar todo contacto con sus clientes a través de diversos medios de comunicación, como son el teléfono, mail, mensajería instantánea, redes sociales, etc.

Los call centers se pueden categorizar por distintos rubros, como ser, por su funcionalidad (mesa de ayuda, emergencias, atención al cliente, telemarketing, etc.), por el tamaño o por las características de las personas encargadas de brindar el servicio, denominados agentes u operadores, (manejo de idiomas, distintos conocimientos, etc.). En cuanto a la organización del mismo existen dos modalidades. Una es denominada plana y refiere a cuando todos los agentes están expuestos a todo tipo de llamadas. La otra modalidad se considera multi-capa, donde cada capa representa un nivel de especialización distinto, dependiendo del servicio que se esté brindando.

1.2. Calidad y eficiencia del servicio

Por lo general, el objetivo de estos centros de llamadas se plantea como la prestación de un servicio con una determinada calidad, sujeto a un presupuesto específico. La calidad de servicio se puede medir en dos dimensiones, una cualitativa y otra cuantitativa. La primera está relacionada a la percepción del cliente (por ejemplo, "estoy satisfecho con la respuesta", "fue un trato cordial", etc.) y está más orientada al marketing; generalmente se capta a través de las encuestas de satisfacción y de opinión. En referencia al aspecto cuantitativo (operativo) se hace foco en distintas medidas de performance, las cuales evalúan el abandono de las llamadas, la velocidad con que son atendidas, la duración del servicio y el rendimiento de los operadores.

El principal indicador relacionado al abandono es el Nivel de Servicio (NS), que busca dar cuenta del tiempo que un cliente está en la cola esperando a ser atendido. Se calcula como la fracción de llamadas que son atendidas antes de un determinado umbral de tiempo, respecto al total de llamadas recibidas. Hay que tener en cuenta

que se toman en consideración únicamente los clientes que esperan más de determinado lapso (habitualmente 5 segundos), las llamadas que abandonan el sistema antes de este lapso son consideradas llamadas fantasmas (*Ghost*) y se descartan. En la industria de la telefonía lo más utilizado es lo que se conoce como regla 80/20, lo que implica que se plantea como objetivo atender por lo menos el 80 % de los clientes antes de que alcancen 20 segundos de espera.

$$NS = \frac{\#\ Llamadas\ Atendidas\ (antes\ del\ umbral)}{\#\ Llamadas\ Recibidas\ (sin\ llamadas\ fantasmas)}$$

El abandono total es cuantificado como el número de clientes que abandonan el sistema antes de ser atendidos, sin considerar las llamadas fantasmas. Por lo tanto, el porcentaje de clientes que fueron atendidos se mide a través del Nivel de Atención (NAT).

$$NAT = \frac{\#\ Llamadas\ Atendidas}{\#\ Llamadas\ Recibidas\ (sin\ llamadas\ fantasmas)}$$

Otra medida de referencia es el tiempo que se destina a la atención de los clientes, denominada Tiempo Medio Operativo (TMO) y se calcula como el promedio de la duración de todas las llamadas atendidas durante el tiempo que se quiera controlar (COPC, 2009). Habitualmente esta medida se controla por día y por mes.

$$TMO = \frac{Duración\ de\ las\ Llamadas}{\#\ Llamadas\ Atendidas}$$

Por otro lado, el rendimiento de los operadores se mide a través de dos indicadores, la Utilización y la Ocupación. El primero es el tiempo que los operadores están hablando o disponibles para hacerlo sobre el total del tiempo trabajado. La segunda de ellas mide el porcentaje promedio de tiempo que los agentes están ocupados en una llamada. Un nivel de ocupación aceptable se encuentra entre 60 % y 80 %, según sugiere en sus manuales asociados a call centers la Corporación Financiera Internacional del Banco Mundial (International Finance Corporation, 2010).

$$\text{Utilización} = \frac{\textit{Tiempo Hablado + Tiempo Disponible}}{\textit{Tiempo Trabajado}}$$

$$\text{Ocupación} = \frac{\textit{Tiempo Hablado}}{\textit{Tiempo Hablado + Tiempo Disponible}}$$

Un dato relevante, es que, a nivel internacional, aproximadamente el 60-70 % de los costos operativos del call center está asociado a los recursos humanos (Gans et al., 2003). Entonces se torna fundamental el equilibrio entre la cantidad de operadores y la calidad del servicio que se brinda. En el ámbito local esta relación de costos fluctúa entre los mismos márgenes.

Los modelos que se utilizan para calcular los diferentes niveles de eficiencia de los agentes y de servicio se basan en la teoría de colas (o líneas de espera) (Bhat, 2008). Estos modelos se basan en datos tales como el número de agentes, la distribución de los arribos y de la duración de las llamadas, el comportamiento de la espera, la modalidad con que son atendidos los clientes, etc.

1.3. Objetivos

El objetivo principal del presente trabajo es encontrar, para un call center específico, el número óptimo de operadores que garantice brindar un servicio determinado, caracterizado por el cumplimiento de determinadas metas en los indicadores de calidad y eficiencia.

Como objetivos específicos, surge la necesidad de modelizar el proceso de arribos de clientes, el tiempo de servicio destinado a las llamadas y el comportamiento de abandono, es decir, el tiempo de espera que tienen dichos clientes para ser atendidos. El problema no se aborda desde un punto de vista analítico sino que, con los insumos anteriores se simula el funcionamiento del call center. Por último se calcula el número óptimo de operadores sujeto a determinadas restricciones.

Lo que resta del trabajo está estructurado en tres capítulos. En el capítulo siguiente se presentan los aspectos principales de las metodologías utilizadas. Luego se dedica un capítulo para el análisis de los datos, la aplicación de las metodologías y los resultados obtenidos. En el último capítulo se presentan las conclusiones que se desprenden del análisis anterior. Finalmente, los apéndices contienen algunas demostraciones de resultados y el código de programación R utilizado.

Capítulo 2

Marco Metodológico

En los desarrollos teóricos que se presentan a continuación se utiliza una línea de razonamiento y nomenclatura similar al libro *An Introduction to Queueing Theory: Modeling and Analysis in Applications.* de U. Narayan Bhat (Bhat, 2008).

2.1. Teoría de Colas

Es importante el estudio de los sistema de colas, ya que permite una descripción del funcionamiento de los call centers. Este funcionamiento refiere a que los clientes llegan de acuerdo a un proceso de arribos para ser atendidos conforme a un proceso de servicio. Quien brinda el servicio es denominado servidor, en un sistema puede haber uno o más. Se asume que cada servidor puede atender a un cliente a la vez y cuando todos los servidores están ocupados, el cliente ingresa a la cola esperando a ser atendido. Cuando se libera un servidor el siguiente cliente a ser atendido es seleccionado de la cola de acuerdo a la disciplina de la misma. Durante el servicio el cliente puede pasar por una o más etapas del servicio antes de abandonar el sistema. La Figura 2.1 muestra una representación esquemática de un sistema de colas.

A continuación se describen los principales componentes de la teoría de colas,

1. **Proceso de arribos**

   Los arribos de los usuarios al sistema se comportan de forma aleatoria. Por lo general se supone un proceso donde la distribución entre las llegadas de dos

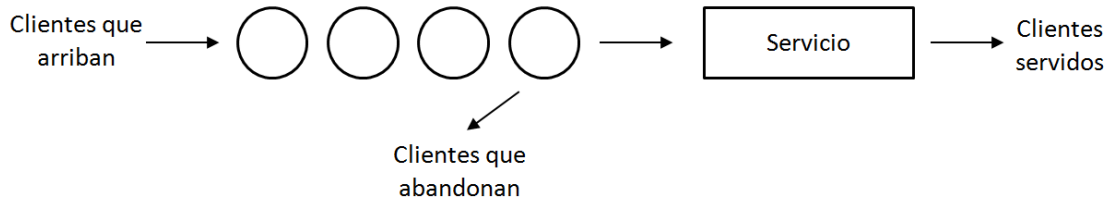

Figura 2.1: Esquema de un sistema de colas

clientes sucesivos son independiente e idénticamente distribuídas (iid). Si se conoce el tiempo exacto que transcurre entre un arribo y otro, este proceso es determinístico. El proceso de llegadas puede variar a lo largo del tiempo, si la tasa de arribos se mantiene constante, se le denomina homogéneo, de lo contrario, es no homogéneo.

2. **Proceso de servicio**

   El tiempo de servicio también suele modelarse como un proceso donde los tiempos entre servicios sucesivos son i.i.d. El tiempo de servicio puede variar con el número de clientes en la cola, siendo este más rápido o más lento. Al igual que el proceso de arribos, puede ser homogéneo o variar con el transcurso del tiempo.

3. **Disciplina de la cola**

   La disciplina de la cola se refiere a la manera en que los clientes son seleccionados dentro de la cola para ser atendidos por el servidor. La modalidad más utilizada es atender primero a quien llega en primer lugar, FIFO (*First In First Out*). Existen también variantes como lo son, el último en llegar primero en ser atendido, LIFO (*Last In First Out*), elección aleatoria entre los clientes que esperan, etc. También existen disciplinas con orden de prioridad, ya sea por el tipo de clientes o por la duración del servicio que requieren.

4. **Capacidad de la cola**

   La capacidad de la cola refiere al número máximo de clientes que pueden estar

esperando a ser atendidos. Si un cliente llega y la cola alcanzó el máximo de su capacidad, se le negará la entrada al sistema. Cuando la capacidad de la cola es grande, se puede asumir que esta es infinita.

5. **Servidores**

   Refiere a la cantidad de operadores con que cuenta el sistema. Se asume que todos estos son idénticos y en paralelo, por lo tanto, un cliente puede ser atendido de la misma manera por cualquiera de ellos. Cuando se habla de servidores en paralelo, se refiere a una única cola que alimenta varios servidores (mono-cola), mientras que el caso de colas independientes es similar a múltiples sistemas con un solo servidor (multi-cola). Estas dos modalidades se esquematiza en la Figura 2.2.

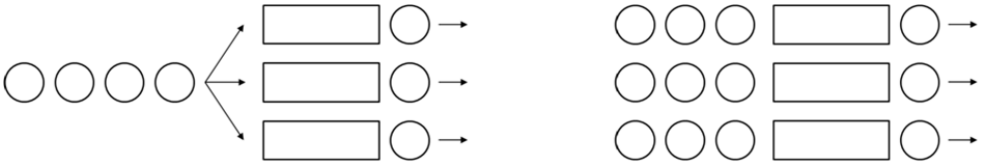

Figura 2.2: Esquema de un servicio mono-cola y multi-cola

6. **Etapas del servicio**

   Un sistema de colas puede tener una o más etapas dependiendo del tipo de servicio. En los sistemas de una etapa, el cliente luego de ser atendido deja el sistema. En los sistemas multietapas, el cliente puede pasar por un número de etapas mayor que uno. En algunos de estos sistemas se puede volver atrás. Un esquema de esto se puede ver en la Figura 2.3.

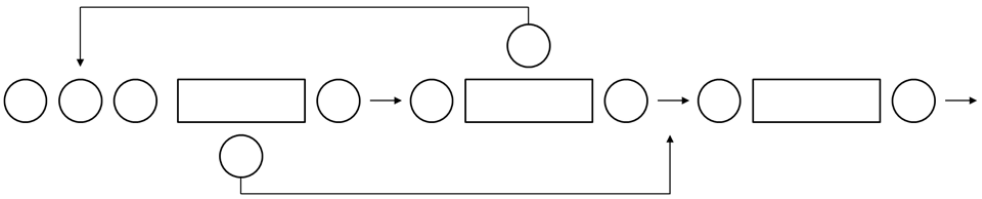

Figura 2.3: Sistema Multi-etapa

Cada una de estas características del sistema de colas se pueden expresar de diversas maneras. David G. Kendall (Kendall, 1953) introdujo una notación para representar la combinación de partes que conforman la estructura del sistema. Esta notación es conocida como notación de Kendall y está conformada por seis componentes:

$$A/S/c/K/N/D$$

donde A representa el proceso de arribos, S denota la distribución del tiempo de servicio, c es el número de servidores, K indica la capacidad del sistema, N es el tamaño de la población de donde provienen los clientes y D refiere a la disciplina de la cola.

A: Proceso de arribos
Indica el tipo de proceso o patrón de llegadas, los cuales pueden ser clasificados según la siguiente nomenclatura,

> **M**, es un proceso *Markoviano* de entradas. Los arribos son aleatorios y provienen de un proceso de Poisson.
>
> **G**, representa una distribución *General* de probabilidad, es decir, que el modelo y sus resultados son aplicables a cualquier distribución estadística.
>
> **D**, *Determinístico* o *Constante*, se fija el tiempo entre los arribos.

S: Distribución del tiempo de servicio.
Esta distribución se clasifica de la misma manera que el proceso de arribos.

c: Número de servidores

K: Capacidad del sistema
Indica la cantidad de clientes que pueden estar en el sistema, ya sea esperando o siendo atendidos. Cuando no hay límite para la cantidad de clientes que pueden esperar en la cola, no se incluye ninguna notación, lo cual implica que es infinita.

N: Tamaño de la población
Es el tamaño de la población del cual provienen los clientes. Cuando se omite, se asume infinito.

D: Disciplina de la cola.

2.2. Procesos Estocásticos

Un proceso estocástico es una familia de variables aleatorias definidas en el mismo espacio de probabilidad (Ω, B, P) e indexadas en un conjunto T que puede pensarse como indicador de tiempo. Por lo tanto, para cada instante de tiempo t (en el caso que el tiempo sea continuo) o en cada época n (tiempo discreto), se tendrá una variable aleatoria distinta,

$$\{X(t), t \in T\}, \qquad T \subset \mathrm{R}$$
$$\{X_n, n \in T\}, \qquad T \subset \mathbb{Z}$$

donde, T es conocido como espacio de parámetros (continuo o discreto) y las variables aleatorias $X(t)$ y X_n toman valores en un conjunto E denominado espacio de estados, el cual puede ser discreto o continuo. Por lo tanto, dependiendo de cómo sean el espacio de parámetros T y el espacio de estados E, se pueden clasificar los procesos estocásticos.

2.3. Cadenas de Markov

Algunos de los modelos más sencillos y utilizados en teoría de colas, están basados en los procesos de Markov. En estos procesos se cumple que el valor que toma la variable aleatoria en determinado instante o época depende únicamente del momento anterior. A esta falta de memoria se la denomina propiedad de Markov.

A continuación se presentan los resultados para cadenas de Markov a tiempo discreto, los cuales son similares en el caso de que el tiempo sea continuo. Entonces, dado el proceso estocástico $\{X_n, n \in T\}$ con espacio de estados discreto, $n \in \mathbb{Z}$ tal

que,

$$\begin{aligned} &P(X_n = j \mid X_{n_1} = i_1, X_{n_2} = i_2, \ldots, X_{n_k} = i_k) \\ &= P(X_n = j \mid X_{n_k} = i_k) \\ &= P_{i_k,j}^{(n_k,n)} \end{aligned}$$

entonces $\{X_n, n = 1, 2, \ldots\}$ es una cadena de Markov.

Aquí se puede observar que el proceso de Markov presenta una dependencia a un solo paso, es decir, el estado de la cadena depende únicamente del evento inmediato anterior.

Las probabilidades de cambio de estados o de transición, es decir la probabilidad de pasar del estado i en el tiempo m al estado j en el tiempo n caracterizan al sistema y anotamos,

$$P_{ij}^{m,n} = P(X_n = j \mid X_m = i), \qquad m < n. \tag{2.1}$$

Las cadenas de Markov cumplen la relación de *Chapman-Kolmogorov*:

$$P_{ij}^{m,n} = \sum_{k \in S} P_{ik}^{m,r} P_{kj}^{r,n}, \qquad m < r < n. \tag{2.2}$$

Se dice que una cadena de Markov X_n, $n = 0, 1, 2, \ldots$ es homogénea en el tiempo, cuando las probabilidades de transición $P_{ij}^{m,n}$ y $P_{ij}^{m+k,n+k}$ son iguales. Sin pérdida de generalidad, se toma $m = 0$, por lo tanto se tiene que:

$$\begin{aligned} P_{ij}^{k,n+k} =& P(X_{n+k} = j \mid X_k = i) \\ =& P(X_n = j \mid X_0 = i) = P_{ij}^{(n)} \qquad \forall k \geq 0 \end{aligned}$$

En forma matricial,

$$\mathbf{P}^{(n)} = \begin{bmatrix} P_{00}^{(n)} & P_{01}^{(n)} & P_{02}^{(n)} & \ldots \\ P_{10}^{(n)} & P_{11}^{(n)} & P_{12}^{(n)} & \ldots \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

Esta matriz de probabilidades tiene las propiedades de que $P_{ij}^{(n)} \geq 0$ y $\sum_{j \in S} P_{ij}^{(n)} = 1$, para todo los valores de n. Por conveniencia, cuando $n = 1$ la probabilidad de transición a un paso se denomina P_{ij} y la matriz de probabilidades de transición $\mathbf{P} = \mathbf{P}^{(1)}$. Por lo tanto, la relación de *Chapman-Kolmogorov*, se puede ver matricialmente:

$$\mathbf{P}^{(n)} = \mathbf{P}^{(r)}\mathbf{P}^{(n-r)}$$

sustituyendo r por 1:

$$\begin{aligned} \mathbf{P}^{(n)} &= \mathbf{P}^{(1)}\mathbf{P}^{(n-1)} \\ &= \mathbf{P}^{(1)}\mathbf{P}^{(1)}\mathbf{P}^{(n-2)} \\ &= \underbrace{\mathbf{P}^{(1)}\mathbf{P}^{(1)}\ldots\mathbf{P}^{(1)}}_{n} \end{aligned}$$

entonces, como $\mathbf{P} = \mathbf{P}^{(1)}$,

$$\mathbf{P}^{(n)} = \mathbf{P}^{n},$$

Consecuentemente, se cumple para $r = 2, \ldots, n$.

Además, cuando $n = 0$ se define $P_{ij}^{(0)}$ como la función delta de Kronecker:

$$P_{ij}^{(0)} = \delta_{ij} = \begin{cases} 0 & \text{si} \quad i \neq j, \\ 1 & \text{si} \quad i = j. \end{cases}$$

Es decir, después de realizar cero pasos la cadena no puede estar en otro lugar más que en su estado de partida.

2.4. Procesos de Poisson

Tanto en la presente sección como en la siguiente, se estudian las cadena de Markov a tiempo continuo. Particularmente, en ésta se expone uno de los ejemplos más importantes de este tipo de modelos, el proceso de Poisson.

Un proceso de Poisson es un proceso de Markov en tiempo continuo que consiste en "contar" eventos que ocurren en un determinado período de tiempo, tal como puede suponerse con el arribo de llamadas en un call center. Por lo tanto es fundamental comprender este tipo de procesos para poder llevar a cabo el estudio del funcionamiento de este centro de llamadas. El tiempo entre cada par de eventos consecutivos tiene una distribución exponencial con parámetro λ, y se asumen independientes entre sí. De esta manera $Z_1, Z_2, \ldots$ son iid, con distribución dada por,

$$\begin{aligned} F(x) &= P(Z_1 \leq x) \\ &= 1 - e^{-\lambda x}, \qquad x > 0, \qquad \lambda > 0 \end{aligned}$$

siendo su función de densidad $\lambda e^{-\lambda x}$, $x > 0$.

Una de las propiedades que caracteriza a la distribución exponencial dentro del conjunto de distribuciones absolutamente continuas es que satisface la propiedad de pérdida de memoria, esto quiere decir, si Z tiene distribución $Exp(\lambda)$, entonces para cualquier s, $t > 0$ se cumple,

$$P(Z > t + s | Z > s) = P(Z > t)$$

Definición 2.4.1. *Se define $X(t)$ el número de eventos ocurridos hasta el tiempo t, de tal manera que los tiempos entre arribos se distribuyen $Exp(\lambda)$. Escribiendo $X(t)$ como un proceso estocástico,*

$$X(t) = \text{máx}\{n : Z_1 + Z_2 + \cdots + Z_n \leq t\}.$$

A este proceso se le llama proceso de Poisson homogéneo, ya que el parámetro λ no depende de t. Una trayectoria típica de este proceso se observa en la Figura 2.4. Los tiempos entre arribos, $Z_1, Z_2, \ldots$, corresponden a los tiempos que transcurren entre un salto del proceso y el siguiente. Como se mencionó anteriormente, estos tiempos son independientes y se distribuyen exponencial con parámetro λ. En consecuencia la variable $W_n = Z_1 + \ldots + Z_n$ tiene una distribución $Erlang(n, \lambda)$, la cual es un caso particular de la distribución Gama cuando n es entero. Esta variable representa el instante de tiempo en que ocurre el *n-ésimo* evento. Aquí se observa la igualdad de eventos $\{X(t) \geq n\} = \{W_n \leq t\}$, lo cual quiere decir que al momento t han ocurrido por lo menos n eventos si, y sólo si, el *n-ésimo* evento ocurrió antes de t.

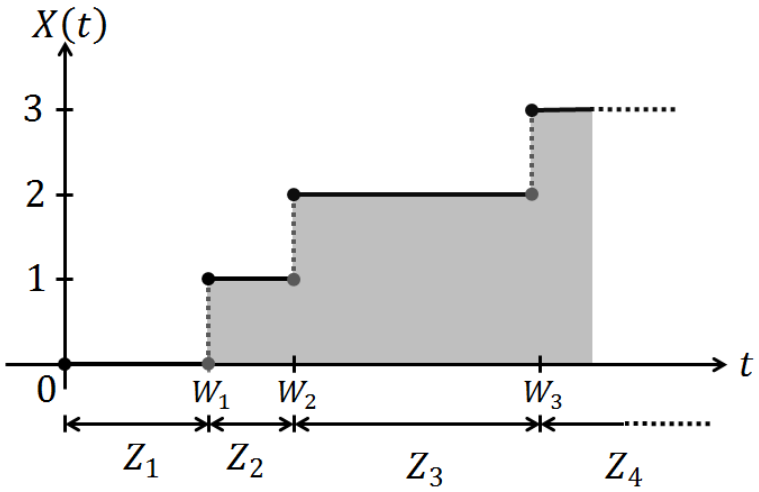

Figura 2.4: Trayectoria de un Proceso de Poisson

Proposición 2.4.1. *El proceso $X(t)$ tiene distribución $Pois(\lambda t)$, para cualquier valor de $t > 0$ y para $n = 0, 1, \ldots$*

$$P(X(t) = n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!} \tag{2.3}$$

Demostración. Como W_n se distribuye $Gamma(n, \lambda)$, su función de distribución para $t > 0$ es,

$$P(W_n \leq t) = 1 - e^{-\lambda t}\sum_{k=0}^{n-1}\frac{(\lambda t)^k}{k!}$$

Entonces para cualquier $t > 0$ y para cada $n = 0, 1, \ldots$

$$\begin{aligned} P(X(t) = n) &= P(X(t) \geq n) - P(X(t) \geq n+1) \\ &= P(W_n \leq t) - P(W_{n+1} \leq t) \\ &= e^{-\lambda t}\frac{(\lambda t)^n}{n!}. \end{aligned}$$

□

Proposición 2.4.2. *Para cualquier tiempo* $0 \leq s < t$*, y para* $n = 0, 1, \ldots$

$$P(X(t) - X(s) = n) = P(X(t-s) = n) = e^{-\lambda(t-s)} \frac{[\lambda(t-s)]^n}{n!} \tag{2.4}$$

Demostración.

$$P(X(t) - X(s) = n) = \sum_{k=0}^{\infty} P(X(t) - X(s) = n | X(s) = k) P(X(s) = k)$$

Dado que en el tiempo s el proceso de Poisson se encuentra en el nivel k, por la propiedad de pérdida de memoria se puede considerar que en ese momento reinicia el proceso de Poisson y en consecuencia la probabilidad del evento $\{X(t) - X(s) = n\}$ es igual a la probabilidad del evento $\{X(t-s) = n\}$. Por lo tanto,

$$\begin{aligned} P(X(t) - X(s) = n) &= \sum_{k=0}^{\infty} P(X(t-s) = n) P(X(s) = k) \\ &= P(X(t-s) = n) \sum_{k=0}^{\infty} P(X(s) = k) \\ &= P(X(t-s) = n). \end{aligned}$$

□

Distribuciones asociadas al proceso de Poisson

Además de las distribuciones exponencial y gamma ya mencionadas existen otras distribuciones de probabilidad que surgen al estudiar ciertas características del proceso de Poisson, como por ejemplo la distribución uniforme.

Proposición 2.4.3. *Dado el evento* $\{X(t) = n\}$*, el vector de tiempos* $(W_1, \ldots, W_n)$ *tiene la misma distribución que el vector de los estadísticos de orden* $(Y_{(1)}, \ldots, Y_{(n)})$ *de una muestra aleatoria* $Y_1, \ldots, Y_n$ *de la distribución* $Unif[0, t]$*, es decir,*

$$f_{W_1,\ldots,W_n|X(t)=n}(w_1, \ldots, w_n|n) = \begin{cases} \frac{n!}{t^n} & 0 < w_1 < \cdots < w_n < t \\ 0 & \textit{otro caso.} \end{cases}$$

Demostración. La función de densidad conjunta de los estadísticos de orden $Y_{(1)}, \ldots, Y_{(n)}$ de una muestra aleatoria $Y_1, \ldots, Y_n$ con función de densidad $f(y)$ es, para $y_1 < \ldots < y_n$,

$$f_{Y_{(1)},\ldots,Y_{(n)}}(y_1, \ldots, y_n) = n! f(y_1) \cdots f(y_n).$$

Cuando la función de densidad $f(y)$ es la uniforme en el intervalo $[0, t]$, esta función de densidad conjunta es la que aparece en el enunciado. Se demuestra que la distribución conjunta de las variables $W_1, \ldots, W_n$, condicionada al evento $X(t) = n$ también tiene esta misma función de densidad. Se utiliza nuevamente la identidad de eventos $\{X(t) \geq n\} = \{W_n \leq t\}$. Para tiempos $0 < w_1 < \cdots < w_n < t$, la función de densidad conjunta condicional $f_{W_1,\ldots,W_n|X(t)=n}(w_1, \ldots, w_n \mid n)$ se puede obtener de la siguiente manera,

$$\tfrac{\partial n}{\partial w_1 \cdots \partial w_n} P\big(W_1 \leq w_1, W_2 \leq w_2, \ldots, W_n \leq w_n \mid X(t) = n\big) =$$

$$\tfrac{\partial n}{\partial w_1 \cdots \partial w_n} P\big(X(w_1) \geq 1, X(w_2) \geq 2, \ldots, X(w_n) \geq n \mid X(t) = n\big) =$$

$$\tfrac{\partial n}{\partial w_1 \cdots \partial w_n} P\big(X(t) - X(w_n) = 0, X(w_n) - X(w_{n-1}) = 1, \ldots$$

$$\ldots, X(w_2) - X(w_1) = 1, X(w_1) = 1\big) / P\big(X(t) = n\big) =$$

$$\tfrac{\partial n}{\partial w_1 \cdots \partial w_n} e^{-\lambda(t-w_n)} e^{-\lambda(w_n - w_{n-1})} \lambda(w_n - w_{n-1}) \cdots$$

$$\cdots e^{-\lambda(w_2 - w_1)} \lambda(w_2 - w_1) e^{-\lambda w_1} \lambda w_1 / [e^{-\lambda t} (\lambda t)^n / n!] =$$

$$\tfrac{\partial n}{\partial w_1 \cdots \partial w_n} n! (w_n - w_{n-1}) \cdots (w_2 - w_1) w_1 / t^n =$$

$$= n! / t^n.$$

□

A continuación se presentan algunas definiciones alternativas del proceso de Pois-

son. La ventaja de contar con definiciones alternativas es que para demostrar que un proceso es de Poisson se puede optar por cualquiera de ellas.

Definiciones alternativas

La Definición 2.4.1 de proceso de Poisson es constructiva pues a partir de los tiempos entre arribos se construye el proceso de conteo correspondiente.

Definición 2.4.2. *Un proceso de Poisson de parámetro $\lambda > 0$ es un proceso a tiempo continuo $X(t), t \geq 0$, con espacio de estados $0, 1, 2, \ldots$, y que cumple con las siguientes propiedades:*

1. $X_0 = 0$
2. *Tiene incrementos independientes y estacionarios*
3. *Para cualquier $t \geq 0$, y cuando $h \to 0$,*
   - *i.* $P(X(t+h) - X(t) \geq 1) = \lambda h + o(h)$
   - *ii.* $P(X(t+h) - X(t) \geq 2) = o(h)$

donde $o(\Delta t)$ es tal que $\frac{o(\Delta t)}{\Delta t} \to 0$, cuando $\Delta t \to 0$.

Esta definición se basa en la descripción local de las características del proceso, la cual presenta ciertas ventajas desde el punto de vista de la interpretación de lo que sucede en un intervalo infinitesimal de tiempo $(t, t+h]$. El proceso comienza en cero y por la propiedad 3 de la definición, la probabilidad de pasar al estado uno al final de un intervalo de tiempo de amplitud h es aproximadamente proporcional a la amplitud del intervalo, o sea, $\lambda h + o(h)$, con $o(h)$ tal que la probabilidad de que el proceso tenga dos o más incrementos en el intervalo es $o(h)$ y por lo tanto la probabilidad de que el proceso no sufra cambio de estado en este intervalo es $1 - \lambda h + o(h)$. Esto significa que en cualquier intervalo infinitesimal solo puede ocurrir un incremento o ninguno. Esta caracterización es de utilidad para la definición de proceso de Poisson no homogéneo.

Definición 2.4.3. *Un proceso de Poisson de parámetro $\lambda > 0$ es un proceso a tiempo continuo $\{X(t), t \geq 0\}$ con espacio de estados $\{0, 1, \ldots\}$, con trayectorias no decrecientes y que cumple las siguientes propiedades:*

1. *X(0) = 0*
2. *Tiene incrementos independientes*
3. *$X(t+s) - X(s)$ se distribuye $Pois(\lambda t)$, para cualquier $s \geq 0$, $t > 0$.*

A partir de esta definición se sabe inmediatamente que la variable $X(t)$ tiene una distribución $Pois(\lambda t)$. La independencia de los incrementos es explícita y la estacionariedad de estos está implícita en el tercer ítem.

2.5. Proceso de Poisson No Homogéneo

Un proceso de Poisson no homogéneo es un proceso de Poisson donde el parámetro λ se sustituye por una función dependiente del tiempo. Este modelo puede ser naturalmente más adecuado para algunas situaciones reales, pero con la particularidad que deja de cumplir la propiedad de Markov.

Definición 2.5.1. *Se define como un proceso de tiempo continuo $\{X(t), t \geq 0\}$, con espacio de estados $\{0, 1, \ldots\}$, con parámetro $\lambda(t)$, una función positiva y localmente integrable, que cumple con las siguientes propiedades:*

1. $X(0) = 0$
2. *Los incrementos son independientes*
3. *Para cualquier $t \geq 0$, y cuando $h \searrow 0$,*
   - *i.* $P(X(t+h) - X(t) = 0) = 1 - \lambda(t)h + o(h)$
   - *ii.* $P(X(t+h) - X(t) = 1) = \lambda(t)h + o(h)$
   - *iii.* $P(X(t+h) - X(t) \geq 2) = o(h)$

La diferencia con un proceso de Poisson homogéneo es que los incrementos dejan de ser estacionarios, es decir, la distribución de probabilidad de la variable incremento $X(t+h) - X(t)$ depende de los valores de la función $\lambda(t)$ en el intervalo $(t, t+h]$. Sin embargo, la variable $X(t)$ continúa siendo Poisson.

Proposición 2.5.1. *La variable $X(t)$ en un proceso de Poisson no homogéneo de parámetro $\lambda(t)$ tiene distribución Pois($\Lambda(t)$), donde:*

$$\Lambda(t) = \int_0^t \lambda(s)\, ds$$

es decir, para $n = 0, 1, \ldots$

$$P(X(t) = n) = e^{-\Lambda(t)} \frac{[\Lambda(t)]^n}{n!}. \tag{2.5}$$

La demostración se encuentra en el Apéndice C.

Un proceso de Poisson no homogéneo y un proceso de Poisson tienen trayectorias no decrecientes y con saltos unitarios, la diferencia es que la frecuencia promedio con la que aparecen los saltos cambia a lo largo del tiempo.

Proposición 2.5.2. *Para el proceso de Poisson no homogéneo, la variable incremento $X(t+s) - X(s)$ tiene distribución Pois($\Lambda(t+s) - \Lambda(s)$).*

La demostración se encuentra en el Apéndice C.

Si la función $\lambda(t)$ es igual a λ y no depende de t, entonces $\Lambda(t) = \lambda t$, y se recupera el proceso de Poisson homogéneo.

2.5.1. Test de Hipótesis

En esta parte se presenta un test utilizado para probar que los arribos provienen de un proceso de Poisson. En primer lugar es conveniente mencionar algunas propiedades:

Propiedad 2.5.1.1. *Si $X \sim Beta(\alpha, 1)$, entonces $-\alpha \log(X) \sim Exp(1)$.*

Demostración.

$$\begin{aligned} P(-\alpha \log(X) \leq t) =& P\left(X \geq e^{-t/\alpha}\right) = \\ =& 1 - \int_0^{e^{-t/\alpha}} t^{\alpha-1} \frac{\Gamma(1+\alpha)}{\Gamma(\alpha)\Gamma(1)} = \\ =& 1 - \frac{\alpha\Gamma(\alpha)}{\Gamma(\alpha)} \frac{t^\alpha}{\alpha}\Big|_0^{e^{-t/\alpha}} = \\ =& 1 - e^{-t}. \end{aligned}$$

□

Propiedad 2.5.1.2. *Sean $X_1, X_2, \ldots, X_N \overset{iid}{\sim} Unif(0, L)$, con $L > 0$, entonces la densidad conjunta de los estadísticos de orden $X_{(k-1)}$ y $X_{(k)}$ viene dada por,*

$$f_{X_{(k-1)},X_{(k)}}(x, y) = \frac{1}{L^2} \frac{N!}{(k-2)!(N-k)!} \left(\frac{x}{L}\right)^{k-2} \left(\frac{1-y}{L}\right)^{N-k}, \quad si\ x < y. \tag{2.6}$$

Demostración. Si $x < y$,

$$\begin{aligned} F_{X_{(k-1)},X_{(k)}}(x, y) =& P\left(X_{(k-1)} \leq x, X_{(k)} \leq y\right) = \\ =& P\left(X_{(k-1)} \leq x\right) - P\left(X_{(k-1)} \leq x, X_{(k)} > y\right) = \\ =& P\left(X_{(k-1)} \leq x\right) - \frac{N!}{(k-1)!(N-k+1)!} \left(\frac{x}{L}\right)^{k-1} \left(\frac{1-y}{L}\right)^{N-k+1} \end{aligned}$$

entonces,

$$\begin{aligned} f_{X_{(k-1)},X_{(k)}}(x, y) =& \frac{\partial^2 F_{X_{(k-1)},X_{(k)}}}{\partial x \partial y}(x, y) = \\ =& \frac{1}{L^2} \frac{N!}{(k-2)!(N-k)!} \left(\frac{x}{L}\right)^{k-2} \left(\frac{1-y}{L}\right)^{N-k}, \quad \text{si } x < y. \end{aligned}$$

□

Propiedad 2.5.1.3. *Sean $X_1, X_2, \ldots, X_N \overset{iid}{\sim} Unif(0, L)$, con $L > 0$, entonces,*

$$\frac{L - X_{(k)}}{L - X_{(k-1)}} \sim Beta\,(N + 1 - k, 1)$$

Demostración. Se realiza la transformación $U = X_{(k-1)}$ y $V = \frac{L-X_{(k)}}{L-X_{(k-1)}}$ partiendo de la distribución conjunta de los estadísticos de orden calculada en 2.5.1.2 y se obtiene la densidad conjunta del vector (U, V),

$$f_{U,V}(u, v) = \frac{1}{L^2} \frac{N!}{(k-2)!(N-k)!} \left(\frac{u}{L}\right)^{k-2} \left(\frac{v(L-u)}{L}\right)^{N-k} (L - u)$$

entonces, intergrando en u

$$\begin{aligned} f_V(v) =& v^{N-k} \frac{N!}{(k-2)!(N-k)!} \int_0^L \left(\frac{u}{L}\right)^{k-2} \left(\frac{1-u}{L}\right)^{N-k+1} \frac{1}{L} du = \\ =& v^{N-k} \frac{N!}{(k-2)!(N-k)!} \frac{(k-2)!(N-k+1)!}{N!} = \\ =& (N-k+1)v^{N-k} \sim Beta\,(N + 1 - k, 1) \end{aligned}$$

□

De esta manera si se desea testear la hipótesis de que el proceso proviene de un proceso de Poisson, se procede de la siguiente manera:

H_0) $X(t)$ es un proceso de Poisson
H_1) $X(t)$ no es un proceso de Poisson

Con esto se prueba de que los arribos pueden provenir de un proceso de Poisson, que no necesariamente tiene que ser homogéneo.

El primer paso para la construcción de este test es fraccionar el intervalo de tiempo a estudiar en pequeños bloques. Por conveniencia se utilizan bloques de igual duración, L, resultando un total de I bloques. Sea T_{ij} el momento en que se produce el j-ésimo suceso ordenado dentro del i-ésimo bloque, $i = 1, \ldots, I$. Por lo tanto $T_{i1} \leq T_{i2} \leq \cdots \leq T_{iJ(i)}$, donde $J(i)$ es el número total de observaciones en el bloque i.

Anotando X_{ij} como la j-ésima observación no ordenada en el bloque i y condicionado en $J(i)$, se cumple que,

$$X_{ij}\Big/ J(i) \overset{iid}{\sim} Unif(0, L)$$

siendo $X_{i(j)} = T_{ij}$ y realizando la transformación a todos los datos,

$$R_{ij} = (J(i) + 1 - j)\left(- log\left(\frac{L - T_{ij}}{L - T_{i,j-1}}\right)\right) \tag{2.7}$$

utilizando la propiedad 2.5.1.3 se tiene que $\frac{L-T_{ij}}{L-T_{i,j-1}}$ se distribuye $Beta(J(i)+1-j, 1)$ y por la propiedad 2.5.1.1, se obtiene que, bajo H_0 cierta $R_{ij} \overset{iid}{\sim} Exp(1)$. Lo cual es equivalente a la prueba original, o sea, basta con realizar la prueba,
$H_0)$ $R_{ij} \overset{iid}{\sim} Exp(1)$
$H_1)$ $R_{ij} \overset{iid}{\nsim} Exp(1)$

2.5.2. Estimación de la función de intensidad

Para estimar la función de intensidad se utilizó la estimación no paramétrica de la densidad por núcleos, también llamado *kernel*. La palabra núcleo refiere a cualquier función K no negativa, que integre a uno y con media cero, (Wasserman, 2006).

Entonces, dada una muestra de n observaciones reales $\{x_1, x_2, \ldots, x_n\}$ se define el estimador de densidad con núcleo,

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right),$$

donde h es un número positivo llamado ancho de la ventana (*bandwith*) y representa el parámetro de suavizado. La elección de este ancho es importante, ya que si son pequeños dan estimaciones rugosas mientras que ventanas grandes dan estimaciones más suavizadas, es decir que tendrán menor sesgo pero con mayor varianza. En la práctica se suele utilizar el núcleo Gaussiano, $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ y el ancho de

ventana utilizado es el llamado *Silverman's rule of thumb*, (Silverman, 1986),

$$h_{SROT} = 0{,}9 \times min\left\{\hat{\sigma}, \frac{I\hat{Q}R}{1{,}34}\right\} \times n^{(-1/5)}$$

donde $\hat{\sigma}$ es el desvío estándar e $I\hat{Q}R$ es el rango intercuartil de la muestra.

2.6. Cadenas de Markov en tiempo continuo

Estos procesos son cadenas de Markov en donde el tiempo es continuo y las variables toman valores enteros, es decir de estados discreto. Sea $\{X(t), t \in T\}$ un proceso a tiempo continuo que inicia en un estado i_1 en el tiempo cero y permanece en este estado un tiempo aleatorio Z_1. Luego salta al estado i_2 distinto del anterior, permaneciendo en este estado por un tiempo aleatorio Z_2, posteriormente pasa a otro estado i_3 distinto del anterior y así sucesivamente. Durante los tiempos aleatorios Z el proceso permanece constante en alguno de sus estados.

En estos procesos la probabilidad de transición está dada por,

$$P_{ij}(s,t) = P[X(t) = j \mid X(s) = i], \qquad s < t. \tag{2.8}$$

Estos procesos satisfacen la propiedad de Markov de pérdida de memoria ya que los tiempos de estadía (Z) en un estado tienen distribución exponencial.

Se supone, al igual que en el caso discreto, que las probabilidades de transición son estacionarias en el tiempo, esto significa que $P_{ij}(0,t)$ y $P_{ij}(s,t+s)$ son iguales, es decir que no dependen del valor s, entonces

$$P_{ij}(t) = P[X(t) = j \mid X(0) = i]$$

En este caso, como el espacio de parámetros es continuo, se utilizan ecuaciones diferenciales para determinar $P_{ij}(t)$, los cuales cumplen con las siguientes propiedades:

1. $P_{ij}(t) \geq 0$

2. $\sum_{j\in S} P_{ij}(t) = 1$
3. $P_{ij}(s+t) = \sum_{k\in S} P_{ik}(s)P_{kj}(t)$
4. P_{ij} es continuo
5. $\lim_{t\to 0} P_{ij}(t) = 1$, si $i = j$ y 0 en otro caso.

Usando series de Taylor, con Δt el incremento infinitesimal en t, se tiene que:

$$P_{ij}(t, t+\Delta t) = P_{ij}(t) + \Delta t P'_{ij}(t) + \tfrac{\Delta t^2}{2} P''_{ij}(t) + \cdots$$

con $t = 0$,

$$P_{ij}(\Delta t) = P_{ij}(0) + \Delta t P'_{ij}(0) + \tfrac{\Delta t^2}{2} P''_{ij}(0) + \cdots$$

tomando el límite cuando $\Delta t \to 0$,

$$\lim_{\Delta t\to 0} \tfrac{P_{ij}(\Delta t)}{\Delta t} = P'_{ij}(0) = \lambda_{ij}, \qquad i \neq j$$

y para cuando $i = j$,

$$\lim_{\Delta t\to 0} \frac{P_{ii}(\Delta t) - 1}{\Delta t} = P'_{ii}(0) = -\lambda_{ii}. \tag{2.9}$$

donde λ_{ij} son tal que,

$$\sum_{j\neq i} \lambda_{ij} = \lambda_{ii}. \tag{2.10}$$

Estas tasas de transición infinitesimales λ_{ij} son consecuencia directa de la propiedad $\sum_{j\in S} P_{ij}(t) = 1$.

Dichas tasas de transición son conocidas como *generadores*, matricialmente queda de la siguiente manera:

$$\mathbf{A} = \begin{bmatrix} -\lambda_{00} & \lambda_{01} & \lambda_{02} & \dots \\ \lambda_{10} & -\lambda_{11} & \lambda_{12} & \dots \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

Esta matriz generadora $\mathbf{A}$ cumple la misma función que la matriz de probabilidades de transición $\mathbf{P}$ (con $n = 1$) para las cadenas de Markov en tiempo discreto.

2.7. Proceso de Nacimiento y Muerte

Un proceso de nacimiento y muerte es un caso particular de un proceso de Markov en tiempo continuo, donde la transición entre estados puede ser solamente de dos tipos: "nacimientos"y "muertes", los cuáles incrementan y disminuyen la población. Es una clase de sistema cuyo estado está completamente determinado por el número de individuos presentes en cada instante de tiempo. Supongamos que siempre que haya n individuos en el sistema:

1. los nuevos arribos al sistema se realizan a una tasa exponencial λ_n

2. los individuos abandonan el sistema a una tasa exponencial μ_n

Es decir, el tiempo hasta la llegada de un nuevo individuo es una variable aleatoria exponencial con parámetro λ_n, y el tiempo hasta la partida de uno de los individuos del sistema es una variable aleatoria exponencial con parámetro μ_n. Siendo estas variables independientes entre sí.

Este tipo de proceso estocástico es de tiempo continuo y espacio de estados discreto, donde $\{\lambda_n\}_{n=0}^{\infty}$ y $\{\mu_n\}_{n=1}^{\infty}$ son las tasas de nacimiento y muerte respectivamente.

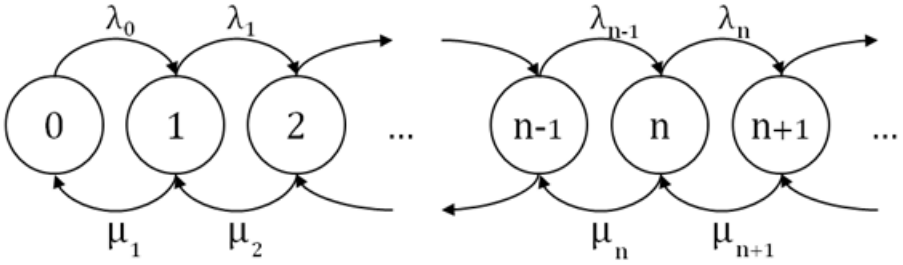

Figura 2.5: Diagrama de transición entre estados

En estos procesos, desde cualquier estado n solamente hay transiciones hacia los estados $n+1$ o $n-1$.

La relación entre las tasas de nacimiento y muerte del proceso con las tasas y probabilidades de transición de la cadena son:

1. λ_0, tasa de transición del estado 0
   $\forall i > 0$, $\lambda_i + \mu_i$ tasa de transición del estado i.

2. probabilidades de transición:

$$p_{01} = 1$$
$$p_{i,i+1} = \lambda_i/(\lambda_i + \mu_i),\ i > 0$$
$$p_{i,i-1} = \mu_i/(\lambda_i + \mu_i),\ i > 0$$

Estos tipos de procesos Markovianos son ampliamente utilizados en modelos poblacionales, donde nacimientos y muertes representan incrementos y decrementos en el tamaño de la población. En sistemas de colas estos eventos corresponden a arribos y partidas.

Cuando la población es de tamaño n, las tasas de transición infinitesimales de nacimiento y muerte son λ_n y μ_n respectivamente. Si se hace referencia a clientes en el sistema, λ_n es la tasa de arribos y μ_n es la tasa de servicio.

Si los arribos provienen de un proceso de Poisson y los tiempos de servicios son exponenciales, las probabilidades de transición durante $(t, t + \Delta t]$, son:

desde el punto de vista de los nacimientos:

1. $P(\text{un nacimiento}) = \lambda_n \Delta t + o(\Delta t)$
2. $P(\text{ningún nacimiento}) = 1 - \lambda_n \Delta t + o(\Delta t)$
3. $P(\text{más de un nacimiento}) = o(\Delta t)$

y del punto de vista de las muertes:

1. $P(\text{una muerte}) = \mu_n \Delta t + o(\Delta t)$
2. $P(\text{ninguna muerte}) = 1 - \mu_n \Delta t + o(\Delta t)$
3. $P(\text{más de una muerte}) = o(\Delta t)$,

En cada uno de los dos casos, los términos $o(\Delta t)$ suman 0, de modo que la probabilidad total de los tres eventos suma 1.

Se considera a $Q(t)$ como el número de clientes en el sistema al momento t y se define:

$$P_{i,n}(t) = P[Q(t) = n|Q(0) = i].$$

Entonces, incorporando las probabilidades de transición durante $(t, t + \Delta t]$, se tiene que:

$$\begin{aligned}
P_{n,n+1}(\Delta t) &= \lambda_n \Delta t + o(\Delta t), && n = 0, 1, 2, \ldots \\
P_{n,n-1}(\Delta t) &= \mu_n \Delta t + o(\Delta t), && n = 1, 2, 3, \ldots \\
P_{n,n}(\Delta t) &= 1 - \lambda_n \Delta t - \mu_n \Delta t + o(\Delta t), && n = 1, 2, 3, \ldots \\
P_{n,j}(\Delta t) &= o(\Delta t), && j \neq n - 1, n, n + 1.
\end{aligned}$$

Con tasas de transición infinitesimal se obtiene la matriz generadora para un proceso de nacimiento y muerte de un sistema de colas:

$$\mathbf{A} = \begin{bmatrix}
-\lambda_0 & \lambda_0 & & & \\
\mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & & \\
 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & \\
 & & & & \\
 & & & &
\end{bmatrix}$$

Esta matriz $\mathbf{A}$, conduce a la denominada *forward Kolmogorov equation* (ver Apéndice B) para $P_{in}(t)$. Para simplificar se denota $P_{in}(t)$ como $P_n(t)$.

$$\begin{aligned}
P_0'(t) &= -\lambda_0 P_0(t) + \mu_1 P_1(t), \\
P_n'(t) &= -(\lambda_n + \mu_n) P_n(t) + \lambda_{n-1} P_{n-1}(t) \\
&\quad + \mu_{n+1} P_{n+1}(t), \qquad n = 1, 2, 3, \ldots
\end{aligned}$$

Si se consideran las transiciones para el proceso $Q(t)$ durante el período $(t, t+\Delta t]$, se tiene:

$$\begin{aligned}
P_0(t+\Delta t) =& P_0(t)[1-\lambda_0\Delta t + o(\Delta t)] + P_1(t)[\mu_1\Delta t + o(\Delta t)],\\
P_n(t+\Delta t) =& P_n(t)[1-\lambda_n\Delta t - \mu_n\Delta t + o(\Delta t)]\\
&+ P_{n-1}(t)[\lambda_{n-1}\Delta t + o(\Delta t)]\\
&+ P_{n+1}(t)[\mu_{n+1}\Delta t + o(\Delta t)]\\
&+ o(\Delta t), \qquad n = 1, 2, 3, \ldots
\end{aligned}$$

Si se le resta $P_n(t), (n = 0, 1, 2, \ldots)$ en ambos miembros y se divide por Δt, se tiene:

$$\begin{aligned}
\frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} =& -\lambda_0 P_0(t) + \mu_1 P_1(t) + \frac{o(\Delta t)}{\Delta t},\\
\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} =& -(\lambda_n + \mu_n)P_n(t)\\
&+ \lambda_{n-1}P_{n-1}(t) + \mu_{n+1}P_{n+1}(t)\\
&+ \frac{o(\Delta t)}{\Delta t}.
\end{aligned}$$

Cuando $\Delta t \to 0$ se obtienen las ecuaciones de Kolmogorov anteriormente mencionadas.

En un estado de equilibrio, el comportamiento del proceso es independiente del parámetro de tiempo y del valor inicial, es decir,

$$\lim_{t\to\infty} P_{i,n}(t) = p_n, \qquad n = 0, 1, 2, \ldots$$

por lo tanto,

$$P_n'(t) \to 0, \qquad t \to \infty$$

Sustituyendo estos resultados en las ecuaciones de Kolmogorov, se obtiene:

$$\begin{aligned}
0 &= -\lambda_0 p_0 + \mu_1 p_1\\
0 &= -(\lambda_n + \mu_n)p_n + \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1}, \qquad n = 1, 2, \ldots
\end{aligned}$$

Reordenando las mismas, se llega a las denominadas ecuaciones de balance:

$$\lambda_0 p_0 = \mu_1 p_1 \tag{2.11}$$

$$(\lambda_n + \mu_n) p_n = \lambda_{n-1} p_{n-1} + \mu_{n+1} p_{n+1}. \tag{2.12}$$

Estas ecuaciones se pueden resolver de forma recursiva. La ecuación (2.11) queda de la siguiente manera,

$$p_1 = \frac{\lambda_0}{\mu_1} p_0.$$

Para $n = 1$, la ecuación (2.12) se resuelve:

$$\begin{aligned}(\lambda_1 + \mu_1) p_1 &= \lambda_0 p_0 + \mu_2 p_2 \\ \lambda_1 p_1 + \mu_1 p_1 &= \lambda_0 p_0 + \mu_2 p_2\end{aligned}$$

Sustituyendo p_1, queda:

$$p_2 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0$$

recursivamente para $n = 2, 3, \ldots$, se tiene que:

$$\mu_n p_n = \lambda_{n-1} p_{n-1}$$

y por lo tanto,

$$p_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} p_0.$$

Entonces, un sistema constituido por n estados cuenta con $n + 1$ ecuaciones de balance. Este sistema de ecuaciones se puede resolver en función de p_0. Luego, si se impone que los p_n sean una distribución de probabilidad, es decir que $\sum p_n = 1$, se puede hallar p_0:

$$p_0 = \left[1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} \right]^{-1}. \tag{2.13}$$

2.8. Modelo M/M/1

Este es el más simple de los modelos de líneas de espera utilizados en la práctica. Se puede pensar como procesos de nacimiento y muerte donde los arribos son los nacimientos y las muertes como la salida de clientes del sistema. Se asume que los arribos provienen de un proceso de Poisson de tasa λ, a esto hace referencia la primer M de la notación de Kendall. Por lo tanto, el número de clientes, $N(t)$, que llegan durante un intervalo de tiempo $(0, t]$ tiene una distribución de Poisson,

$$P[N(t) = j] = e^{-\lambda t}\frac{(\lambda t)^j}{j!}, \qquad j = 0, 1, 2, \ldots$$

Ello significa que el tiempo entre arribos se distribuye de manera exponencial de parámetro λ. La segunda M en la notación de Kendall supone que el tiempo de servicio también se distribuye exponencial con parámetro μ. El tercer componente del modelo refiere a que existe un solo servidor. Cuando un cliente llega al sistema y éste está libre, es atendido, de lo contrario ingresa a la cola, la cual se comporta FIFO.

Bajo estos supuestos, se tiene que

$$E[tiempo\,entre\,arribos] = \frac{1}{\lambda}$$
$$E[tiempo\,de\,servicio] = \frac{1}{\mu}$$

La relación entre la tasa de arribos y la tasa de servicio juega un papel importante en la medición del rendimiento del sistema. Se define ρ como el volumen de tráfico,

$$\rho = \frac{tasa\,de\,arribos}{tasa\,de\,servicio}.$$

En el caso del modelo $M/M/1$, donde $\rho = \lambda/\mu$, se tiene el caso especial de los procesos de nacimiento y muerte con tasas $\lambda_n = \lambda$ y $\mu_n = \mu$, quedando la matriz generadora de la siguiente manera,

$$\mathbf{A} = \begin{bmatrix} -\lambda & \lambda & & \\ \mu & -(\lambda+\mu) & \lambda & \\ & \mu & -(\lambda+\mu) & \lambda \\ & & & \\ & & & \end{bmatrix}$$

Las correspondientes *forward Kolmogorov equation* para $P_n(t)$ son

$$\begin{aligned} P_0'(t) &= -\lambda P_0(t) + \mu P_1(t), \\ P_n'(t) &= -(\lambda+\mu)P_n(t) + \lambda P_{n-1}(t) \\ &\quad + \mu P_{n+1}(t), \qquad n = 1, 2, 3, \ldots \end{aligned}$$

con $P_n(0) = 1$ cuando $n = i$ y 0 en otro caso.

Para calcular la probabilidad de que haya n clientes en el sistema, p_n, se utilizan las ecuaciones de balance,

$$\begin{aligned} \lambda p_0 &= \mu p_1 \\ (\lambda+\mu)p_n &= \lambda p_{n-1} + \mu p_{n+1}, \qquad n = 1, 2, \ldots \end{aligned}$$

despejando en función de p_0,

$$\begin{aligned} p_1 &= \frac{\lambda}{\mu} p_0 \\ p_n &= \frac{\lambda}{\mu} p_{n-1} = \left(\frac{\lambda}{\mu}\right)^n p_0, \qquad n = 1, 2, \ldots \end{aligned}$$

Por ser p_n una distribución de probabilidad se tiene que $\sum_{n=0}^{\infty} p_n = 1$, entonces:

$$\sum_{n=0}^{\infty} p_n = \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n p_0 = p_0 \frac{1}{1-\lambda/\mu} = 1.$$

Para que esto ocurra es necesaria la condición de convergencia $\lambda/\mu < 1$, la cual es también la condición de estabilidad de este sistema de colas. Resolviendo esta

ecuación, se obtiene $p_0 = 1 - \lambda/\mu$, por lo tanto

$$p_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right), \qquad n \geq 0 \tag{2.14}$$

como $\rho = \lambda/\mu < 1$, se tiene

$$p_n = \rho^n(1-\rho), \qquad n = 0, 1, 2, \ldots$$

En este caso el volumen de tráfico $\rho = 1 - p_0$, es llamado *factor de utilización*. Así como se denota a $Q(t)$ como el número de clientes en el sistema en el momento t, se escribe $Q(\infty) = Q$ para representar a la cantidad de clientes cuando el sistema alcanzó el estado estacionario y se le llama Q_q al número de clientes en la cola, excluyendo al que está siendo atendido.

También se define el número medio de clientes en el sistema como $L = E(Q)$ y el número medio de clientes en la cola como $L_q = E(Q_q)$.

$$L = \sum_{n=1}^{\infty} np_n = \sum_{n=1}^{\infty} n\rho^n(1-\rho) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda} \tag{2.15}$$

y donde L_q es

$$\begin{aligned} L_q &= \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty} np_n - \sum_{n=1}^{\infty} p_n \\ &= L - \rho = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}. \end{aligned} \tag{2.16}$$

Se puede interpretar que el número medio de clientes en el sistema es la suma de la cantidad de clientes promedio en la cola más el número medio de clientes en servicio.

Tiempo de espera del Cliente

Otro elemento importante, en la caracterización de una línea de espera, es el tiempo que transcurre mientras un cliente no es atendido. Existen dos característi-

cas importante, que son el tiempo de espera en la cola y el tiempo que el cliente permanece en el sistema. Este último es igual tiempo en la cola más el tiempo de servicio. Cuando el sistema está en equilibrio, estos tiempos se denotan T y T_q respectivamente.

Cuando hay n clientes en el sistema y el tiempo de servicio es exponencial con parámetro μ, el tiempo total de servicio de los n clientes tiene como función de distribución *Erlang*, con función de densidad

$$f_n(x) = \mu e^{-\mu x} \frac{(\mu x)^{n-1}}{(n-1)!}.$$

Sea $F_q(t) = P(T_q \leq t)$, la función de distribución del tiempo de espera T_q, donde,

$$F_q(0) = P(T_q = 0) = P(Q = 0) = 1 - \rho$$

Debido a la propiedad de pérdida de memoria de la distribución exponencial, el tiempo de servicio restante de los clientes que están siendo atendidos es también exponencial con parámetro μ. Escribiendo $dF_q(t) = P(t < T_q \leq t + dt)$, para $t > 0$,

$$\begin{aligned} dF_q(t) =& \sum_{n=1}^{\infty} p_n e^{-\mu t} \frac{\mu^n t^{n-1}}{(n-1)!} dt \\ =& (1-\rho) \sum_{n=1}^{\infty} \rho^n e^{-\mu t} \frac{\mu^n t^{n-1}}{(n-1)!} dt \\ =& \lambda(1-\rho) e^{-\mu(1-\rho)t} dt \end{aligned}$$

Por la discontinuidad en 0 de la distribución de T_q,

$$\begin{aligned} F_q(t) =& P(T_q = 0) + \int_0^t dF_q(t) \\ =& 1 - \rho e^{-\mu(1-\rho)t} \end{aligned}$$

Se define el tiempo medio de clientes en la cola como $E(T_q) = W_q$,

$$W_q = E(T_q) = \frac{\rho}{\mu(1-\rho)} = \frac{\lambda}{\mu(\mu-\lambda)} \tag{2.17}$$

Como el tiempo total dentro del sistema es la suma entre el tiempo en la cola y el tiempo de servicio, entonces el tiempo medio de un cliente en el sistema, W, es,

$$W = E(T) = E(T_q) + E(tiempo\, servicio)$$

$$W = \frac{\lambda}{\mu(\mu - \lambda)} + \frac{1}{\mu} = \frac{1}{\mu - \lambda} \tag{2.18}$$

De las ecuaciones (2.15) y (2.18) se desprende,

$$L = \frac{\lambda}{\mu - \lambda} = \lambda W. \tag{2.19}$$

Análogamente se aplica para el caso del número medio de clientes en la cola,

$$L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \lambda W_q.$$

La ecuación (2.19) es la denominada *Ley de Little*, la cual establece que el número promedio de clientes en un sistema (L) es igual a la tasa promedio de arribo de los clientes al sistema (λ) por el tiempo promedio que un cliente permanece en el sistema (W).

Proceso de Salida

El producto entre el proceso de arribos y el servicio es lo que se conoce como el proceso de salida. Si el servidor está continuamente ocupado, este coincide con el proceso de servicio. Cuando existe tiempo ocioso hay una pausa en el proceso. Sin embargo, cuando el proceso está en equilibrio, se pueden derivar propiedades sin hacer referencia a las llegadas y al servicio.

Sea $t_1, t_2, \ldots$ los momentos de las partidas del sistema y se define $T_n = t_{n+1} - t_n$, el intervalo de tiempo entre la última partida y la próxima. Cuando la cola está en equilibrio, es decir, cuando la intensidad de tráfico es menor a uno ($\rho < 1$), se considera a la variable aleatoria como T.

Sea $Q(x)$ el número de clientes en el sistema x cantidad de tiempo después de la salida,

$$F_n(x) = P[Q(x) = n, T > x].$$

En donde la distribución límite del proceso $Q(t)$ se mantiene igual para cualquier t, ya sea el momento en que se produce un arribo, una salida o cualquier instante t. Por lo tanto, independientemente del valor de x, se tiene

$$P[Q(x) = n] = (1 - \rho)\rho^n, \qquad n = 0, 1, 2, \ldots$$

Se puede determinar $F_n(x)$ como,

$$F(x) = P(T > x) = \sum_0^\infty F_n(x).$$

Para un n específico, debido a la propiedad de Markov de este tipo de procesos, la variable aleatoria T depende únicamente de n y no de los intervalos entre arribos anteriores. Para establecer la relación entre $Q(x)$ y T, y poder determinar la distribución de T, se comienza considerando las transiciones en el intervalo de tiempo $(x, x+\Delta x]$. Esto quiere decir que se debe considerar la posibilidad de que sólo existan arribos durante $(x, x + \Delta x]$,

$$\begin{aligned} F_0(x, x + \Delta x) &= F_0(x)[1 - \lambda\Delta x] + o(\Delta x) \\ F_n(x, x + \Delta x) &= F_n(x)[1 - \lambda\Delta x - \mu\Delta x] \\ &\quad + F_{n-1}(x)\lambda\Delta x + o(\Delta x), \qquad n = 1, 2, \ldots \end{aligned}$$

Al dividir por Δx y con $\Delta x \to 0$, se obtiene

$$\begin{aligned} F_0'(x) &= -\lambda F_0(x) \\ F_n'(x) &= -(\lambda + \mu)F_n(x) + \lambda F_{n-1}(x) \qquad n = 1, 2, \ldots \end{aligned}$$

Como $F_n(0) = P[Q(0) = n] = p_n$, la primera de estas ecuaciones se resuelve

teniendo en cuenta que

$$\frac{d}{dx} ln F_0(x) = \frac{F_0'(x)}{F_0(x)} = -\lambda$$

por lo tanto,

$$\ln F_0(x) = -\lambda x + C$$
$$F_0(x) = p_0 e^{-\lambda x}$$

Por inducción se tiene que,

$$F_{n-1}(x) = p_{n-1} e^{-\lambda x}, \qquad n = 1, 2, \ldots$$

Finalmente, sustituyendo este resultado en,

$$F_n'(x) = -(\lambda + \mu) F_n(x) + \lambda F_{n-1}(x)$$

se obtiene,

$$F_n'(x) + (\lambda + \mu) F_n(x) = \lambda p_{n-1} e^{-\lambda x}.$$

Multiplicando a ambos lados por $e^{(\lambda+\mu)x}$, integrando y utilizando que $F_n(0) = p_n$, se llega a la forma general:

$$F_n(x) = p_n e^{-\lambda x}, \qquad n = 1, 2, 3, \ldots$$

Resultando la distribución,

$$F(x) = \sum_{n=0}^{\infty} p_n e^{-\lambda x} = e^{-\lambda x}, \tag{2.20}$$

la misma es igual a la distribución de los tiempos entre arribos.

Este importante resultado, lleva a concluir que el proceso de salida del $M/M/1$ en equilibrio tiene la misma distribución de Poisson que el proceso de arribos. Por lo tanto, el número esperado de clientes atendidos en un período de largo t, es igual

a λt, (Bhat, 2008).

2.9. Modelo M/M/c

Este es un modelo de colas, que a diferencia del anterior cuenta con c servidores. Estos modelos son los más utilizados en el análisis de aquellos sistemas con más de un servidor, como bancos, cajas de supermercados, check-in de aeropuertos, call centers, etc. Como en el modelo anterior, se asume que los clientes arriban según un proceso de Poisson y forman una sola cola, el tiempo de servicio tiene una distribución exponencial y existen c servidores independientes entre sí. Si un servidor queda disponible, inmediatamente ingresa el primer cliente que arribó a la cola sin generar tiempo ocioso.

Sea λ la tasa de arribos y μ la tasa de servicio, ello implica que el tiempo entre arribos y el tiempo de servicio se distribuyen de forma exponencial con parámetros λ y μ, respectivamente. La tasa de servicio μ es igual para todos los servidores. Con el fin de utilizar el modelo de nacimiento y muerte, es necesario establecer los valores de λ_n y μ_n, cuando hay n clientes en el sistema. La tasa de arribos no varía con el número de clientes en el sistema, es decir, que se mantiene constante. Por otro lado, la tasa de servicio se ve afectada por la cantidad de clientes en el sistema.

Se supone n $(n = 1, 2, \ldots, c)$ servidores ocupados al momento t. Durante el intervalo de tiempo $(t, t + \Delta t]$, la probabilidad que un servidor se desocupe es $\mu\Delta t + o(\Delta t)$. Como hay n servidores ocupados en t, la probabilidad de que algunos de ellos complete su servicio durante este intervalo puede ser determinado usando la distribución binomial,

$$
\begin{aligned}
&= \binom{n}{1}[\mu\Delta t + o(\Delta t)][1 - \mu\Delta t + o(\Delta t)]^{n-1} \\
&= n\mu\Delta t + o(\Delta t)
\end{aligned}
$$

De una manera similar, la probabilidad de que un número r $(r > 1)$ de servidores

ocupados completen el servicio durante el intervalo $(t, t+\Delta t]$ es,

$$\begin{aligned} &= \binom{n}{r}[\mu\Delta t + o(\Delta t)]^r[1 - \mu\Delta t + o(\Delta t)]^{n-r} \\ &= o(\Delta t) \end{aligned}$$

Por lo tanto, cuando hay n servidores ocupados al momento t, el único evento que puede ocurrir en $(t, t+\Delta t]$ que reduzca el número de clientes en el sistema, es la finalización de un servicio y su probabilidad es $n\mu\Delta t + o(\Delta t)$. Por esto la tasa de servicio en ese momento es $n\mu$. Desde el punto de vista de un proceso de nacimiento y muerte, se tiene,

$$\lambda_n = \lambda \quad n = 0, 1, 2, \ldots$$
$$\mu_n = \begin{cases} n\mu & n = 1, 2, \ldots, c-1, \\ c\mu & n = c, c+1, \ldots. \end{cases}$$

La matriz generadora $\mathbf{A}$ para este proceso es,

$$\mathbf{A} = \begin{array}{c} 0 \\ 1 \\ \vdots \\ c \\ c{+}1 \\ \vdots \end{array} \begin{bmatrix} -\lambda & \lambda & & & & & \\ \mu & -(\lambda+\mu) & \lambda & & & & \\ & & \ddots & & & & \\ & & & c\mu & -(\lambda+c\mu) & \lambda & \\ & & & & c\mu & -(\lambda+c\mu) & \lambda \\ & & & & & \ddots & \end{bmatrix}$$

Sea $Q(t)$ el número de clientes en el sistema al momento t y $P_n(t) = P[Q(t) = n|Q(0) = i]$. La probabilidad límite para $p_n = \lim_{t\to\infty} P_n(t)$, se tiene escribiendo $PA = 0$,

$$\begin{aligned} \lambda p_0 &= \mu p_1 \\ (\lambda + n\mu)p_n &= \lambda p_{n-1} + (n+1)\mu p_{n+1}, && 0 < n < c \\ (\lambda + c\mu)p_n &= \lambda p_{n-1} + c\mu p_{n+1}, && c \leq n < \infty \end{aligned}$$

Como se vió en el modelo $M/M/1$, se llega a la solución recursivamente,

$$\begin{aligned} n\mu p_n &= \lambda p_{n-1}, && n = 1, 2, \ldots, c, \\ c\mu p_n &= \lambda p_{n-1}, && n = c+1, c+2, \ldots \end{aligned}$$

Por lo tanto,

$$\begin{aligned} p_n &= \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0, && 0 \le n \le c, \\ p_{c+r} &= \left(\frac{\lambda}{c\mu}\right)^r p_c, && r = 0, 1, 2, \ldots, \\ p_n &= \left(\frac{\lambda}{c\mu}\right)^{n-c} p_c, && n = c, c+1, \ldots \end{aligned}$$

Escribiendo $\frac{\lambda}{c\mu} = \rho$ y simplificando, se obtiene,

$$\begin{aligned} p_n &= \frac{1}{n!}(c\rho)^n p_0, && 0 \le n \le c, \\ &= \frac{1}{c!}(c\rho)^c \rho^{n-c} p_0, && c \le n < \infty, \end{aligned}$$

Si en esta ecuación se utiliza la condición que $\sum_0^\infty p_n = 1$, se tiene,

$$p_0 = \left[\sum_{r=0}^{c-1} \frac{(c\rho)^r}{r!} + \frac{(c\rho)^c}{c!(1-\rho)}\right]^{-1}, \tag{2.21a}$$

$$\begin{aligned} p_n &= \frac{(c\rho)^n}{n!} p_0, && 0 \le n \le c, \\ &= \frac{c^c \rho^n}{c!} p_0, && c \le n < \infty, \end{aligned} \tag{2.21b}$$

con $\frac{\lambda}{c\mu} = \rho < 1$. Como $c\mu$ es la tasa máxima de servicio, se puede considerar a ρ como la intensidad de tráfico en el sistema. La ecuación (2.21b) puede ser escrita como,

$$p_n = \rho^{n-c} p_c, \qquad n \ge c,$$

cuando el número de clientes en el sistema es mayor o igual a c, el sistema se comporta como un $M/M/1$ con una tasa de servicio $c\mu$. Por conveniencia se escribe $\alpha = \frac{\lambda}{\mu}$, por lo que $\alpha/c = \rho$ y se reescribe p_0 utilizando α,

$$\begin{aligned} p_0 &= \left[\sum_{r=0}^{c-1} \frac{(c\rho)^r}{r!} + \frac{(c\rho)^c}{c!(1-\rho)}\right]^{-1}, \\ p_n &= \frac{\alpha^n}{n!} p_0, && 0 \le n \le c, \\ &= \frac{\alpha^c}{c!}\left(\frac{\alpha}{c}\right)^{n-c} p_0, && c \le n \le \infty. \end{aligned}$$

Cualquier cliente que arribe al sistema tendrá que esperar a ser atendido sólo si el número de clientes es mayor a c. La probabilidad de que este evento ocurra viene dada por $\sum_{n=c}^{\infty} p_n$, y por lo tanto

$$P(espera) = C(c, \alpha)$$

$$P(espera) = \frac{\alpha^c}{c!}\left(1 - \frac{\alpha}{c}\right)^{-1}\left[\sum_{r=0}^{c-1} \frac{\alpha^r}{r!} + \frac{\alpha^c}{c!}\left(1 - \frac{\alpha}{c}\right)^{-1}\right]^{-1} \tag{2.22}$$

Esta fórmula es conocida como la *fórmula de espera de Erlang* o *segunda fórmula de Erlang.*

Escribiendo L y L_q como el número medio de clientes en el sistema y en la cola respectivamente y utilizando $c\rho = \alpha$, se define,

$$\begin{aligned} L = \sum_{n=1}^{\infty} n p_n &= p_0 \left[\sum_{n=1}^{\infty} n \frac{1}{c!} \alpha^c \rho^{n-c}\right] \\ &= p_0 \left[\sum_{n=1}^{c} n \frac{\alpha^n}{n!} + \sum_{n=c+1}^{\infty} n \rho^{n-c} \frac{\alpha^c}{c!}\right] \\ &= p_0 \left[\alpha \sum_{n=1}^{c} \frac{\alpha^{n-1}}{(n-1)!} + \frac{\alpha^c}{c!} \sum_{n=c+1}^{\infty} n \rho^{n-c}\right] \end{aligned}$$

aplicando el cambio de variable de $r = n - c$,

$$
\begin{aligned}
=&p_0\left[\alpha\sum_{r=0}^{c-1}\frac{\alpha^r}{r!}+\frac{\alpha^c}{c!}\sum_{r=1}^{\infty}(r+c)\rho^r\right]\\
=&p_0\left[\alpha\sum_{r=0}^{c-1}\frac{\alpha^r}{r!}+\frac{\alpha^c}{c!}\Big(\sum_{r=1}^{\infty}r\rho^r+\sum_{r=1}^{\infty}c\rho^r\Big)\right]\\
=&p_0\left[\alpha\sum_{r=0}^{c-1}\frac{\alpha^r}{r!}+\frac{\alpha^c}{c!}\Big(\frac{\rho}{(1-\rho)^2}+\frac{c\rho}{(1-\rho)}\Big)\right]\\
=&\frac{p_0\alpha^c\rho}{c!(1-\rho)^2}+\alpha p_0\underbrace{\left[\sum_{r=0}^{c-1}\frac{\alpha^r}{r!}+\frac{\alpha^c}{c!(1-\rho)}\right]}_{p_0^{-1}}
\end{aligned}
$$

Por lo tanto,

$$
\begin{aligned}
L =&\frac{p_0\alpha^c\rho}{c!(1-\rho)^2}+\alpha, \qquad p_c = p_0\tfrac{\alpha^c}{c!}\\
=&\frac{p_c\rho}{(1-\rho)^2}+\alpha
\end{aligned}
$$

y el número medio de clientes en el cola puede verse como,

$$
\begin{aligned}
L_q &= \sum_{n=c+1}^{\infty}(n-c)p_n\\
&= \sum_{n=c+1}^{\infty}(n-c)\frac{\alpha^c}{c!}\rho^{n-c}p_0\\
&=\frac{\alpha^c}{c!}p_0\sum_{n-c=1}^{\infty}(n-c)\rho^{n-c}\\
&=\frac{\alpha^c}{c!}p_0\sum_{r=1}^{\infty}r\rho^r\\
&=\frac{p_0\alpha^c\rho}{c!(1-\rho)^2}\\
&=\frac{p_c\rho}{(1-\rho)^2}
\end{aligned}
$$

Comparando los resultados obtenidos para L y L_q, se supone que el término $c\rho$

representa el número esperado de servidores ocupados. También respresenta el factor de utilización de los c servidores. La utilización de cada uno de los servidores se puede ver como:

$$\sum_{n=1}^{c-1} \frac{n}{c} p_n + \sum_{n=c}^{\infty} p_n$$

utilizando la expresión de p_n en esta ecuación se llega a que el factor de utilización individual de los servidores es igual a ρ.

Tiempo de espera

Cuando el número de clientes en el sistema es mayor que c, el tiempo entre salidas es exponencial con tasa $c\mu$. Sea T_q el tiempo de espera del cliente cuando $t \to \infty$ y $F_q(t) = P[T_q \leq t]$, se aprecia que,

$$\begin{aligned} F_q(0) =& P[T_q = 0] = P(Q < c) \\ =& \sum_{n=0}^{c-1} p_n \\ =& p_0 \sum_{n=0}^{c-1} \frac{\alpha^n}{n!} \end{aligned}$$

De la expresión de p_0 se deduce,

$$\sum_{n=0}^{c-1} \frac{\alpha^n}{n!} = \frac{1}{p_0} - \frac{\alpha^c}{c!}(1-\rho)^{-1}$$

por lo tanto

$$F_q(0) = 1 - \frac{\alpha^c p_0}{c!(1-\rho)}$$

Como se vió para la cola $M/M/1$, pero en este caso con múltiples servidores, se tiene

$$\begin{aligned} dF_q(t) =& \sum_{n=c}^{\infty} p_n e^{-c\mu t} \frac{(c\mu t)^{(n-c)}}{(n-c)!} c\mu dt \\ =& p_c e^{-c\mu t} \sum_{n=c}^{\infty} \rho^{n-c} \frac{(c\mu t)^{(n-c)}}{(n-c)!} c\mu dt \\ =& c\mu p_c e^{-c\mu(1-\rho)^t} dt \\ =& \frac{c\mu\alpha^c}{c!} p_0 e^{-c\mu(1-\rho)^t} dt. \end{aligned}$$

Como el valor de la función de distribución del tiempo de espera de los clientes, evaluado en cero, no contribuye al valor esperado de dicho tiempo, T_q, se tiene que

$$\begin{aligned} W_q =& \int_0^{\infty} t dF_q(t) = \int_0^{\infty} c\mu p_c t e^{-c\mu(1-\rho)^t} dt \\ =& \frac{p_c}{c\mu(1-\rho)^2} \end{aligned}$$

multiplicando por λ, se llega a la expresión L_q vista anteriormente. Nuevamente se verifica la fórmula de *Little*, $L_q = \lambda W_q$.

Proceso de Salida

El procedimiento mencionado en la cola $M/M/1$ también es aplicable al caso $M/M/c$, llegando al mismo resultado para la distribución del proceso de salida, siendo

$$F_n(x) = p_n e^{-\lambda x}, \qquad n = 1, 2, 3, \ldots$$

y

$$F(x) = \sum_{n=0}^{\infty} p_n e^{-\lambda x}$$

$$F(x) = e^{-\lambda x}. \tag{2.23}$$

2.10. Simulación

Los resultados obtenidos en las secciones anteriores son utilizados habitualmente en la práctica y los mismos se basan en dos grandes supuestos, que los arribos provienen de un proceso de Poisson y que el tiempo de servicio viene de una familia exponencial. Hay situaciones en la que estos supuestos o alguno de ellos no se cumplen, lo que determina que las soluciones analíticas sean muy complejas o no sean posibles de determinar. Dadas estas dificultades analíticas, en este trabajo se optó por utilizar métodos de simulación que permite describir con exactitud el comportamiento de un sistema complejo (Devroye, 1986).

2.10.1. Simulación de Procesos de Poisson

2.10.1.1. Simulación de Procesos de Poisson Homogéneos

El objetivo es simular en el período $[0, K]$ un proceso de Poisson homogéneo con parámetro λ.

Método 1: Espaciamiento exponencial

Este método se basa en la definición 2.4.1,

1. Inicia el tiempo en 0, por lo tanto $Z \leftarrow 0$.
2. Inicia el número de sucesos en 0, $k \leftarrow 0$.
3. Se genera una variable exponencial $Exp(1)$, E.
4. Se contabilizan los sucesos, $k \leftarrow k + 1$.
5. $Z \leftarrow Z + E/\lambda$.
6. $Z_k \leftarrow Z$.
7. Se continúa con el algoritmo hasta que $Z_1 + \ldots + Z_k$ supere a K.

Método 2: Distribución condicional

Este método se basa en el Teorema 2.4.3

1. Se genera una variable aleatoria N con distribución $Pois(\lambda K)$.
2. Se generan N variables aleatorias $X_1, \ldots, X_N$ iid uniformemente distribuidas en [0,K].

2.10.1.2. Simulación de Procesos de Poisson No Homogéneos

Método 1: Método de la densidad

1. Se genera una variable aleatoria N con distribución $Pois\left(\Lambda(K)\right)$.
2. Se generan N variables aleatorias $X_1, \ldots, X_N$ iid con distribución

$$\frac{\Lambda(t)}{\Lambda(K)}, \qquad 0 \leq t \leq K$$

Se puede observar que si el proceso es homogéneo coincide con el método de la distribución condicional.

2.10.2. Recocido Simulado

En el presente trabajo el supuesto que los tiempos de servicio son exponenciales no se cumple, notándose claramente una distribución bimodal, que se modela paramétricamente por una mezcla de normales, como la estimación por máxima verosimilitud no tiene una solución cerrada, se debe utilizar algún algoritmo de optimización. Por ejemplo, el algoritmo de Recocido Simulado (Simulated Annealing), ver por ejemplo (Haggstrom, 2002). Otra posibilidad podría ser el algoritmo esperanza−maximización o algoritmo EM.

El funcionamiento de este algoritmo procede del proceso físico del templado de metales. En los metales, para conseguir que la estructura molecular del mismo tenga

las propiedades deseadas de resistencia o flexibilidad, es necesario controlar la velocidad del proceso de templado (enfriamiento). Si se hace adecuadamente, el estado final del metal es un estado de mínima energía. Si el descenso de la temperatura se hace rápidamente, no se alcanza el estado de mínima energía, y en su lugar se llega a un estado con mayor energía, (Brooks y Morgan, 1995).

La idea del algoritmo es simular una cadena de Markov, con espacio de estados S, cuya única distribución estacionaria concentre la mayoría de la probabilidad en un conjunto de estados $s \in S$, que minimiza (o en algunos casos, maximiza) una función $f(s)$. Si se simula la cadena por un período largo de tiempo, se alcanza el equilibrio. Entonces se simula una nueva cadena cuya distribución estacionaria se concentra más aún en los s que den lugar a valores pequeños de f y así sucesivamente.

La primera etapa del algoritmo es elegir una temperatura inicial T_1 y fijar una secuencia de temperaturas decrecientes $T_1 > T_2 > \ldots$, con T_i tendiendo a 0 cuando $i \to \infty$, y una secuencia de números naturales $N_1, N_2, \ldots$ que representan los pasos a simular de la cadena para cada uno de los valores de temperatura, estos N tienen que ser suficientemente grandes como para que cada una de las cadenas alcance el estado estacionario.

Luego se simula el algoritmo de Metrópolis con la distribución de Boltzmann en S y T_1, durante N_1 pasos. Después se vuelve a simular el algoritmo en S y T_2, durante N_2 pasos y así sucesivamente hasta que el sistema llegue al punto de "mínima energía", es decir, que encuentra el óptimo.

La implementación de este algoritmo sigue los siguientes pasos:

1. Dada una temperatura inicial T_1, se sortea un estado inicial x_a y se calcula $f(x_a)$.

2. Se selecciona aleatoriamente un posible candidato x_p cercano al estado original y se calcula $f(x_p)$.

3. Se comparan los dos candidatos utilizando el método de Metrópolis.

$$U \leq e^{(f(x_p)-f(x_a))/T_i}$$

con $U \sim Unif(0,1)$. Si se cumple esta desigualdad la cadena acepta el estado x_p.

4. Independientemente si la cadena se mueve de estado o no, se repiten los pasos 2 y 3 durante N_i pasos.

5. Una vez alcanzado el equilibrio para la temperatura T_i, la temperatura disminuye a T_{i+1}, según la secuencia de temperaturas planteadas. El proceso retorna al paso 2.

Capítulo 3

Resultados

En este capítulo se hace un acercamiento al funcionamiento del call center, se describe cómo se obtuvieron los datos asociados a cada llamada y se estudian los distintos procesos asociados a la teoría de colas, los cuales son utilizados más adelante para llegar al objetivo principal del estudio, la optimización de los recursos.

Como se mencionó anteriormente en la sección 2.9, en la práctica el modelo más empleado por los call centers para estimar el rendimiento del sistema es el $M/M/c$, también conocido como *Erlang-C*. Es utilizado principalmente por la simplicidad que presenta en el cálculo de la probabilidad de que todos los servidores se encuentren ocupados y por lo tanto para determinar el número total de agentes necesarios. Los supuestos de este modelo son que el proceso de arribos y el de servicio se distribuyen exponenciales (M), que el sistema cuenta con c servidores y los clientes tienen paciencia "infinita", lo cual hace que no exista abandono. En general algunos de estos supuestos no se cumplen y/o directamente no son testeados. La comprobación o refutación de estos supuestos motivan el presente estudio.

Antes de comenzar con el estudio de los procesos de llegada y de servicio, se empieza describiendo el recorrido que podía realizar una llamada cuando ingresaba al sistema.

3.1. Recorrido que realiza una llamada

Al ingresar una llamada al sistema se derivaba a una contestadora automática, denominada *IVR* (*Interactive Voice Response*), la cual le brindaba información general al cliente. Los que permanecían en el sistema demostraban la necesidad de hablar con un agente y por lo tanto ingresaban a la cola, en otro caso eran consideradas llamadas "fantasmas". Si al momento de ingresar a la cola un agente se encontraba disponible, automáticamente se le asignaba dicha llamada para iniciar la conversación, de lo contrario, el cliente continuaba en la cola. Estos clientes eran atendidos bajo la disciplina FIFO y sin prioridades de atención. El recorrido que realizaba la llamada se puede ver en forma esquemática en la Figura 3.1.

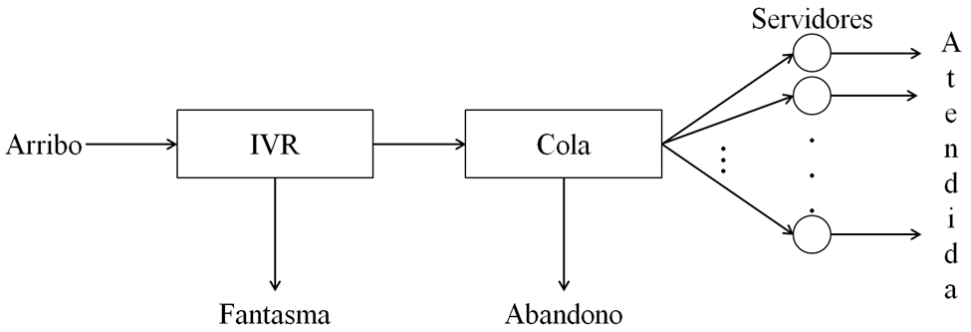

Figura 3.1: Recorrido de una llamada que ingresa al call center

Todas las llamadas que abandonaban en el IVR o en los primeros 5 segundos en la cola no eran factibles de ser atendidas, por lo tanto eran consideradas llamadas fantasmas y no se tenían en cuenta. Todas las llamadas que ingresaban a la cola tenían dos posibles destinos: el cliente abandonaba el sistema sin ser atendido o era atendido por un operador.

Para aquellas llamadas en donde el cliente abandonaba el sistema luego de estar en la cola, se registraba el horario de ingreso y el momento en que dejaba el sistema, y por tanto el tiempo que permaneció esperando. Por otro lado, para las llamadas que fueron atendidas se registraban los mismos tiempos que el caso anterior y además los tiempos vinculados a la atención, el momento en que fue atendida y el tiempo total de atención.

Para el análisis del funcionamiento del call center, se buscaron meses donde el comportamiento del mismo no se viera afectado por eventos particulares. Estos eventos refieren a feriados, vacaciones escolares, problemas técnicos del servicio y otros imponderables que afectaban y/o alteraban los factores fundamentales del proceso.

3.2. Análisis exploratorio

Los datos analizados provenían de un call center que brindaba un servicio asociado a la educación, siendo sus principales usuarios jóvenes y niños. Aunque, eventualmente este servicio comenzó a funcionar en el mes de marzo del año 2010, fue aumentando su alcance con el correr de los años y culminó en diciembre de 2013.

A lo largo de su existencia, el total de llamadas recibidas fue superior a 2.000.000. En el Cuadro 3.1 se presentan algunas medidas descriptivas mensuales y la evolución se muestra en la Figura 3.2. Aquí se aprecia una cierta estacionalidad con menor tráfico de llamadas en los meses de verano con un incremento del mismo sobre mitad del año. Este comportamiento era esperable debido a la fuerte vinculación del servicio con el año lectivo.

	2010	**2011**	**2012**	**2013**
Media	49.424	57.436	66.398	45.241
Desvío	9.988	15.051	26.760	6.655
Mínimo	35.786	34.883	35.333	33.856
Máximo	69.807	84.777	114.727	56.869

Cuadro 3.1: Datos descriptivos de llamadas recibidas mensuales

A mediados del año 2012 se implementó un cambio en la operativa de la empresa que contrataba el servicio del call center que afectó a un número importante de clientes, el cual generó un gran aumento en el volumen de llamadas recibidas. A través de la implementación de nuevos procesos y la adaptación de los clientes a los mismos, se logró alcanzar cierta estabilidad en la operativa. Por este motivo

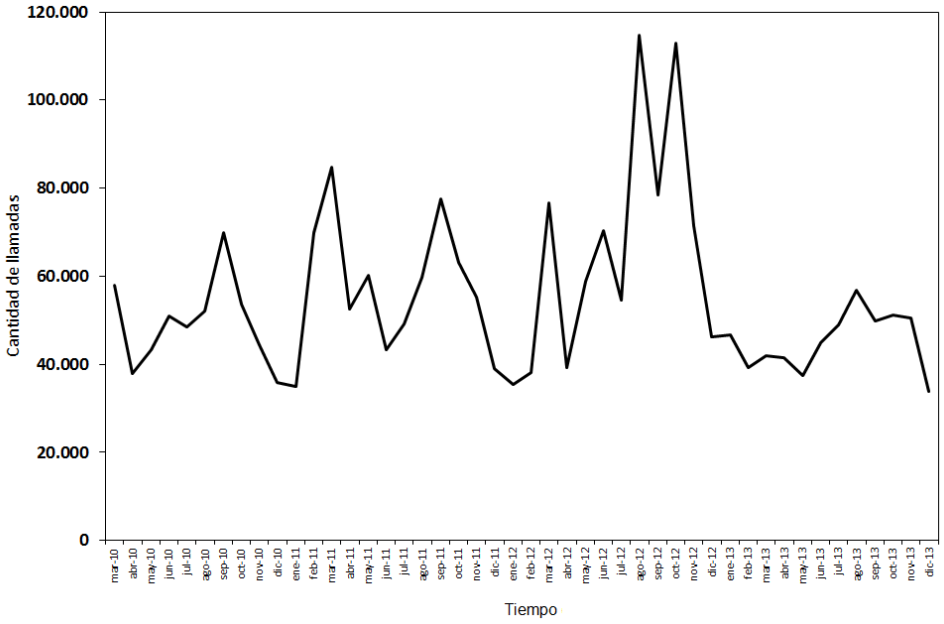

Figura 3.2: Evolución mensual de llamadas, 2010 - 2013

el estudio se enfocó en el año 2013 y su evolución mensual se muestra en la Figura 3.3.

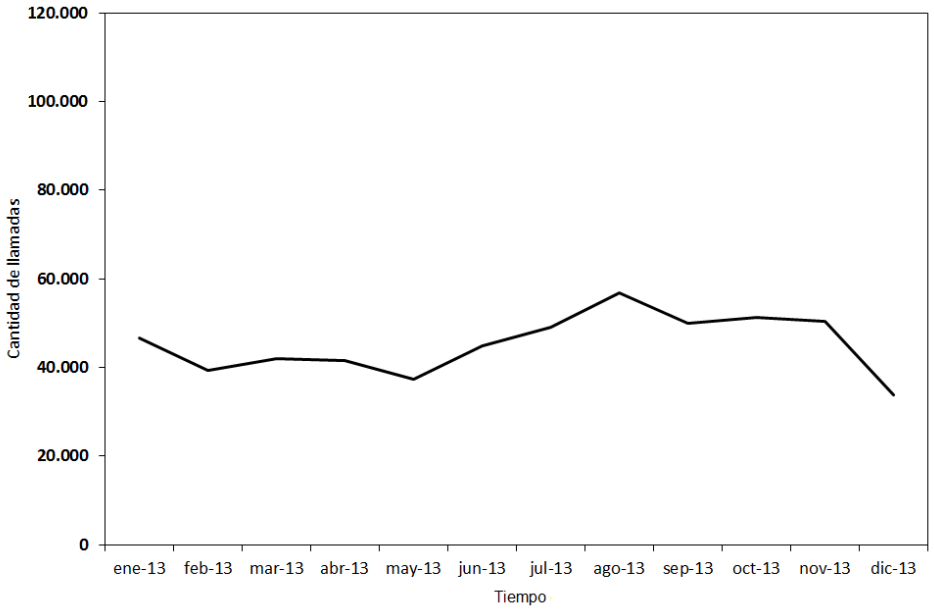

Figura 3.3: Evolución mensual de llamadas del año 2013

Como principales datos descriptivos del año 2013 se tiene que el promedio mensual de llamadas fue aproximadamente 45.000, presentando en diciembre y agosto el mínimo y el máximo respectivamente.

Para analizar si existió estacionalidad semanal (lunes a viernes), es que se cambió la óptica de estudio hacia el volumen diario. La Figura 3.4 muestra el tráfico diario de llamadas recibidas para este año.

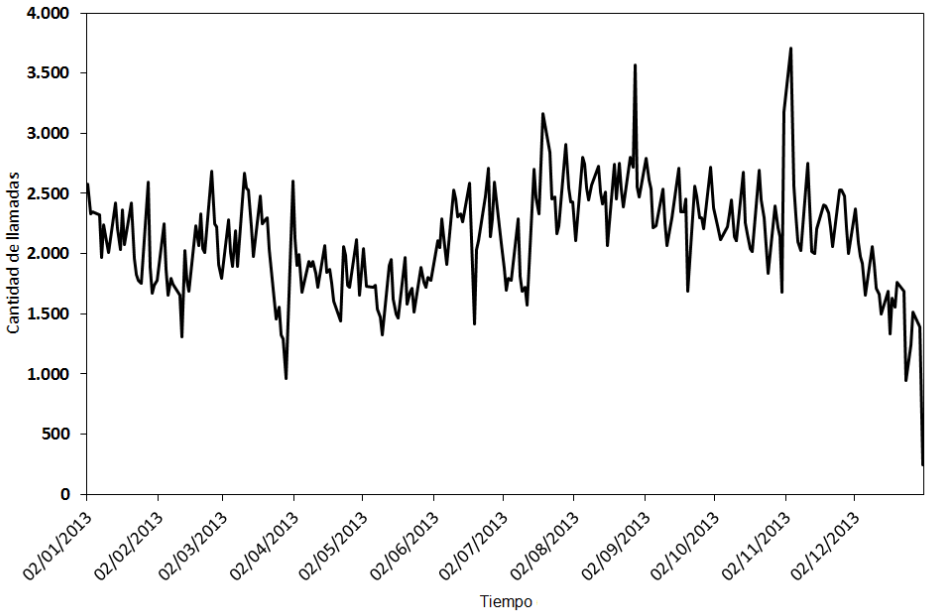

Figura 3.4: Evolución diaria de llamadas del año 2013

Analizando esta serie de datos se puede observar que para el año 2013 la media diaria estaba en el entorno de las 2.000 llamadas y que el día de menor tráfico ocurrió a fines del año. También se nota cierta variabilidad en los datos por lo que para una mejor visualización se muestran en la Figura 3.5 los diagramas de caja para cada mes.

Al analizar en mayor detalle este gráfico, se observan 9 datos atípicos[1] en cuanto

[1]Se entiende por observaciones atípicas a aquellas que están por debajo de $(Q_1 - 1,5 \times IQR)$ y por encima de $(Q_3 + 1,5 \times IQR)$, donde Q_1 y Q_3 son el primer y tercer cuartil respectivamente e $IQR = (Q_3 - Q_1)$ el rango intercuartil.

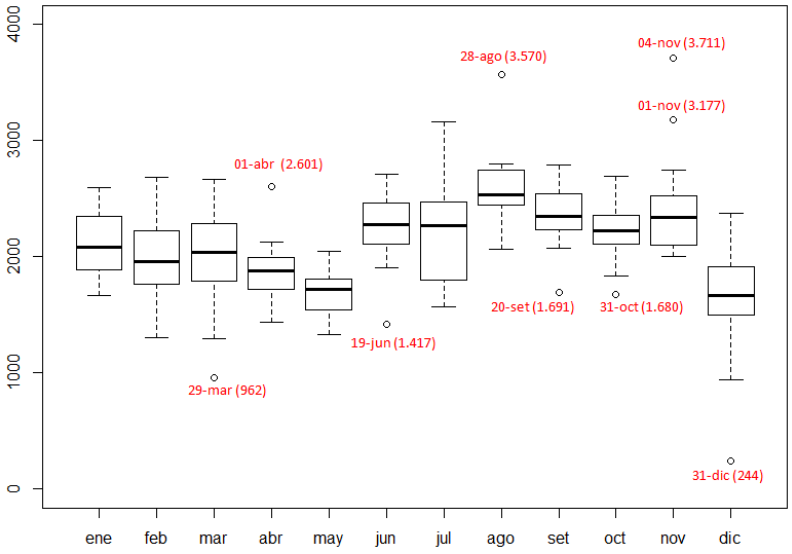

Figura 3.5: Diagrama de caja de llamadas por mes del año 2013

a lo que el volumen de llamadas respecta. En el Apéndice A se detalla los eventos que generaron estos registros.

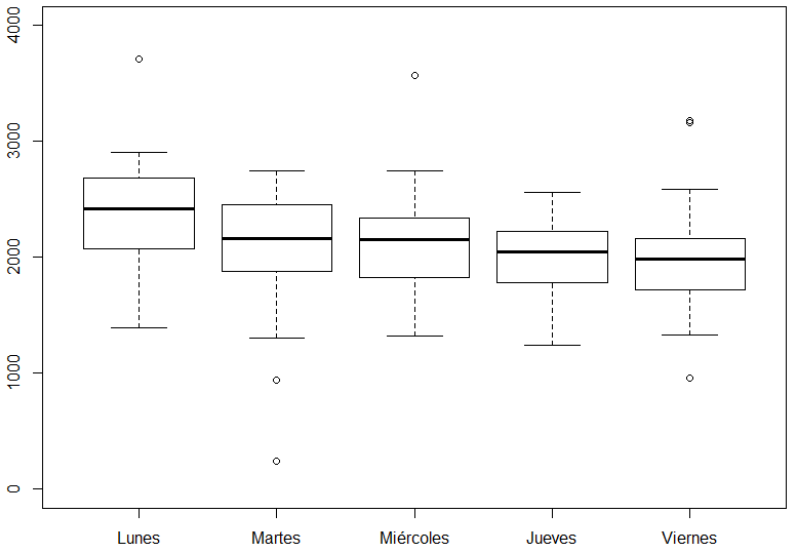

Figura 3.6: Diagrama de caja de llamadas recibidas por día de semana del año 2013

Para analizar si existía alguna relación entre el volumen de llamadas recibidas

en cada día de la semana, se tomó en cuenta las llamadas recibidas durante todo el año, donde la media de los lunes fue de 2.358 llamadas, mientras que en los martes y miércoles sus medias eran similares entre sí con 2.098 y 2.120 llamadas respectivamente y donde los jueves y viernes presentaron las medias más bajas con 2.018 y 1.967 llamadas respectivamente.

Al observar estos datos y sus diagramas de caja, representados en la Figura 3.6, no es claro que los días de la semana tuvieran un comportamiento similar en volumen. Por tanto se utilizó la prueba t de diferencias de medias, comparando dos a dos todos los días de la semana. Los valores p para dichas pruebas se presentan en el Cuadro 3.2.

	Lunes	**Martes**	**Miércoles**	**Jueves**	**Viernes**
Lunes	1	0,0249	0,0019	0,0000	0,0000
Martes		1	0,3331	0,0378	0,0011
Miércoles			1	0,2749	0,0212
Jueves				1	0,2007
Viernes					1

Cuadro 3.2: Valores p de las pruebas t

Estas pruebas indicaron que, con un nivel de significación de 0,05, existía una diferencia significativa entre las medias de los días lunes con el resto de los días de la semana. Por otro lado, las diferencias no fueron significativas entre las medias de los días martes-miércoles, miércoles-jueves y jueves-viernes.

Las medidas de performance del call center, aunque puedan calcularse para intervalos pequeños, a la hora de la toma de decisiones son analizadas de forma mensual. Por este motivo, a modo de ejemplo, la Figura 3.7 muestra las llamadas recibidas por día del mes de octubre y la Figura 3.8 muestra para el mismo mes el promedio de arribos cada media hora.

Como se puede ver en esta última figura, el volumen de llamadas recibidas parecería no ser constante a lo largo del día. Para verlo en mayor detalle, en el Cuadro 3.3

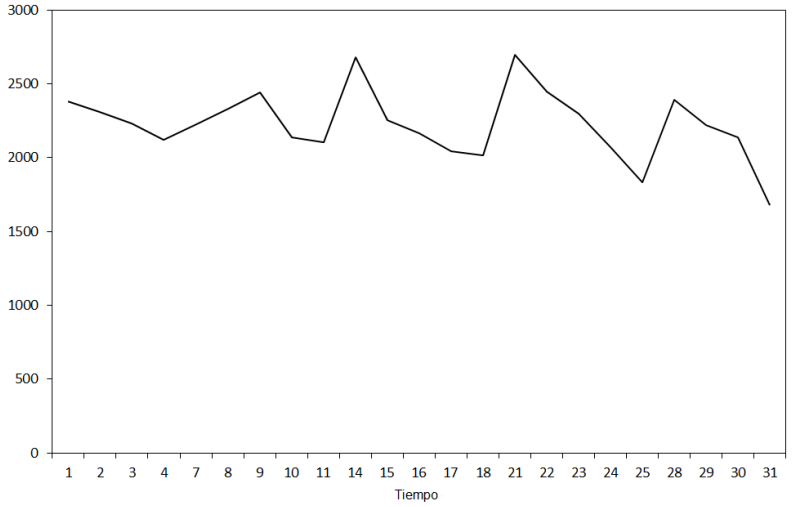

Figura 3.7: Arribos por día trabajado del mes de octubre de 2013

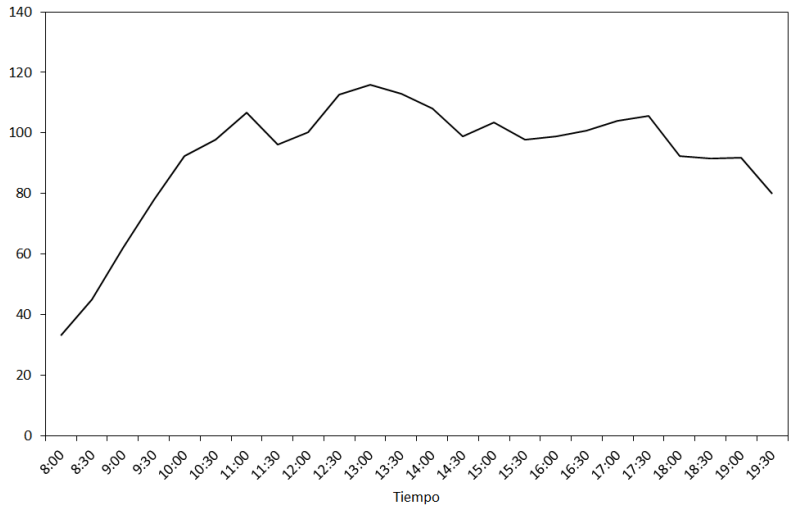

Figura 3.8: Promedio de arribos cada media hora del mes de octubre de 2013

se presenta el promedio de arribos para los intervalos de media hora de este mes en particular. En este cuadro se aprecia que existen diferencias entre los distintos intervalos, habiendo un mínimo de 33 llamadas en la primer media hora del día y un máximo de 116 llamadas en el horario de 13:00 a 13:30. Es por este motivo que en la sección siguiente se evaluó si los arribos provenían de un proceso de Poisson no homogéneo.

Hora	Arribos	Hora	Arribos	Hora	Arribos
8:00 - 8:30	33	12:00 - 12:30	100	16:00 - 16:30	99
8:30 - 9:00	45	12:30 - 13:00	113	16:30 - 17:00	101
9:00 - 9:30	62	13:00 - 13:30	116	17:00 - 17:30	104
9:30 - 10:00	78	13:30 - 14:00	113	17:30 - 18:00	106
10:00 - 10:30	92	14:00 - 14:30	108	18:00 - 18:30	92
10:30 - 11:00	98	14:30 - 15:00	99	18:30 - 19:00	92
11:00 - 11:30	107	15:00 - 15:30	104	19:00 - 19:30	92
11:30 - 12:00	96	15:30 - 16:00	98	19:30 - 20:00	80

Cuadro 3.3: Promedio de arribos cada media hora de octubre de 2013

El segundo proceso de importancia a analizar es el tiempo de servicio. Como primer dato descriptivo, se observó que se atendieron 512.889 llamadas durante todo el 2013 y tuvieron un tiempo medio operativo (TMO) de 218 segundos.

La Figura 3.9 muestra un diagrama de caja para cada mes, donde se aprecia la presencia de datos atípicos extremos[2]. Estas llamadas atípicas no fueron tenidas en cuenta ya que sus duraciones superaban los 15 minutos, considerado excesivo para este tipo de servicio, a lo cual se agrega que sólo representaban un volumen menor al 1 % del total. Por otro lado como criterio *ad hoc* se consideró que las llamadas con duraciones muy cortas no correspondían a la necesidad del servicio. Por este motivo se resolvió no considerar las llamadas menores a 6 segundos, un $0,2\,\%$ del total.

Una vez quitados estos atípicos la media anual del tiempo de servicio se redujo a 209 segundos. En el Cuadro 3.4 se resume la media y el desvío estándar de este tiempo para cada uno de los meses del año 2013. La media muestra una evolución estacional con bajos valores en los meses de principios y fines de año debido a las vacaciones escolares y los valores más altos concentrados en el año lectivo.

[2]Atípico extremo es la observación que se encuentra por encima de $(Q_3 + 3 \times IQR)$ y por debajo de $(Q_1 - 3 \times IQR)$.

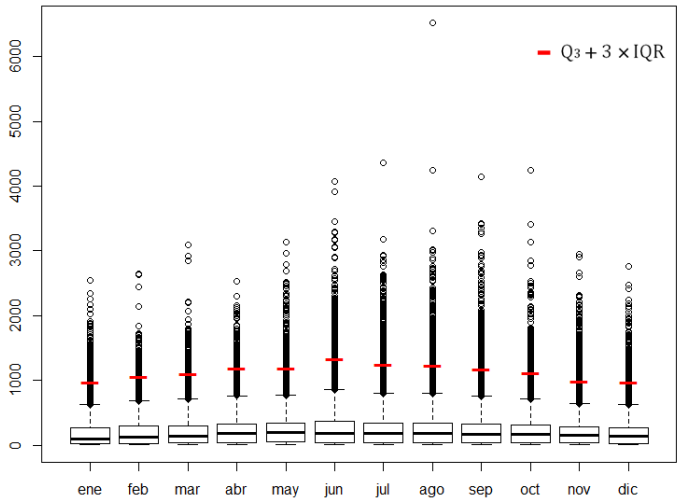

Figura 3.9: Diagrama de caja del tiempo de servicio por mes del año 2013

Mes	Media	Desvío Estandar
Enero	167	181
Febrero	183	188
Marzo	194	194
Abril	221	203
Mayo	227	203
Junio	244	238
Julio	228	225
Agosto	235	226
Setiembre	218	212
Octubre	208	199
Noviembre	196	180
Diciembre	177	176

Cuadro 3.4: Datos descriptivos del tiempo de servicio del año 2013

3.3. Proceso de arribos

Como se mencionó anteriormente, los modelos utilizados en los call centers generalmente asumen que el proceso de arribos es Poisson con tasa constante. En la práctica lo que se hace es fraccionar el día en bloques de medias horas y con esto generar para cada uno de los bloques un modelo de cola distinto, obteniendo de es-

ta manera medidas de performance estimadas que pueden variar abruptamente de un bloque al siguiente. Una mejor práctica sería analizarlo con cierta continuidad, generando bloques de menor tamaño, lo que hace pensar en un proceso de Poisson homogéneo dentro de los bloques. Para comprobar esto se aplicó el test desarrollado en la sección 2.5.1, para la hipótesis nula de que los arribos provenían de un proceso de Poisson.

El primer paso para la construcción de este test implica fraccionar el día en intervalos de tiempo relativamente cortos. Por conveniencia se utilizan bloques de igual duración de tiempo, L, resultando un total de I bloques. En este caso se utilizaron bloques lo suficientemente pequeños, de 6 minutos de duración, con el fin de poder asumir una tasa constante en cada intervalo.

Bajo la hipótesis nula de que la tasa de arribos es constante dentro de cada uno de estos bloques, pero no necesariamente igual entre bloques, los R_{ij} son variables exponenciales independientes de tasa 1.

Un problema que surge en el cálculo de los R_{ij} es que como el tiempo esta medido en segundos, puede ocurrir que dos o más llamadas arriben en el mismo instante, lo cual genera que T_{ij} sea igual a $T_{i,j-1}$, causando que el R_{ij} sea 0. Para evitar esta situación a cada T_{ij}, se le sumó una realización de una variable aleatoria $Unif(0,1)$, agregando así un ruido a cada una de estas llamadas.

Para trabajar con estos bloques, como sugiere el artículo (Brown et al., 2005), se puede seleccionar bloques sucesivos dentro de un mismo día o un mismo tramo horario de bloques para varios días, con la lógica de ver el proceso como un todo y poder testear la exponencialidad dentro del grupo seleccionado.

Para probar la exponencialidad de los R_{ij} se utilizó el test de Kolmogrov-Smirnov, el cual utiliza el estadístico,

$$KS_n = sup_{x \geq 0} |F_n(x) - (1 - exp(-x))|,$$

donde F_n es la función de distribución empírica de los datos dividido su promedio.

La implementación de este test fue realizada en el software estadístico R (R Core Team, 2015), con el paquete *exptest* (Novikov et al., 2013).

De este modo se pueden generar muchos grupos de bloques. Para ilustrar el procedimiento se presenta la salida del test para el día 11 de octubre de 2013 de 8:00 a 20:00, dando un valor del estadístico de Kolmogrov-Smirnov, $KS = 0{,}0193$ con un respectivo valor p de 0,1962. Por lo tanto, bajo un nivel de significación de 0,05, no hubo evidencia para rechazar la hipótesis nula de que los arribos provenían de un proceso de Poisson. Esto se puede ver también en el gráfico $Q - Q$ de cuantiles exponenciales, que muestra la Figura 3.10. La salida del software se puede apreciar en el Apéndice A.

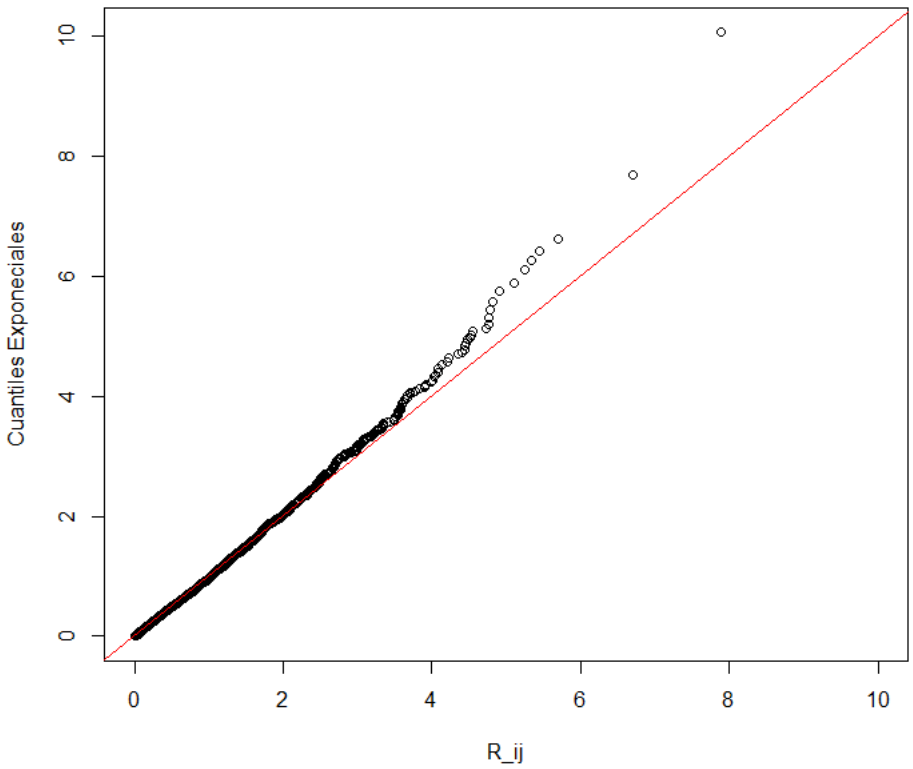

Figura 3.10: Gráfico Q-Q de los Rij del 11 de octubre de 2013

Como se mencionó anteriormente, se probó la hipótesis nula de exponencialidad para varios bloques de distintos días. A modo de ejemplo se presenta la salida del

test para el tramo de media hora que comenzaban a las 14:00 para todo el mes. El valor del estadístico para este test fue $KS = 0{,}0179$ (valor $p = 0{,}1909$), nuevamente no hay evidencia para rechazar la hipótesis nula. Con el gráfico $Q-Q$ de los cuantiles exponenciales, de la Figura 3.11, se puede afirmar este resultado.

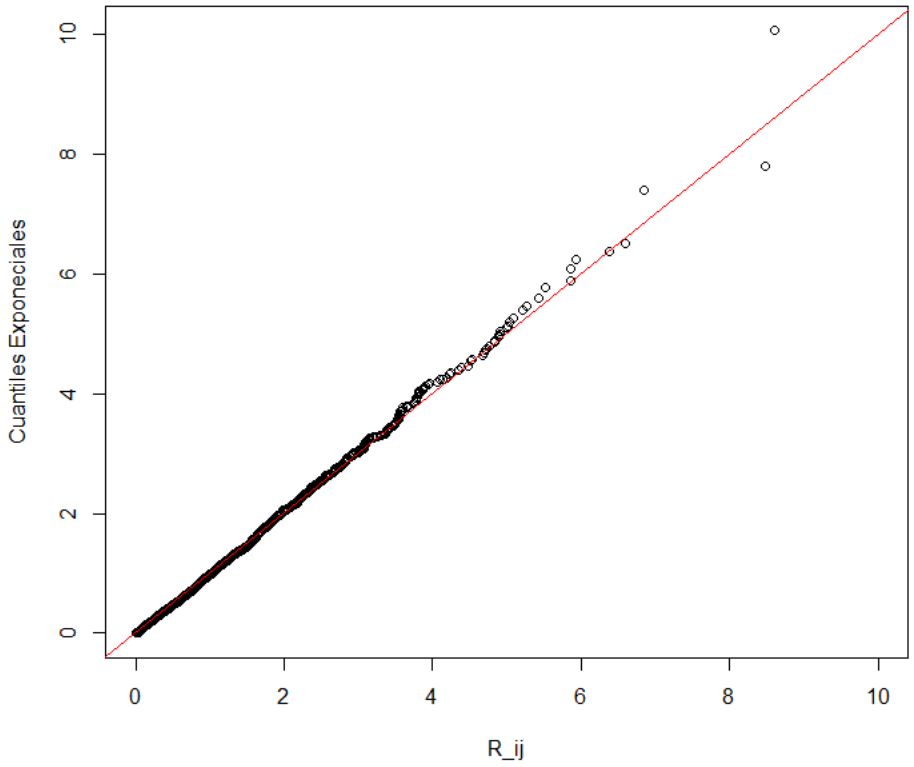

Figura 3.11: Gráfico Q-Q de los Rij del tramo de 14:00 a 14:30 de los días de octubre de 2013

Estos resultados son dos ejemplos de varios que se obtuvieron para distintas selecciones de bloques. Del total de las pruebas realizadas, el 94.8 % no rechaza la hipótesis nula de que los arribos provenían de un proceso de Poisson. Adicionalmente se aplicó este método a las 51,220 llamadas del mes de octubre, en la Figura 3.12 se presenta el Gráfico $Q-Q$ entre los cuantiles exponenciales ($\lambda = 1$) y los cuantiles de estos R_{ij}, donde se puede apreciar que no existen diferencias entre ambas distribuciones.

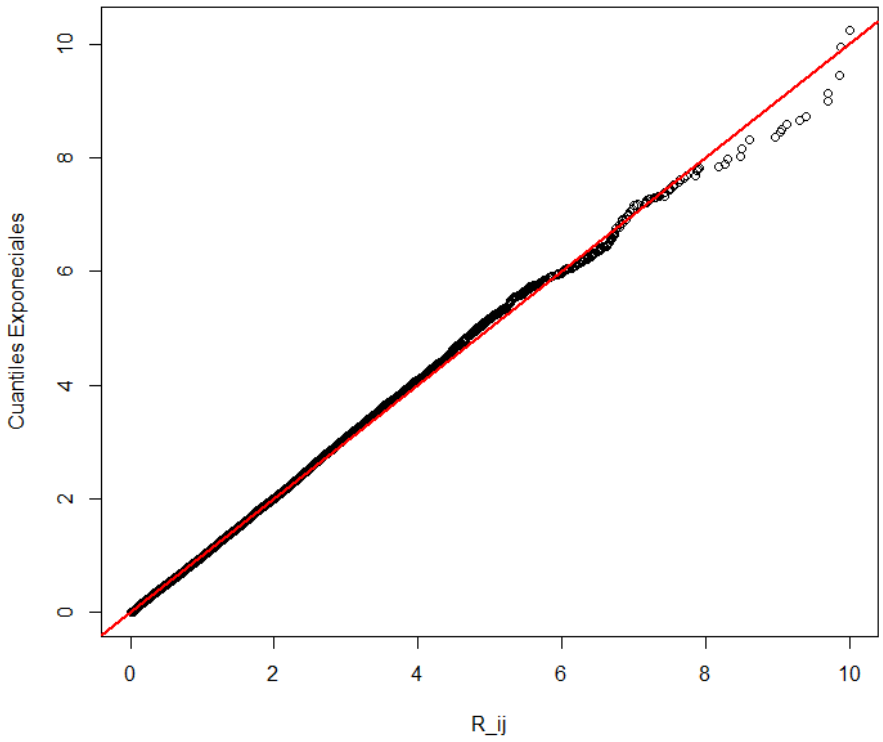

Figura 3.12: Gráfico Q-Q de los Rij del mes de octubre de 2013

Es de hacer notar que en el procedimiento anterior se realizaron múltiples pruebas de hipótesis sobre los mismos datos, el cual pudo llevar al problema de pruebas múltiples ("*multiple testing*"). Existen herramientas para enfrentar este problema, pero en este estudio no se hizo foco en esto, ya que los resultados obtenidos no parecen indicar que el problema sea tan grave.

Luego de concluir que no se rechazó la hipótesis que los arribos provienen de un proceso de Poisson, se estimó la función de intensidad del proceso. Como se vió anteriormente, las pruebas realizadas no descartan que existan diferencias para el volumen de llamadas entre los distintos días de la semana, pareciendo oportuno realizar estimaciones distintas para cada uno de estos días.

En la Figura 3.13 se observa el promedio de las intensidades estimadas para cada día de la semana, las cuales fueron utilizadas posteriormente como insumo en la función de simulación del funcionamiento del call center. Estas parecerían ser muy

similares pero de distinto volumen, lo cual es coherente con los resultados de las pruebas t. Para asegurar las metas del call center, al momento de optimizar se utilizó únicamente el día lunes ya que contaba con mayor volumen de llamadas.

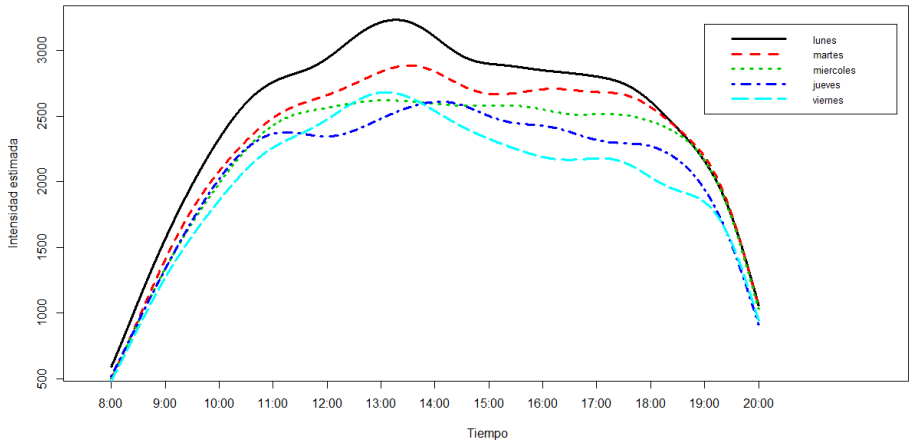

Figura 3.13: Función de intensidad estimada de lunes a viernes de los meses de setiembre a noviembre de 2013

Para realizar estas estimaciones se utilizó el método de densidad por núcleos y se consideraron los datos de los meses de setiembre, octubre y noviembre de 2013. En este caso se empleó un núcleo Gaussiano, pero estos estimadores no son sensibles a dicha elección. Para el ancho de la ventana h, hay que ser más cuidadosos porque la estimación sí es sensible a este aspecto. Se optó por la regla sencilla de Silverman, obteniendo estimaciones relativamente suaves. Con este ancho de ventana se realizó, a través de la función *density* del software R, la estimación y luego se ponderó por el número de llamadas de ese día de manera que la integral coincidiera con el volumen de llamadas diario.

3.4. Tiempo de servicio

La mayoría de los modelos de teoría de colas utilizados por los call centers asumen que el tiempo de servicio se ajusta a una distribución exponencial. En ello influye principalmente la facilidad de su análisis cuando se combina con el supuesto de que

el proceso de arribos es Poisson.

Para el análisis se tomó en consideración un sólo mes por contar con un volumen suficiente de llamadas, en este caso se estudió el mes de octubre del 2013. En el Cuadro 3.5 pueden verse algunas medidas descriptivas de dicho mes, una vez quitados los datos atípicos extremos.

	Mediana	**Media**	**Desvío Estándar**
Tiempo (segs.)	161	208	199

Cuadro 3.5: Datos descriptivos del tiempo de servicio del mes de octubre de 2013

La Figura 3.14 muestra el histograma del tiempo de servicio. En el mismo se puede apreciar una distribución bimodal y por lo tanto suponer que existen dos tipos de llamadas, unas de corta duración que pueden estar vinculadas a consultas básicas, llamadas equivocadas o incluso llamadas molestas, y otro grupo de llamadas de más larga duración vinculadas a las consultas de la problemática diaria e incluso llamadas complejas, que requerían varios minutos para su correcta resolución.

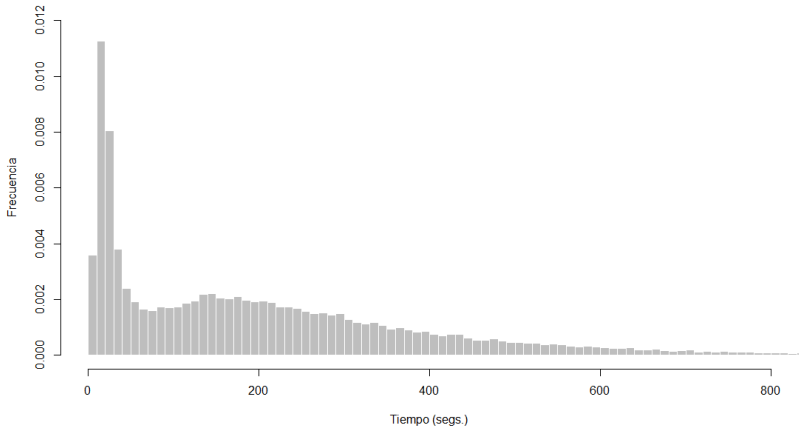

Figura 3.14: Histograma del tiempo de servicio

Mediante la aplicación del test de hipótesis de Kolmogorov-Smirnov para exponencialidad, no hay evidencia para asumir que el tiempo de servicio proviene de una

distribución exponencial, la salida de este test puede verse en el Apéndice A.

Cuando el supuesto de exponencialidad no se cumple hay antecedentes que proponen utilizar el logarítmo de este tiempo (Brown et al., 2005). La Figura 3.15 muestra el histograma del logaritmo del tiempo de servicio donde se ve más claramente la bimodalidad mencionada.

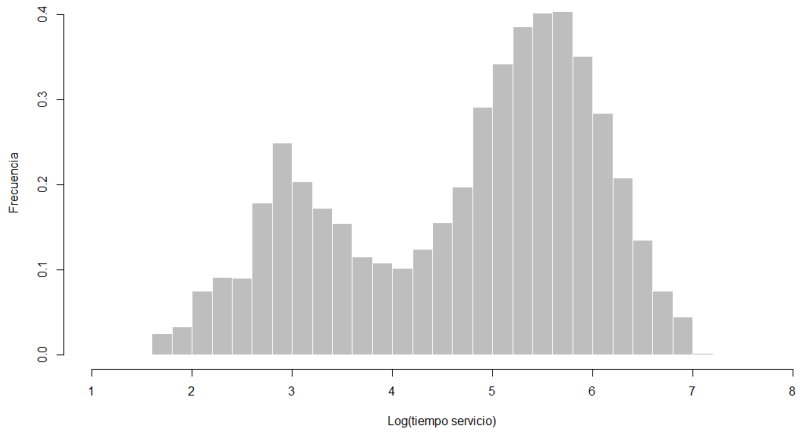

Figura 3.15: Histograma del logarítmo del tiempo de servicio

En vista de la bimodalidad mencionada, una mezcla de distribuciones normales parece ser una aproximación razonable. En consecuencia se estimó dicha distribución por máxima verosimilitud. Esto se hizo minimizando el opuesto del logarítmo de la función de verosimilitud con respecto a los cinco parámetros, los dos de posición (media de las normales), los dos de escala (desvío estándar de las normales) y el parámetro de mezcla. Para la resolución numérica se utilizó el algoritmo de *Recocido Simulado*.

Luego de aplicado este algoritmo, los parámetros estimados y su densidad correspondiente (Figura 3.16) quedaron de la siguiente manera,

$$\hat{\mu_1} = 3{,}003; \qquad \hat{\sigma_1}^2 = 0{,}371$$
$$\hat{\mu_2} = 5{,}504; \qquad \hat{\sigma_2}^2 = 0{,}422$$
$$\hat{\tau} = 0{,}330.$$

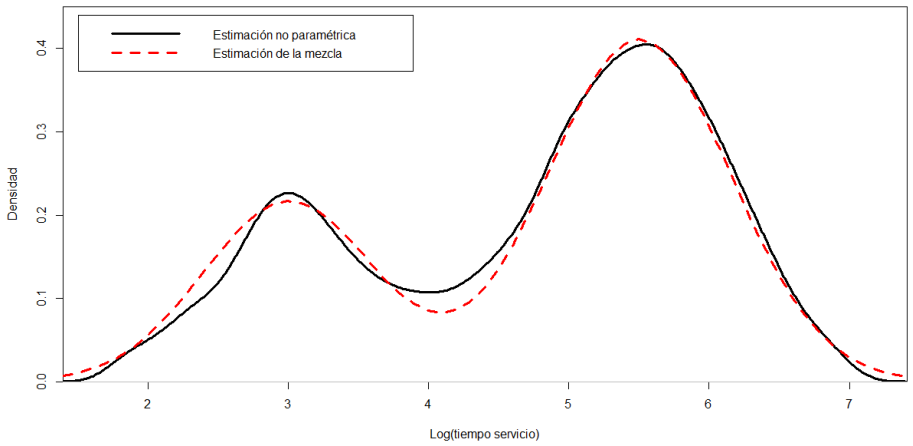

Figura 3.16: Comparativo de las estimaciones del tiempo de servicio

Las estimaciones paramétricas y no paramétricas respaldan el modelo de mezcla de normales por el cual finalmente se opta dada la simplicidad a la hora de simular.

3.5. Tiempo de espera del cliente

El tercer componente fundamental en la teoría de colas es el tiempo de espera. Aquí se debe hacer una distinción entre el tiempo que esperan los clientes que fueron atendidos y el de los que abandonaron. Otra diferencia a tener en cuenta es entre el tiempo que el cliente "necesita" esperar antes de ser atendido, frente al tiempo que está "dispuesto" a esperar antes de abandonar el sistema. El primero hace referencia a un tiempo de espera virtual, porque equivale al tiempo que un cliente con paciencia infinita hubiese esperado hasta ser atendido. El segundo se refiera a la paciencia del cliente.

El tiempo que el cliente está dispuesto a esperar antes de ser atendido es observable únicamente para los casos en que el cliente abandona la espera, quedando censurado en el momento que es atendido.

En el Cuadro 3.6 se muestra la cantidad de llamadas, la media y la mediana del tiempo de espera de todas las llamadas recibidas, atendidas y abandonadas para cada mes del año 2013. Se puede observar que el tiempo de espera total fue muy bajo, 10 segundos en promedio. Esto se debió a que más del 80 % de las llamadas fueron atendidas directamente, es decir que no tuvieron tiempo de espera, como se ve en el histograma del tiempo de espera de las llamadas recibidas (Figura 3.17).

		Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov	Dic	Total
Total	**Media**	4	11	8	13	15	11	15	10	7	7	10	7	10
	Mediana	3	3	3	3	3	3	3	3	3	3	3	3	3
	# Llamadas	46.638	39.292	41.940	41.474	37.416	44.845	49.029	56.869	49.883	51.220	50.430	33.856	542.892
Atendidas	**Media**	4	7	7	10	11	9	11	8	6	6	8	4	8
	Mediana	3	3	3	3	3	3	3	3	3	3	3	3	3
	# Llamadas	46.257	36.648	40.701	38.890	34.049	41.779	43.584	53.538	48.138	49.447	47.341	32.517	512.889
Abandonadas	**Media**	35	60	51	52	56	42	44	35	30	33	39	58	45
	Mediana	26	52	44	43	50	32	32	25	22	23	29	36	33
	# Llamadas	381	2.644	1.239	2.584	3.367	3.066	5.445	3.331	1.745	1.773	3.089	1.339	30.003

Cuadro 3.6: Tiempo de espera en segundos

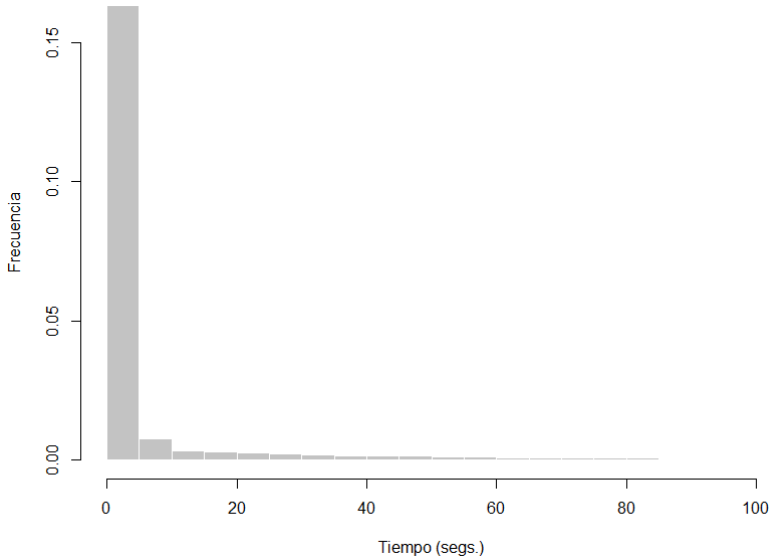

Figura 3.17: Histograma del tiempo de espera de las llamadas recibidas del año 2013

Dentro de las llamadas que abandonaron se evidenció un comportamiento de poca paciencia de los clientes, ya que el 50 % de ellos lo hacen antes de los 33 segundos, dicho comportamiento se ve en la Figura 3.18.

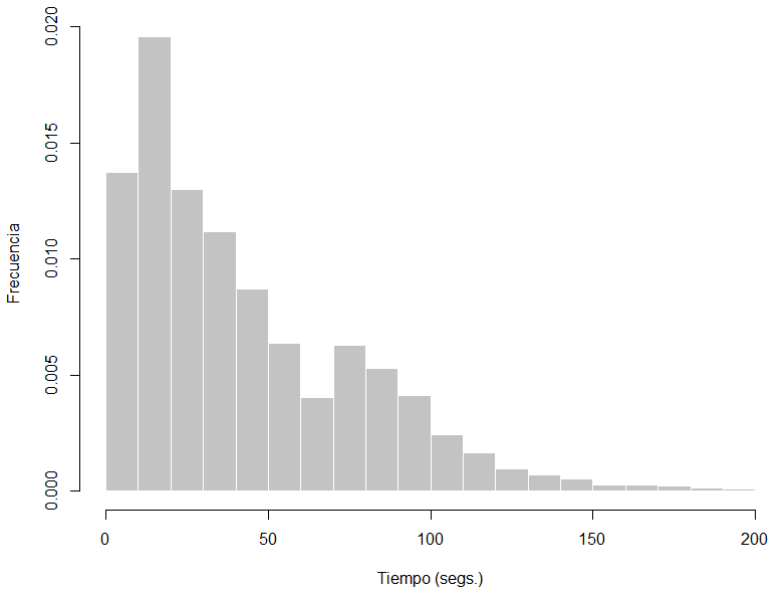

Figura 3.18: Histograma del tiempo de espera de las llamadas abandonadas del año 2013

Debido a lo anteriormente mencionado, el tiempo de espera se asumió como determinístico. Siendo conservadores, para los análisis siguientes este tiempo fue fijado en 45 segundos, es decir, el promedio de espera anual de las llamadas que abandonaron el sistema. Un tratamiento más sofisticado debería tomar en cuenta la censura y quedó fuera del alcance de este trabajo.

3.6. Optimización de recursos

Como se mencionó previamente el costo asociado a los operadores en un call center se supone en el entorno del 70 % del costo total. Por este motivo es importante conocer cual es el número óptimo de operadores para brindar el servicio deseado.

En busca de este objetivo se construyeron dos funciones, la primera de ellas para representar el funcionamiento del call center y la segunda para optimizar la cantidad

de operadores de acuerdo a los requerimientos deseados.

En función que los días de la semana presentaban una diferencia respecto al volumen total de llamadas recibidas y que se contaba con un staff de operadores fijo de lunes a viernes, se decidió trabajar con el día lunes, ya que es el día de la semana que arribaba el mayor flujo de llamadas y de este modo asegurar el alcance de las metas en los siguientes días de la semana.

3.6.1. Simulación del funcionamiento del Call Center

En lo que sigue se describe la construcción de una función que simula el comportamiento del Call center.

Se trabaja con los mismos tres turnos de agentes existentes cuando se relevaron los datos. De 8:00 a 14:00, de 14:00 a 20:00 y un turno que se superponía con los anteriores de 11:00 a 17:00. En estos turnos el número de operadores es variable.

Los arribos de llamadas se simularon de un proceso de Poisson no homogéneo con las intensidades estimadas. El tiempo de servicio se simuló basado en la mezcla de distribuciones normales con los parámetros estimados en la sección anterior.

La función se programó con el software estadístico R y opera de la siguiente manera. En primer lugar a cada turno de atención se le asigna un número determinado de operadores. Luego se simula el total de llamadas recibidas en el día a través de una distribución de Poisson, tomando como tasa el promedio del volumen diario recibido en los últimos tres meses para el correspondiente día de la semana. Seguidamente se distribuyen estos arribos a lo largo del día por medio de la función de intensidad estimada, de este modo se tiene el instante en que las llamadas arriban al sistema. Posteriormente, a cada uno de estos arribos se le asigna un tiempo de servicio, generado aleatoriamente mediante la distribución de mezcla estimada.

Una vez que se obtienen los arribos y sus tiempos de servicio comienza la simulación del funcionamiento del call center. Para cada llamada se evalúa al momento de arribar si hay algún operador disponible, en caso de que lo haya, esta llamada

es atendida inmediatamente. Por el contrario, de no haber un operador disponible, dicho cliente se une al final de la cola y será atendido cuando se libere un operador, o sea, cuando hayan sido atendidas (o abandonadas) todas las llamadas que arribaron antes. Si un cliente alcanza los 45 segundos de espera, se lo quita del sistema y se registra un abandono. Al finalizar la jornada laboral a las 20:00 horas, no se permite el ingreso de nuevas llamadas, y se continua atendiendo hasta que no queden mas clientes en la cola.

Con los datos simulados se calculan las distintas medidas de performance según las definiciones adoptadas al inicio del trabajo. Estas medidas son: nivel de atención, nivel de servicio, ocupación, tiempo medio en el sistema (tiempo promedio que un cliente permanece en el sistema), tiempo medio de espera (tiempo promedio que un cliente está en la cola) y tiempo medio operativo. Para poder inferir sobre estas medidas es necesario realizar este procedimiento un número suficiente de veces.

Los argumentos que utiliza como insumos esta función son:

1. *oper*, cantidad de operadores en cada uno de los tres turnos de 6 horas.
2. *h_entr*, horario de entrada de cada turno.
3. *lim_esp*, límite de espera hasta abandonar el sistema.
4. *ts_obj*, umbral de tiempo para el nivel de servicio, es decir, si la llamada es atendida antes de este umbral se contabiliza para el nivel de servicio.
5. *dia_sem*, dia de la semana que se quiere simular.
6. *iter*, cantidad de iteraciones realizadas.
7. *plot*, argumento lógico si se desea graficar el funcionamiento del call center para cada iteración.

Por otro lado el resultado de la función devuelve dos objetos, el primero es denominado *"metricas"* y el segundo *"media"*. El primero muestra los resultados de cada una de las iteraciones y el segundo el promedio de los resultados de todas las iteraciones. Las medidas en cuestión son las siguientes:

1. *recibidas*, total de llamadas recibidas.
2. *ocup*, nivel de ocupación.
3. *tmo*, tiempo medio operativo.
4. *t_sis*, tiempo promedio que los clientes estuvieron en el sistema.
5. *t_esp*, tiempo promedio que los clientes esperaron en la cola.
6. *nat*, nivel de atención.
7. *ns*, nivel de servicio.

A modo de ejemplo se ejecutó la función para el día lunes, con una disposición de operadores de 9 en el primer turno, 6 en el segundo y 11 en el último. Aquí se asignó 20 segundos como el umbral del nivel de servicio. Al igual que como se planteó anteriormente, se consideró que cuando un cliente alcanzaba los 45 segundos de espera éste abandonaba el sistema.

A continuación se presenta el código utilizado y los resultados de una ejecución con los parámetros mencionados y un gráfico de una de las iteraciones para ilustrar el funcionamiento del mismo (Figura 3.19).

```
> call.center(opers = c(9,6,11), h_entr = c(8,11,14), lim_esp = 45,
ts_obj = 20, dia_sem = "lunes", iter = 200, plot = TRUE)
metricas
      recibidas   ocup tmo t_sis t_esp    nat     ns
 [1,]      2551 0.8258 207   194  12.0 0.8800 0.7162
 [2,]      2635 0.8631 217   198  14.4 0.8467 0.6558
 [3,]      2598 0.8407 206   193  11.3 0.8822 0.7267
 [4,]      2620 0.8475 208   195  13.1 0.8733 0.6924

.
.
.
```

```
[197,]       2618 0.8502 211   196  13.3 0.8640 0.6864
[198,]       2526 0.8215 204   194  10.9 0.8935 0.7458
[199,]       2593 0.8408 215   196  13.7 0.8461 0.6676
[200,]       2623 0.8463 212   195  13.8 0.8563 0.6748

media
recibidas       ocup        tmo     t_sis     t_esp       nat        ns
2582.3700     0.8392   211.0450  195.3850   12.8700    0.8650    0.6945
```

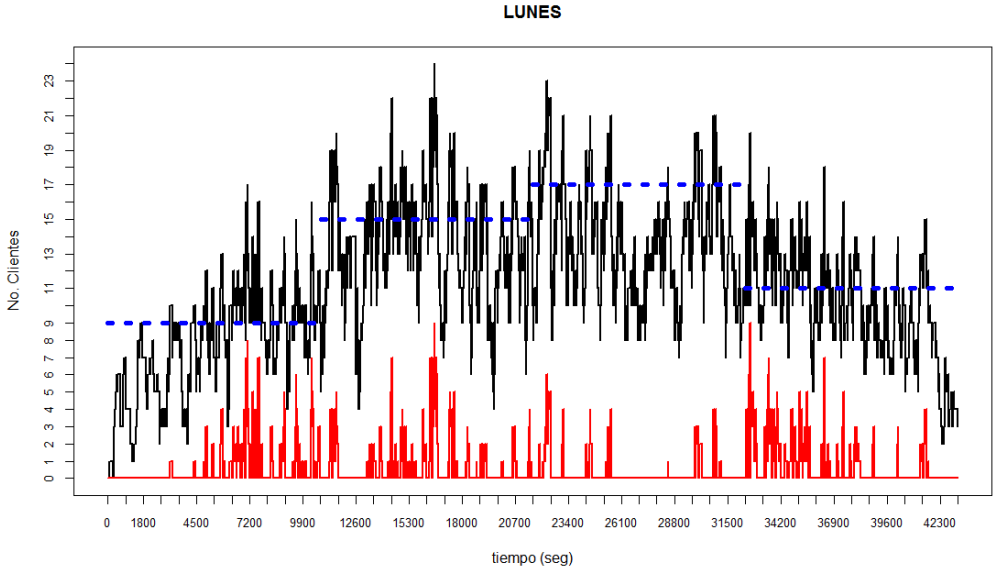

Figura 3.19: Funcionamiento del call center con *opers=c(9,6,11)*

La línea oscura de la Figura 3.19 representa la cantidad de clientes en el sistema en cada instante de tiempo, la línea clara la cantidad de clientes en la cola y las líneas horizontales son la cantidad de operadores que hay en los distintos momentos del día.

En este ejemplo, con 26 operadores, la media del nivel de atención (*nat*) de todas las iteraciones es aproximadamente de 87 % y el promedio del nivel de servicio (*ns*) está en el entorno del 70 %. Ambos niveles no cumplían con los objetivos de 95 % para el nivel de atención y de 80 % para el nivel de servicio. Esto significa que la cantidad de operadores no es suficiente para brindar el servicio deseado. La falta

de operadores también se reflejaba en el alto porcentaje de ocupación, superando la cota superior recomendada.

Con el fin de cumplir con los objetivos de atención y de servicio, se incrementó de manera arbitraria el número total de operadores una cantidad suficiente que garantice sobrepasar holgadamente los niveles. La nueva cantidad se estipuló en 42 operadores, distribuidos en los turnos también arbitrariamente de la siguiente manera: 18 en el primero, 8 en el segundo y 16 en el último turno. Luego se corrió nuevamente la función, dejando el resto de los argumentos iguales al caso anterior. Bajo estas condiciones se consiguieron los siguientes resultados:

```
> call.center(opers = c(18,8,16), h_entr = c(8,11,14), lim_esp = 45,
ts_obj = 20, dia_sem = "lunes", iter = 200, plot = TRUE)
media
recibidas      ocup       tmo     t_sis     t_esp       nat        ns
2582.3700    0.5951  211.0600  210.3750    1.3025    0.9906    0.9699
```

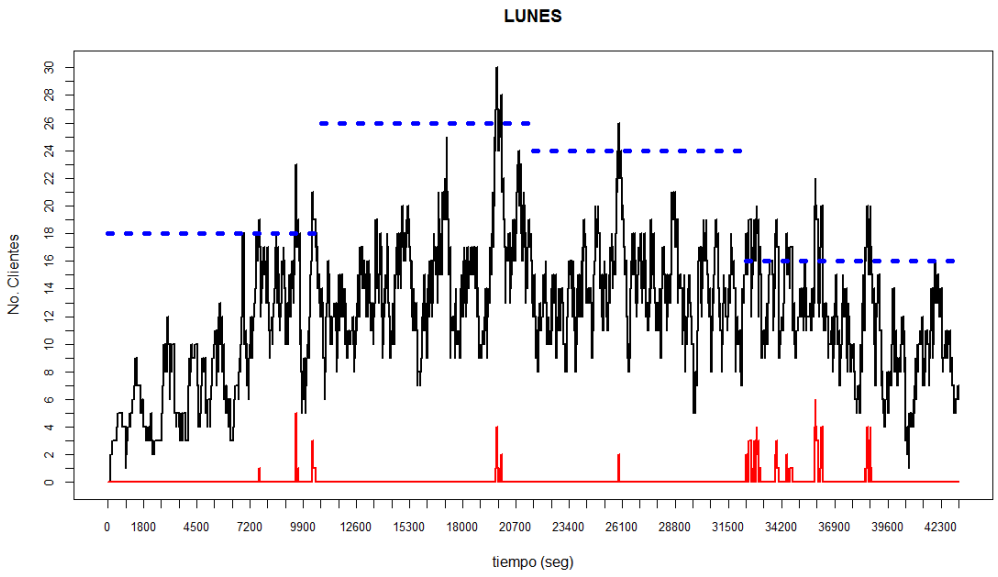

Figura 3.20: Funcionamiento del call center con *opers=c(18,8,16)*

Estos valores se corresponden con la Figura 3.20.

Con este incremento en la cantidad de operadores, el nivel de atención se ubicó en el entorno del 99 % y el nivel de servicio cercano al 97 % superando así los objetivos planteados en 4 y 17 puntos porcentuales respectivamente. Por otro lado el nivel de ocupación descendió a 59 %, en contraposición con el caso anterior este indicador quedó por debajo de los niveles recomendados. Con estos resultados obtenidos se evidencia un sobredimensionamiento en el número de operadores. Para buscar, de manera sistemática, una asignación con el menor número de operadores pero que cumpla con las metas establecidas para los indicadores se programó una nueva función.

3.6.2. Optimización del funcionamiento del Call Center

El objetivo de esta sección fue encontrar el número óptimo de operadores que cumplieran con que el nivel de abandono no fuera mayor al 5 % y que se atendiera al menos el 80 % de las llamadas antes de los 20 segundos. Ambos niveles son argumentos de la función. En la búsqueda del óptimo se evaluó la función anterior bajo distintos escenarios de operadores.

Los argumentos de esta función son los siguientes:

1. *min_serv*, cantidad mínima de operadores/servidores para cada uno de los tres turnos.

2. *max_serv*, cantidad máxima de operadores/servidores para cada uno de los tres turnos.

3. *h_entr*, horario de entrada de cada turno.

4. *lim_esp*, límite de espera hasta abandonar el sistema.

5. *ts_obj*, umbral de tiempo para el nivel de servicio.

6. *NS_obj*, objetivo del nivel de servicio.

7. *NAT_obj*, objetivo del nivel de atención.

8. *cump_obj*, porcentaje objetivo de iteraciones que cumplan los niveles, es decir, para que un escenario sea considerado como candidato a óptimo, debe cumplir los niveles planteados por lo menos con este porcentaje del total de las iteraciones.

9. *dia_sem*, dia de la semana que se quiere optimizar.

10. *iter*, cantidad de iteraciones realizadas.

Con los dos primeros argumentos de la función (cantidad mínima y máxima de operadores por turno) se generan todas las combinaciones posibles de operadores para los tres turnos. Para cada uno de estas combinaciones se simula el funcionamiento del call center, mediante la función de la sección anterior (*call.center*), tantas veces como el número de iteraciones definidas. Para cada simulación se registra el promedio del nivel de atención y de servicio alcanzados. Al finalizar las iteraciones se registra que porcentaje de estas cumplieron con ambos niveles y se compara con el objetivo planteado (*cump_obj*), si es mayor o igual se considera este escenario como candidato al óptimo, sino es descartado.

Entre los candidatos, se considera como óptimo a la combinación que requiera la mínima cantidad de operadores totales. De haber más de una combinación con este número de operadores, se selecciona la que genera el mayor nivel de servicio promedio.

Para conocer el número óptimo de operadores necesarios para alcanzar los niveles deseados, lo ideal sería evaluar todas las combinaciones de operadores posibles, pero a efectos prácticos se acotó el análisis en rangos razonables. Por este motivo se decidió restringir el rango de operadores a evaluar, quedando entre 9 y 18 operadores para el primer y tercer turno y entre 2 y 11 para el segundo, generando así 1000 combinaciones distintas. La cantidad de operadores del segundo turno fue menor a la de los otros dos debido a que se solapaba por tres horas con cada uno de los otros turnos.

Se decidió trabajar con un mínimo de 9 operadores debido a que, por el análisis previo de los arribos, en los primeros 180 minutos de los lunes se recibían en promedio 470 llamadas aproximadamente, con un tiempo medio de servicio en el entorno

de los 210 segundos. Asumiendo estos valores arribaban 2,61 llamadas por minuto y multiplicando por la duración promedio de las llamadas se obtiene un equivalente de 9,14 operadores, lo cual generaba un indicio del mínimo de operadores a utilizar.

Los resultados obtenidos indicaron que un 37,9 % de las combinaciones de operadores simuladas cumplían con los objetivos planteados, por lo que fueron consideradas candidatas al óptimo. Del total de combinaciones, hubo 7 combinaciones que requirieron 33 operadores, la menor cantidad dentro de todos los candidatos al óptimo (listadas en el Apéndice D). De estas 7 combinaciones, la que generó el máximo nivel de servicio fue la combinación que asignaba 14 operadores en el primer turno, 5 en el segundo y 14 en el tercero. Esta combinación fue considerada la *óptima*.

Bajo el número óptimo de operadores se esperaría tener en promedio un nivel de atención superior al 96 % y un nivel de servicio mayor al 88 %. La otra medida importante es el nivel de ocupación, el cual se ubicó en el entorno del 73 %, encontrandose dentro de los márgenes sugeridos (60-80 %). También se puede apreciar que en promedio los clientes esperarían alrededor de 5 segundos en la cola.

En la Figura 3.21 se muestra un ejemplo de como sería el comportamiento del funcionamiento del call center bajo esta combinación de operadores. En ella se observa como las colas son esporádicas y distribuidas de manera relativamente uniforme a lo largo del día, teniendo un máximo de 7 clientes en espera simultáneamente.

Con la asignación hallada para los días lunes, se evaluó con este mismo número de operadores, debido a que ellos trabajan toda la semana, cuales serían los niveles esperados para los restantes días de la semana. En el Cuadro 3.7 se presentan las principales medidas de performance y en la Figura 3.22 un ejemplo del funcionamiento del call center para cada uno de estos días.

En dicho cuadro se observa como a medida que avanza la semana los niveles de atención y de servicio van aumentando, lo cual era esperable debido al descenso del volumen de las llamadas anteriormente mencionado. También se aprecia como se redujo el nivel de ocupación, llegando a un promedio de 61,5 % los días viernes. Este valor es bajo, aunque está en el límite de los valores recomendados y puede

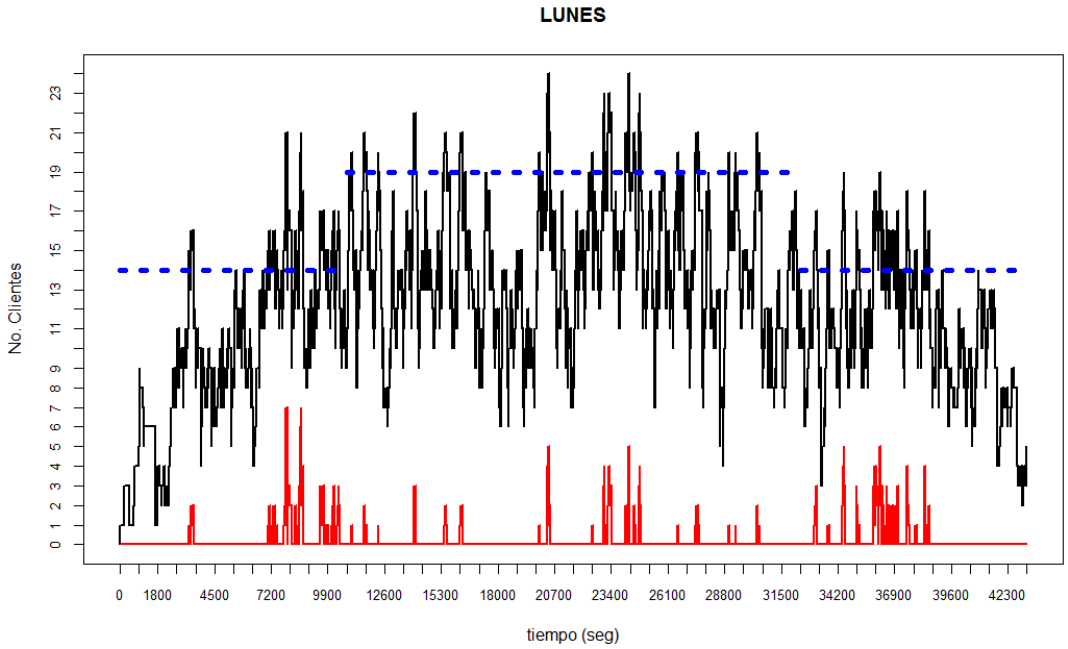

Figura 3.21: Funcionamiento óptimo del call center, *opers=c(14,5,14)*

Día	**Nivel de Atención**	**Nivel de Servicio**	**Ocupación**
Lunes	96,27 %	88,58 %	73,50 %
Martes	97,33 %	92,05 %	68,97 %
Miércoles	97,79 %	93,51 %	66,30 %
Jueves	98,43 %	95,31 %	63,56 %
Viernes	98,93 %	96,58 %	61,52 %

Cuadro 3.7: Niveles para todos los días de la semana

considerarse como un indicio de que habría un sobredimensionamiento de personal.

Una alternativa a esto es asignar el número óptimo de operadores de otro día de la semana, lo cual llevaría a disminuir el número de operadores, incumpliendo los niveles los días lunes, pero reduciendo los costos asociados a los recursos humanos. En definitiva hay una evaluación costo beneficio que dependerá de las penalidad o multa por no alcanzar estos niveles.

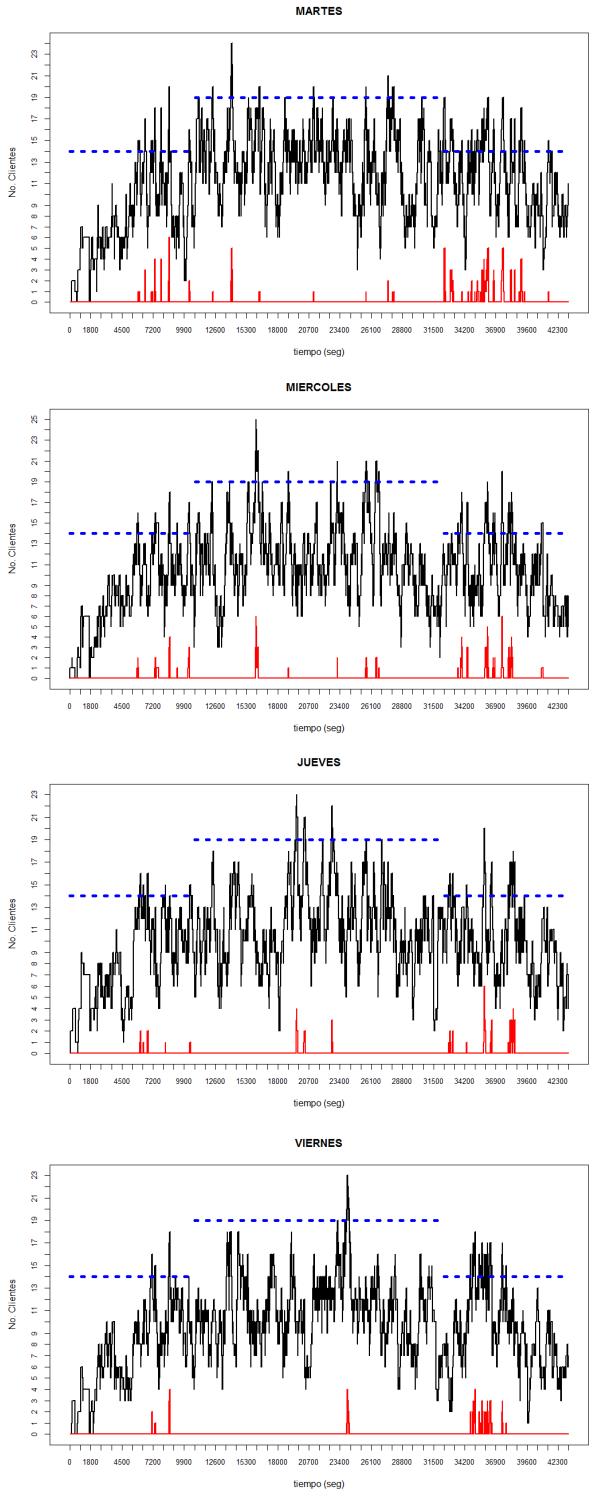

Figura 3.22: Ejemplo de funcionamiento del call center de martes a viernes

Capítulo 4

Conclusiones

En primer lugar, se debe tener en cuenta que se utilizó la información de un call center con una dimensión relativamente grande el cual brindaba su servicio de call center a diversas empresas.

En particular, se estudió en profundidad al información referida a una de estas empresas, para comprobar la veracidad de algunos supuestos sobre los cuales se basan los cálculos de los call centers para fijar sus objetivos y estrategias. En la industria de la telefonía lo más utilizado son los resultados elementales de la teoría de colas, específicamente el modelo $M/M/c$, el cual facilita el cálculo de las distintas mediadas del rendimiento del sistema.

Dicho modelo asume en primera instancia que los arribos de llamadas, uno de los principales procesos a considerar, se comportan acorde a un proceso de Poisson homogéneo, en tramos relativamente amplios de tiempo, por ejemplo, media hora. Se realizaron pruebas para comprobar esta suposición, llegando a la conclusión que el comportamiento de los mismos se ajusta de manera más natural a un proceso Poisson no homogéneo, con una función de intensidad relativamente "bien portada".

En segundo lugar, se analizó el tiempo de servicio de los clientes, otro proceso clave en el funcionamiento de los call center, que sin analizar su validez se supone tiene una distribución exponencial. El análisis realizado mostró que la distribución exponencial no sería apropiada para el tiempo de servicio, ya que presentaba una

distribución bimodal. A partir de ello, se procedió a modelizar este proceso de forma paramétrica suponiendo una mezcla de distribuciones normales.

En relación a los datos obtenidos sobre el tiempo de espera, es decir la paciencia de los clientes, no se analizaron en profundidad. Para casos como este, en que los clientes que abandonan tienen muy poca paciencia, habría que analizar dicho comportamiento, para luego trabajar con técnicas de datos censurados. Para la función de simulación del funcionamiento del call center, de modo conservador, se consideró que si un cliente alcanzaba los 45 segundos en la cola, abandonaba el sistema.

Con estos desarrollos previos se construyó una función con el fin de simular el funcionamiento del call center. Esta función, mediante la fijación de ciertos valores para las variables exógenas, tales como número de operadores, horas de entrada, límite de espera, umbrales de tiempo para el nivel de servicio y día de la semana, generaba los niveles de las principales medidas de performance del sistema.

Estos avances permitieron pasar a la etapa final que se centró en el desarrollo de la función que optimiza el número total de operadores. Esta función generaba distintas combinaciones de operadores para los distintos turnos. En base a esto y mediante la función anterior, modelizaba el comportamiento del call center sujeto a distintas restricciones que aseguraban brindar los niveles de servicio requeridos.

De esta manera se llegó al objetivo principal del estudio, optimizar el número de operadores del call center, debido a la importancia económica que esto tiene en el costo operativo del mismo. Siendo este dimensionamiento de 33 operadores, distribuidos en 14 operadores en el primer turno, 5 en el segundo y 14 en el tercero. Este número fue buscado para los días lunes por ser el día de mayor tráfico de llamadas y así garantizar los niveles de toda la semana.

En este caso, como los operadores trabajaban de lunes a viernes, de mantenerse el número óptimo de operadores de los días lunes, implicaría tener recursos ociosos el resto de la semana. En su defecto podría optimizarse otro día de la semana, así reducir el número total de operadores, lo cual implicaría no alcanzar los niveles los días lunes pero reducir los costos de personal.

Una oportunidad de mejora para este tipo de servicios sería poder contar con flexibilidad en la jornada laboral de los operadores, es decir, no tener la restricción de que trabajen de lunes a viernes. De este modo se podría optimizar el dimensionamiento para cada uno de los días de la semana.

Bibliografía

U. Narayan Bhat. *An Introduction to Queueing Theory: Modeling and Analysis in Applications*. Birkhauser, 2008.

S.P. Brooks y B.J.T. Morgan. Optimization using simulated annealing. *Journal of the Royal Statistical Society. Series D (The Statistician)*, Vol.44, No.2:241–257, 1995.

Lawrence Brown, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, y Linda Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100:36–50, 2005.

COPC. *COPC-2000: Customer Service Provider Standard*, 2009.

Luc Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag New York Inc., 1986.

Noah Gans, Ger Koole, y Avishai Mandelbaum. Telephone call center: Tutorial, review, and research prospect. *Manufacturing and Service Operations Managment*, 5:79–141, 2003.

Olle Haggstrom. *Finite Markov Chains and Algorithmic Applications*. Cambridge University Press, 2002.

World Bank Group International Finance Corporation. *Mobile Money Toolkit*. 2010.

David G. Kendall. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *The Annals of Mathematical Statistics*, Vol.24, No.3:338–354, 1953.

Alexey Novikov, Ruslan Pusev, y Maxim Yakovlev. *exptest: Tests for Exponentiality*, 2013. URL http://CRAN.R-project.org/package=exptest. R package version 1.2.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL http://www.R-project.org/.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

Larry Wasserman. *All of Nonparametric Statistics*. Springer Science + Business Media, 2006.

Apéndice A

Apéndice de resultados

1. Este cuadro muestras las fechas que hubieron eventos particulares que generaron un dato atípico:

Fecha	Volumen recibido	Evento
29/03/2013	962	Feriado de Semana Santa
01/04/2013	2.601	Día posterior a Semana Santa
19/06/2013	1.417	Feriado
28/08/2013	3.570	Problemas en la operativa
20/09/2013	1.691	Feriado (vacaciones escolares)
31/10/2013	1.680	Ningún evento en particular
01/11/2013	3.177	Problemas en la operativa
04/11/2013	3.711	Problemas en la operativa
31/12/2013	244	Medio horario de atención

Cuadro A.1: Eventos generadores de atípicos

Los problemas en la operativa refieren a problemas en el servicio que brindaba la empresa que contrató al call center, los cuales generaron aumentos en las llamadas.

2. Salida del software para el test de hipótsis de Kolmogorov-Smirnov, para un día en particular y con 10.000 réplicas para la simulación de Monte Carlo:

```
> ks.exp.test(x=dia11, nrepl=10000)

Kolmogorov-Smirnov test for exponenciality
data: dia11
KSn = 0.0193,  p-value = 0.1962
```

3. Salida del software para el test de hipótsis de Kolmogorov-Smirnov, para el tiempo de servicio, con 2.000 réplicas para la simulación de Monte Carlo:

```
> ks.exp.test(x=serv_oct, nrepl=2000)

Kolmogorov-Smirnov test for exponenciality
data: serv_oct
KSn = 0.0966,  p-value < 2.2e-16
```

Apéndice B

Ecuaciones de Kolmogorov

Sea $\{X(t), t \in \mathtt{T}\}$ un proceso de Markov homogéneo,

$$P_{ij}(t) = P[X(t) = j \mid X(0) = i].$$

Existen dos tipos de ecuaciones diferenciales para determinar $P_{ij}(t)$ en un proceso de Markov. Estas son *forward kolmogorov equations* y *backward kolmogorov equations*. En estos procesos la relación de Chapman-Kolmogorov, es:

$$P_{ij}(t+s) = \textstyle\sum_{k \in S} P_{ik}(t) P_{kj}(s)$$

Sea $s = \Delta t$, entonces

$$P_{ij}(t+\Delta t) = \textstyle\sum_{k \in S} P_{ik}(t) P_{kj}(\Delta t)$$

Restando $P_{ij}(t)$ en ambos lados de la ecuación y dividiendo por Δt

$$\frac{P_{ij}(t+\Delta t) - P_{ij}(t)}{\Delta t} = \sum_{k \neq j} \frac{P_{ik}(t) P_{kj}(\Delta t)}{\Delta t} + P_{ij}(t) \frac{P_{jj}(\Delta t) - 1}{\Delta t}$$

Con $\Delta t \to 0$, se tiene

$$P'_{ij}(t) = -\lambda_{jj} P_{ij}(t) + \sum_{k \neq j} \lambda_{kj} P_{ik}(t) \tag{B.1}$$

Esta es la *forward kolmogorov equations* para $i, j \in S$. Matricialmente queda de la siguiente manera:

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{A}$$

Las probabilidades de transición $P_{ij}(t)$ se pueden determinar resolviendo estas ecuaciones diferenciales, junto con la condición inicial $\mathbf{P}(0) = \boldsymbol{I}$.

De manera similar se puede obtener las *backward kolmogorov equations*, comenzando por:

$$P_{ij}(\Delta t + t) = \sum_{k \in S} P_{ik}(\Delta t) P_{kj}(t) \tag{B.2}$$

En forma matricial,

$$\mathbf{P}'(t) = \mathbf{A}\mathbf{P}(t).$$

Apéndice C

Proceso de Poisson No Homogéneo: demostración

C.1. Demostración Teorema 1

Se define $p_n(t) = P(X(t) = n)$ y se considera cualquier $h > 0$. Se denota por $p_n(t, t+h]$ a la probabilidad $P(X(t+h) - X(t) = n)$. Por la hipótesis de independencia, para $t \geq 0$ y cuando $h \searrow 0$,

$$\begin{aligned} p_0(t+h) &= p_0(t)p_0(t, t+h] \\ &= p_0(t)(1 - \lambda(t)h + o(h)). \end{aligned}$$

Calculando las derivadas se obtiene la ecuación diferencial $p_0'(t) = -\lambda(t)p_0(t)$, cuya solución es $p_0(t) = ce^{-\Lambda(t)}$, donde la constante c es uno debido a la condición inicial $p_0(0) = 1$. Ahora se encuentra $p_n(t)$ para $n \geq 1$.

Nuevamente por independencia,

$$\begin{aligned} p_n(t+h) &= p_n(t)p_0(t, t+h] + p_{n-1}(t)p_1(t, t+h] + o(h) \\ &= p_n(t)(1 - \lambda(t)h + o(h)) + p_{n-1}(t)(\lambda(t)h + o(h)) + o(h). \end{aligned}$$

Entonces $p_n'(t) = -\lambda(t)p_n(t) + \lambda(t)p_{n-1}(t)$, con condición inicial $p_n(t) = 0$ para $n \geq 1$. Si se define $q_n(t) = e^{\Lambda(t)}p_n(t)$ la ecuación diferencial se transforma en $q_n'(t) =$

$\lambda(t)q_{n-1}(t)$, con condiciones iniciales $q_n(0) = 0$ y $q_0(t) = 1$. Esta ecuación se resuelve iterativamente primero para $q_1(t)$, después $q_2(t)$, y así sucesivamente, en general

$$q_n(t) = \frac{(\Lambda(t))^n}{n!}$$

y de aquí se obtiene $p_n(t)$.

C.2. Demostración Teorema 2

Se define $p_n(t) = P(X(t) = n)$ y se considera cualquier $h > 0$. Se denota por $p_n(t, t+h]$ a la probabilidad $P(X(t+h) - X(t) = n)$. Por la hipótesis de independencia, para $t \geq 0$ y cuando $h \searrow 0$,

$$\begin{aligned} p_0(t+h) &= p_0(t)p_0(t, t+h] \\ &= p_0(t)(1 - \lambda(t)h + o(h)). \end{aligned}$$

Se escribe $X(t+s) = X(s) + (X(t+s) - X(s))$, en donde por el axioma de incrementos independientes, en el lado derecho aparece la suma de dos variables aleatorias independientes. Recordando que la función generatriz de momentos de la distribución de $Poisson(\lambda)$ es $M(r) = exp[\lambda(e^r - 1)]$, y al aplicar este resultado a la ecuación anterior se obtiene,

$$exp[\Lambda(t+s)(e^r - 1)] = exp[\Lambda(s)(e^r - 1)]M_{X(t+s)-X(s)}(r).$$

Por lo tanto $M_{X(t+s)-X(s)}(r) = exp[(\Lambda(t+s) - \Lambda(s))(e^r - 1)]$.

Apéndice D

Resultado de la Función optimización

A continuación se presenta el comando utilizado para ejecutar la función, los candidatos al óptimo que necesitan la menor cantidad de operadores total y el escenario óptimo para este call center.

```
prueba_op <- call.optim(min_serv=c(9,2,9), max_serv=c(18,11,18), horas=c(8,11,14), lim_esp=45, ts_obj=20,
                        NS_obj=0.8, NAT_obj=0.95, cump_obj=0.8, dia_sem="lunes", iteraciones=100)

     Turno_1 Turno_2 Turno_3 Tot_opers NAT_Esperado NS_Esperado Ocupacion Tiempo_espera Pasan
446       13       6      14        33       95.71%      87.67%    73.42%          5.28  0.80
518       14       3      16        33       96.07%      87.82%    73.24%          5.26  0.86
527       14       4      15        33          96%       87.7%    73.47%          5.28  0.86
536       14       5      14        33       96.27%      88.58%     73.5%          4.93  0.91
608       15       2      16        33       96.05%       87.5%    73.31%          5.40  0.88
617       15       3      15        33          96%      87.37%    73.77%          5.42  0.85
626       15       4      14        33       95.88%      87.63%     73.4%          5.32  0.83

[[2]]
        Turno_1 Turno_2 Turno_3 Tot_opers NAT_Esperado NS_Esperado Ocupacion Tiempo_espera Pasan
Optimo:      14       5      14        33       96.27%      88.58%     73.5%          4.93  0.91
```

Apéndice E

Código en R

E.1. Lectura de los datos

```
##########################################
############ LECTURA DE DATOS ############
##########################################

datos2013 <- read.csv2("Datos2013.csv")

##se renombran las variables
colnames(datos2013) <- c("Campaa","Type","Agent","ANI","StartTime","InitTime","EndTime",
"DurationTime","Ringing","AttentionTime","DispositionTime","WrapupTime","HoldTime","svl",
"isghost","outSched","isTransf")
head(datos2013)
names(datos2013)

##se cambia el formato, para poder trabajar con fechas
datos2013$StartTime <- ymd_hms(datos2013$StartTime)
datos2013$EndTime <- ymd_hms(datos2013$EndTime)
datos2013$InitTime <- ymd_hms(datos2013$InitTime)
datos2013$DispositionTime <- hms(datos2013$DispositionTime)
datos2013$DurationTime <- hms(datos2013$DurationTime)
datos2013$WrapupTime <- hms(datos2013$WrapupTime)
datos2013$HoldTime <- hms(datos2013$HoldTime)
datos2013$AttentionTime <- hms(datos2013$AttentionTime)

##columna con el da del ao
Dia <- day(datos2013$StartTime)
Mes <- month(datos2013$StartTime)
datos2013 <- cbind(datos2013, Dia, Mes)

##se unifican los niveles
levels(datos2013$isghost) <- c("No","Yes","Yes","Yes")
```

```
levels(datos2013$svl) <- c("No","Yes","Yes","Yes")
levels(datos2013$outSched) <- c("No","Yes","Yes")
levels(datos2013$isTransf) <- c("No","Yes","Yes")

##se eliminan las llamadas ghost y las fuera de horario y se renombra como "datos"
datos <- subset(datos2013, datos2013$isghost == "No")
datos <- subset(datos, datos$outSched == "No")
datos <- subset(datos, hour(datos$StartTime) >= 8)

################################################
############ FIN DE LECTURA DE DATOS ############
################################################
```

E.2. Estimación de intensidades

```
#######################################################
############ ESTIMACION DE INTENSIDADES ############
#######################################################

##se trabaja con los meses de setiembre a noviembre
lunes <- subset(datos$StartTime, wday(datos$StartTime)==2 & month(datos$StartTime) %in% c(9,10,11))
martes <- subset(datos$StartTime, wday(datos$StartTime)==3 & month(datos$StartTime) %in% c(9,10,11))
miercoles <- subset(datos$StartTime, wday(datos$StartTime)==4 & month(datos$StartTime) %in% c(9,10,11))
jueves <- subset(datos$StartTime, wday(datos$StartTime)==5 & month(datos$StartTime) %in% c(9,10,11))
viernes <- subset(datos$StartTime, wday(datos$StartTime)==6 & month(datos$StartTime) %in% c(9,10,11))


##se estandarizan los datos, llevandolos a la escala a 0-1
trnf<-function(x){   (x-0.3333)/ (0.8333-0.3333)}
intensidad<-function(v){
h_srot <- 0.9 * min(sd(v), IQR(v)/1.34) * length(v)^(-1/5)
a<-density(v, from = 0, to = 1, bw= h_srot , n=43200)
length(v)*(a$y)
}

############ INTENSIDAD LUNES ############
dia_ano <- as.numeric(names(table(yday(lunes))))
tpp<- list()

for (i in 1:length(dia_ano)){
    tpp[[i]] <- hour(lunes[yday(lunes) == dia_ano[i]])/24
+ minute(lunes[yday(lunes) == dia_ano[i]])/24/60
+ second(lunes[yday(lunes) == dia_ano[i]])/24/60/60
}
a <- density (tpp[[1]], from = 0, to = 1, n=43200)
xx <- a$x

tpp_lun1<-lapply(tpp,trnf)
MM<-lapply(tpp_lun1,intensidad)
```

```
tpp_lun1<-tpp_lun1[-10]  ##se quita la observacion 10 por ser atipica
MM<-lapply(tpp_lun1,intensidad)
xxx<-8+12*xx
inten<-do.call(rbind,MM)
int_lunes<-apply(inten,2,mean)


############ INTENSIDAD MARTES ############
dia_ano <- as.numeric(names(table(yday(martes))))
tpp<- list()

for (i in 1:length(dia_ano)){
    tpp[[i]] <- hour(martes[yday(martes) == dia_ano[i]])/24
+ minute(martes[yday(martes) == dia_ano[i]])/24/60
+ second(martes[yday(martes) == dia_ano[i]])/24/60/60
}
tpp_mar1<-lapply(tpp,trnf)
MM<-lapply(tpp_mar1,intensidad)
inten<-do.call(rbind,MM)
int_martes<-apply(inten,2,mean)

############ INTENSIDAD MIERCOLES ############
dia_ano <- as.numeric(names(table(yday(miercoles))))
tpp<- list()

for (i in 1:length(dia_ano)){
    tpp[[i]] <- hour(miercoles[yday(miercoles) == dia_ano[i]])/24
+ minute(miercoles[yday(miercoles) == dia_ano[i]])/24/60
+ second(miercoles[yday(miercoles) == dia_ano[i]])/24/60/60
}
tpp_mie1<-lapply(tpp,trnf)
MM<-lapply(tpp_mie1,intensidad)
inten<-do.call(rbind,MM)
int_miercoles<-apply(inten,2,mean)

############ INTENSIDAD JUEVES ############
dia_ano <- as.numeric(names(table(yday(jueves))))
tpp<- list()

for (i in 1:length(dia_ano)){
    tpp[[i]] <- hour(jueves[yday(jueves) == dia_ano[i]])/24
+ minute(jueves[yday(jueves) == dia_ano[i]])/24/60
+ second(jueves[yday(jueves) == dia_ano[i]])/24/60/60
}
tpp_jue1<-lapply(tpp,trnf)
MM<-lapply(tpp_jue1,intensidad)
tpp_jue1 <- tpp_jue1[-9]  ##se quita la observacion 9 por ser atipica
MM<-lapply(tpp_jue1,intensidad)
inten<-do.call(rbind,MM)
int_jueves<-apply(inten,2,mean)
```

```
############ INTENSIDAD VIERNES ############
dia_ano <- as.numeric(names(table(yday(viernes))))
tpp<- list()

for (i in 1:length(dia_ano)){
    tpp[[i]] <- hour(viernes[yday(viernes) == dia_ano[i]])/24
+ minute(viernes[yday(viernes) == dia_ano[i]])/24/60
+ second(viernes[yday(viernes) == dia_ano[i]])/24/60/60
}
tpp_vie1<-lapply(tpp,trnf)
MM<-lapply(tpp_vie1,intensidad)
tpp_vie1 <- tpp_vie1[-c(3,9)]  ##se quitan las observaciones 3 y 9 por ser atipicas
MM<-lapply(tpp_vie1,intensidad)
inten<-do.call(rbind,MM)
int_viernes<-apply(inten,2,mean)

##########################################################
############ FIN ESTIMACION DE INTENSIDADES ############
##########################################################
```

E.3. Estimación del Tiempo de Servicio

```
##########################################################
############ ESTIMACION TIEMPO DE SERVICIO ############
##########################################################

##se leen las llamadas abandonadas y las atendidas
abandonadas <- subset(datos, is.na(datos$InitTime))
atendidas <- datos[-which(rownames(datos)%in%rownames(abandonadas)),]

##tiempo de servicio para las llamadas atendidas
servt <- atendidas$AttentionTime + atendidas$WrapupTime
servt <- hour(servt)*60*60 + minute(servt)*60 + second(servt)

#  se buscan y quitan los outliers extremos de cada mes (Q3+(Q3-Q1)*3)
out2<-0
servt_lim <-list()

for (i in 1:12){
    mes <- subset(servt ,month(atendidas$StartTime) == i)
    a <- length(mes)
    extremos <- quantile(mes, 0.75) + 3*(IQR(mes))
    ext <- extremos

    while (sum(mes > ext) > 0) {
        mes <- subset(mes, mes < ext)
```

```
        ext <- quantile(mes, 0.75) + 3*(IQR(mes))
    }

    servt_lim[[i]] <- mes ## mes limpio de outliers

    aux <- servt_lim[[i]]
    servt_lim[[i]] <- aux[aux>=6]

    out2[i] <- a - length(servt_lim[[i]]) # cantidad de outlieres por mes
}

## se selecciona el mes de Octubre y se aplica el logaritmo
servt_oct <- servt_lim[[10]]
lservt_oct <- log(servt_oct)
y <- lservt_oct

## funcion Log Verosimilitud, la cual se busca minimizar
logv <- function(param,x){
    m1 <- param[1]; m2<-param[2]
    s1 <- param[3]; s2<-param[4]
    tau <- param[5]
    L <- 0
    n <- length(x)
    for (i in 1:n) L <- L + log(tau*dnorm(x[i],m2,sqrt(s2)) + (1-tau)*dnorm(x[i],m1,sqrt(s1)))
    return(-L)
}

## funcion para estimar parametros
mezcla<-function(T=3, FUN, n=40, d=0.01, d1=list(mu=c(0,20), sigma=c(0,20)), r=c("FALSE","TRUE")){
#T=Temperatura inicial; FUN=Funcion a minimizar; n=numero de simulaciones hasta cambio de temperatura;
d=decaimiento de la temperatura; d1=rangos de los parametros mu y sigma;
r=indicador para mostrar pasos intermedios
    r <- match.arg(r)
    h <- match.fun(FUN)
    k <- 0 # contador cambio de temperatura
    iter <- 1 # numero total de iteraciones
    # punto de partida de los parametros
    mu1 <- mean(y)
mu2 <- mu1
    sig1 <- var(y)
sig2 <- sig1
    tau <- runif(1) #se selecciona aleatoriamente
  param.p <- param.a <- c(mu1, mu2, sig1, sig2, tau)

while(T >= 0){
        j <- sample(1:5, 1)
        if (j < 3) param.p[j] <- runif(1, d1$mu[1], d1$mu[2])
        if (j > 2 && j < 5) param.p[j] <- runif(1, d1$sigma[1], d1$sigma[2])
        if (j == 5) param.p[j] <- runif(1,0,1)
        if (runif(1) < exp((h(param.a,y) - h(param.p,y)) / T)) {
            param.a <- param.p}
```

```
        k <- k+1
        if (r == "TRUE") cat("k = ",k,"T = ",T,"param = ",param.a,'\n')
        if (k == n){
            T <- T-d #baja la temperatura
            k <- 0
        }
        iter<-iter+1}
    return(param.a)
}


############################################################
############ FIN ESTIMACION TIEMPO DE SERVICIO #############
############################################################
```

E.4. Funcionamiento del Call center

```
#############################################
############ FUNCION CALL CENTER ############
#############################################

call.center <- function(opers=c(8,8,8), h_entr=c(8,11,14), lim_esp=45, ts_obj=20,
                        dia_sem="lunes", iter=3, ploteo=F){

    horas <- 12*60*60 ## jornada laboral vista en segundos (con la escala de intensidad)

    t <- c(0:horas)  ## Duracion del periodo de tiempo

    if(length(opers)!=3) stop("Debe definir la cantidad de operadores para cada turno.")
    Opers <- floor(opers)  ## Numero de servidores (operadores)

    if(!iter%in%seq(10000)) stop("Debe redefinir la cantidad de iteraciones (1 a 10000).")
    num_iter <- iter

    recibidas <- NA
    util <- NA
    ocup <- NA
    tmo <- NA
    t_sis <- NA
    t_esp <- NA
    ns <- NA
    nat <- NA

    dia_sem <- toupper(dia_sem)

    if (!dia_sem%in%c('LUNES','MARTES','MIERCOLES','JUEVES','VIERNES')) {
```

```
    stop('Debe ingresar un dia de Lunes a Viernes.')
} else if (dia_sem == 'LUNES'){
    intens <- int_lunes
    tpp_final <- tpp_lun1
} else if (dia_sem == 'MARTES'){
    intens <- int_martes
    tpp_final <- tpp_mar1
} else if (dia_sem == 'MIERCOLES'){
    intens <- int_miercoles
    tpp_final <- tpp_mie1
} else if (dia_sem == 'JUEVES'){
    intens <- int_jueves
    tpp_final <- tpp_jue1
} else if (dia_sem == 'VIERNES'){
    intens <- int_viernes
    tpp_final <- tpp_vie1
}

for (k in 1:num_iter){
    arribos_tot <- rpois(1, mean(as.numeric(as.vector(lapply(tpp_final, length)))))
    arribos <- sort(sample(xx, size=arribos_tot, replace=TRUE, prob=intens))
    arribos <- arribos + runif(length(arribos), -0.0001, 0.0001)
    arribos <- (arribos - min(arribos)) / (max(arribos) - min(arribos))
    arribos <- sort(arribos*horas)

    matriz <- data.frame("entrada" = arribos)

    # Estimacion del tiempo de servicio Mezcla de distribuciones (mu1,mu2,sig1,sig2,tau)
    bimodal <- c(5.5039601, 3.0028497, 0.4221683, 0.3706169, 0.3304126)
    bino <- rbinom(nrow(matriz), 1, (1-bimodal[5]))
    duracion <- exp(bino*rnorm(nrow(matriz),bimodal[1], sqrt(bimodal[3]))
                    +(1-bino) * rnorm(nrow(matriz),bimodal[2],sqrt(bimodal[4])))

    matriz <- cbind(matriz, duracion)

    n_s <- 0   # numero de clientes en el sistema
    n_q <- 0   # numero de clientes en la cola
    corta <- 0 # numero de clientes que cortan

    atendido <- matriz[1,1] # tiempo de llegada (y atencion) del primer arribo
    salida <- atendido + matriz[1,2] # tiempo que va a salir el primer arribo

    for (i in 2:nrow(matriz)) {
        n_s[i] <- sum(salida[1:i-1] > matriz[i,1])
        if (matriz[i,1] <= (h_entr[2]-8)*60*60){
            c_aux <- Opers[1]
        } else if (matriz[i,1] <= (h_entr[3]-8)*60*60){
            c_aux <- Opers[1] + Opers[2]
        } else if(matriz[i,1] <= (h_entr[2]+6-8)*60*60){
            c_aux <- Opers[2] + Opers[3]
        } else {
```

```
            c_aux <- Opers[3]
        }
        if (n_s[i] < c_aux) { # miro si al momento de entrar la llamada, hay servidores libres
            atendido[i] <- matriz[i,1]
            n_q[i] <- 0
            corta[i] <- 0
            salida[i] <- atendido[i] + matriz[i,2]
        } else if (min(salida[which(salida[1:(i-1)]>atendido[i-1])]) - matriz[i,1] < lim_esp){
            # Hay cola y menor a lo que espera
            atendido[i] <- min(salida[which(salida[1:(i-1)]>atendido[i-1])])
            # va a ser atendido cuando corte el primero de los que estan en el sistema
            n_q[i] <- n_s[i] - c_aux
            corta[i] <- 0
            salida[i] <- atendido[i] + matriz[i,2]
        } else {
            atendido[i] <- matriz[i,1] + lim_esp
            corta[i] <- 1
            salida[i] <- atendido[i]
            n_q[i] <- n_s[i] - c_aux
        }
    }


    if (ploteo == T){
        plot(n_s ~ matriz[,1], typ="s",lwd=2,xlab="tiempo (seg)",
             ylab="No. Clientes", main=paste0(dia_sem),axes=F)
        axis(1,seq(0,max(t),60*15),cex.axis=0.8)
        axis(2,seq(0,max(n_s),1),cex.axis=0.8)
        box()
        lines(n_q ~ matriz[,1], typ="s", lwd=2, col=2)
        lines(rep(Opers[1],sum(matriz[,1]<(h_entr[2]-8)*60*60)) ~
                  matriz[which(matriz[,1]<(h_entr[2]-8)*60*60),1], col=4, lwd=5, lty=15)
        lines(rep(Opers[1]+Opers[2],sum((matriz[,1]>=(h_entr[2]-8)*60*60) &
             (matriz[,1]<(h_entr[1]-8+6)*60*60))) ~ matriz[which((matriz[,1]>=(h_entr[2]-8)*60*60)&
             (matriz[,1]<(h_entr[1]+6-8)*60*60)),1], col=4, lwd=5, lty=15)
        lines(rep(Opers[2]+Opers[3],sum((matriz[,1]>=(h_entr[3]-8)*60*60) &
             (matriz[,1]<(h_entr[2]-8+6)*60*60))) ~ matriz[which((matriz[,1]>=(h_entr[3]-8)*60*60)&
             (matriz[,1]<(h_entr[2]+6-8)*60*60)),1], col=4, lwd=5, lty=15)
        lines(rep(Opers[3],sum(matriz[,1]>=(h_entr[2]+6-8)*60*60)) ~
                  matriz[which(matriz[,1]>=(h_entr[2]+6-8)*60*60),1], col=4, lwd=5, lty=15)
    }

    recibidas[k] <- dim(matriz)[1]
    ocup[k] <- sum(matriz$duracion[which(corta==0)]) / (sum(Opers)*6*60*60)
    tmo[k] <- mean(matriz$duracion[which(corta==0)])
    t_sis[k] <- mean(salida - matriz$entrada)
    t_esp[k] <- mean(atendido - matriz$entrada)
    nat[k] <- sum(corta==0)/recibidas[k]
    ns[k] <- sum((atendido[which(corta==0)] - matriz$entrada[which(corta==0)]) < ts_obj)/recibidas[k]
}
metricas <- cbind(recibidas, "ocup"=round(ocup,4), "tmo"=round(tmo,0), "t_sis"=round(t_sis,0),
```

```
                    "t_esp"=round(t_esp,1), "nat"=round(nat,4), "ns"=round(ns,4))
    resul <- return(list(metricas=metricas,media=apply(metricas,2,mean)))
}

################################################
############ FIN FUNCION CALL CENTER ############
################################################
```

E.5. Optimización de los recursos

```
################################################
############ FUNCION OPTIMIZACION ############
################################################

call.optim <- function(min_serv=c(12,6,10), max_serv=c(14,8,12), horas=c(8,11,14), lim_esp=45, ts_obj=20,
                       NS_obj=0.8, NAT_obj=0.95, cump_obj=0.8, dia_sem="lunes", iteraciones=1){
    vector_ns <- c(NA,NA,NA,NA,NA,NA,NA,NA)
    noper_dif_1 <- c(min_serv[1] : max_serv[1])
    noper_dif_2 <- c(min_serv[2] : max_serv[2])
    noper_dif_3 <- c(min_serv[3] : max_serv[3])
    noper_largo <- length(noper_dif_1)
    num_oper <- cbind(rep(noper_dif_1,rep(noper_largo^2,noper_largo)),
                      rep(rep(noper_dif_2,rep(noper_largo,noper_largo)),noper_largo),
                      rep(noper_dif_3,noper_largo^2))

    for(i in 1:nrow(num_oper)){
        funci <- call.center(num_oper[i,],h_entr=horas, lim_esp, ts_obj, dia_sem, iteraciones)
        pasan <- sum(funci$metricas[,6]>=NAT_obj & funci$metricas[,7]>=NS_obj)
        vector_ns <- rbind(vector_ns,
                           c(num_oper[i,1],
                             num_oper[i,2],
                             num_oper[i,3],
                             round(funci$media[6]*100,2),
                             round(funci$media[7]*100,2),
                             round(funci$media[2]*100,2),
                             round(funci$media[5],2),
                             pasan/iteraciones)
                           )
    }
    resul <- data.frame("Turno_1"=vector_ns[-1,1],
                        "Turno_2"=vector_ns[-1,2],
                        "Turno_3"=vector_ns[-1,3],
                        "Tot_opers"=vector_ns[-1,1]+vector_ns[-1,2]+vector_ns[-1,3],
                        "NAT_Esperado"=paste0(vector_ns[2:(nrow(num_oper)+1),4],"%"),
                        "NS_Esperado"=paste0(vector_ns[2:(nrow(num_oper)+1),5],"%"),
                        "Ocupacion"=paste0(vector_ns[2:(nrow(num_oper)+1),6],"%"),
                        "Tiempo_espera"=(vector_ns[2:(nrow(num_oper)+1),7]),
```

```
                    "Pasan"=vector_ns[2:(nrow(num_oper)+1),8])
    tot_opers <- resul[,4]
    pasan_NS_obj <- which(vector_ns[-1,8] >= cump_obj)
    options(warn=-1)
    opt_opers <- pasan_NS_obj[which(tot_opers[pasan_NS_obj]==min(tot_opers[pasan_NS_obj]))]
    opt_opers2 <- which.max(vector_ns[opt_opers+1,5])
    optimo <- resul[opt_opers[opt_opers2],]
    resultados <- list()
    resultados[[1]] <- resul
    if (nrow(optimo)>0){
        resultados[[2]] <- optimo
        row.names(resultados[[2]]) <- "Optimo:"
    }  else {
        resultados[[2]] <- "No se alcanza el numero optimo de operadores para los nieveles requeridos"
    }
    return(resultados)
}
##################################################
############ FIN FUNCION OPTIMIZACION ############
##################################################
```