PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

ESTUDIO DEL MODELO DE COLAS PARA UN MEJORAMIENTO DE LA EFICIENCIA EN UN CALL CENTER

Tesis para optar el grado académico de Magíster en Ingeniería Industrial con mención en Logística que presenta

DÍAZ RUÍZ, OSCAR RAUL

Dirigido por

DR. FERNÁNDEZ PÉREZ, MIGUEL ANGEL

San Miguel, 2020

RESUMEN EJECUTIVO

Este trabajo tiene como finalidad analizar y comparar modelos utilizados en el mercado para el cálculo de la capacidad de atención en un *Call center*, tomando como punto de partida modelos que se adhieran a la demanda pronosticada y mantengan el nivel de servicio meta. Siendo el modelo que principalmente utiliza el sector de *Call center*, el modelo Erlang C, que considera los ratios de arribo, servicio y abandono como atributos fijos para el cálculo de la cantidad de asesores, cuando en la práctica estos atributos tiene un grado de variabilidad. Tomando como base el citado modelo, se evaluarán los resultados obtenidos con otros modelos para la determinación de la capacidad de atención en un *Call center*.

La presente tesis tiene como objetivo calcular el número óptimo de asesores para la atención de llamadas en un *Call center*, de modo que los costos operativos se minimicen y se garantice un nivel de servicio determinado por la cantidad de llamadas atendidas con respecto al total de llamadas. El presente trabajo nace a raíz de la necesidad de buscar una metodología alternativa que permita facilitar el cálculo de la capacidad instalada en un *Call center*. La evaluación de diversos escenarios que contemplen la variabilidad de la demanda es la principal dificultad para el cálculo de la capacidad de atención por medio de asesores telefónicos. Por ello es necesario buscar una metodología que permita una real reducción en la cantidad de asesores considerando el cumplimiento de los lineamientos de calidad como el nivel de servicio meta establecido en el mercado en 80%[1], ya que este indicador es uno de los más importantes para la gestión del *Call center*.

[1] 80% de las llamadas son atendidas dentro de los 20 segundos de ingresadas al call center.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ANEXOS

INTRODUCCIÓN

Desde el invento del teléfono y su posterior proliferación, el público comenzó a depender del servicio de proveedores de telecomunicaciones. Debido a la creciente base suscriptores, las compañías telefónicas enfrentan diversos problemas de planificación de recursos para brindar un buen servicio a los usuarios (Fluss, 2005). Entre estos problemas se pueden mencionar:

- Relación entre cantidad de asesores disponibles y los niveles de servicio, al existir mayor número de asesores en un *Call center* es posible atender la demanda de llamadas, reduciendo el número de clientes en espera y por ende se reduce el abandono, pero es posible que los costos de estas atenciones se eleven debido a que mantener el staff de asesores durante los tiempos muertos en los cuales la demanda está por debajo de la capacidad de atención existente (Saltzman & Mehrotra, 2001).
- Relación entre modelo y resultados, a pesar de que el modelo *Erlang* C sea el más difundido, este asume condiciones que llegan a simplificar demasiado el modelo, asumiendo por ejemplo que las llamadas llegan según un proceso de Poisson con una tasa promedio conocida, y que son atendidas por un número definido de agentes estadísticamente idénticos, con tiempos de servicio que siguen una distribución exponencial; estas condiciones distorsionan los resultados para el cálculo de la capacidad de asesores requerida en un *Call center*, por lo cual es necesario considerar aquellas variables que si tengan un impacto significativo en los modelos matemáticos elaborados para resolver este problema (Robbins, Medeiros, & Harrison, 2010).

Una gran pregunta era cuántos teléfonos y operadores eran necesarios para ejecutar la operación y qué nivel de servicio serían inaceptables para las personas que llaman. Para las compañías telefónicas, contar con una elevada cantidad de recursos llega a ser ineficiente y

generan altos costos para los suscriptores. Uno de los primeros en abordar la problemática de la gestión de recursos en *Call centers* fue A. K. Erlang, un ingeniero de la *Copenhagen Telephone Company,* quien en 1917 desarrolló la fórmula de colas Erlang C para determinar el requerimiento de personal. Actualmente, esta fórmula es ampliamente utilizada por *Call Centres* (Sharp, 2003).

Según Brown et al. (2005), una de las grandes prioridades de los *Call centers* es la maximización de la eficiencia del personal, ya que representa un gasto muy significativo en la estructura general de costos. Sin embargo, es un error cuando la productividad se convierte en el único objetivo. Si bien es importante optimizar el desempeño de los agentes del centro de contacto, es fundamental encontrar un equilibrio entre productividad, rendimiento, calidad y satisfacción del cliente (Brown et al. 2005), también es necesario considerar la repetición de llamadas, llamadas innecesarias, escalada de llamadas y quejas a la alta gerencia, devoluciones de llamadas, etc. (Sharp, 2003). Los *Call centers* que recompensan solo la productividad encontrarán que esa calidad y el rendimiento se ven afectados, lo que resulta en insatisfacción del cliente y, en última instancia, desgaste. Por lo tanto, es esencial que los *Call centers* encuentren combinación correcta de los siguientes componentes (Fluss, 2005):

- productividad: llamadas por hora, tiempo promedio de conversación, correos electrónicos por hora;
- rendimiento: tasas de ahorro, ventas cerradas;
- calidad: adhesión del agente a las políticas y procedimientos;
- la satisfacción del cliente: es un conjunto detallado de reglas y objetivos requeridos por el cliente para lograr un resultado un servicio aceptable.

Un desafío central en el diseño y gestión de un servicio en general, y un *Call center* en particular, es lograr un equilibrio entre la eficiencia operativa y calidad de servicio. En general,

las operaciones del centro de llamadas se pueden clasificar en dos categorías (Aktekin & Ekin,2016): inbound y outbound. Los *Call centers Inbound* son aquellos que reciben llamadas de clientes (llamar a un banco, servicio al cliente para ayuda, etc.); mientras que en los *Call centers Outbound,* los agentes llaman a clientes (empresas de cobro y telemarketing que están promoviendo nuevos productos y características).

La base para realizar el cálculo de la cantidad de asesores necesarios en un *Call center* es contar con el pronóstico de la demanda. El pronóstico de la demanda es una predicción de acontecimientos futuros que se utiliza con propósitos de planificación basados en modelos matemáticos que utilizan los datos históricos disponibles, en métodos cualitativos que aprovechan la experiencia administrativa y los juicios de los clientes, o en una combinación de las dos cosas (Krajewski, Ritzman, & Malhotra, 2008).

El pronóstico de la demanda requiere elegir las técnicas matemáticas o estadísticas adecuadas para generar predicciones adecuadas. Según Bowersox, Closs, & Cooper (2007) existen tres categorías de técnicas de pronósticos: (1) cualitativa, (2) de series de tiempo, y (3) causal. Una técnica cualitativa emplea datos como las opiniones de los expertos y la información especial para predecir el futuro y puede o no considerar el pasado; mientras que la técnica de series de tiempo se concentra por completo en los esquemas históricos y sus cambios para generar las predicciones. Una técnica causal, como una regresión, emplea información refinada y específica de las variables para desarrollar una relación entre un evento en curso y una actividad prevista.

Tabla 1

Aplicaciones de pronóstico de la demanda

	Horizonte de tiempo		
Aplicación	**Corto plazo (0 a 3 meses)**	**Mediano plazo (3 meses a 2 años)**	**Largo plazo (más de 2 años)**
Cantidad pronosticada	Productos o servicios individuales	Total de ventas Grupos o familias de productos o servicios	Total de ventas
Área de decisión	Administración de inventario Programación del ensamblaje final Programación de horarios de trabajo Programación maestra de la producción	Planificación de personal Planificación de la producción Programación maestra de la producción Compras Distribución	Localización de instalaciones Planificación de la capacidad Administración de procesos
Técnica de pronóstico	Series de tiempo Causal De juicio	Causal De juicio	Causal De juicio

Fuente: (Krajewski, Ritzman, & Malhotra, 2008)

Según Krajewski, Ritzman, & Malhotra (2008) para elegir adeacuadamente la técnica de pronóstico es necesario considerar el horizonte de tiempo, la cantidad a pronosticar y el área de decisión (ver Tabla 1); como también lo explica Brown et al. (2005), la utilización eficaz de las técnicas para la elaboración de un pronóstico depende de las características de la situación y las capacidades de la técnica de pronóstico. Según Bowersox, Closs & Cooper (2007), los principales criterios para evaluar la viabilidad de la aplicación de una técnica de pronósticos son: (1) la precisión; (2) el horizonte de tiempo de la predicción;(3) el valor de la predicción; (4) la disponibilidad de datos; (5) el tipo de esquema de datos; y (6) la experiencia de quién predice. Cada técnica de predicción debe evaluarse en lo cualitativo y en lo cuantitativo en relación con estos seis criterios.

La teoría de colas es el estudio de la espera en distintas modalidades y utiliza modelos para representar sistemas que surgen en la práctica (Hillier & Lieberman, 2010). Teoría de colas es aplicable a diversas empresas de servicio y de manufacturera, recolectando información del patrón de llegada de los clientes y características del procesamiento del sistema (Krajewski, Ritzman, & Malhotra, 2008). En el caso particular de los *Call centers Inbound* son ejemplos de sistemas de colas donde las llamadas llegan, esperan en una línea virtual y

luego son atendidas por un agente. Estos son usualmente modelados utilizando la cola M / M / S, o en la terminología estándar de la industria: el modelo Erlang C. Una herramienta muy útil para resolver problemas de teoría de colas es la simulación porque permite estimar el desempeño de un proceso diseñado en base a los lineamientos y consideraciones de la teoría de colas, y mediante una cantidad adecuada de réplicas para obtener resultados del comportamiento y la posibilidad de generar escenarios que permitan una mejor evaluación (Mehrotra & Fama, 2003).

El presente trabajo nace a raíz de la necesidad de buscar una metodología alternativa que permita facilitar el cálculo de la capacidad instalada en un *Call center* y genere eficiencia en la asignación de recursos, sin disminuir los niveles de servicio que esperan los clientes, ya que, dentro de la estructura de costos de este rubro de negocio, el pago de salarios y horas extras es de gran impacto económico. La principal dificultad observada durante la elaboración de la presente tesis está relacionada con el cálculo de la capacidad de atención por medio de asesores telefónicos, la cual es realizada en la empresa en estudio de manera manual y empleando generalmente un prolongado tiempo. Esta dificultad no permite evaluar diversos escenarios que contemplen la variabilidad de la demanda (picos de demanda, ausentismo, etc.). Este trabajo busca encontrar un modelo para el cálculo de la capacidad de atención por medio de asesores telefónicos tomando como punto de partida la comparación de dos modelos propuestos por diferentes autores que buscan un equilibrio entre contar con un modelo que sea posible de replicar y a su vez mejore la adhesión del modelo a la demanda planificada con respecto a los resultados de un modelo Erlang C, manteniendo el nivel de servicio meta. Para este fin, ambos modelos son testeados con datos históricos de un *Call center* real que brinda servicios de Inbound.

La presente tesis ha sido estructurada de la siguiente manera, en el Capítulo I se analiza la literatura relacionada y proporciona un marco teórico necesario para el análisis de ambos

modelos a investigar. El Capítulo II presenta la problemática en los *Call center* para la determinación de la cantidad de asesores, partiendo de la metodología utilizada actualmente. El Capítulo III presenta las herramientas involucradas en la metodología: los dos modelos de cálculo de Recursos para un *Call center* (el de Inferencia Bayesiana y otro para SPC[2]), las características de ambos y sus modelos de optimización. El Capítulo IV presenta un estudio de caso en el que la metodología propuesta se aplica a un problema del mundo real relacionado con un *Call center* existente en funcionamiento; las metodologías propuestas se aplican a una prueba de muestra obtenida del *Call center* en estudio. Finalmente, el Capítulo V deriva las conclusiones de este trabajo y las direcciones futuras de la investigación.

Este trabajo se centrará en las operaciones de un *Call center Inbound* y tiene como objetivo general aumentar la eficiencia en el uso de un *Call center Inbound* por medio de un modelo de optimización que determine la capacidad instalada de asesores. El objetivo específico consiste en comparar dos modelos de optimización con enfoques distintos, que permitan determinar la capacidad instalada, así como analizar sus diferencias con la metodología que actualmente aplica la empresa, mediante la cual se calcula el requerimiento de asesores semanal por horas utilizando la calculadora Erlang C, empleando como inputs un pronóstico de llamadas, el nivel de servicio esperado y un tiempo de atención constante.

[2] Solución de Personal Consecutiva, de Woo Kim & Ho Ha (2010)

CAPÍTULO I: MARCO TEÓRICO

1.1 Enfoque de la tesis

En el presente trabajo buscamos encontrar un enfoque adecuado para tratar el tema del cálculo de la cantidad de asesores requeridos en un *Call center* Inbound, debido a que tanto en la práctica como en la literatura revisada observamos que, si bien el modelo Erlang C está muy difundido, la simplificación de sus variables ocasiona distorsiones que impactan en el cálculo. Esta búsqueda nos ha orientado a revisar otras fuentes y hemos hallado dos metodologías, la del modelo de colas con abandono de Aktekin & Soyer (2012) y la de Solución de Personal Consecutiva (SPC) de Woo Kim & Ho Ha (2010), que a través de considerar otros enfoques adicionales logran mejoras sustanciales en el cálculo de la cantidad necesaria de asesores.

En cuanto a la primera metodología a evaluar, tenemos al modelo que desarrolla los detalles de la inferencia bayesiana para las colas con abandono más conocido como el Modelo M / M / s + M (*Erlang*-A), trabajo realizado por Aktekin & Soyer (2012), cuyo objetivo es desarrollar un análisis bayesiano de las colas M / M / s + M, considerando que las colas, distribución de llegadas, tiempos de servicio y tiempo de abandono siguen una distribución exponencial y el sistema es atendido por servidores idénticos. Este modelo investiga cuestiones como: (a) incertidumbre de modelo sobre la llegada, el servicio y tasas de abandono; (b) evalua las distribuciones de las características operativas como la cantidad de clientes en el sistema; (c) implicaciones de personal y nivel de servicio. Emplearemos este estudio debido a que aborda cuestiones de inferencia estadística en los modelos de colas bayesianas con clientes impacientes, haciendo énfasis en las operaciones del centro de llamadas y con potencial para estudios posteriores.

Por tro lado, tenemos la segunda metodología a evaluar, el modelo de Solución de Personal Consecutiva (SPC), usando simulación en el *Call center*. Dicho modelo fue

desarrollado por Woo Kim & Ho Ha (2010) basado en el trabajo de Whitt (1999) y propone calcular la cantidad de personal en función del tiempo, contando con una cantidad de asesores que permita responder las llamadas de forma inmediata lo cual deja de lado la preocupación habitual del análisis de rendimiento sobre el impacto de espera antes de ser atendido o el bloqueo de llamadas debido a que no hay agentes disponibles. En este contexto, Whitt (1999) indica que los requisitos para el cálculo de asesores necesarios en un futuro próximo intervalo de tiempo son dos: (a) el número de llamadas actuales que permanecerán en progreso en el intervalo siguiente; y (b) el número de llamadas nuevas que llegarán y permanecerán en servicio.

1.1.1 Los *Call center*s

Los *Call centers* o centro de llamadas son un grupo de recursos (personas y tecnologías) que brindan un servicio telefónico, y existen diversas formas de este servicio dependiendo del tipo y la forma de servicio que ofertan, los más comunes son:

- *Inbound Call center* – denominado también centro de llamadas entrantes, este tipo de servicio se dedican exclusivamente a llamadas entrantes iniciadas por el cliente.
- *Outbound Call center* – denominado también centro de llamadas salientes, este tipo de servicio exclusivamente realizan las llamadas a los clientes o potenciales clientes.
- *Blended Call center* – este tipo de servicio combina llamadas entrantes y salientes.
- *Contact Center* – son centros que utilizan las empresas para la gestión de sus contactos con sus clientes por medio de diversos medios de comunicación tales como el teléfono, e-mail, mensajería instantánea, redes sociales, etc.

Los *Call centers* brindan servicios en diferentes rubros: por su funcionalidad – mesa de ayuda, emergencias, atención al cliente, telemarketing, etc; por su tamaño o por las características de los agentes u operadores – manejo de idiomas, distintos conocimientos, etc.

En cuanto a la organización de los *Call centers* existen dos modalidades:

- Plana – Los agentes están expuestos a todo tipo de llamadas.
- Multi-capa – Cada capa representa un nivel de especialización diferente, dependiendo del servicio que brinda.

1.1.2 Calidad y eficiencia del servicio

Barrenechea, G. (2016) señala que el objetivo de los *Call centers* es brindar un servicio con una determinada calidad, sujeto a un presupuesto específico. La calidad del servicio se puede medir en dos dimensiones, una cualitativa y otra cuantitativa. La primera determina la percepción del cliente (satisfacción del servicio) y está focalizada al marketing y se obtienen a través de las encuestas de satisfacción y de opinión. Y lo cuantitativo (operación) se focaliza en diferentes medidas de rendimiento, como el abandono de llamadas, el tiempo de espera para ser atendido, duración del servicio y el rendimiento de los operadores.

El indicador importante referente al abandono, es el Nivel de Servicio (NS), y se calcula como la fracción de llamadas atendidas antes de un determinado umbral de tiempo, respecto al total de llamadas recibidas; no se consideran las llamadas que abandonan el sistema dentro de los 5 segundos (llamadas fantasmas). Normalmente los *Call centers* se trazan como objetivo atender como mínimo al 80% de los clientes antes de 20 segundos de espera, o regla 80/20.

$$NS = \frac{\#\text{Llamadas Atendidas (antes del umbral)}}{\#\text{LLamadas recibidas (sin } llamadas\ fantasmas)} \tag{1.1}$$

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

$$NAT = \frac{\text{\#Llamadas Atendidas}}{\text{\#LLamadas recibidas (sin } llamadas\ fantasmas)} \tag{1.2}$$

El Tiempo Medio Operativo (TMO) se calcula como el promedio de la duración de todas las llamadas atendidas durante el tiempo que se requiera controlar, generalmente se controla por día y por mes.

$$TMO = \frac{\text{Duración de las llamadas}}{\text{\#LLamadas atendidas}} \tag{1.3}$$

No obstante, el rendimiento de los operadores se mide a través de dos indicadores, la utilización y la ocupación. La utilización es el tiempo que los operadores están hablando menos los reductores internos entre el total de horas habladas. Los reductores son todo lo que aleje a un agente de su capacidad de tomar contacto con los clientes, siendo los reductores internos los que pueden producirse durante la jornada laboral: reuniones de equipo, entrenamiento, formación, fallas del sistema, imprevistos, descansos, entre otros. La ocupación mide el porcentaje promedio de tiempo que los agentes están ocupados atendiendo una llamada. El nivel de ocupación aceptable esta entre el 60% y 80%.

$$Utilización = \frac{\text{Tiempo Hablado} - \text{Reductores internos}}{\text{Tiempo Hablado}} \tag{1.4}$$

$$Ocupación = \frac{\text{Tiempo Hablado}}{\text{Tiempo Hablado} + \text{Tiempo Disponible}} \tag{1.5}$$

1.1.3 Teoría de Colas.

Para Taha (2012) y Hillier & Lieberman, (2010) el proceso de esperar por una atención es parte cotidiana de la vida, que puede ser observado tanto en seres humanos como en los diferentes procesos existentes. El problema que se busca resolver en estos casos es el de disminuir todo lo posible la espera, donde la espera tiene una relacion inversa a la disponibilidad de recursos, con lo cual es necesario encontrar un balance entre ambos para evitar o tener largas esperas o gasto inútil de recursos. Siendo los elementos encontrados en un sistema de colas los siguientes: (a) un insumo, o población de clientes, que genera clientes potenciales; (b) una fila de espera formada por los clientes; (c) la instalación de servicio, constituida por una persona (o una cuadrilla), una máquina (o grupo de máquinas) o ambas cosas, si así se requiere para proveer el servicio que el cliente solicita; (d) una regla de prioridad para seleccionar al siguiente cliente que será atendido por la instalación de servicio

La **Error! Reference source not found.** ilustra los elementos básicos de un modelo de colas. Los triángulos, círculos y cuadrados dentro de la población sirven para ilustrar la diversas necesidades de los de los clientes. El sistema de servicio está constituido por las filas y las instalaciones de servicio, las cuales poseen una distribución de probabilidad para el tiempo de atención de clientes. Una vez que se ha prestado el servicio, los clientes atendidos salen del sistema.

Figura 1:Elementos básicos de los modelos de colas
Fuente Hillier & Lieberman (2010)

1.1.3.1 Modelos de Teoría de Colas

Según Fitzsimmons & Fitzsimmons (2008) existen diferentes modelos en las Teorías de Colas (ver **Error! Reference source not found.**), una forma de clasificar los modelos con servidores paralelos empleando la notación *A/B/C* donde hay tres puntos principales a ser identificados: *A* representa la distribución del tiempo entre arribos; *B* la distribución de los tiempos de servicio; y *C* el número de servidores en paralelos. La notación empleada para describir los modelos es el siguiente:

- M: indica que la distribución de los arribos o del tiempo de servicio es exponencial;
- D: indica que la distribución de los arribos o del tiempo de servicio es constante;
- E_k: indica una distribución Erlang con parámetro k;
- G: indica una distribución general con media y varianza

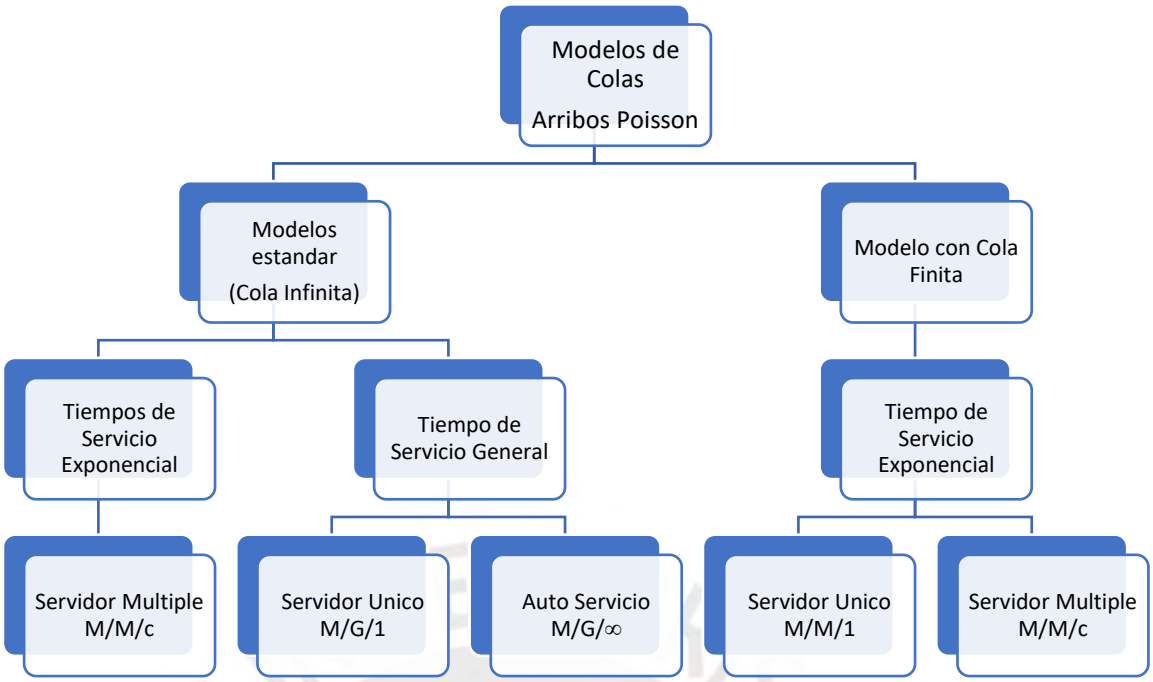


Figura 2:Clasificación de los Modelos de Colas

Fuente: Fitzsimmons & Fitzsimmons (2008)

1.1.4 Modelo Erlang C

Los *Call centers* son ejemplos de sistemas de colas cuyo modelo de colas es el M / M / N, o más conocido como el modelo Erlang C. Este modelo, ampliamente utilizado por profesionales y académicos, hace muchas suposiciones que son cuestionables ya que supone que las llamadas llegan a un proceso de Poisson con una tasa promedio conocida y que son atendidas por un número definido de agentes estadísticamente idénticos con tiempos de servicio que siguen una distribución exponencial asumiendo que todas las personas que llaman esperan tanto como sea necesario para el servicio sin abandonar, es decir, colgar (Robbins, Medeiros, & Harrison, 2010).

El modelo Erlang C (cola M / M / N) es un multiservidor muy simple de colas como lo describe Hillier & Lieberman (2010) a continuación:

- Longitud de la fila: El número de clientes que forman una fila de espera.
- Número de clientes en el sistema: El número de clientes que reciben el servicio.

- Tiempo de espera en la fila: El tiempo que espera un cliente hasta ser atendido.
- Tiempo total en el sistema: El tiempo total transcurrido desde la entrada al sistema hasta la salida del mismo.
- Utilización de las instalaciones de servicio: La utilización colectiva de instalaciones de servicio refleja el porcentaje de tiempo que éstas permanecen ocupadas.

Según Gans, Koole & Mandelbaum (2003), las llamadas llegan con una distribución Poisson con un ratio de arribo λ, considerando que los eventos Poisson cuentan con tiempos entre llegadas independientes y distribuidos exponencialmente de forma idéntica. La cola que se forma es infinita y tiene una lógica FIFO; los clientes son atendidos por *N* asesores con iguales características, que tiene una ratio de atención de $N\mu$ y un tiempo de servicio exponencial de μ. Robbins, Medeiros & Harrison (2010), presentan una serie de ecuaciones que permiten calcular lo siguiente:

Intensidad de tráfico: $R \triangleq \frac{\lambda}{\mu}$

Utilización del sistema: $\rho \triangleq \frac{\lambda}{N\mu} = \frac{R}{N}$

Probabilidad de que los N asesores estén ocupados:

$$P\{Espera > 0\} = \frac{1-(\sum_{m=0}^{N-1}\frac{R^m}{m!})}{((\sum_{m=0}^{N-1}\frac{R^m}{m!})+\left(\frac{R^m}{N!}\right)\left(\frac{1}{1-\frac{R}{N}}\right))} \quad (1.6)$$

Velocidad promedio de respuesta (ASA):

$$ASA \triangleq E[Espera] = P\{Espera > 0\} * E[espera \mid espera > 0]$$

$$ASA = P\{Espera > 0\} * \left(\frac{1}{N}\right)\left(\frac{1}{\mu}\right)\left(\frac{1}{1-\rho}\right) \quad (1.7)$$

Nivel de Servicio:

$$Nivel\ de\ Servicio\ \triangleq P\{Espera \leq T\}$$

$$Nivel\ de\ Servicio\ = 1 - P\{Espera > 0\}\, P[espera\ |espera > 0]$$

$$Nivel\ de\ Servicio = 1 - C(N,R) * \ e^{-N\mu(1-\rho)T} \qquad (1.8)$$

Debido a que el modelo no considera el abandono de la llamada, el cálculo de este indicador no está contemplado en las ecuaciones arriba mencionadas.

Para las fórmulas arriba citadas, se considera lo siguiente:

λ : Promedio del número de llamadas en un periodo de tiempo

μ: Promedio del número de llamadas atendidas en un periodo de tiempo

R : Intensidad del tráfico

N: Cantidad de asesores en un periodo de tiempo

La intensidad del tráfico (R) es el periodo de tiempo que tomarían todas las llamadas telefónicas si se ordenaran de extremo a extremo. Entonces, si tenemos 200 llamadas con un tiempo promedio de 3 minutos, tendríamos un total de 600 minutos de llamadas o 10 horas de llamadas. La unidad técnica para horas de llamadas se denomina Erlang, entonces, para el caso la intensidad del tráfico sería de 10 Erlangs.

Ahora necesitamos estimar la cantidad bruta de agentes necesarios para manejar esta intensidad de tráfico. Sabemos que tenemos 10 Erlangs (10 llamadas de horas de tráfico por hora). Esto significa que el número mínimo de agentes que necesitaríamos en el centro de llamadas sería de 10 agentes. Esta cifra de 10 agentes supondría que todas las llamadas llegan exactamente en el momento en que finalizó una llamada anterior y que no se forman colas. Entonces, comenzamos agregando 1 a la intensidad del tráfico.

Estimación 1: N = R + 1 = 10 + 1 = 11 agentes

Luego alimentamos la Intensidad de tráfico (R) y el Número de agentes (N) en la fórmula Erlang C (fórmula 1.6) para ver cuál es la probabilidad de que una llamada espere y luego calcular el Nivel de servicio utilizando la fórmula 1.8.

Si el nivel de servicio calculado es menor a la meta, aumentamos el número de agentes hasta que se alcance el nivel de servicio.

Para determinar la cantidad real de asesores (Nr) habría que considerar el grado de ausentismo del personal (G) por diversos motivos, siendo este factor utilizado en la industria de 30% (The Leading Contact Centre Magazine, 2016), entonces dicha cantidad real sería: $Nr = \frac{N}{(1-\frac{G}{100})}$.

1.1.5 Modelo Erlang A

El modelo Erlang A, es una extensión del modelo Erlang C, el cual considera el factor abandono. Según Robbins, Medeiros, & Harrison (2010), recién en el año de 1946 este modelo fue presentado por el matemático Conny Palm, que toma en cuenta la cantidad de personas que abandonan sus llamadas antes de comunicarse con un asesor. Según Gans, Koole, & Mandelbaum (2003) y Mandelbaum & Zeltyn (2005), la capacidad de espera de los clientes tiene una media ϴ, si el tiempo de espera excede esta valor ϴ se presenta el abandono de la llamada. El cálculo de esta métrica emplea una función Gamma incompleta:

$$\gamma\,(x,y) \triangleq \int_0^y t^{x-1} * e^{-t} dt; x > 0, y \geq 0 \qquad (1.9)$$

Mandelbaum & Zeltyn (2005) proponen las siguientes ecuaciones para representar el modelo Erlang A:

$$J = \frac{e^{-\frac{\lambda}{\theta}}}{\theta} * \left(\frac{\theta}{\lambda}\right)^{\frac{n\mu}{\theta}} * \gamma(\frac{n\mu}{\theta} * \frac{\lambda}{\theta}) \qquad (1.10)$$

$$\varepsilon = \frac{\sum_{j=0}^{n-1} \frac{1}{j!} * \left(\frac{\lambda}{\mu}\right)^{j}}{\frac{1}{(n-1)!} * \left(\frac{\lambda}{\mu}\right)^{n-1}} \tag{1.11}$$

Probabilidad de espera:

$$P\{Espera > 0) = \frac{\lambda J}{\varepsilon + \lambda J} * (1 - \theta) \tag{1.12}$$

Operador J y Operador ε: los autores los emplean como una simplificación para las ecuaciones anteriores;

n: Número de asesores bruto

λ: número de llamadas por intervalo de tiempo

μ: número de llamadas atendidas por unidad de tiempo

ɵ: número de abandonos por unidad de tiempo (impaciencia)

Conny Palm incorporó el parámetro de "paciencia" dentro de la ecuación de Erlang, asumiendo una probabilidad de abandono media para cada tiempo de espera, de esta manera se pone un tope a la paciencia media. La paciencia promedio es la tasa a la que abandonará el 50% de las personas, que se puede calcular fácilmente en un *Call center* al trazar el porcentaje de llamadas abandonadas contra el tiempo. Un cliente abandona y sale de la cola cuando su tiempo de espera excede su tiempo de paciencia.

En la Figura 3 se muestra el % de llamadas abandonadas vs el tiempo en que este ocurre, dependiendo del tipo de *Call center*, ya sea ventas o soporte técnico, la paciencia promedio será diferente, debido a las diferencias en la intención de los clientes y su disposición a esperar (The Leading Contact Centre Magazine, 2019).

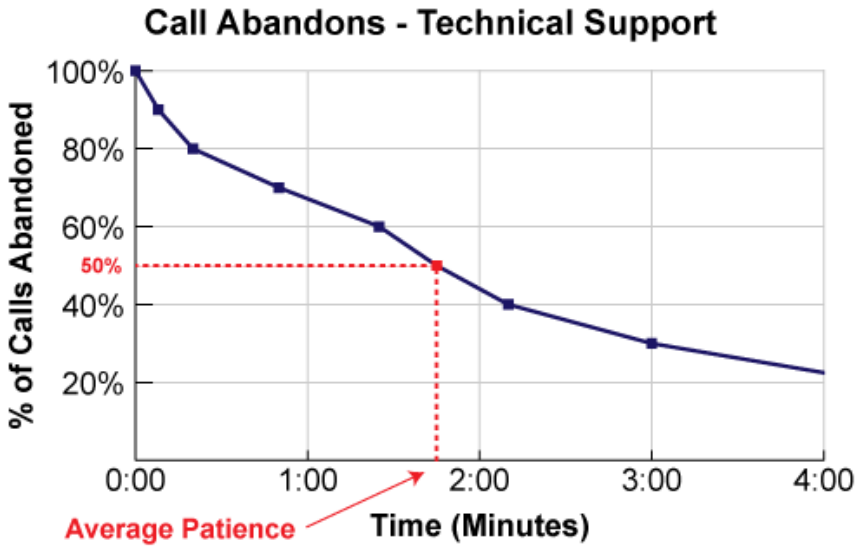


Figura 3:% de llamadas abandonadas vs tiempo en que ocurre el abandono

Fuente: (The Leading Contact Centre Magazine, 2019)

Es importante remarcar que el parámetro de paciencia es muy sensible y un poco engañoso cuando lo estimamos a partir de los datos de nuestra operación, dado que no sabemos cuál es la paciencia promedio de las llamadas que no se abandonan.

Un error muy común es medir el tiempo promedio de abandono real y usar ese dato sin ajustar en la fórmula, lo que generaría una sobreestimación del abandono y por ende un sub-dimensionamiento de la cantidad de recursos necesarios (el problema opuesto a Erlang C).

1.1.6 Pautas para determinar los objetivos de nivel de servicio

Según Sharp (2003) existen varios métodos para determinar los objetivos de nivel de servicio, pero los siguientes cuatro enfoques se han extraído de la experiencia colectiva de gerentes de *Call center*:

- Minimizar el abandono: Donde se busca elegir un objetivo de nivel de servicio consiste en mantener lo más bajo posible los tiempos de espera en cola sin perder llamadas.
- Estándar en la industria: El objetivo 80/20 ha sido citado en algunos manuales como un "estándar de la industria" bastante común porque para muchos *Call centers* es un

equilibrio razonable entre las expectativas de las personas que llaman y la practicidad de tener suficiente personal para cumplir el objetivo.

- Relacionarse con la competencia: Comparando la gestión de los competidores u otras organizaciones similares y luego usar esta información como punto de partida.
- Realizar una encuesta al cliente: Un cuarto método para elegir el nivel de servicio es realizar una encuesta a los clientes analizando la tolerancia de la persona que llama y de esta forma conocer que esperan las personas que llaman.

1.1.7 Análisis Bayesiano aplicada a la teoría de colas

El análisis de modelos de colas desde una perspectiva Bayesiana parece remontarse a principios de los años setenta, donde los primeros trabajos abordan el problema de la estimación puntual de la tasa de llegadas y de servicio en sistemas de colas Markovianos mediante estimadores Bayes (Muddapur (1972); Reynolds (1973)). En la misma época, Bagchi y Cunningham (1972) proponen algunos procedimientos Bayesianos para el diseño óptimo de sistemas de colas con un único servidor, con el fin de encontrar la mejor tasa de servicio y capacidad del sistema según unos costes preestablecido.

Según Ausín Olivera (2003) la metodología Bayesiana permite un procedimiento sencillo de incorporar la incertidumbre resultante de las estimaciones asociadas al proceso de llegadas y de servicio en la predicción de estas cantidades de interés.

Aktekin y Soyer (2012) presentan un trabajo, basado en los antecedentes relevados por Brown, et al. (2005) y Mandelbaum & Zeltyn (2005), que tuvo como objetivo desarrollar el análisis Bayesiano del modelo de colas M/M/ s + M que considera los ratios de arribo, servicio y abandono como varibles en vez de ser fijos como en el modelo Erlang.

1.1.8 Solución de Personal Consecutiva (SPC)

En general, los modelos de colas simples suponen que la tasa de llegada de llamadas entrantes de los centros de llamadas son constantes y los tiempos de servicio se distribuyen de manera idéntica, sin embargo, las tasas de llegada a los centros de llamadas reales varían en el transcurso de un día, por lo cual Whitt (1999) propone calcular la cantidad de personal en función del tiempo, contando con una cantidad de asesores que permita responder las llamadas de forma inmediata lo cual deja de lado la preocupación habitual del análisis de rendimiento sobre el impacto de espera antes de ser atendido o el bloqueo de llamadas debido a que no hay agentes disponibles. Con este contexto Whitt (1999) indica que los requisitos para el cálculo de asesores necesarios en un futuro próximo intervalo de tiempo son dos: (a) el número de llamadas actuales que permanecerán en progreso en el intervalo siguiente y (b) el número de llamadas nuevas llamadas que llegarán y permanecerán en servicio. Según Woo Kim & Ho Ha (2010) emplear una Solución de Personal Consecutiva (SPC) es beneficioso ya que considera llamadas incompletas y la cantidad determinada de dotación de personal es menor que bajo un enfoque Erlang C, lo cual conlleva a una disminución en el total costos operativos para el *Call center* manteniendo el nivel de servicio. Otro beneficio es emplear intervalos de planificación más cortos, obteniendo una planificación precisa y actualización de los niveles de personal durante un horizonte de planificación único.

CAPÍTULO II: PROBLEMÁTICA DEL CASO

Como se revisó en el marco teórico, desde los inicios de las operaciones de los *Call center* el cálculo de la cantidad de asesores disponibles para atender a los clientes manteniendo los correspondientes niveles de servicio, según el rubro, han sido un problema perene en este tipo de organizaciones. Este problema ha retomado fuerza en los últimos tiempos debido a la existencia de nuevas tecnologías que permiten facilitar los cálculos y la necesidad de acompañar el crecimiento del sector debido al arribo de tecnologías para contar con multi-sites o el uso de IVRs incrementando la complejidad de la realidad (Gans, Koole, & Mandelbaum, 2003). En cualquier solución que se proponga es necesario considerar los aspectos de ambiente laboral en este problema (Mandelbaum & Zeltyn, 2009) por un lado es necesario cumplir los niveles de servicio meta, como indica Sharp (2003) también es importante considerar la calidad del ambiente de trabajo donde laboran los agentes.

En este trabajo, buscamos comparar el enfoque tradicional en la resolución del problema del cálculo de la cantidad de asesores en un *Call center* por medio del modelo Erlang C contrastándolo con los modelos propuestos por Aktekin & Soyer (2012), empleando un enfoque de optimización, en base a las restricciones detalladas en su trabajo, y por Woo Kim & Ho Ha (2010), que emplea un enfoque de simulación dependiente del tiempo. En todos los casos para evaluar con los mismos inputs emplearemos información real de un *Call center* que cuenta actualmente con 142 ubicaciones y atiende en promedio 353,800 llamadas mensuales; compuesto por 9 equipos de 20 asesores en promedio distribuidos a lo largo del día.

De esta manera, el objetivo general es determinar el número óptimo de asesores en un rango horario, de modo que los costos operativos se minimicen y se garantice un nivel de servicio determinado como explica Sharp (2003) y considerar que a menor nivel de servicio mayor es la espera del cliente, y donde está presente la repetición de llamadas, llamadas innecesarias, escalada de llamadas y quejas a la alta gerencia, devoluciones de llamadas, etc.

2.1.Definición del problema

Luego de la revisión de los diferentes autores que han trabajado sobre el tema de teoría de colas aplicada a un *Call center*, definimos como problema a resolver el encontrar una metodología para el cálculo de la cantidad de asesores necesarios para atender llamadas en un *Call center* cumpliendo con los niveles de servicio correspondientes, considerando aspectos tales como el abandono de llamadas sin sobrecargar a los asesores.

2.2.Proceso de atención en un *Call center* y terminología relacionada

Según Fluss (2005) los *Call centers* tiene como misión clásica el tema del servicio, ventas o una combinación de ambos, al menor costo posible. En los últimos 30 años, hacer un buen trabajo en servicio y ventas fue suficiente, pero actualmente ya no es suficiente; hoy en día, las empresas no pueden permitirse ignorar los millones de dólares en oportunidades de ingresos sin explotar que fluyen a través de sus centros de contacto en clientes no estructurados discusiones y correos electrónicos.

Para Sharp (2003), ante el requisito de generar ganancias, muchas empresas enfrentan un problema importante como el aumento de los costos de personal, por lo cual en los próximos años la gestión del personal será importante por:

- El *Call center* promedio gasta entre el 60 y el 70% de su presupuesto anual en salario del personal;
- A nivel mundial, las tasas de rotación de agentes promedian 22% y se acercan al 50% en algunas industrias;
- El ausentismo laboral está aumentando y podría llegar hasta un 17% en la industria del cuidado de la salud, 10% en telecomunicaciones y consumo, así como en los mercados de productos, y 9% en promedio en todos los mercados verticales;

- Más del 80% de las empresas utilizan anuncios externos para buscar agentes y el 72% utilizan agencias de contratación, con costos significativos;
- Existe una grave escasez de personal calificado en lugares y el reclutamiento se está volviendo muy difícil;
- Un rápido aumento en el crecimiento de la industria de *call / contact center*;
- El crecimiento de la interacción CRM y multimedia requerirá agentes calificados y experimentados, y los costos de capacitación aumentarán en consecuencia.

Según Woo Kim & Ho Ha (2010) la gestión de operaciones de *Call center* más tradicional se realiza por medio de un Workforce Management (WFM) o Administración de los recursos, que se define como aquellos procesos empleados por los *Call centers* para asegurar el número correcto de agentes, con las habilidades adecuadas, en el momento adecuado y en el correcto estado de ánimo para ofrecer la experiencia deseada del cliente. Es necesario cumplir, según Sharp (2003), que los niveles de servicio mientras se administran los costos es un ciclo iterativo que requiere que se completen varios procesos clave. Los sistemas de WFM deben ofrecer las siguientes funcionalidades para respaldar la empresa moderna centrada en el cliente:

- Programación para cumplir con los niveles de servicio;
- Adherencia entre el modelo y el pronóstico real;
- reporteria para la gestión e información para los pronósticos;
- Contar con escenarios que permitan flexibilidad al modelo;
- Contar con una atención multisitio;
- Soporte legal;
- Soporte tecnológico.

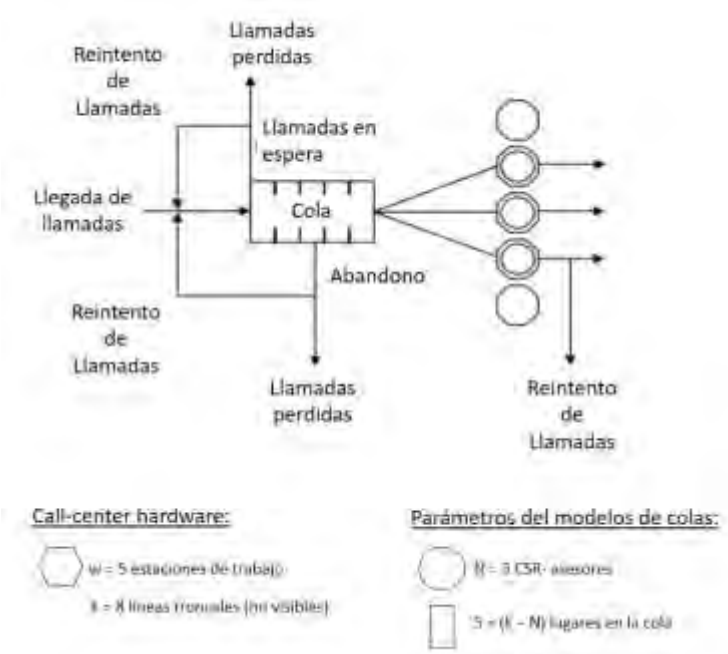


Figura 4:Esquema operativo de un Call center básico

Fuente: Gans, Koole, & Mandelbaum (2003)

En la **Error! Reference source not found.** se muestra el esquema de un *Call center* basico, según Gans, Koole & Mandelbaum (2003), este cuenta con un conjunto de k líneas troncales de ingreso de llamadas al *Call center*, donde hay $w \leq k$ estaciones de trabajo, a menudo referidas como posiciones (modulos donde se ubicarán los asesores), con un grupo de agentes $N \leq w$ que atiende las llamadas entrantes. Estas últimas, en caso encuentren todas las k líneas troncales ocupadas recibe una señal de ocupado, se bloquea la entrada al sistema y apenas haya una línea disponible será ocupada por una de las llamadas que estaba en espera. Si hay menos de N agentes ocupados, la llamada que ingresa es atendida inmediatamente por un asesor, en cambio si hay más de N agentes ocupados pero menos de k llamadas en el sistema, la llamada entrante espera en cola, con un orden FIFO, hasta que un agente esté disponible; en caso los clientes que se impacientan cuelgan o abandonan antes ser atendidos. Una vez que una llamada sale del sistema, libera los recursos (línea troncal, estación de trabajo y asesor) para volver a estar disponibles para las llamadas entrantes. Existe una fracción de las llamadas que al no recibir atención reintentan ingresar al servicio; un grupo restante de llamadas bloqueadas

o abandonadas se pierden. Los clientes también pueden volver al sistema cuyos reingresos pueden ser para atenciones de servicios adicionales.

Continuando con lo indicado por Gans, Koole & Mandelbaum (2003), para cualquier N fijo (número de asesores), se utilizan ampliamente en la gestión de los *Call centers* estos modelos de teorias de colas, cuyo modelo más simple y más utilizado es el de Cola M / M / N, también conocida como Erlang C; sin embargo, en muchas aplicaciones este modelo es una simplificación excesiva ya que ignora las señales de ocupado, cliente impaciente y servicios que abarcan múltiples visitas. En la práctica, el proceso de servicio esbozado anteriormente es a menudo mucho más complicado, por ejemplo, la incorporación de un IVR, con el cual los clientes interactúan antes de unirse a la cola de los agentes, crea dos estaciones en tándem: un IVR seguido de pull de asesores.

La figura N°5 muestra el promedio de volumen diario de llamadas atendidas en un *Call center* para el periodo agosto – septiembre del año 2019, considerando periodos de 45 minutos, reflejando una mayor densidad de llamadas desde las 08:45 a las 19:15 horas del día.

Asimismo, la figura N°6 muestra el promedio de llamadas abandonadas en un *Call center* para el periodo agosto – septiembre del año 2019, considerando periodos de 45 minutos, evidenciando una mayor densidad entre las 10:15 y las 18:30 horas del día, llegándose a un máximo de 72 abandonos a las 17:45 horas.

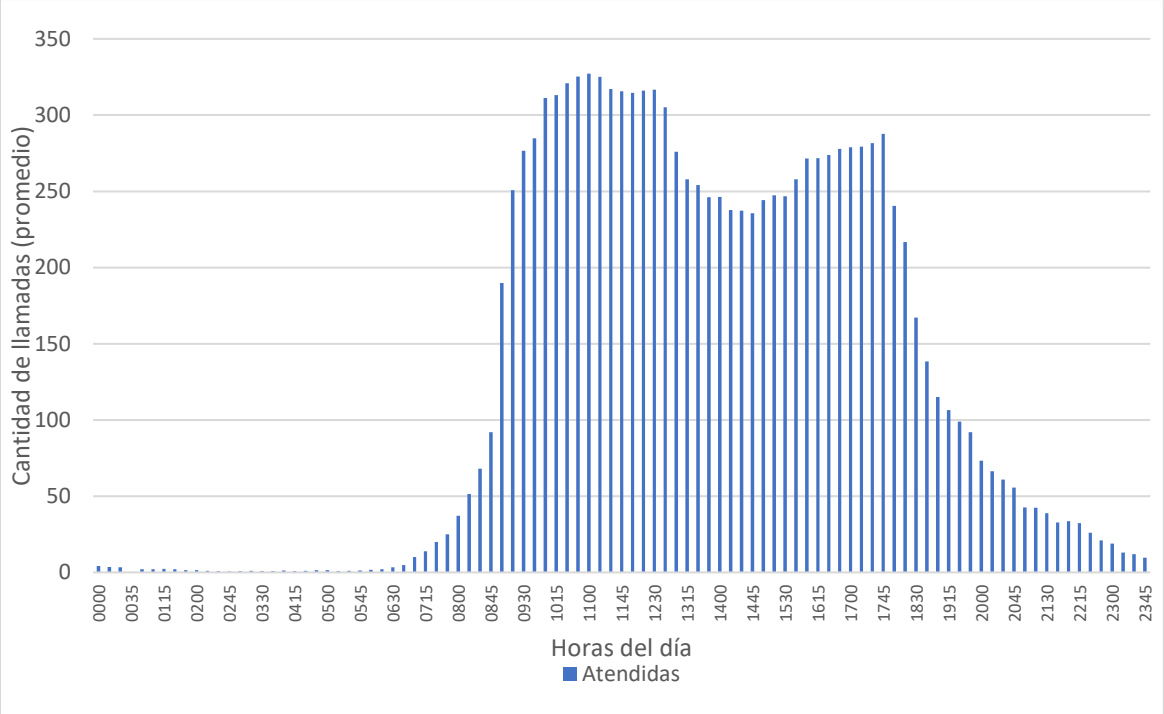


Figura 5:Volumen de llamadas atendidas en el Call center promedio Ago-Sep 2019

Fuente: Empresa en estudio (2019)

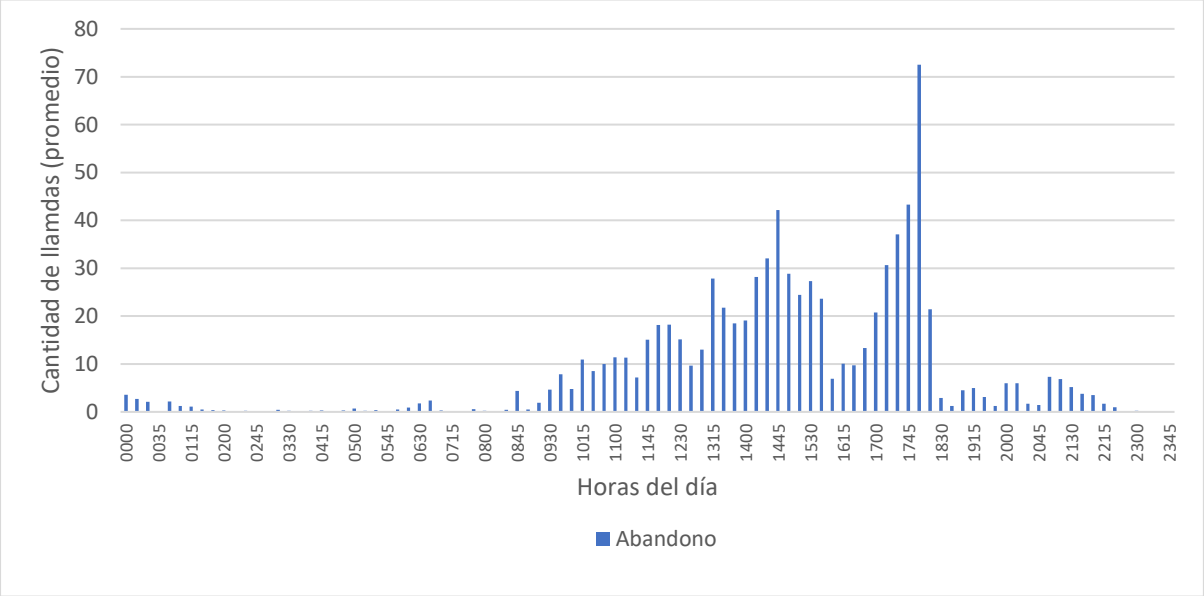


Figura 6:Volumen de llamadas abandonadas en el Call center promedio Ago-Sep 2019.

Fuente: Empresa en estudio (2019)

2.3. Descripción del problema

En la atención de un *Call center*, como indica Sharp (2003), existe el indicador nivel de servicio que permite medir de forma limitada el rendimiento general del *Call center* ya que

indica que "cuán pocas personas tuvieron que esperar más que una cierta cantidad de segundos antes de ser atendidos por un asesor". Esta medición no proporciona información sobre la calidad en la atención (debido a que esta se da en la interacción asesor-cliente), pero proporciona un panorama sobre el funcionamiento del *Call center*. Un buen nivel de servicio facilita la capacidad de respuesta de llamadas por parte de la organización, mientras que uno deficiente incrementa las quejas de los clientes y del tiempo de atención por las disculpas que deberá realizar el asesor. Estas deficiencias tienen como consecuencia una disminución en la calidad del servicio ofrecido, ya que tener asesores sobrecargados constantemente pueden volverse menos "amigables con el cliente", con lo cual los clientes no resuelven sus dificultades y los asesores cometen más errores y contribuyen a la repetición de llamadas, llamadas innecesarias, escalada de llamadas y quejas a la alta gerencia, devoluciones de llamadas, etc., todo lo cual impulsa a que el nivel de servicio siga bajando en un círculo vicioso.

2.4. Cálculo de asesores basado en un modelo tradicional

En sus investigaciones Gans, Koole & Mandelbaum (2003) detallan el procedimiento para el cálculo de la cantidad de asesores necesarios para el manejo operativo de un *Call center*. Para este cálculo se emplea el modelo Erlang C, asumiendo que los periodos evaluados son cortos (entre 30 a 60 minutos). Las consideraciones del modelo son:

- Contar con la demanda en intervalos de 30 o 60 minutos, el cual debe ser constante en el periodo y presentar una distribución Poisson;
- Tener un nivel de servicio constante;
- El tiempo de servicio debe contar con una distribución Exponencial y ser independiente entre ellos;
- Contar con una lógica FIFO para la atención de los clientes;
- Ignorar el abandono, bloqueos y reintentos de llamadas.

A continuacion se ilustra un ejemplo desarrollado en Gans, Koole & Mandelbaum (2003) para el modelo Erlang C. La **Error! Reference source not found.**7 presenta el patrón de arribos de llamadas con periodos de 30 minutos. En este ejemplo, evaluaremos un escenario que nos permitirá desarrollar toda la serie de ecuaciones y supuestos básicos del modelo que serán empleados. A continucación detallamos la información presentada en cada columna:

- Time: periodo de tiempo donde fue realizada la medición
- Recvd: cantidad de llamadas recibidas
- Answ: cantidad de llamadas atendidas
- Abn%: Porcentaje de abandono
- ASA: tiempo promedio para contestar una llamada
- AHT: tiempo promedio de atención de los asesores
- Occ%: porcentaje de ocupación
- On Prod%: porcentaje de asesores disponibles
- On Prod FTE: cantidad de FTE disponibles (FTE acrónimo de Full time Equivalent, medida de equivalencia empleada para poder homologar personal tiempo completo con personal tiempo parcial)
- SCH open FTE: cantidad de FTE programados
- SCH Avail %: cantidad de FTE programados en porcentaje

Charlotte - Center

Time	Recvd	Answ	Abn %	ASA	AHT	Occ %	On Prod%	On Prod FTE	Sch Open FTE	Sch Avail %
0	20,577	19,860	3.5%	30	307	95.1%	85.4%	222.7	234.6	95.0%
8:00	332	308	7.2%	27	302	87.1%	79.5%	59.3	66.9	88.5%
8:30	653	615	5.8%	58	293	96.1%	81.1%	104.1	111.7	93.2%
9:00	866	796	8.1%	63	308	97.1%	84.7%	140.4	145.3	96.6%
9:30	1,152	1,138	1.2%	28	303	90.8%	81.6%	211.1	221.3	95.4%
10:00	1,330	1,286	3.3%	22	307	98.4%	84.3%	223.1	229.0	97.4%
10:30	1,364	1,338	1.9%	33	296	99.0%	84.1%	222.5	227.9	97.6%
11:00	1,380	1,280	7.2%	34	306	98.2%	84.0%	222.0	223.9	99.2%
11:30	1,272	1,247	2.0%	44	298	94.6%	82.8%	218.0	233.2	93.5%
12:00	1,179	1,177	0.2%	1	306	91.6%	88.6%	218.3	222.5	98.1%
12:30	1,174	1,160	1.2%	10	302	95.5%	93.6%	203.8	209.8	97.1%
13:00	1,018	999	1.9%	9	314	95.4%	91.2%	182.9	187.0	97.8%
13:30	1,061	961	9.4%	67	306	100.0%	88.9%	163.4	182.5	89.5%
14:00	1,173	1,082	7.8%	78	313	99.5%	85.7%	188.9	213.0	88.7%
14:30	1,212	1,179	2.7%	23	304	96.6%	86.0%	206.1	220.9	93.3%
15:00	1,137	1,122	1.3%	15	320	96.9%	83.5%	205.8	222.1	92.7%
15:30	1,169	1,137	2.7%	17	311	97.1%	84.6%	202.2	207.0	97.7%
16:00	1,107	1,059	4.3%	46	315	99.2%	79.4%	187.1	192.9	97.0%
16:30	914	892	2.4%	22	307	95.2%	81.8%	160.0	172.3	92.8%
17:00	615	615	0.0%	2	328	83.0%	93.6%	135.0	146.2	92.3%
17:30	420	420	0.0%	0	328	73.8%	95.4%	103.5	116.1	89.2%
18:00	49	49	0.0%	14	180	84.2%	99.1%	5.8	1.4	416.2%

Figura 7:Ejemplo de informe resumido en lapsos de 30 min.
Fuente: Gans, Koole, & Mandelbaum (2003)

A partir de la información que se muestra en la Figura 7, se desprende la cantidad de asesores programados (Sch Open FTE) en cada segmento horario (cada 30 minutos) para la atención de la demanda de llamadas ingresadas a un *Call Center* (Recvd), cálculo obtenido a partir del uso de la metodología Erlang C desarrollado en el numeral 1.1.4.

2.5. Desarrollo de ejemplo empleando Erlang C

Para el caso, consideraremos en un *Call center* determinado, un intervalo de tiempo de 1 hora, siendo la cantidad de llamadas atendidas que tomaremos en este intervalo de 973.

El Average Handling Time (AHT – tiempo promedio de atención) es de 302 segundos o 5.03 minutos.

- Nivel de servicio: 80%

- Tiempo objetivo de atención (en segundos): B=20
- Ocupación Máxima: 85%
- Factor de ausentismo (asesores adicionales por ausencias): G= 30%

La aplicación del modelo consiste en 6 pasos:

Paso 1: calcular la intensidad del tráfico de llamadas. El tráfico de llamadas mide la utilización de los recursos a través del tiempo. Este cálculo se traduce en horas de llamadas, que en forma técnica se le conocen como Erlang y no tienen unidades. Este cálculo se obtiene a partir de la cantidad de llamadas y del AHT. La fórmula empleada es la siguiente:

$$Intensidad = A = \frac{LH * AHT}{60} \quad (2.1)$$

Donde LH = 973 llamadas y AHT = 5.033 minutos. Reemplazando los datos obtendremos que A= 81.62 Erlang

Paso 2: para determinar el número inicial de asesores necesarios para poder atender la intensidad calculada en el paso 1, consideramos que si tenemos 81.62 Erlang es necesario contar al menos con dicha cantidad de asesores por lo cual consideraremos lo siguiente:

$$Numero\ inicial\ de\ asesores = N = A + 1 \quad (2.2)$$

Por lo cual el resultado de N seria 83, es necesario redondear el resultado ya que no es posible contar con una fracción de un asesor.

Paso 3: en el modelo Erlang C, el siguiente paso es hallar la probabilidad de espera de las llamadas para ser atendidas (recordemos que en este modelo no existe el abandono por lo cual la llamada podría esperar indefinidamente). Esta probabilidad de espera representa la espera de la llamada para ser atendida en el caso que todos los asesores estén ocupados, definida por la siguiente ecuación:

$$\mathrm{Pw} = \left(\left(\frac{A^N}{N!}\right) * \left(\frac{N}{N-A}\right)\right) / \left(\sum_{i=0}^{N-1}\left(\frac{A^i}{i!}\right)\right) + \left(\left(\frac{A^N}{N!}\right) * \left(\frac{N}{N-A}\right)\right) \quad (2.3)$$

Donde reemplazaremos los valores de A y N calculados en el paso 2. El resultado es 0.9625.

En la fórmula 2.3 se está adaptando la fórmula 1.6 en función de A, siendo que A=R.

Paso 4: a partir del cálculo realizado en el paso 3 va a ser posible hallar el nivel de servicio (S):

$$S = 1 - \left[\mathrm{Pw} * \ e^{-\left[(N-A)*\left(\frac{B}{AHT}\right)\right]}\right] \tag{2.4}$$

Se está adaptando la fórmula 1.8 en función de A y AHT, siendo que AHT es la duración promedio de una llamada en minutos.

Al realizar el cálculo con los valores iniciales, el valor de S solo llega a ser 5%, lo cual nos indica que solo un 5% de las llamadas atendidas esperan 20 segundos o menos. Por ende, este nivel de servicio asociado a un N= 83 asesores no es suficiente para cumplir con el lineamiento de un nivel de servicio del 80 %. Para cubrir este requerimiento es necesario incrementar el valor de N hasta lograr un nivel de servicio mínimo del 80%. En la Tabla 2 se muestran los resultados obtenidos al incrementar gradualmente la cantidad de agentes disponibles y podemos observar que con un N= 95 logramos obtener un S=80%.

Tabla 2

Resultados del Nivel de servicio según el número de asesores disponibles

N	Pw	S
83	93%	5%
84	76%	17%
85	63%	27%
86	52%	35%
87	43%	42%
88	36%	49%
89	30%	55%
90	24%	60%
91	20%	65%
92	17%	69%
93	14%	73%
94	11%	77%
95	9%	80%
96	7%	82%

Paso 5: Revisar la máxima ocupación de los asesores, ya que este indicador nos permite saber cuan saturados de trabajo se encuentran, ya que debemos considerar que no debe sobrepasar un 85% para evitar fatiga y bajo performance de los asesores. Este indicador se calcula de la siguiente manera:

$$Máxima\ ocupación = O = \frac{A}{N} * 100 \qquad (2.5)$$

En este caso, en el paso anterior fue obtenido que para un nivel de servicio del 80% es necesario contar con N = 95 asesores, el cual le corresponde un valor de O = 87 %. En la

Tabla 3 se muestra cómo disminuye la ocupación según se incremente el número de asesores disponibles, con lo cual sería necesario llegar a un N= 97.

Tabla 3

Resultados de la Ocupación según el número de asesores disponibles

N	Pw	S	O
83	93%	5%	100%
84	76%	17%	98%
85	63%	27%	97%
86	52%	35%	96%
87	43%	42%	95%
88	36%	49%	94%
89	30%	55%	93%
90	24%	60%	92%
91	20%	65%	91%
92	17%	69%	90%
93	14%	73%	89%
94	11%	77%	88%
95	9%	80%	87%
96	7%	82%	86%
97	6%	85%	85%

Paso 6: Incorporar al número final de asesores un factor de seguridad que permita cubrir eventualidades como el ausentismo de los asesores o las vacaciones, de la siguiente forma:

$$N_R = \frac{N}{\left[1-\left(\frac{G}{100}\right)\right]} \tag{2.6}$$

En este ejemplo, el número de asesores a emplear es N=97, aplicando un factor de ausentismo G de 33%, se obtiene un número de asesores final N_R de 145.

CAPÍTULO III: MODELOS DE COLAS PRESENTES EN LA LITERATURA

Este capítulo presenta las metodologías propuestas por Aktekin & Soyer (2012) y Woo Kim & Ho Ha (2010) para determinar la cantidad de asesores requeridos para la atención de un *Call center*. En la Sección 3.1 se presentan los modelos matemáticos para determinar la cantidad de asesores necesarios con las metodologías presentadas en la investigación *Bayesian Analysis of Queues with Impatient Customers: Applications to Call centers* (BAQ) de Aktekin & Soyer (2012); y el trabajo de Woo Kim & Ho Ha (2010), desarrollado en el artículo *Consecutive Staffing Solution Using Simulation in the Contact Center* (SPC). En la Sección 3.2 se explica la importancia de considerar un stock de seguridad debido al impacto de las ausencias de los asesores; y finalmente, en la Sección 3.3 se comparan ambos modelos.

En este capítulo se procederá a evaluar los modelos estudiados en las dos primeras etapas a partir de la información obtenida del entorno actual de la gestión de *Call center* a estudiar; detallando las características que deben considerarse en ambos casos y contrastando con el modelo empleado actualmente. Asimismo, se identificarán buenas prácticas existentes sobre los distintos niveles de servicio que existen como estándares internacionales en la atención en este tipo de rubro.

3.1. Modelos matemáticos para determinar la cantidad de asesores necesarios

Para las organizaciones en general el principal objetivo es gestionar sus operaciones con una cantidad medida de recursos garantizando lograr los objetivos de calidad de producción o servicio. En la gestión de *Call centers* se menciona que en la estructura de costos el tema de los salarios representa entre un 60 – 70 % (Gans, Koole, & Mandelbaum, 2003), por lo cual las actividades relacionadas a cuantificar la capacidad instalada requerida son muy importantes. En principio, la organización tiene dos caminos al establecer las políticas sobre este tema y el de contar con un staff constante o uno variable que tome en consideración la demanda de

llamadas. Esto último permite buscar soluciones óptimas en diferentes momentos del día siempre considerando los picos y valles de la demanda, que son la base del presente estudio.

Los dos modelos en los cuales se basa la presente tesis son:

(1) *Bayesian Analysis of Queues with Impatient Customers: Applications to Call centers* (BAQ) de Aktekin & Soyer (2012); y (2) el trabajo de Woo Kim & Ho Ha (2010), desarrollado en el artículo *Consecutive Staffing Solution Using Simulation in the Contact Center* (SPC). Ambos trabajos apuntan a poder contar modelos que permitan describir de mejor forma la realidad en los *Call centers* y no ser tan estáticos como el modelo Erlang C. Las directrices de estos modelos son:

- Modelar con ciertas condiciones de incertidumbre en temas tales como la demanda, tiempos de servicio y abandono.
- Considerar dentro del sistema el impacto de contar con llamadas en proceso sobre todo en las distribuciones a emplear.
- El impacto de estas mejoras en el cálculo de la cantidad de asesores y su impacto en el nivel de servicio ofrecido.

A continuación, en la Sección 3.1.1 se proporciona la notación para los modelos matemáticos. En la Sección 3.1.2 y la Sección 3.1.3 se revisan las características particulares de cada modelo.

3.1.1 Notación

Los modelos emplean los siguientes parámetros y notaciones:

Índice

j: Índice de clientes que permanecen en la cola de atención

Parámetros

s: Número de asesores bruto

λ: ratio de arribo

μ: ratio de servicio

ɵ: ratio de abandono

c: costo de un asesor disponible

a: costo de abandono

3.1.2 Modelo matemático - Bayesian Analysis of Queues (BAQ)

Este enfoque presenta dos temas importantes, primero es su tratamiento de la incertidumbre en λ (t), μ y θ que crea un modelo estocástico, servicio, y procesos de abandono. En los modelos de colas tradicionales, se supone que estas tasas de entrada son fijas. En nuestra configuración, la incertidumbre en las tasas se modela con una distribución gamma independiente variables aleatorias. El enfoque de modelado considerado es la inferencia bayesiana de las colas M / M / s + M.

Las ecuaciones que definen este modelo serían las siguientes (Mandelbaum & Zeltyn, 2009):

F.O. Minimizar

$$E\{C(s,\lambda,\mu,\theta)\} = \int_{\lambda}\int_{\mu}\int_{\theta}\{cs + a\lambda \Pr(\mathrm{Ab}|\lambda,\mu,\theta)\}p(\lambda,\mu,\theta)d\lambda\, d\mu\, d\theta. \quad (3.1)$$

S.T.

$$A\,(x,y) = \frac{xe^{y}}{y^{x}}\tau(x,y) \quad (3.2)$$

$$P_n = \Pr(N = n|\lambda,\mu,\theta) = \begin{cases} P_s \frac{s!}{n!s^{n-s}} & \text{for} \quad 0 \le n \le s, \\ P_s \frac{\left(\frac{\lambda}{\theta}\right)^{n-s}}{\prod_{k=1}^{n-s}\left(\frac{s\mu}{\theta}+k\right)} & \text{for} \quad n \ge s+1, \end{cases} \quad (3.3)$$

$$P_s = \frac{E_{1,s}}{1 + \left[A\left(\frac{s\mu}{\theta}, \frac{\lambda}{\theta}\right) - 1\right] E_{1,s}}, \tag{3.4}$$

$$E_{1,s} = \frac{\frac{r^s}{s!}}{\sum_{j=0}^{s} \frac{r^j}{j!}}, \tag{3.5}$$

$$r = \lambda/\mu \tag{3.6}$$

$$P_{T_q} = \Pr(T_q > 0|\lambda, \mu, \theta) = \sum_{n=s}^{\infty} \Pr(N = n|\lambda, \mu, \theta), \tag{3.7}$$

(3.7)

$$\Pr_j(\mathrm{Ab}|\mu, \theta) = \frac{(j+1)\theta}{s\mu + (j+1)\theta}, \; j \geq 0. \tag{3.8}$$

$$P_{\mathrm{Ab}|T_q} = \Pr(\mathrm{Ab}|T_q > 0, \lambda, \mu, \theta) = \frac{1}{\rho A\left(\frac{s\mu}{\theta}, \frac{\lambda}{\theta}\right)} + 1 - \frac{1}{\rho}, \tag{3.9}$$

$$\rho = \lambda/s\mu \tag{3.10}$$

$$P_{\mathrm{Ab}} = \Pr(\mathrm{Ab}|\lambda, \mu, \theta) = \left(\frac{1}{\rho A\left(\frac{s\mu}{\theta}, \frac{\lambda}{\theta}\right)} + 1 - \frac{1}{\rho}\right) \times \sum_{n=s}^{\infty} \Pr(N = n|\lambda, \mu, \theta). \tag{3.11}$$

(3.11)

La ecuación 3.1 es la función objetivo (F.O) de minimizar el costo promedio de operación (en cada unidad de tiempo), debido a que λ, μ y ɵ van a tomar valores aleatorios según lo establecido en el modelo Bayesiano, es necesario expresar la F.O. como una integral triple en función a los tres parámetros. La restricción (3.2) define el estado de listo, donde el

sistema ya llego a un nivel de estabilidad en el modelo, donde $x, y > 0$ y $\gamma = (x, y)$ es una función acumulada de gamma. Una vez definida la ecuación (3.2) la probabilidad de tener n clientes en el sistema, condicionado a λ, μ y ϴ se obtiene con la ecuación (3.3), que a su vez está compuesta por las ecuaciones (3.4), (3.5) y (3.6). El termino E_1, es también llamado probabilidad de bloqueo, causado por las llamadas pérdidas a causa del bloqueo del sistema por encontrarse totalmente ocupado. La probabilidad que el cliente espere en cola está definida en (3.7). La probabilidad de abandono está definida en (3.8) y la probabilidad de abandono de una llamada que no es atendida apenas llegue se define en (3.9). A partir de la definición de (3.8), (3.9) y (3.10) es posible calcular la probabilidad del abandono durante el estado de listo como se presenta en (3.11).

Para resolver este modelo es necesario hacer un muestreo y aplicar la aproximación de Montecarlo, tal como se describe en el citado paper de esta sección.

3.1.3 Modelo matemático: Consecutive staffing solution using simulation in the contact center (SPC)

Según Woo Kim & Ho Ha (2010), para intervalos de planificación cortos, es muy difícil que se logre un estado estable en la mayoría de los casos. En este contexto, el enfoque Consecutive staffing solution (CSS) es ideal, ya que su enfoque depende del tiempo y considera aquellas llamadas incompletas. Para aplicar el CSS es necesario emplear las siguientes funciones:

- $N_t(m,c,\lambda,\mu)$: El número promedio de clientes en el sistema (en atención), con tasa de llegada ƛ, duración promedio del servicio μ^{-1} en el momento t donde haya *m* clientes en el momento 0 y se asignaron c agentes para responder llamadas entrantes durante el intervalo [0; t].

- SL_t (m,c,λ,μ): El nivel de servicio promedio del sistema con una llegada tasa λ y duración promedio del servicio μ^{-1} durante el intervalo de tiempo [0; t].

Ahora, Considere intervalos de planificación consecutivos 1, 2, 3, . . . , n, con longitud idéntica (15 minutos en este caso) y pronósticos de tasas de llegada y duraciones promedio de servicio para cada intervalo es $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n$ y $\mu_1, \mu_2, \mu_3, \ldots, \mu_n$, que se estiman de datos históricos. Entonces, los niveles de personal para los intervalos de planificación pueden ser determinado por los siguientes pasos:

- Para el primer intervalo de planificación, la cantidad de asesores c_1 se puede determinar cómo número entero mínimo que satisface $SL_{15}(0, c_1, \lambda_1, \mu_1) < SL^*$, en el que SL * es el límite superior del nivel de servicio.
- Cuando comience el segundo intervalo, habrá $N_{15}(0, c_1, \lambda_1, \mu_1)$ clientes en el sistema. Por lo tanto, se puede determinar el nivel de personal para el segundo intervalo, c_2 como el mínimo entero que satisface $SL_{15}(N_{15}(0, c_1, \lambda_1, \mu_1), c_2, \lambda_2, \mu_2) < SL^*$. Cuando comienza el tercer intervalo, habrá clientes $N_{15}(N_{15}(0, c_1, \lambda_1, \mu_1), c_2, \lambda_2, \mu_2)$ en el sistema. Así sucesivamente para el resto de los niveles.

Los niveles de personal para los siguientes intervalos consecutivos se pueden determinar de la misma manera. Note que formular ecuaciones $N_t(m,c,\lambda,\mu)$ y $SL_t(m,c,\lambda,\mu)$ podría ser complejo y puede llegar a tomar un alto tiempo computacional. Sin embargo, dichas restricciones se pueden aproximar usando simulación (Mehrotra & Fama, 2003). El enfoque de personal basado en la simulación tiene importantes ventajas como generalización y validación automática (Atlason, Epelman, & Henderson, 2008; Feldman et al., 2008).

3.2. Stock de seguridad – ajuste del resultado

Debido a que los asesores requieren descansos, toman vacaciones y tienen ausencias por algún tipo de causa, es necesario contar con un factor que permita ajustar en cierta medida

la cantidad de asesores y de esta forma contar con un stock de seguridad. Este factor agrega realismo a los requerimientos de personal al tener en cuenta los descansos, el absentismo, la capacitación y el trabajo no telefónico, donde los horarios deberían permitir contar con las personas adecuadas en los lugares correctos en los tiempos correctos (Sharp, 2003). En el problema de programación en los *Call centers* de mayor tamaño es más fácil contar con un backup, por ejemplo, un centro de llamadas de 1,000 asesores que usa solo 50 horarios factibles agregaría entre 0 y 50 asesores adicionales para poder cubrir temas de vacaciones. Eso es un intervalo de [0% – 5%] por encima de lo establecido (Gans, Koole, & Mandelbaum, 2003).

CAPÍTULO IV: CASO DE ESTUDIO CALL CENTER

El pronóstico de las ventas de servicios en los *Contact Centers*, dentro de los cuales están considerados los *Call center* se incrementen en más de 9% anual según Guy Ford en la entrevista concedida a Bruno Ysla (2017). Por esa razón, se han realizado grandes inversiones en los últimos años para la ampliación de servicios y la integración tecnológica que permitirá el contacto a través de chats, *Facebook o Whatsapp*. En el año 2016 las ventas en el negocio de *Contact Centers* estaban alrededor de 445 millones de dólares americanos (ver figura 8).

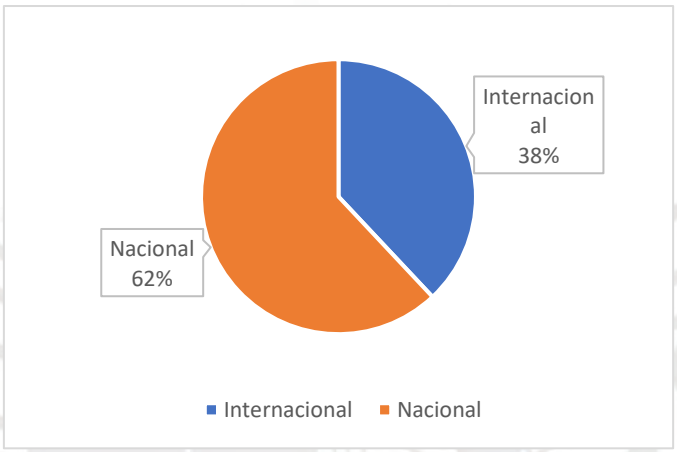


Figura 8:Ventas del mercado de Contact Centers 2016.
Fuente: Ysla (2017)

Las mejoras en espacios físicos van acompañadas por la búsqueda de métodos más eficientes para utilizar los recursos, en este sentido, la metodología desarrollada en esta tesis ayudará a aprovechar al máximo futuras implementaciones de nuevas posiciones y utilizar correctamente este incremento de capacidad y utilizar eficientemente los recursos. Por lo tanto, se busca aplicar las metodologías revisadas en el capítulo anterior probando con la información real obtenida en un *Call center*.

4.1. Realidad en el *Call center*

El funcionamiento real de un *Call center* es bastante más complejo que las metodologías empleadas en general para resolver el problema del cálculo de la cantidad de asesores necesarios para atender la demanda de llamadas. La simplificación más importante que se realiza es la de no considerar el abandono como parte del sistema ya que esta puede llegar a ser un volumen importante de clientes no atendidos y obviados por el sistema creando un sesgo en las mediciones realizadas (Mandelbaum & Zeltyn, 2005). A continuación, revisaremos la data real y empleando los modelos de cálculo de cantidad de asesores procuraremos revisar cuál de ellos obtiene un mejor resultado cumpliendo los niveles de servicio requeridos por los clientes.

4.1.1 Arribos y Abandonos de llamadas

Los grandes *Call centers* logran generar gran cantidad de data, lo cual dificulta su almacenamiento. Según Gans, Koole, & Mandelbaum (2003) los *Call centers* tienen por practica agrupar las llamadas en intervalos de 15 o 30 minutos de tal forma que se promedia la información y permite reducir los costos de almacenamiento de la información. En este caso el Call center a estudiar emplea intervalos de 15 minutos para acumular la información obtenida en las llamadas (ver **Error! Reference source not found.** y **Error! Reference source not found.**). La información analizada fue recopilada durante los meses de agosto y septiembre de 2019, como se puede observar en estas gráficas en el caso de las llamadas atendidas se presentan picos en el volumen de llamadas tanto en las mañanas entre las 10:45 am hasta el mediodia y durante la tarde entre las 05:00 pm hasta las 06:00 pm. En el caso de las llamadas abandonadas podemos observar un pico marcado a las 06:00 pm, donde probablemente la cantidad de asesores disponibles no cubre la demanda de llamadas y es posible que muchos de los clientes prefieran abandonar antes que seguir esperando por la atención. Para gestionar de forma adecuada un Call center es necesario que los indicadores tomen en cuenta aquellos

clientes que abandona, incluso es preferible considerar como abandono aquellos clientes que explicitamente declaran que por el servicio ofrecido no vale la pena esperar.

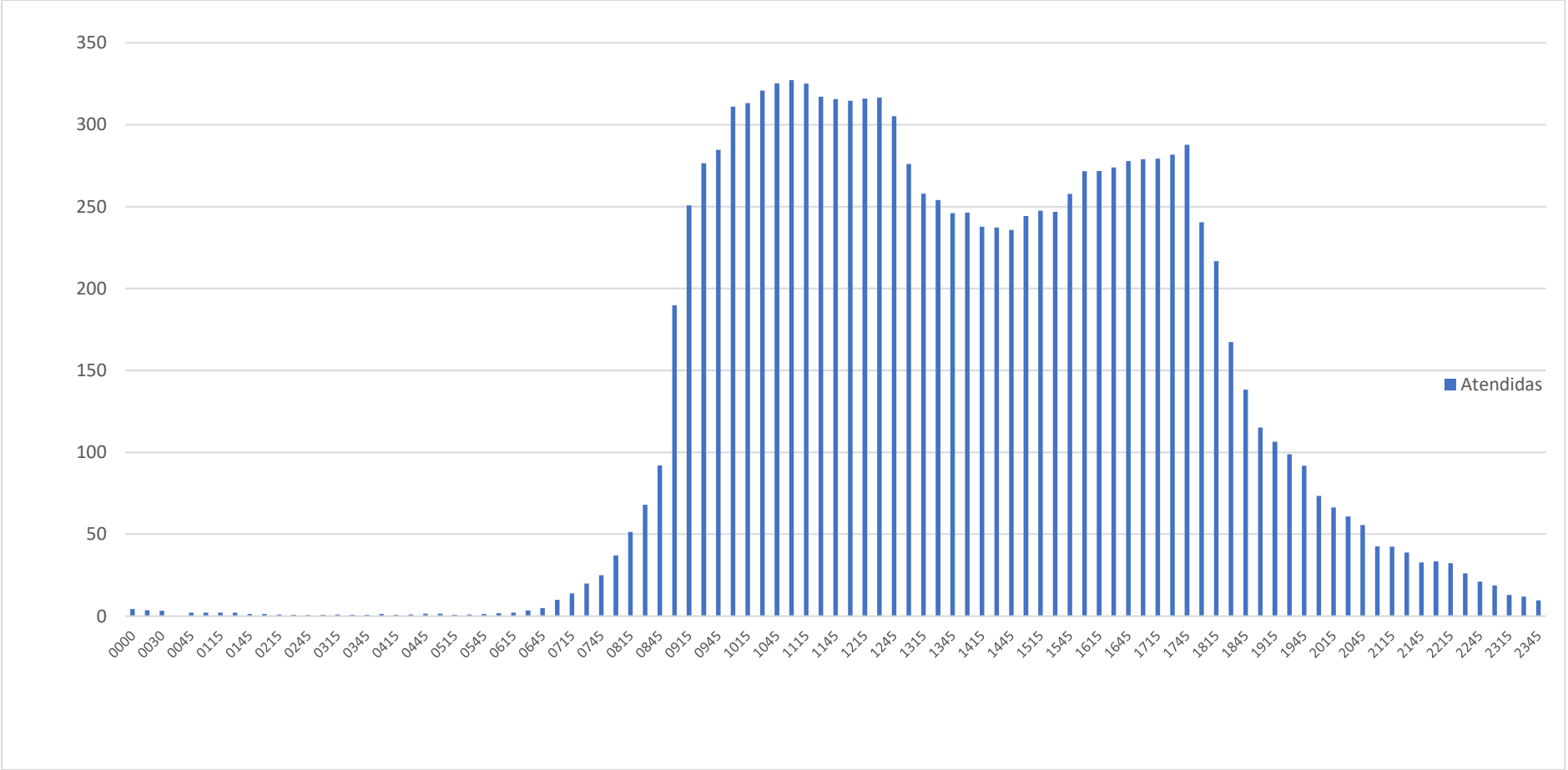


Figura 9:Volumen de llamadas atendidas en el Call center promedio Ago-Sep 2019.

Fuente: Empresa en estudio (2019)

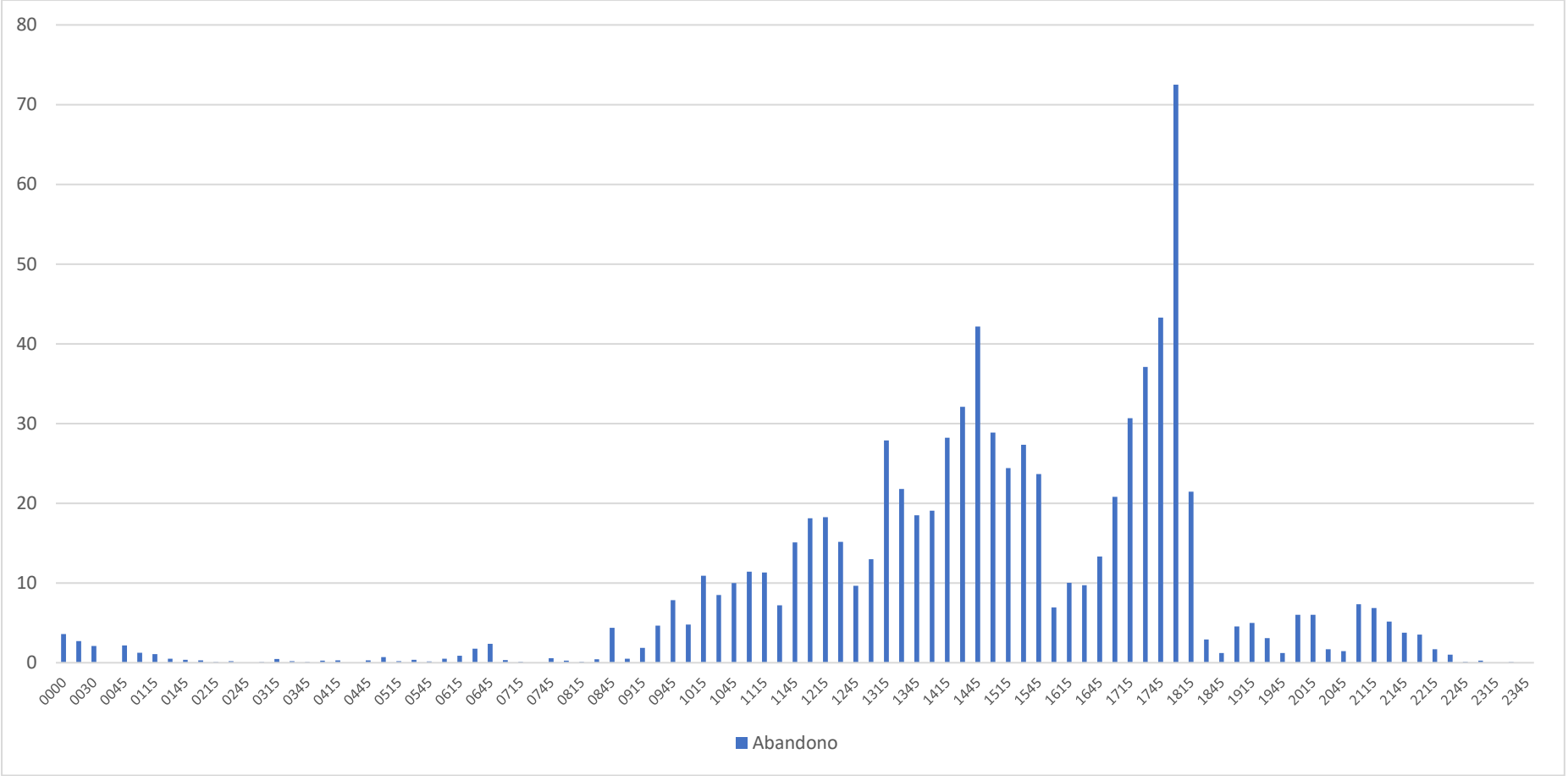


Figura 10:Volumen de llamadas abandonadas en el Call center promedio Ago-Sep 2019.

Fuente: Empresa en estudio (2019)

4.1.2 Tiempo de atención de las llamadas

Como parte de la simplificación del modelo de *Call center,* en la práctica se considera al tiempo de atención como una constante, y en los modelos más avanzados como un valor promedio, cuando en realidad dicho tiempo es variable. Como se mencionó anteriormente, la mayoría de los modelos de *Call center* suponen que el tiempo de atención de las llamadas se distribuyen exponencialmente, pero Mehrotra & Fama (2003) recomiendan usar información distribucional más precisa siempre que sea posible, ya que los autores explican que la razón principal por la que la industria del centro de llamadas acepta la suposición de que los tiempos de atención son exponenciales es porque los datos históricos son almacenados en intervalos donde solo se manejan promedios.

4.2.Parámetros y consideraciones en los modelos a estudiar

A continuación, presentaremos un breve resumen de las características de los modelos a evaluar y que aspectos consideran cada uno de ellos:

	Modelo BAQ	**Modelo SPC**	**Modelo Erlang C**
Indicador objetivo que debe ser cumplido	Nivel de servicio, tiempo de espera y % abandono	% utilización	Nivel de Servicio y tiempo de espera
Método de cálculo empleado	Erlang A añadiendo variabilidad en los tiempos entre llegadas, atención y espera	Modelo M/M/S	Erlang C
Distribuciones asumidas	Tiempos entre llegadas, atención y espera con distribuciones Gamma para generar la variabilidad	Distribuciones Poisson para las llegadas y Exponenciales para los tiempos de atención.	Distribuciones Poisson para las llegadas y tiempos de atención constantes

Figura 11:Comparativos de las características de los modelos a estudiar

Fuente: Woo Kim & Ho Ha (2010) y Aktekin & Ekin, (2016)

En el presente trabajo consideraremos lo siguiente:

- Distribución de arribos: se tomará la muestra relevada (un equipo de trabajo durante los días lunes del mes de septiembre de 2019)
- Periodo de tiempo evaluado: 1 hora
- Disciplina de la cola: FIFO
- Tiempo promedio atención: se tomará de la muestra relevada (un equipo de trabajo durante los días lunes del mes de septiembre de 2019)
- Nivel de servicio: 80% del total de atenciones deben haber sido contestadas antes de los 20 segundos de espera
- Porcentaje de utilización: no debe superar el 85%

4.3. Caso de estudio

En base a una muestra promedio de llamadas de los días lunes de septiembre de 2019 (información de un equipo de trabajo), acumuladas en intervalos de una hora, donde revisaremos primero como se desenvuelven los promedios de tiempos entre arribos, tiempos de espera y tiempos de atención (ver Figuras 12, 13 y 14). En el caso del tiempo de espera está muy por encima de la espera meta de 20 segundos. Realizando las consultas respectivas nos indican que los días lunes siempre hay un incremento en el volumen de llamadas y de forma adicional también un incremento del ausentismo entre los asesores por lo cual la combinación de ambos factores genera problemas en la atención.

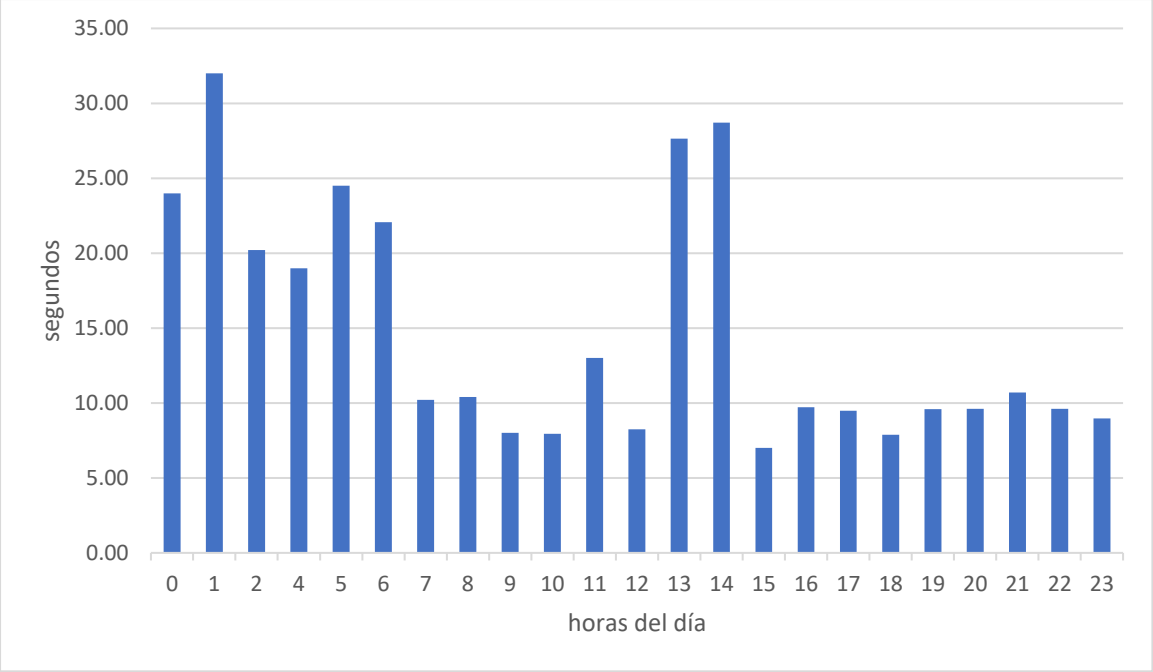


Figura 12:Tiempo promedio de tiempo entre arribos para los días lunes.

Fuente: Empresa en estudio (2019)

Figura 13:: Tiempo promedio de tiempo espera para los días lunes

Fuente: Empresa en estudio (2019)

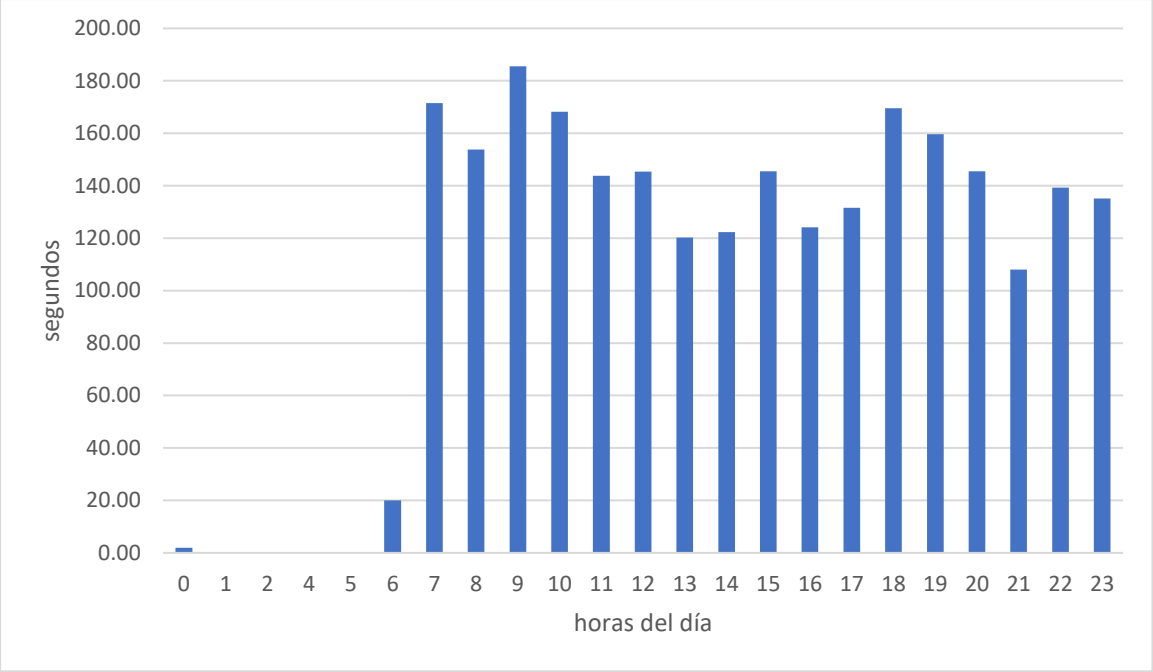


Figura 14:Tiempo promedio de tiempo de atención para los días lunes.
Fuente: Empresa en estudio (2019)

En el primer caso se realizará la evaluación de esta información con el modelo BAQ. Si bien el *Call center* se encuentra en funcionamiento las 24 horas del día, es posible observar que a partir de las 7 am hasta la medianoche el comportamiento del *Call center* es distinto con respecto a las horas en la madrugada. En base a la información muestral obtenida del *Call center* se han armado rangos por hora para poder calcular el promedio de tiempos entre arribos, promedio de tiempos de espera y promedios de tiempo de atención. Empleando el modelo propuesto y utilizando la Calculadora Erlang A[3] (ver Anexo 2) para determinar la cantidad de asesores por cada hora, presentamos los resultados que se muestran en la Tabla 4.

Para dicha tabla se toma en consideración la siguiente data de entrada:

- Ratio de arribo (llamadas atendidas por hora)
- Promedio de tiempo de servicio (promedio de atención en minutos)

[3] http://erlang.chwyean.com/erlang/erlangA.html#instructions

- Tiempo promedio de paciencia (se considera 2 minutos para este modelo)
- Tiempo de espera aceptable = 20 segundos (según condiciones del servicio 80/20)
- Capacidad de cola = 1000 (utilizado en el mercado)
- Nivel de servicio = 80% (según condiciones del servicio)

La calculadora Erlang A, resuelve la función objetivo del modelo BAQ expresado en (3.1), considerando las restricciones expresadas en las fórmulas 3.2 a la 3.11, arrojando los siguientes resultados:

- Cantidad de asesores
- Nivel de servicio (%)
- Probabilidad de llamada en espera (%)
- Abandono (%)
- Tiempo promedio de espera (en minutos)

Tabla 4

Resultados de la evaluación con el modelo BAQ

Rango hora	Promedio tiempo entre arribos (en segundos)	Promedio espera (en segundos)	Promedio atención (en segundos)	Llamadas atendidas	Asesores	Nivel de Servicio (%) SLD	Probabilidad de llamada en espera (%)	Abandono (%)	Tiempo promedio de espera (minutos)
0	24.0	6.0	1.9	31	1.0	100%	22.72%	0.35%	0.0
1	32.0	0.0	0.0	8	1.0	100%	22.24%	0.34%	0.0
2	20.2	0.0	0.0	5	1.0	100%	22.24%	0.34%	0.0
4	19.0	0.0	0.0	4	1.0	100%	22.24%	0.34%	0.0
5	24.5	0.0	0.0	2	1.0	100%	22.24%	0.34%	0.0
6	22.1	0.0	20.1	14	1.0	100%	22.24%	0.34%	0.0
7	10.2	14.5	171.5	129	7.0	80.83%	34.09%	8.6%	0.2
8	10.4	28.4	153.9	242	11.0	80.28%	39.55%	8.25%	0.2
9	8.0	70.6	185.6	402	21.0	81.04%	42.43%	7.67%	0.2
10	8.0	90.0	168.2	531	25.0	82.17%	44.43%	7.2%	0.1
11	13.0	86.6	143.9	469	19.0	80.97%	45.48%	7.68%	0.2
12	8.2	83.2	145.4	422	18.0	84.31%	39.13%	6.4%	0.1
13	27.6	64.6	120.2	454	16.0	83.87%	41.68%	6.6%	0.1
14	28.7	81.8	122.3	539	19.0	83.96%	43.54%	6.58%	0.1
15	7.0	160.1	145.5	455	19.0	82.83%	42.28%	6.97%	0.1
16	9.7	129.5	124.2	429	16.0	85.74%	37.86%	5.87%	0.1
17	9.5	73.2	131.6	378	15.0	85.13%	37.17%	6.09%	0.1
18	7.9	111.5	169.6	319	16.0	83.44%	36.93%	6.78%	0.1
19	9.6	36.5	159.7	201	10.0	83.03%	34.07%	7.17%	0.1
20	9.6	44.0	145.5	233	11.0	86.57%	30.19%	5.58%	0.1
21	10.7	37.5	108.0	286	10.0	86.48%	33.47%	5.59%	0.1
22	9.6	53.0	139.3	256	11.0	83.58%	35.71%	6.8%	0.1
23	9.0	44.3	135.2	171	8.0	86.84%	28.04%	5.55%	0.1

Con la metodología SPC planteada por Woo Kim & Ho Ha (2010) se busca atender la mayor cantidad de llamadas y solo tener una mínima cantidad de llamadas que pasen de un periodo de tiempo a otro (en cola). En base a la misma información empleando el modelo de colas M/M/S para lo cual se utilizará la calculadora de la teoría de colas M/M/C[4] (ver Anexo 3), y como indican los autores, se toma en cuenta tener como objetivo solo 25 llamadas en el sistema entre periodos y un máximo de 85.5% de utilización de los asesores, tendriamos los siguientes resultados (ver Tabla 5):

Para dicha tabla se toma en consideración la siguiente data de entrada:

- Ratio de arribo λ (llamadas arribadas por hora)
- Ratio de servicio μ (llamadas atendidas por hora)

[4] https://www.supositorio.com/rcalc/rcalclite.htm

- Estimación de número de servidores.

La calculadora de la teoría de colas M/M/C, resuelve la metodología para simular la cantidad de asesores del modelo SPC expresado en (3.1.3), siendo que para cada hora se realiza la iteración con una estimación del número de servidores hasta que se cumpla con las condiciones del servicio, es decir, el factor de utilización de asesores sea menor a 85.5%, arrojando los siguientes resultados:

- Cantidad de asesores
- Cantidad de llamadas en el sistema
- Promedio de llamadas en cola
- Tiempo promedio en el sistema (en horas)
- Tiempo de espera promedio en cola (en horas)
- % de utilización de asesores.

Tabla 5

Resultados de la evaluación con el modelo SPC

Rango hora	Ratio arribos	Ratio servicio	Asesores	Cantidad de llamadas en el sistema	Promedio de llamadas en cola	Tiempo promedio en el sistema (en horas)	Tiempo espera promedio en cola (en horas)	Utilización de Asesores (%)
0	150.00	1891.53	1	0	0.01	0.00	0.00	8.00%
1	112.50	0.00	1	0	0.00	0.00	0.00	5.95%
2	178.22	0.00	1	0	0.00	0.00	0.00	5.95%
4	189.47	0.00	1	0	0.00	0.00	0.00	5.95%
5	146.94	0.00	1	0	0.00	0.00	0.00	5.95%
6	163.11	179.36	1	0	0.00	0.00	0.00	5.95%
7	352.35	20.99	20	19	1.85	0.05	0.01	83.93%
8	346.13	23.40	18	16	1.53	0.05	0.00	82.18%
9	449.30	19.40	27	25	2.07	0.06	0.00	85.78%
10	452.77	21.40	25	23	1.79	0.05	0.00	84.63%
11	276.56	25.02	13	14	2.71	0.05	0.01	85.03%
12	436.80	24.76	21	19	1.82	0.04	0.00	84.01%
13	130.26	29.94	6	5	1.00	0.04	0.01	72.51%
14	125.37	29.43	5	8	3.79	0.06	0.03	85.20%
15	513.16	24.75	25	22	1.35	0.04	0.00	82.93%
16	370.36	28.99	15	15	2.59	0.04	0.01	85.17%
17	379.37	27.35	17	15	1.46	0.04	0.00	81.59%
18	456.80	21.23	26	23	1.27	0.05	0.00	82.76%
19	375.70	22.54	20	18	1.69	0.05	0.00	83.34%
20	374.46	24.75	18	17	2.01	0.05	0.01	84.05%
21	336.14	33.34	12	13	2.46	0.04	0.01	84.02%
22	374.18	25.85	17	17	2.43	0.05	0.01	85.15%
23	401.57	26.63	18	17	1.93	0.04	0.00	83.78%

Empleando la metodología actual empleada en el Call center, basada en un modelo Erlang C, y utilizando la Calculadora Erlang C5 (ver Anexo 4) para determinar la cantidad de asesores por cada hora, presentamos los resultados que se muestran en la Tabla 6.

Para dicha tabla se toma en consideración la siguiente data de entrada:

- Promedio del tiempo de atención (segundos); estándar=210 seg
- Ratio de servicio μ (llamadas atendidas por hora);
- Nivel de servicio (%) y;
- Tiempo de respuesta de atención (segundos).

La calculadora Erlang C resuelve la metodología para simular la cantidad de asesores del modelo Erlang C expresado en las fórmulas (2.1) a la (2.6), siendo que para cada hora se realiza la iteración con una estimación del número de servidores hasta que se cumpla con las condiciones del servicio expresadas en las fórmulas (2.4) referida al nivel de servicio de 80%, y la fórmula (2.5) referida a la utilización de asesores que no debe ser mayor de 85%., arrojando los siguientes resultados:

- Cantidad de asesores;
- Utilización de asesores (%);
- Nivel de servicio (%)
- Tiempo promedio en el sistema (en horas);
- Tiempo de espera promedio en cola (en horas) y;
- % de utilización de asesores.

[5] https://www.callcentrehelper.com/tools/erlang-calculator/

Tabla 6

Resultados de la evaluación con el modelo actual

Rango hora	Tiempo promedio de atención (segundos)	Llamadas atendidas	Asesores	Utilización de Asesores (%)	Nivel de Servicio (%)
0	210	31	4	45.2%	89.4%
1	210	8	2	23.4%	92.4%
2	210	5	2	14.6%	96.8%
4	210	4	2	11.7%	97.9%
5	210	2	1	11.7%	89.2%
6	210	14	3	27.2%	95.5%
7	210	129	11	68.4%	87.1%
8	210	242	18	78.4%	83.1%
9	210	402	28	83.8%	82.1%
10	210	531	37	83.7%	87.8%
11	210	469	33	82.9%	87.2%
12	210	422	29	84.9%	80.2%
13	210	454	32	82.8%	86.9%
14	210	539	37	85.0%	85.2%
15	210	455	32	82.9%	86.5%
16	210	429	30	83.4%	84.3%
17	210	378	27	81.7%	85.6%
18	210	319	23	80.9%	83.8%
19	210	201	16	73.3%	88.3%
20	210	233	18	75.5%	87.5%
21	210	286	21	79.4%	84.5%
22	210	256	19	78.6%	83.9%
23	210	171	14	71.3%	88.3%

De la Figura 15 se observa que desde las 0 a 6 hrs. se mantiene constante con un número mínimo de asesores; en el periodo desde las 6 a 10 hrs. se incrementa sustancialmente en los tres modelos, sin embargo, el modelo BAQ utiliza el menor número de asesores, llegando a ser de 30 el más alto por el modelo EC; de 10 a 15 hrs. se reduce la cantidad de asesores, se mantiene el modelo BAQ en mínimo con respecto a los asesores empleado en el modelo EC; de las 15 a 23 hrs. se muestra un cambio radical en cuanto se refiere a los asesores del modelo SPC, ya que obtiene la más alta cantidad con respecto a los otros dos, obteniendo por ejemplo a las 18 hrs más de 23 asesores, pero el modelo BAQ se mantienen con el mínimo de asesores con respeto a los otros dos modelos.

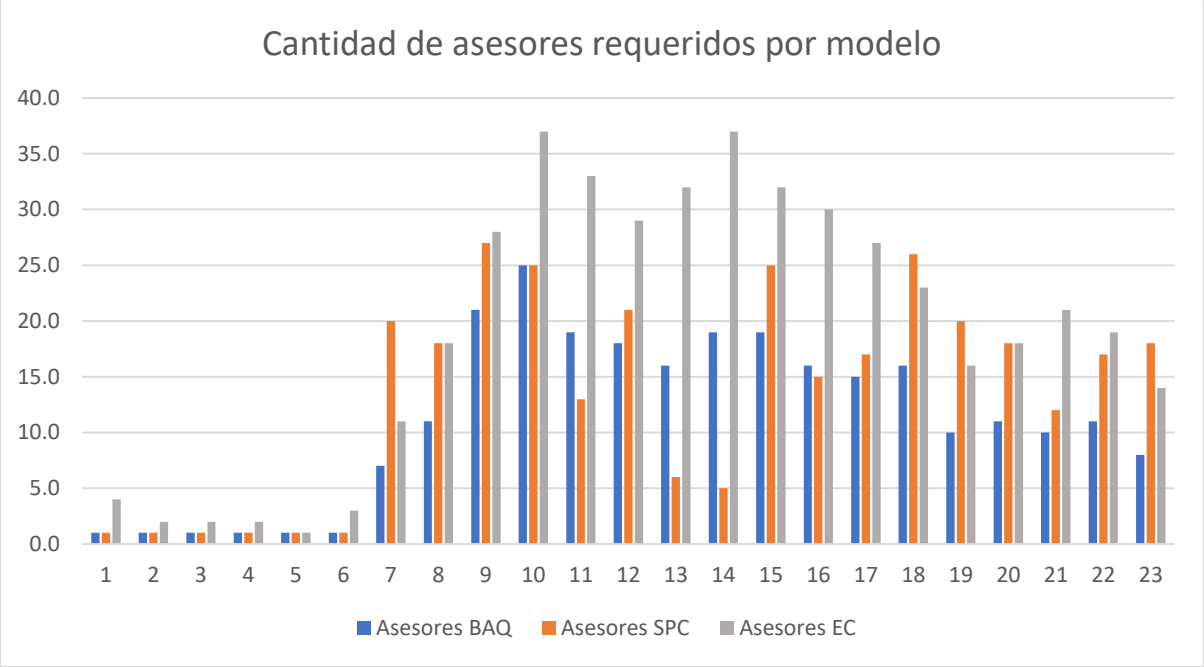


Figura 15:Comparativo de cantidad de agentes promedio requeridos para un equipo durante los días lunes del mes de septiembre 2019 – Según diferentes modelos aplicados
Fuente: Empresa en estudio (2019)

4.4.Hallazgos importantes

En este caso, por medio de la información muestral hemos podido analizar bajo tres diferentes métodos la cantidad de asesores requeridas para un mismo periodo de tiempo. De la figura 15 podemos observar claramente una menor cantidad de asesores necesarios bajo la metodología BAQ en relación con la metodología Erlang C (de 6 a 23 horas), tomando en cuenta que la metodología BAQ considera el abandono de llamadas cuando el cliente ha llegado a su límite de paciencia, a diferencia de la metodología Erlang C que no considera el abandono, asumiendo que ningún cliente abandonará la llamada y en consecuencia asume que atenderá a todos los clientes, lo que conlleva a requerir mayor cantidad de asesores.

Las consideraciones para el modelo BAQ permiten que haya una mejora con respecto al modelo Erlang C empleado actualmente en el *Call center*, debido a que al considerar el

abandono la cantidad de recursos requeridos sería menor, y al añadir la variabilidad en los tiempos de atención entre llegadas de llamadas y espera en cola, contaremos con un resultado que permita determinar la cantidad necesaria de asesores en los momentos del día donde realmente se requieran.

En cuanto a la comparación de los resultados obtenidos entre la metodología BAQ y SPC, en los rangos de 6 a 10 horas y 16 a 23 horas, la metodología BAQ arroja una cantidad de asesores menor a la metodología SPC, sin embargo, para el rango de 10 a 16 horas la cantidad de asesores con la metodología SPC es menor al obtenido con la metodología BAQ, este efecto se debe a que utilizando la metodología SPC se requiere cumplir con la condición de tener una mínima cola, esto conlleva a que la cantidad de asesores requeridos durante los periodos de alta afluencia de llamadas se incremente para cumplir con el requerimiento del modelo, lo cual se puede observar comparando los promedios de arribos de llamadas por hora en cada rango citado, es decir, en los rangos de arribos de 6 a 10 horas el promedio de arribos es de 353, y en el rango de 16 a 23 horas, el promedio de arribos es de 385, sin embargo, para el rango de 10 a 16 horas, el promedio de arribos es de 309.

En el modelo SPC podemos observar que los resultados muestran variabilidad con respecto a los demás resultados, debido a la orientación del modelo de contar con un límite en la cantidad de llamadas que permanecen en el sistema de un intervalo a otro, en este caso al presentarse una variabilidad en la cantidad de arribos en determinados rangos de tiempo, para los rangos de tiempo en que se incremente este promedio de arribos se incrementará también la necesidad de contar con más recursos para atender la demanda en ese periodo de tiempo.

4.5.Evaluación económica

4.5.1. Análisis de los resultados obtenidos del número de asesores

De la tabla 7, podemos apreciar que el modelo BAQ es el que requiere menor cantidad de asesores para un día determinado, por tanto, siendo el más conveniente a utilizar para la determinación de la cantidad de asesores en un *Call center*, materia de evaluación en el presente documento. Sin embargo, para el dimensionamiento de recursos humanos, se requiere evaluar la asignación de dicho requerimiento horario de personal, para un horario regular de 8 horas. En ese sentido, utilizaremos el cuadro de asignación de recursos, donde se establecerá el mejor esquema de asignación horaria respetando el horario de 8 horas y los descansos de 1 hora por día para el refrigerio, para lo cual se aplicará el SOLVER de EXCEL como herramienta de optimización.

Tabla 7

Número de asesores calculado para 01 día de labores, según modelo

Rango hora	BAQ	SPC	ERLANG C
0	1	1	4
1	1	1	2
2	1	1	2
3	1	1	2
4	1	1	2
5	1	1	1
6	1	1	3
7	7	20	11
8	11	18	18
9	21	27	28
10	25	25	37
11	19	13	33
12	18	21	29
13	16	6	32
14	19	5	37
15	19	25	32
16	16	15	30
17	15	17	27
18	16	26	23
19	10	20	16
20	11	18	18
21	10	12	21
22	11	17	19
23	8	18	14
TOTAL DÍA	**259**	**310**	**441**

Fuente: Elaboración propia

En la figura 16 se muestra la asignación de recursos del *Call center* para el resultado de la demanda horaria (de lunes a viernes) del modelo BAQ, requiriendo un plantel de personal ascendente a 57 personas con un horario de 8 horas (ver celda A13)

Para la obtención de dicho resultado, en la columna L (celdas L15:L38) se insertó la demanda calculada en el numeral 4.3. que se muestra en la tabla 10, luego se asignaron en 9 grupos de trabajo (G1 al G9) (ver celdas B13:J13), y un grupo para el horario de amanecida. A partir de ello se realizó la distribución de horarios durante el día (celdas B15:K38), según los requerimientos de la demanda (celdas L15:L38), de tal manera de obtener la menor cantidad de personal excedente en cada franja horaria, cuyo resultado se aprecia en la columna U (celdas U15:U38).

	A	B	C	D	E	F	G	H	I	J	K	L	M	U
12	TOTAL RRHH	GRUPOS DE TRABAJO												
13	57	G1	G2	G3	G4	G5	G6	G7	G8	G9	Aman.	PERSONAS		Verifica cumplimiento demanda
14	Hora	9	4	10	15	5	0	0	0	13	1	Real	Demanda	
15	00 - 01										1	1	1	0
16	01 - 02										1	1	1	0
17	02 - 03										1	1	1	0
18	03 - 04										1	1	1	0
19	04 - 05										1	1	1	0
20	05 - 06										1	1	1	0
21	06 - 07										1	1	1	0
22	07 - 08	1										7	7	0
23	08 - 09	1	1									11	11	0
24	09 - 10	1	1	1								21	21	0
25	10 - 11	1	1	1	1							36	25	11
26	11 - 12	1	1	1	1	1						41	19	22
27	12 - 13				1	1	1					18	18	0
28	13 - 14	1	1	1			1	1				21	16	5
29	14 - 15	1	1	1	1	1	1	1	1			41	19	22
30	15 - 16		1	1	1	1	1	1	1			32	19	13
31	16 - 17			1	1	1	1	1	1			28	16	12
32	17 - 18				1	1	1	1	1	1		31	15	16
33	18 - 19					1	1	1	1	1		16	16	0
34	19 - 20							1	1	1		11	10	1
35	20 - 21								1	1		11	11	0
36	21 - 22									1		11	10	1
37	22 - 23									1		11	11	0
38	23 - 00									1		11	8	3
39											TOTAL		259	106

Figura 16:Asignación de recursos humanos – modelo BAQ (de lunes a viernes)

Fuente: Elaboración propia.

En la figura 17 se muestra la asignación de recursos del *Call center* para el resultado de la demanda horaria (de lunes a viernes) del modelo SPC, requiriendo un plantel de personal ascendente a 75 personas con un horario de 8 horas (ver celda A13)

Para la obtención de dicho resultado, en la columna L (celdas L15:L38) se insertó la demanda calculada en el numeral 4.3. que se muestra en la tabla 10, luego se asignaron en 9 grupos de trabajo (G1 al G9) (ver celdas B13:J13), y un grupo para el horario de amanecida. A partir de ello se realizó la distribución de horarios durante el día (celdas B15:K38), según los requerimientos de la demanda (celdas L15:L38), de tal manera de obtener la menor cantidad de personal excedente en cada franja horaria, cuyo resultado se aprecia en la columna U (celdas U15:U38).

	A	B	C	D	E	F	G	H	I	J	K	L	M	U
12	TOTAL RRHH	GRUPOS DE TRABAJO												
13	75	G1	G2	G3	G4	G5	G6	G7	G8	G9	Aman.	PERSONAS		Verifica cumplimiento demanda
14	Hora	22	0	7	17	6	0	2	0	20	1	Real	Demanda	
15	00 - 01										1	1	1	0
16	01 - 02										1	1	1	0
17	02 - 03										1	1	1	0
18	03 - 04										1	1	1	0
19	04 - 05										1	1	1	0
20	05 - 06										1	1	1	0
21	06 - 07										1	1	1	0
22	07 - 08	1										20	20	0
23	08 - 09	1	1									20	18	2
24	09 - 10	1	1	1								27	27	0
25	10 - 11	1	1	1	1							44	25	19
26	11 - 12	1	1	1	1	1						50	13	37
27	12 - 13				1	1	1					21	21	0
28	13 - 14	1	1	1			1	1				29	6	23
29	14 - 15	1	1	1	1	1	1	1	1			52	5	47
30	15 - 16		1	1	1	1	1	1	1			30	25	5
31	16 - 17			1	1	1	1	1	1			30	15	15
32	17 - 18				1	1	1	1	1	1		43	17	26
33	18 - 19					1	1	1	1	1		26	26	0
34	19 - 20							1	1	1		20	20	0
35	20 - 21								1	1		18	18	0
36	21 - 22									1		18	12	6
37	22 - 23									1		18	17	1
38	23 - 00									1		18	18	0
39											TOTAL		310	181

Figura 17:Asignación de recursos humanos – modelo SPC (de lunes a viernes)

Fuente: Elaboración propia.

En la figura 18 se muestra la asignación de recursos del *Call center* para el resultado de la demanda horaria (de lunes a viernes) del modelo Erlang C, requiriendo un plantel de personal ascendente a 92 personas con un horario de 8 horas (ver celda A13)

Para la obtención de dicho resultado, en la columna L (celdas L15:L38) se insertó la demanda calculada en el numeral 4.3. que se muestra en la tabla 10, luego se asignaron en 9 grupos de trabajo (G1 al G9) (ver celdas B13:J13), y un grupo para el horario de amanecida. A partir de ello se realizó la distribución de horarios durante el día (celdas B15:K38), según los requerimientos de la demanda (celdas L15:L38), de tal manera de obtener la menor cantidad de personal excedente en cada franja horaria, cuyo resultado se aprecia en la columna U (celdas U15:U38).

	A	B	C	D	E	F	G	H	I	J	K	L	M	U
12	TOTAL RRHH	GRUPOS DE TRABAJO												
13	92	G1	G2	G3	G4	G5	G6	G7	G8	G9	Aman.	PERSONAS		Verifica cumplimiento demanda
14	Hora	13	7	10	31	0	4	0	0	23	4	Real	Demanda	
15	00 - 01										1	4	4	0
16	01 - 02										1	4	2	2
17	02 - 03										1	4	2	2
18	03 - 04										1	4	2	2
19	04 - 05										1	4	2	2
20	05 - 06										1	4	1	3
21	06 - 07										1	4	3	1
22	07 - 08	1										11	11	0
23	08 - 09	1	1									18	18	0
24	09 - 10	1	1	1								28	28	0
25	10 - 11	1	1	1	1							59	37	22
26	11 - 12	1	1	1	1	1						59	33	26
27	12 - 13				1	1	1					33	29	4
28	13 - 14	1	1	1			1	1				32	32	0
29	14 - 15	1	1	1	1	1	1	1	1			63	37	26
30	15 - 16		1	1	1	1	1	1	1			50	32	18
31	16 - 17			1	1	1	1	1	1			43	30	13
32	17 - 18				1	1	1	1	1	1		56	27	29
33	18 - 19					1	1	1	1	1		25	23	2
34	19 - 20							1	1	1		21	16	5
35	20 - 21								1	1		21	18	3
36	21 - 22									1		21	21	0
37	22 - 23									1		21	19	2
38	23 - 00									1		21	14	7
39											TOTAL		441	169

Figura 18:Asignación de recursos humanos – modelo Erlang C (de lunes a viernes)

Fuente: Elaboración propia.

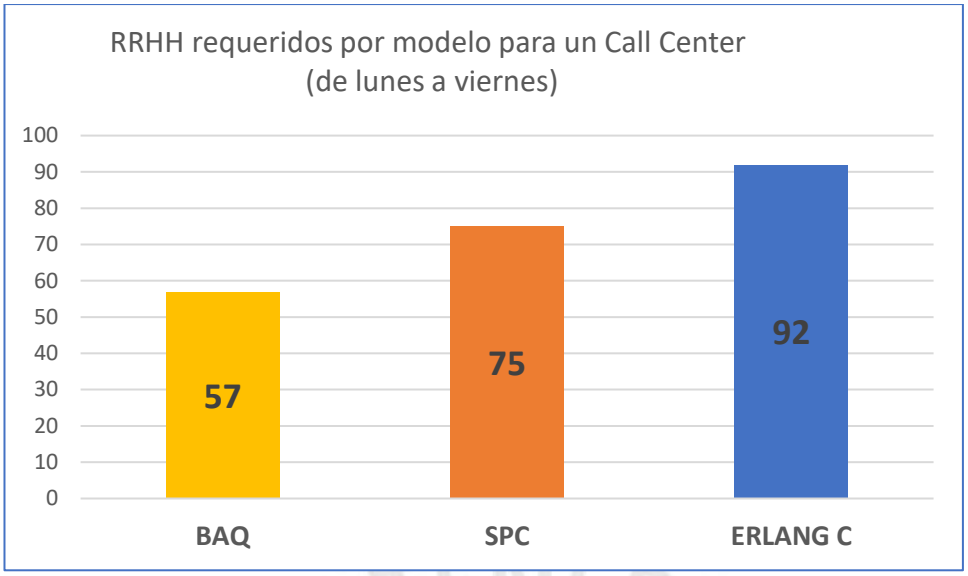


Figura 19:Recursos humanos requeridos para un Call center, por modelo de evaluación (horario de lunes a viernes)

Fuente: Elaboración propia.

Según los resultados obtenidos de la asignación de recursos humanos, considerando los tres modelos de evaluación de la presente tesis, el modelo BAQ arroja el mejor resultado con un grupo de 57 asesores, a diferencia de los otros modelos de evaluación con 75 y 92 asesores respectivamente, tal como se puede apreciar en la Figura 19.

CONCLUSIONES

- En el presente trabajo se propone el uso de una metodología diferente para optimizar el problema del cálculo de la cantidad de asesores requeridos en un *Call center* para la atención de llamadas, minimizando los costos operativos de manera que se garantice un nivel de servicio determinado. A partir de otros dos modelos presentados (BAQ y SPC) se busca obtener un resultado que mejore los cálculos actuales que se realizan con el modelo Erlang C, el cual simplifica muchos aspectos de la realidad. Los modelos analizados toman en cuenta aspectos tales como la consideración del abandono y variabilidad (caso BAQ) y la utilización de los recursos (caso SPC) de tal forma que pueden reflejar y obtener menores resultados que con el empleo único del Erlang C.
- De las restricciones de utilización y factores de variación se observa que el modelo BAQ tiene un mejor desempeño que el modelo SPC, debido a que el tipo de consideraciones empleadas por el modelo BAQ, el abandono y la variabilidad en los tiempos entre llegadas de las llamadas, tiempos de atención y tiempos de espera reflejan la realidad que ocurre en un sistema común de *Call center*.
- El modelo SPC, presenta problemas en operaciones donde exista variabilidad en la tasa de llegada de llamadas debido a que uno de sus parámetros es contar con un límite en la cantidad de llamadas que continúan en atención entre los intervalos de tiempo (en este caso 1 hora), tendiendo a requerir una mayor cantidad de asesores para cubrir la demanda de llamadas y no fallar en el cumplimiento del límite de cantidad de llamadas que pasan de un intervalo a otro.
- La metodología BAQ tiene como objetivo equilibrar el nivel de servicio y su costo operativo relacionado, mientras el modelo SPC se centran en la utilización de los asesores e indirectamente del nivel de servicio, ya que al tener como lineamiento tener

un límite en la cantidad de llamadas que deben seguir en atención entre intervalos de tiempo, indirectamente el modelo se orienta a que tener suficientes asesores para cubrir casi por completo la demanda.

- Los datos obtenidos en el presente trabajo servirán para evaluar y mejorar el sistema existente para el cálculo del requerimiento de asesores en un *Call center*. La operación de cada *Call center* tiene aspectos muy particulares y una metodología BAQ brinda flexibilidad para adaptarse a cualquier escenario, sobre todo con su orientación a la variabilidad en los tiempos entre llegadas, atención y espera.
- Se puede profundizar en el estudio del modelo matemático BAQ aplicando métodos de simulación de datos para la obtención de la cantidad de asesores y así obtener otra aproximación de resultados, de tal manera que nos permita realizar un análisis de sensibilidad de dichos resultados.

BIBLIOGRAFÍA

Aktekin, T., & Ekin, T. (2016). Stochastic Call Center Staffing with Uncertain Arrival, Service and Abandonment Rates: A Bayesian Perspective. *Naval Research Logistics*, 460 - 478.

Aktekin, T., & Soyer, R. (2012). Bayesian Analysis of Queues with Impatient Customers: Applications to Call Centers. *Naval Research Logistics*, 441- 456.

Atlason, J., & Epelman, M. A. (2004). Call Center Staffing with Simulation and Cutting Plane Methods. *Annals of Operations Research*, 333–358.

Atlason, J., Epelman, M., & Henderson, S. (2008). Optimizing Call Center Staffing Using Simulation and Analytic Center Cutting-Plane Methods. *Management Science*, 295–309.

Ausín Olivera, M. (2003). *Análisis Bayesiano De Sistemas De Colas.* Getafe: Universidad Carlos III de Madrid.

Baccelli, F., & Hebuterne, G. (1981). On queues with impatient customers. *INRIA*, 1 -21.

Barrenechea, G. y Gonzáles, G. (2016). *Optimización del número de operadores de un Call Center.* Montevideo: FCEA. Universidad de la República. Disponible en: http://www.iesta.edu.uy/wp-content/uploads/2016/03/Informe-de-Pasant%C3%ADa-Barrenechea-Gonz%C3%A1lez-v.f.pdf.

Bowersox, D., Closs, D., & Cooper, M. (2007). *Administración y Logística en la Cadena de Suministros.* México: McGraw-Hill/Interamericana Editores.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., & Zhao, L. (2005). Statistical analysis of a telephone call center: a queuing science perspective. *Journal of the American Statistical Association*, 36-50.

Castillo, E., Conejo, A., Pedregal, P., García, R., & Alguacil, N. (2002). *Formulación y Resolución de Modelos de Programación Matemática en Ingeniería y Ciencia.* Ciudad Real: Universidad De Castilla-La Mancha.

Feldman, Z., Mandelbaum, A., Massey, W., & Whitt, W. (2008). Staffing of Time-Varying Queues to Achieve Time-Stable Performance. *Management Science*, 324–338.

Fitzsimmons, J., & Fitzsimmons, M. (2008). *Service Management.* Nueva York: Mc Graw Hill.

Fluss, D. (2005). *The Real-Time Contact Center.* Nueva York: American Management Association.

Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management* , 79-141.

Gay, D. (2014). The AMPL Modeling Language: An Aid to Formulating and Solving Optimization Problems. *Numerical Analysis and Optimization*, 95 - 116.

Hillier, F., & Lieberman, G. (2010). *Introducción a la Investigación de Operaciones.* México: McGraw-Hill.

Krajewski, L., Ritzman, L., & Malhotra, M. (2008). *Administración de operaciones.* México: Pearson Educatión.

Mandelbaum, A., & Zeltyn, S. (2005). The Palm/Erlang-A Queue, with Applications to Call Centers. *Faculty of Industrial Engineering & Management Technion*, 1 - 39.

Mandelbaum, A., & Zeltyn, S. (2009). Staffing Many-Server Queues with Impatient Customers: Constraint Satisfaction in Call Centers. *Operations Research*, 1189–1205.

Mehrotra, V., & Fama, J. (2003). Call Center Simulation Modeling: Methods, Challenges, and Opportunities. *Proceedings of the 35th Conference on Winter Simulation*, 135-143.

Robbins, T., Medeiros, D., & Harrison, T. (2010). Evaluating the Erlang C and Erlang A Models for Call Center Modeling. 1- 40.

Rowbotham, F., Galloway, L., & Azhashemi, M. (2007). *Operations Management in Context.* Oxford: Macmillan Company.

Saltzman, R., & Mehrotra, V. (2001). A Call Center Uses Simulation to Drive Strategic Change. *Interfaces*, 87–101.

Sharp, D. (2003). *Call Center Operation: Design, Operation, and Maintenance.* Burlington: Digital Press.

Smith, A., & Gelfand, A. (1992). Bayesian Statistics without Tears: A Sampling-Resampling Perspective. *The American Statistician,*, 84-88.

Taha, H. (2012). *Investigación de operaciones.* México: Pearson Education.

The Leading Contact Centre Magazine. (2016, 08 17). *How to Calculate Contact Centre Shrinkage*. Retrieved from callcentrehelper.com: https://www.callcentrehelper.com/how-to-calculate-contact-centre-shrinkage-90353.htm

The Leading Contact Centre Magazine. (2020). A Beginner's Guide to the Erlang A Formula. *The Leading Contact Centre Magazine*, 1.

Whitt, W. (1999). Dynamic staffng in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 205–212.

Woo Kim, J., & Ho Ha, S. (2010). Consecutive staffing solution using simulation in the contact center. *Industrial Management & Data Systems*, 718-730.

Ysla, B. (08 de Diciembre de 2017). *Apeco: "La nueva ley no nos va a dar un salto en las ventas"*. Obtenido de Semana Económica: https://semanaeconomica.com/sectores-empresas/servicios/257514-la-nueva-ley-no-nos-va-a-dar-un-salto-en-las-ventas

ANEXO

Anexo 1: Teorías relacionadas

Distribución de Poisson

Según India Arroyo, Luis C. Bravo M., Dr. Ret. Nat. Humberto Llinás, Msc. Fabián L. Muñoz (2014), la Distribución de Poisson Llamada así en honor a Simeón-Denis Poisson, quien la describió por primera vez a finales del siglo XIX Dentro de su trabajo: Recherches sur la probabilité des jugements en matièrescriminelles et matièrecivile.

La distribución de Poisson es una distribución de probabilidad discreta que expresa, a partir de una frecuencia de ocurrencia media λ, la probabilidad que ocurra un determinado número de eventos durante un intervalo de tiempo dado o una región específica.

Definición: Sea una variable aleatoria que representa el número de eventos aleatorios independientes que ocurren a una rapidez constante sobre el tiempo o el espacio. Se dice entonces que la variable aleatoria tiene una distribución de Poisson con función de probabilidad:

$$p(k,\lambda) := f(k) = \begin{cases} \dfrac{e^{-\lambda}\lambda^{k}}{k!}, & \text{si k} = 0,1,2,\ldots;\ \lambda > 0; \\ 0, & \text{de otra manera.} \end{cases} \tag{1.13}$$

La función de distribución acumulativa de Poisson, la cual permite determinar la probabilidad de que una variable aleatoria de Poisson sea menor o igual a un valor específico, tiene la siguiente forma:

$$P(X \le k) = F(k,\lambda) = \sum_{i=0}^{k} \frac{e^{-\lambda}\lambda^{i}}{i!} \tag{1.14}$$

Función Gamma $\Gamma(\alpha)$

Función Gamma es una función que extiende el concepto de factorial a los números complejos. Fue presentada, en primera instancia, por Leonard Euler entre los años 1730 y 1731. La función gamma $\Gamma(\alpha)$ se define como:

Sea $\Gamma: (0, \infty) \rightarrow R$, donde

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} \, dx, \text{ para } \alpha > 0 \qquad (1.15)$$

Algunas propiedades adicionales de $\Gamma(\alpha)$ son:

$\Gamma(n)$ = (n-1)!

$\Gamma(n+1)$ = n!, si n es un entero positivo

Distribución Gamma

Se le conoce, también, como una generalización de la distribución exponencial, además de la distribución de Erlang y la distribución Ji-cuadrada. Es una distribución de probabilidad continua adecuada para modelizar el comportamiento de variables aleatorias con asimetría positiva y/o los experimentos en donde está involucrado el tiempo.

Definición: Una variable aleatoria X tiene una distribución gamma si su función de densidad está dada por:

$$f(x,\alpha,\beta) = \begin{cases} \dfrac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, & \text{para } x > 0; \alpha, \beta > 0; \\ 0, & \text{de otra manera.} \end{cases} \qquad (1.16)$$

Distribución de Erlang

Esta distribución fue desarrollada para examinar el número de las llamadas telefónicas que se pudieron efectuar, al mismo tiempo, a los operadores de las estaciones de conmutación. Recibe su nombre en honor al científico danés Agner Krarup Erlang, quien la introdujo por primera vez a principios del año 1900.

La distribución de Erlang sucede cuando el parámetro α, en la distribución gamma, es un entero positivo. Es decir,

$$f(x,\alpha,\beta)=\begin{cases}\dfrac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta}, & \text{para } x>0;\alpha\in Z^{+},\beta>0;\\ 0, & \text{de otra manera.}\end{cases} \tag{1.17}$$

Es la función de densidad de probabilidad para la variable aleatoria X que tiene una distribución de Erlang.

Distribución exponencial

Se ha observado que la distribución gamma, cuando el parámetro toma un valor entero positivo, se conoce como distribución de Erlang. Ahora bien, cuando ese entero positivo es igual a uno, esto es α=1, la distribución de *Erlang* se reduce a la conocida distribución exponencial, siendo así la distribución exponencial un caso especial de la distribución gamma. Debido a lo mencionado anteriormente y al hecho de que ésta distribución se deriva de la distribución de Poisson, su descubrimiento se le atribuye a Agner Krarup Erlang y Siméon-Denis Poisson. La distribución exponencial es utilizada para determinar la probabilidad de que en cierto tiempo suceda un determinado evento.

Definición: Una variable aleatoria tiene una distribución exponencial si su función de densidad está dada por:

$$f(x,\lambda)=\begin{cases}\lambda e^{-\lambda x}, & \text{si } x>0;\ \lambda>0;\\ 0, & \text{de otra manera.}\end{cases} \tag{1.18}$$

Donde λ es igual a $1/\beta$

De este modo, se puede interpretar a la distribución exponencial como el lapso que transcurre hasta el primer evento de Poisson. De hecho, las aplicaciones más relevantes de la distribución exponencial son situaciones en donde se aplica el proceso de Poisson. La relación entre la

distribución exponencial y la distribución de Poisson la podemos observar de la siguiente manera. Recordemos que la distribución de Poisson es una distribución con un solo parámetro λ donde λ representa el número medio de eventos por unidad de tiempo. Consideremos ahora la variable aleatoria *X* descrita por el tiempo que se requiere para que ocurra el primer evento. Haciendo uso de la distribución de Poisson, encontramos que la posibilidad de que no ocurra algún evento, en el periodo hasta el tiempo *t* está dada por:

$$p(0,\lambda t) = P(X=0) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t} \tag{1.19}$$

Podemos ahora utilizar lo anterior y hacer que sea el tiempo para el primer evento de Poisson. La probabilidad de que la duración del tiempo hasta el primer evento exceda *x* es la misma que la probabilidad de que no ocurra algún evento de Poisson en *x*. Esto último, por supuesto, está dado por $e^{-\lambda x}$ como resultado, P (X≥ x) = $e^{-\lambda x}$.

Así la función de distribución acumulada para está dada por:

$$P(0 \le X \le x) = 1 - e^{-\lambda x} \tag{1.20}$$

Ahora bien, a fin de que reconozcamos la presencia de la distribución exponencial, podemos diferenciar la función de distribución acumulada anterior para obtener la función de densidad:

$$f\ (x) = \lambda e^{-\lambda x} \tag{1.21}$$

que es la función de densidad de la distribución exponencial con $\lambda = 1/\beta$.

Método Montecarlo

Según Galán Martín (2017), las técnicas o métodos de Monte Carlo son requeridos en partes muy importantes de muchos problemas científicos. En los últimos años, en el campo de

procesado estadístico de la señal, el procesado de señal bayesiano ha ido ganando popularidad. Este modo de procesado requiere calcular distribuciones de observaciones desconocidas y momentos de ellas. Desgraciadamente, estas distribuciones normalmente son imposibles de ser calculadas analíticamente en la práctica de los problemas, y como alternativa surgen los métodos de Monte Carlo, los cuales aproximan distribuciones objetivo con medidas aleatorias compuestas por muestras que tienen un peso asociado. En muchos de los problemas científicos a los que nos enfrentaremos, nos puede ser necesario el cálculo de una integral con la forma:

$$X_1, X_2, X_3, ..., X_m \sim \pi(x)$$

$$I = \int_D f(x)\pi(x)dx \tag{1.22}$$

donde D es el dominio de π(x) y f(x) es nuestra función objetivo de interés.

Si se pudiesen tomar uniformemente una serie de muestras aleatorias desde la región D que fuesen independientes e idénticamente distribuidas (i.i.d) X1, X2, ..., Xm, una aproximación de I sería:

$$\hat{I}_m = \frac{1}{m}\left[f(x_1) + ... + f(x_m)\right]. \tag{1.23}$$

Por tanto, si pretendemos aproximar la esperanza de la estimación, tendrá la forma:

$$E_\pi\left[\hat{I}_m\right] = E_\pi\left[\frac{1}{m}\sum_{m=1} f(x_m)\right]. \tag{1.24}$$

Haciendo el cálculo, obtenemos que

$$E_\pi \left[\frac{1}{m} \sum_{m=1} f(x_m) \right] = E_\pi \left[f(x_m) \right],$$

(1.25)

Por tanto, se trata de un estimador que será insesgado. En cuanto a la varianza de este estimador, sería de la forma:

$$Var_\pi \left[\hat{I}_m \right] = Var_\pi \left[\frac{1}{m} \sum_{m=1} f(x_m) \right] = \frac{1}{m^2} \sum_{m=1} Var_\pi \left[f(x_m) \right] = \frac{1}{m} Var_\pi \left[f(x_m) \right]$$

(1.26)

y de ello se puede deducir que es de la forma cte/m y es un estimador consistente.

Según el teorema central del límite, la media de un número lo suficientemente grande de variables aleatorias independientes que tengan una media común y una varianza finita tiende a estabilizarse en su media común, es decir, que con probabilidad 1 tenemos que:

$$\lim_{m \to \infty} \hat{I}_m = I$$

(1.27)

Una de las ventajas de este tipo de aproximaciones, frente a otras que puedan darnos a simple vista un error algo menor, es que su error, teóricamente, decae asintóticamente conforme crece m, independientemente de las dimensiones de la región D, situación que provocaba uno de los grandes defectos de los métodos deterministas que se basan en aproximaciones numéricas como por ejemplo las reglas de Simpson o Newton-Cotes, que no eran capaces de escalar correctamente a medida que la dimensión de D crecía, a pesar de que estos métodos para dimensiones pequeñas de la región D nos proporcionaban un error menor que una aproximación de Monte Carlo.

La proporción del error en el esquema de integración de una aproximación de Monte Carlo sigue teniendo los mismos problemas cuando se está trabajando en regiones que tengan un número elevado de dimensiones. A pesar de esto, surgen dos dificultades intrínsecas:

1. Cuando la región sobre la que trabajamos D es grande en un espacio con una gran cantidad de dimensiones, la varianza σ2, la cual nos mide cuanto de uniforme es la función en la región D, puede ser muy grande.

2. Puede ser que no podamos producir muestras aleatorias y uniformes desde π(x).

Para superar estas dificultades, los investigadores normalmente utilizan la idea del muestreo de importancia. Pero en este método se generan muestras aleatorias x1, x2, ..., xm desde una distribución no uniforme a la cual llamaremos q(x) o proposal, que genera mayor masa de probabilidad en las partes "importantes" del espacio de la región D. Nuestra aproximación de la integral I quedaría como:

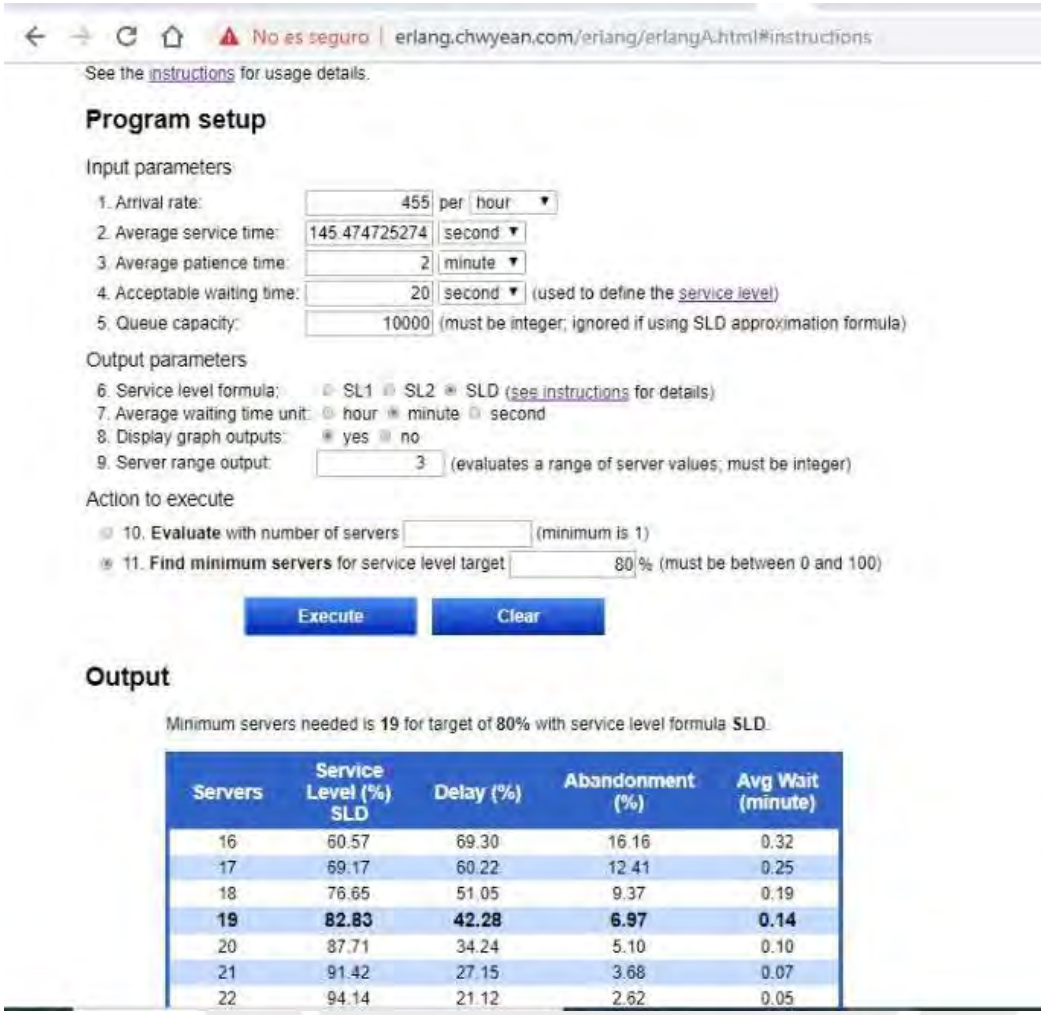
$$\hat{I} = \frac{1}{m} \sum_{j=1}^{m} \frac{f(x^j)}{q(x^j)} \pi(x_j)$$

(1.28)

cuya varianza σ2 π será la varianza tomada desde la proposal q de f(xi)/q(xi).

En el mejor de los casos, se puede escoger una q(x) ∝ π(x) | f(x) | con I siendo finita, lo que da como resultado una estimación exacta de I, pero, por desgracia, no es una situación habitual. De modo más objetivo, podemos esperar una buena función q que pueda explorar más en regiones donde el valor de f sea mayor, y es en esta situación de tener que generar muestras aleatorias desde q donde surgiría el problema.

Anexo 2: Aplicación de la Calculadora Erlang A (modelo BAQ) para la hora 15



No es seguro | erlang.chwyean.com/erlang/erlangA.html#instructions

See the instructions for usage details.

Program setup

Input parameters

1. Arrival rate: 455 per hour
2. Average service time: 145.474725274 second
3. Average patience time: 2 minute
4. Acceptable waiting time: 20 second (used to define the service level)
5. Queue capacity: 10000 (must be integer; ignored if using SLD approximation formula)

Output parameters

6. Service level formula: SL1 SL2 SLD (see instructions for details)
7. Average waiting time unit: hour minute second
8. Display graph outputs: yes no
9. Server range output: 3 (evaluates a range of server values; must be integer)

Action to execute

10. **Evaluate** with number of servers (minimum is 1)
11. **Find minimum servers** for service level target 80 % (must be between 0 and 100)

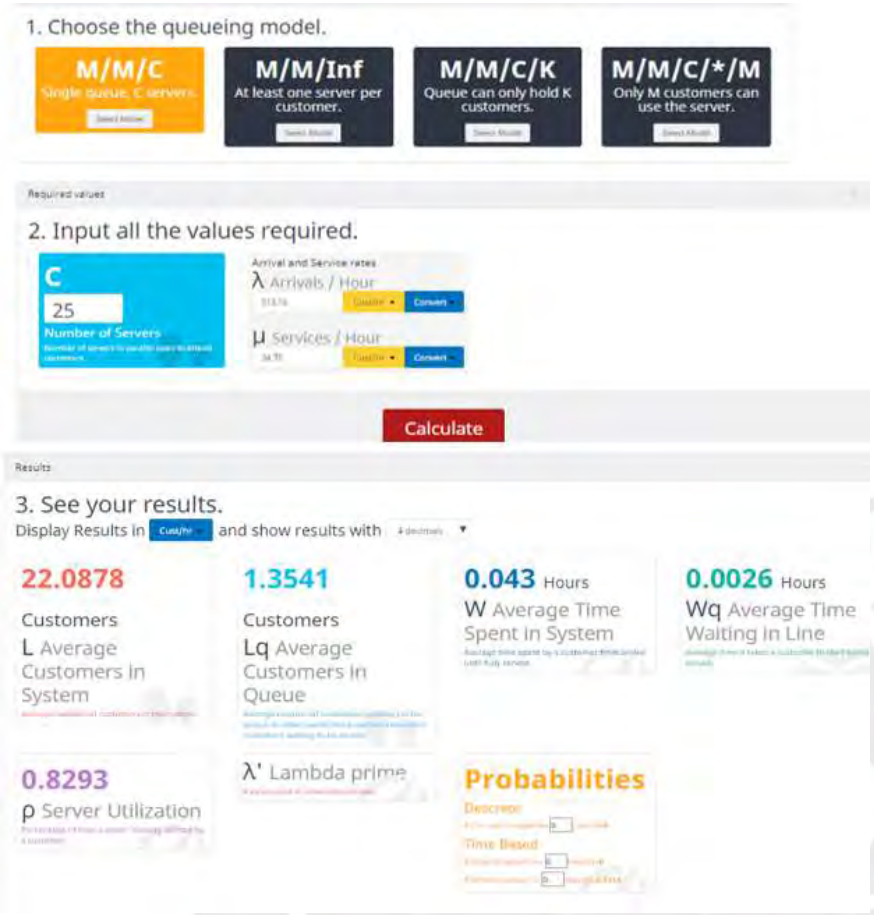
Execute Clear

Output

Minimum servers needed is **19** for target of **80%** with service level formula **SLD**.

Servers	Service Level (%) SLD	Delay (%)	Abandonment (%)	Avg Wait (minute)
16	60.57	69.30	16.16	0.32
17	69.17	60.22	12.41	0.25
18	76.65	51.05	9.37	0.19
19	**82.83**	**42.28**	**6.97**	**0.14**
20	87.71	34.24	5.10	0.10
21	91.42	27.15	3.68	0.07
22	94.14	21.12	2.62	0.05

Anexo 3: Aplicación de la Calculadora Teoría de colas M/M/C (modelo SPC) para la hora 15



Anexo 4:Aplicación de la Calculadora Erlang C (modelo Erlang C) para la hora 15