Modelización, Simulación y Optimización del personal operativo en la administración de Call/Contact Center

Ángel Rubén Barberis[1] y Lorena Del Moral Sachetti[1]

[1] Universidad Nacional de Salta. Facultad de Ciencias Exactas.
Departamento de Informática
{barberis, lorena_dms}@unsa.edu.ar

Resumen. El presente trabajo expone una descripción del proceso de la estimación óptima de los requerimientos de agentes, y se analiza a través de la simulación el problema de la programación de turnos de trabajos, desde la perspectiva de la programación matemática no lineal, para la administración efectiva y eficiente del recurso humano en los Centros de Llamadas o Contactos Telefónicos (Telephone Call/Contact Centers). La simulación que se implementa lleva a cabo dos procesos de optimización: 1) La determinación óptima de los recursos humanos necesarios para la atención de los clientes, que se materializa con la resolución de un modelo de programación lineal. 2) La determinación de la política más adecuada para la asignación de turnos, bajo la restricción de mantener un cierto nivel de servicio. Para éste último, se utiliza el método de las Combinaciones Lineales basadas en función objetivo no lineal y restricciones lineales convexas. El software de simulación que se implementó, está basado en la técnica de simulación de eventos discretos permitiendo un diseño visual de los modelos de análisis. El aporte significativo y novedoso del trabajo, se orienta hacia la utilización de técnicas sencillas de la Programación Matemática No Lineal en la programación de turnos, y el uso de la herramienta de simulación para hacer una predicción precisa de los requerimientos de personal y la asignación de turnos.

Palabras Claves: Optimización de Agentes en Call Center, Programación de Turnos en Centros de Contactos, Métodos de Combinaciones Lineales, Call Center.

1 Introducción

Durante los últimos veinte años en todo el mundo, y especialmente en la última década en Argentina, los servicios de atención telefónica han crecido rápidamente, tanto para el mercado doméstico como para la exportación de esos servicios a otros países. De esta forma, se ha transformado en una de las industrias más activas en la generación de empleo (especialmente de primer empleo), en contraposición con las tradicionales industrias manufactureras que han aumentado la productividad mediante la automatización de la producción, reduciendo la cantidad de empleados por cantidad de productos producidos.
En este contexto, la estimación de la cantidad necesaria de personas, el diseño de los turnos y la asignación del personal, son los principales problemas que enfrenta la

administración de los Centros de Llamadas o Contactos Telefónicos (Call/Contact Centers). El objetivo del administrador de estos Centros de Atención Telefónica es asignar y actualizar dinámicamente el personal para que las llamadas entrantes sean contestadas en el menor tiempo posible, bajo la restricción de ciertos niveles de servicio fijados por la gerencia, las posibilidades ofrecidas por la tecnología disponible, el diseño de los procesos de atención y una amplia variedad de cuestiones humanas relacionadas con el personal asignado. Es por ello que el diseño y dimensionamiento de estos tipos de organizaciones constituyen un área de gran interés, ya que se deben identificar los parámetros adecuados que logren un equilibrio entre la eficiencia operacional (minimización de costos) y la calidad del servicio (accesibilidad a los agentes).

La combinación de los factores costo, calidad y satisfacción no es trivial, por lo que la comunidad científica mundial, en la última década, ha comenzado a estudiar con mayor profundidad diversos modelos matemáticos y probabilísticos para optimizar los recursos, determinar proyecciones de desempeño, y poder conocer así aspectos cuantitativos de los niveles operacionales de la organización. Los modelos que dan soporte a la gerencia operacional son típicamente analíticos y sobre éstos se centra un subconjunto de modelos que, en general, se originan en el área de la Investigación Operativa, y en la Teoría de Colas en particular.

Por otra parte, la simulación de sistemas ha contribuido a desarrollar aún más el estudio de los sistemas de colas, proporcionando soluciones dinámicas para distintos escenarios, contribuyendo en la experimentación y en el diseño de los Call/Contact Centers. El presente trabajo describe la modelización y simulación relacionada con la minimización de costo en la contratación de personal, y la maximización de la calidad del servicio de atención al cliente, materializado en una adecuada programación de turnos, para una mayor accesibilidad a los agentes. La herramienta de simulación desarrollada permite el diseño visual de distintos escenario para el análisis y optimización del recurso humano. La técnica de simulación adoptada en la implementación de la herramienta es la impulsada por eventos, en la que el sistema que se modela, cambia de estado según la ocurrencia de un evento de entrada o salida discreta.

1.1 Estado del Arte

En la literatura actual se pueden encontrar varios trabajos en los que se expone el proceso de dimensionamiento y optimización de Centros de Llamadas Telefónicas [1], [2], [4]. El enfoque paso a paso presentado por Buffa et al. (1976) [4], constituyó la base para muchos estudios relacionados con la administración de personal. Inicialmente se propuso un proceso de 4 pasos: 1) estimación de la tasa de llamada (llamada telefónica entrante) para cada período en que se dividía el día laboral; 2) determinación de los requerimientos mínimos de personal para alcanzar un nivel de servicio especificado para cada período; 3) programación de los turnos diarios; y 4) asignación del personal a los turnos programados. En estudios más recientes, Atlason et al. (2008) [1] ratifican el proceso de dimensionamiento y proponen una metodología general, basada en el método del plano de corte de Kelley Jr. (1960), para optimizar la programación de agentes en un Call Center básico (single skill),

bajo restricciones de cumplimiento de cierto nivel de servicio. Atlason [1] usa el método del plano de corte para resolver un problema de programación matemática lineal (entera) para encontrar la cantidad de personal, luego utiliza esta información como entrada en un modelo de simulación, para estimar así el nivel de servicio. Si el nivel de servicio no es el deseado, incorpora nuevas restricciones al problema de programación lineal y vuelve a iterar. De esta manera combina la simulación con técnicas de la Investigación de Operaciones para resolver el problema de la asignación de recursos.

Recientemente, el avance de la tecnología permitió a los Centros de Llamadas Telefónicas convertirse en Centros de Contactos Telefónicos al expandir sus servicios de simples llamadas telefónicas a intercambios multimediales, envío y recepción de fax, correo electrónico, etc., incorporando adicionalmente la administración de múltiples habilidades (multi-skill) de los agentes que atienden estos contactos. Como consecuencia de esto último, algunos investigadores trabajan en el modelado y optimización de Centros de Contactos con multi-skills. Cezik y L'Ecuyer (2008) [5] describen una generalización del método de Atlason et al. [1] en el contexto de los Centros de Contacto con multi-habilidades. En la misma línea de investigación, Bhulai et al. (2008) [2] y Pot et al. (2008) [3] proponen un método de dos pasos para optimizar turnos y personal en el contexto de las multi-habilidades. Bhulai divide el día laboral en períodos de tiempos regulares y determina el nivel de personal óptimo para cada grupo de habilidades y en cada período. En el segundo paso, determina los turnos de tal forma que se satisfaga la optimalidad de los agentes. Por su lado, propone un nuevo método para determinar la cantidad de personal al describir un algoritmo que puede aproximar la carga de trabajo prevista y la capacidad laboral, teniendo en cuenta la aleatoriedad del proceso de llegada de las llamadas.

En cuanto a las técnicas de optimización de Call Centers, en la literatura actual se pueden encontrar las investigaciones de Ger Koole y Erik van der Sluis (2003) [6] en la que abordan la problemática de la programación de turnos, explotando el concepto de la multi-modularidad. Al introducir la multi-modularidad y bajo ciertas condiciones, los autores proponen el uso de un algoritmo de búsqueda local para alcanzar el óptimo global según las restricciones del nivel de servicio. También se pueden encontrar varios enfoques basados en la Programación Lineal (LP) combinados con otras técnicas. Caprara et al. (2003) [7] utilizan Programación Entera, Programación Dinámica, y la heurística para determinar la mejor política de turnos, teniendo en cuenta los días de descanso, para el mínimo grupo de personas en un Call Center de Emergencias. Los trabajos de Pierre L'Ecuyer et al. (2006) [8] y Cezik et al. (2008) [5] hacen uso de la Programación Entera para optimizar el costo de personal, y combinan sus resultados con la simulación para evaluar el nivel de servicio. Alex Fukunaga et al. (2002) [9] hacen uso de técnicas de búsqueda con algoritmos de inteligencia artificial para optimizar la planificación de personal con una aplicación de software.

2 Modelo de Funcionamiento de un Call/Contact Center

Un esquema de funcionamiento de un Centro de Llamadas Telefónicas puede observarse en la figura 1. Los clientes que llaman e ingresan al sistema son puestos en una "cola de espera" de acuerdo al tipo de servicio que se debe brindar, y son atendidos, típicamente, en una modalidad FIFO (First-In, First-Out). La segmentación entre servicios puede ser dada en base al número discado por el cliente o usuario (DNIS, Servicio de Identificación del Número Marcado) o con algún menú de pre-atención (por ejemplo "marque 1 para soporte, marque 2 para ventas"). Si cuando llega una llamada hay un agente libre, la llamada es presentada directamente al agente. Si cuando llega una llamada, todos los agentes se encuentran ocupados, la llamada es encolada. En estos casos, típicamente se proporcionan mensajes y música de espera a los clientes. Si el cliente decide esperar, finalmente la llamada es presentada a un agente. Si el cliente corta, la llamada se considera "abandonada". [13, 25].

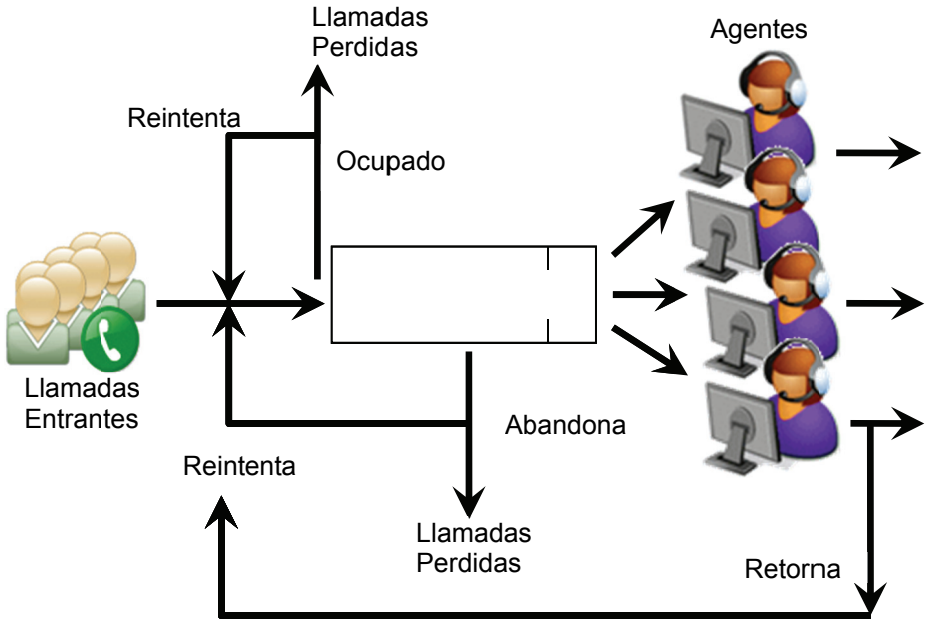

Figura 1. Esquema de Funcionamiento de Centro de Contacto.

Si todas las líneas urbanas del Centro de Contactos Telefónicos se encuentran ocupadas, el cliente recibirá tono de ocupado proporcionado por la red pública. Generalmente se dispone de más líneas urbanas que agentes.

Los clientes que abandonaron la cola de espera, y los que recibieron tono de ocupado de la red pública, es posible que traten de contactarse nuevamente, generando "reintentos". Por otro lado, si la llamada fue atendida por un agente, pero el servicio solicitado no pudo ser completado, es posible que el cliente también intente contactarse nuevamente, generando nuevas llamadas.

De esta manera se configura un ambiente virtual en donde los clientes y los agentes son invisibles entre sí. Los clientes que entraron al sistema, esperan en la cola hasta que sucede una de dos situaciones: un agente es asignado para ser atendido, o bien, llegan a ser impacientes y abandonan la cola. El flujo de eventos que refleja esta situación se muestra en la figura 2.

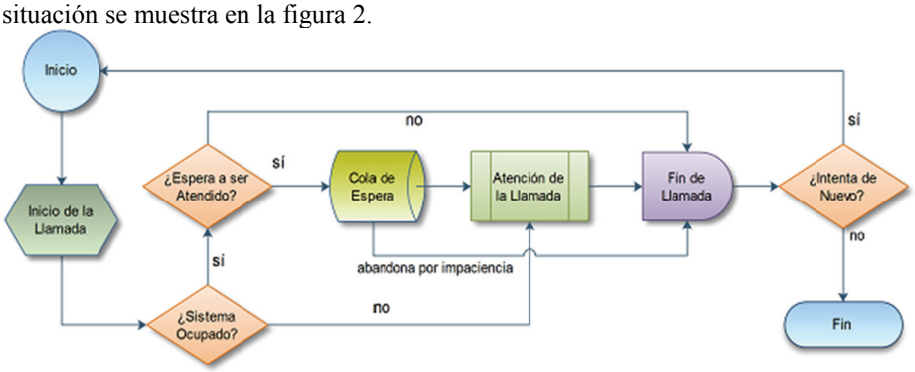

Figura 2: Flujo de eventos de un modelo básico de un Centro de Llamadas

2.1 Proceso General de Optimización

Sobre el esquema de funcionamiento de la figura 1, se desarrolla el proceso de optimización, que consta cuatro etapas a saber (para más detalle mirar [1]):

1. *Pronóstico*: Predicción del volumen de llamadas de clientes (cargas del sistema) que accederán al sistema en cada período de planeación en una jornada. Típicamente, las estimaciones se realizan para una o dos semanas de la planificación. El período de planeación son intervalos de tiempo medidos en minutos en que se subdivide una jornada laboral.
2. *Análisis de Capacidad (Dimensionamiento)*: Se determinan los Requerimientos de agentes necesarios para cada período, asegurando al mismo tiempo, un nivel de servicio de satisfacción al cliente. El servicio se mide típicamente en términos de tiempos de espera del cliente y/o tasas de abandono de la cola. A esta etapa se la conoce como Modelo de Scheduling.
3. *Construcción/Programación de Turnos*: La programación de turnos consiste en el la selección de la política más adecuada de turnos de trabajos que se asignarán al personal, para cubrir los requerimientos necesarios. Este problema, generalmente, se resuelve determinando los turnos tal que se maximicen los niveles de servicios de atención al cliente, lo que se logra mediante la resolución de un problema de optimización por programación entera y/o simulación.
4. *Rostering*: asignación de los empleados a los turnos programados. Contratación y despido (Planificación a largo plazo)

La determinación de los agentes requeridos y su respectiva programación de turnos, constituye dos de las tareas más importante en la administración de recursos humanos en los Call/Contact Center. La estimación del número de agentes requeridos

es un proceso continuo y dinámico que demanda un monitoreo permanente de las circunstancias cambiantes por la que atraviesa diariamente un Centro de Llamadas Telefónicas. Las revisiones del flujo de llamadas entrantes pueden ser hechas en períodos regulares de 30 o 60 minutos durante la jornada laboral. Lo más usual es que se evalúe en cada período de una hora el nivel de llamadas que recibe el Call Center, teniendo en cuenta las variaciones en horas pico, y analizar los factores que causan el colapso, si es que lo hubiere, sobre todos cuando se lanza por algún medio de difusión masiva (por ejemplo, radio, televisión, etc) alguna campaña de mercadotecnia. La información que se necesita conocer al momento de estimar el número de agentes requeridos en un período particular de la jornada laboral son:

1. Número de llamadas recibidas
2. Duración promedio de esas llamadas
3. Tiempo promedio de espera que se estima aceptable para que una llamada entrante permanezca en la cola antes de ser atendida por un agente. Esto es conocido como "el tiempo de espera aceptable" AWT (Acceptable Waiting Time).

Los ítems 1 y 2 permiten determinar el nivel de tráfico de entrada, y el ítem 3 es considerado como una medida de desempeño. Estos tres ítems alimentan a la etapa de *pronóstico*, que a su vez, provee los parámetros de rendimiento estimado a la etapa de *dimensionamiento*.

La programación de turnos para los agentes en un Call/Contact Center constituye la última tarea preponderante de la administración, que pretende operar eficientemente, y mantener al mismo tiempo un cierto nivel de satisfacción del cliente que se comunica telefónicamente. La preocupación de estos Centros de Contacto Telefónico radica en la necesidad de satisfacer adecuadamente la demanda con los recursos disponibles. Los recursos son los agentes para quienes se programan los turnos, teniendo en cuenta el nivel de servicio que se quiere alcanzar como objetivo. El nivel de servicio puede ser definido como el grado de satisfacción de los clientes con el servicio ofrecido. Generalmente el nivel de servicio involucra diversos aspectos como los relacionados con la calidad de las respuestas, el tiempo de espera de las personas, entre otros. Algunos son difíciles de cuantificar y otros mucho más fáciles de determinar.

Las medidas de rendimientos comúnmente usadas como nivel de servicio son el *Factor de Servicios Telefónico* (TSF) y la velocidad media de respuesta (ASA - Average Speed of Answer). El TSF es el porcentaje de llamadas que son atendidas en un tiempo menor al tiempo de espera prefijado (AWT). El ASA es usada como métrica del tiempo de espera [26], también conocida como *Tiempo Medio de Espera*.

El cálculo del TSF como el ASA depende del modelo de cola que se esté usando para el Centro de Llamadas o Contacto Telefónico. Los modelos de Cola más usados son los conocidos como Erlang-C y Erlang-A. El indicador del nivel de servicio usado en éste trabajo es el dado por el Factor de Servicios Telefónicos, basado en modelo de Erlang-C.

2.2 Modelo de Erlang-C (M/M/k)

El modelo de cola más simple y ampliamente utilizado en la administración Call/Contact Centers es el sistema M/M/n, también llamado modelo de Erlang-C. Dada la tasa de arribo λ, la duración promedio del servicio μ^{-1} y k servidores trabajando en paralelo, la fórmula de Erlang-C, describe en forma teórica la fracción duradera del tiempo en que todos los k servidores están ocupados, o dicho de otra manera, la fracción del tiempo en que el cliente espera en la cola antes de ser atendido por un agente. El modelo de Erlang-C es muy restrictivo. Asume, entre otras cosas, recursos infinitos (cola de espera infinita y paciencia del cliente infinita), un ambiente en estado estacionario, en el cual las llegadas se conforman según un proceso de Poisson, la duración de los servicios es exponencialmente distribuida, los clientes y los servidores son estadísticamente idénticos y actúan independientemente el uno del otro. No reconoce, entre otras cosas, el comportamiento de abandono de los clientes, parámetros dependientes del tiempo, y la heterogeneidad de los clientes, cuando el sistema real si los soporta. Desde el punto de visto estadístico, el modelo de Erlang-C mide la probabilidad de que en estado estable todos los operarios estén ocupados cuando una llamada telefónica ingresa al sistema, entonces haciendo $r = \lambda/\mu$ se tiene que,

$$ErlangC(k,r) = P(W > 0) = 1 - \frac{\sum_{m=0}^{k-1} \frac{r^m}{m!}}{\sum_{m=0}^{k-1} \frac{r^m}{m!} + \frac{r^k}{k!} \cdot \frac{1}{1-\frac{r}{k}}} \qquad (1)$$

El TSF bajo el modelo de Erlang-C, que mide la probabilidad de que una llamada se encuentre en la cola de espera un tiempo menor a un cierto AWT fijado por la administración, será:

$$TSF(k,r,AWT) = P(W \leq AWT) = 1 - ErlangC(k,r) \cdot e^{-\mu(k-r)\,AWT} \qquad (2)$$

2.2 Modelo de Erlang-A (M/M/k+M)

Los modelos que no tienen en cuenta el fenómeno de abandono distorsionan las medidas de rendimientos estadísticos, y por ende, no pueden proporcionar información adecuada sobre la Carga del Sistema, que es muy importante para los administradores de los Centros de Contacto. Ignorar las estadísticas de dichos abandonos puede causar una sub o sobre dimensionamiento en la contratación del personal. Por una parte, si el nivel de servicio se mide solamente para aquellos clientes que fueron atendidos, el resultado sería injustamente optimista, el efecto de un abandono es de menor paciencia para los que están más atrás en la cola, así como para las llamadas futuras. Esta situación llevaría a la falta de personal. Por otra parte, usar las herramientas de gestión de la mano de obra que no tienen en cuenta el abandono daría lugar a una sobre contratación de personal, cuando realmente pocos agentes son necesarios para resolver la mayoría de los objetivos de servicio.

El Erlang-A es un modelo que contempla adecuadamente los indicadores de abandono en las llamadas. El modelo básico parte del Erlang-C y le asocia a cada llamada que ingresa al sistema un tiempo de paciencia distribuida exponencialmente con media θ^{-1}. Cuando el cliente ingresa al sistema se encuentra con un tiempo de espera ofrecido AWT. Si el tiempo ofrecido excede el tiempo de paciencia del cliente, entonces la llamada se pierde al abandonar el sistema, o bien, espera pacientemente hasta recibir servicio. El parámetro de paciencia θ no es más que la tasa de abandono individual. La teoría de cola identifica al modelo como una cola tipo M/M/n+M para referirse al Modelo de Erlang-A (A en alusión al Abandono), en donde el +M se refiere a la distribución exponencial del tiempo de paciencia. Básicamente el Erlang-A combina las principales características de los Modelos de Erlang-C y Erlang-B. Véase [27].

Para determinar la probabilidad de espera bajo el modelo de Erlang-A, es necesario expresar la fórmula de Erlang-B, que permite determinar cantidad líneas troncales que necesitará una central telefónica.

$$ErlangB(r,n) = \frac{\frac{r^n}{n!}}{\sum_{i=0}^{n} \frac{r^i}{i!}} \; ; donde \; r = \frac{\lambda}{\mu} \qquad (3)$$

En la expresión (3), el parámetro *r* es el tráfico ofrecido y, *n* la cantidad de servidores requeridos. Esta expresión calcula la probabilidad de que la nueva llamada que se produce en el sistema sea bloqueada, que es igual a la probabilidad de que ninguno de los *n* servidores esté libre.

Teniendo en cuenta la expresión (3), la probabilidad de que una llamada entrante a todos los agentes ocupado y tenga que esperar es:

$$P(W > 0) = \frac{A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) \cdot ErlangB(r,n)}{1 + \left[A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) - 1\right] \cdot ErlangB(r,n)}, \quad donde \; r = \frac{\lambda}{\mu} \qquad (4)$$

La probabilidad de abandono representa la fracción de los que abandonan el servicio. Esta probabilidad se calcula como un simple producto entre la probabilidad de abandonar dado que hay demora, por la probabilidad de demora en el sistema. Esto es:

$$P\{Ab\} = P(Ab \mid W > 0) \cdot P(W > 0) \qquad (5)$$

3 Simulación y Optimización

El uso de la simulación de modelos sistémicos y/o matemáticos como instrumentos de evaluación de alternativas está teniendo una importancia cada vez mayor en el terreno de la evaluación económica de servicios y tecnologías de los Centros de Llamadas o Contacto Telefónico, con un papel cada vez más relevante como ayuda en la toma de decisiones.

3.1 Estructura del Simulador

Con el objeto de investigar y analizar diversos escenarios con distintos parámetros de exigencias, se ha desarrollado un software de simulación de Centros de Llamadas o Contactos Telefónicos, programado en Borland Delphi. El software está basado en el modelo de los sistemas de colas en el que se integran los sub modelos de reintento de llamadas, estimación de agentes necesarios por períodos, optimización de agentes a contratar por jornada (optimización mediante programación matemática lineal), y la optimización de la grilla de turnos (optimización mediante programación matemática no lineal). Todos estos módulos se combinan y se sincronizan para llevar a cabo una simulación y optimización dinámica que le permite a la herramienta hacer una predicción óptima de los recursos en función de la carga de sistema.

El modelo básico que se buscó implementar, es el que se muestra en la figura 3.

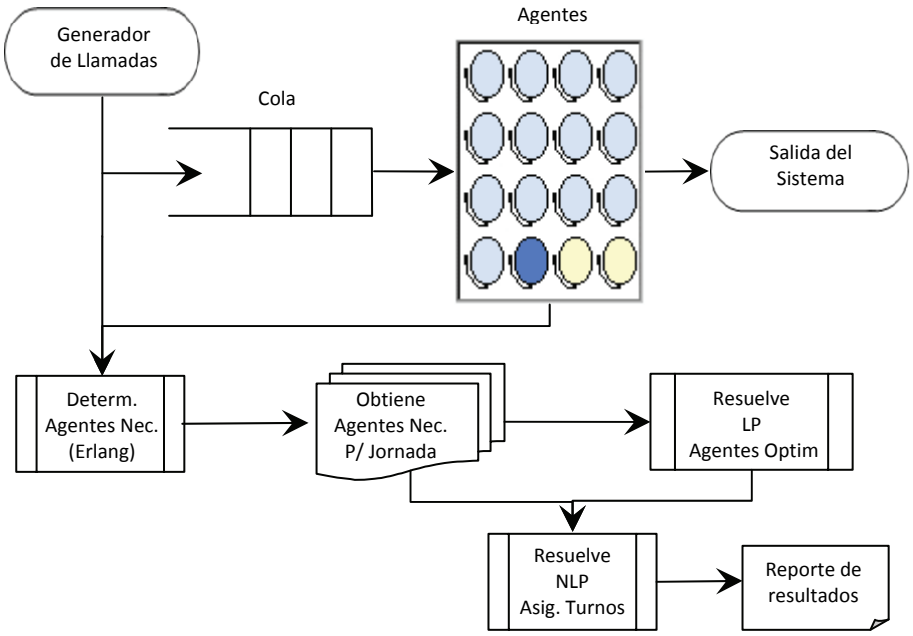

Figura 3. Esquema Básico del Simulador

El simulador implementado está basado en la técnica de la simulación orientada a objetos e impulsada por eventos discretos, bajo la supervisión de un Despachador. Básicamente el simulador, consta de dos módulos principales: el Despachador y el kernel. El Despachador es un objeto que controla el reloj de la simulación, como así también la sincronización de la producción de eventos que se realizan durante la actividad de kernel. También provee otros objetos que le sirven para llevar un control y sincronización adecuada de la Simulación.

La simulación discreta se ocupa de los sistemas cuya dinámica se puede considerar (debido al nivel de abstracción) como secuencia de evento en los puntos

del tiempo discreto. El punto clave de un lenguaje de simulación discreta es la manera en que controla la secuencia apropiada de actividades en el modelo.

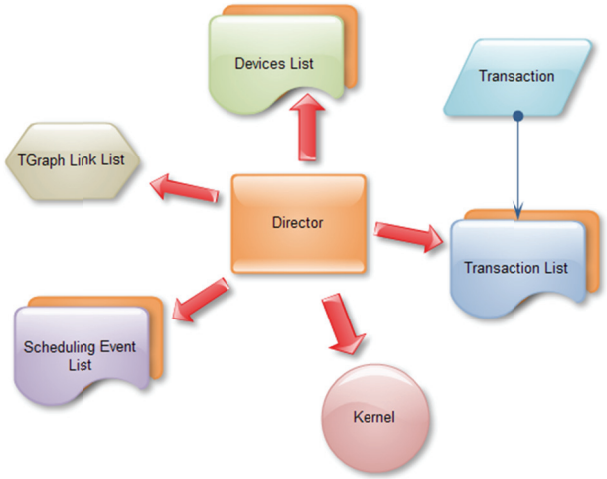

Figura 4. Esquema del Despachador

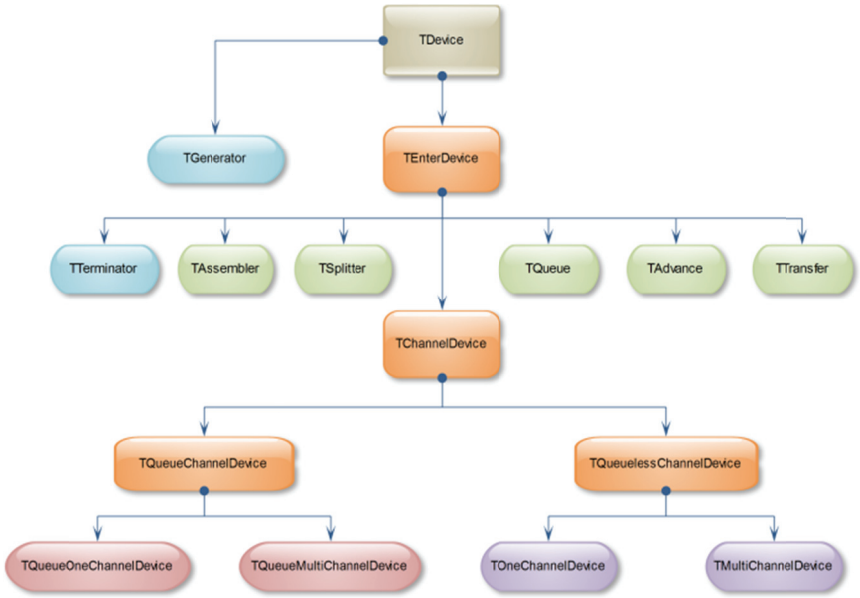

Figura 5. Estructura de objetos del kernel

La "*Simulación impulsada por eventos*" se caracteriza por la programación y cancelación directa de eventos futuros. El enfoque es muy general. Los usuarios miran la dinámica del sistema simulado como una secuencia de eventos relativamente independientes. Cada evento puede ser programado o cancelado, y puede servir para programar una operación en un instante concreto, e incluso para programar el final de la simulación. El kernel del simulador (rutina principal del sistema) debe llevar

registro de los eventos programados. Esto se debe a que cada evento es representado por un objeto llamado "Transacción"; en la que cada transacción contiene la hora de creación, el tipo de evento y otros datos requeridos por el usuario. El kernel mantiene a las transacciones en una pila, ordenadas por hora de programación del evento. Después de la terminación de una rutina de evento, el sistema quita la transacción con el tiempo más bajo de la pila, actualiza el tiempo del modelo, y comienza nuevamente la rutina correspondiente. Esto se repite hasta que la pila de eventos se vacía o el simulador se detenga por alguna otra razón. Al enfoque basado en expresiones explícitas de eventos en el que el sistema cambia de estado en cantidades numerables de instantes de tiempo, se lo denomina Simulación de Eventos Discretos.

Una vez definido el funcionamiento del kernel, se diseñan los dispositivos que intervendrán en la simulación. Estos dispositivos, no son más que objetos que derivan de una misma clase abstracta TDevice. Luego se especializan de acuerdo a la función específica de cada dispositivo. Esto se muestra en la figura 5.

En todos sus niveles, los dispositivos registran estadísticas de la transacción que procesan. Esto le permite al Despachador generar reportes del estado del sistema simulado.

Es importante destacar una característica esencial en el simulador desarrollado de la que no disponen otras herramientas de simulación de Call Center encontradas en el mercado, y es la habilidad del objeto generador de generar llamadas según un proceso de Poisson y no según una cuantía de Poisson. Como así también, la de permitir el diseño de distribuciones personalizadas bimodales como una variable aleatoria puramente estocástica o controlada. Esta característica posibilita un estudio mucho más amplio y realista de los escenarios de Call Center simulados.

3.2 Optimización de la Cantidad de Agentes a contratar

La herramienta simulará varias secuencias de jornadas laborales, y para cada una de estas jornadas, se resolverá un problema de programación lineal mediante el algoritmo Simplex que determinará la cantidad mínima de agentes que garantice un nivel de servicio preestablecido. Cabe aclarar que el modelo es en realidad un problema de Programación Lineal Entera, con la característica de que la matriz con dimensión $(J\,x\,J-T+1)$ de restricciones es totalmente unimodular[1], por lo tanto, los extremos del poliedro generado son enteros, y el algoritmo del simplex devuelve soluciones óptimas enteras. Para más detalle se recomienda leer [29] y [30]. Básicamente, se resolverá:

$$LP\begin{cases} Minimizar\ Z = \sum_{p=1}^{J-T+1} x_p \\ sujeto\ a \sum_{i=m_p}^{M_p} x_i \geq R_p, \end{cases} \quad (6)$$

[1] Una matriz es totalmente si toda submatriz cuadrada tiene determinante -1, 0 ó 1.

Donde J = Número de períodos, T = Duración de los turnos, R_p = Agentes requeridos en período p.

Los x_p representan el número de agentes que comienzan su turno en el período p. Dichos agentes estarán en servicio en los períodos $p, p + 1, \ldots, p + T - 1$. El último período en el cual pueden comenzar es en $J - T + 1$. El objetivo es minimizar el número de agentes contratados, representados por la función objetivo Z. Debemos garantizar que el número de agentes en el período p sea mayor o igual a los requeridos R_p. En un período típico p están trabajando los contratados en ese período más los contratados en $p - 1$, $p - 2$,… $p - T$. Es decir, el número de agentes en el período p se puede escribir como $\sum_{i=m_p}^{M_p} x_i$, donde $m_p = max(1, p\text{-}T+1)$ y $M_p = mín(p, J\text{-}T+1)$. Luego de resuelto este problema lineal, y recordando que una jornada laboral se divide en períodos regulares de tiempo, y teniendo en cuenta que se dispone de los agentes necesarios período por período para garantizar el nivel de servicio impuesto como objetivo, se obtendrá la cantidad de agentes necesarios para una jornada, tal que, al final de la misma, se logre la meta del nivel de servicio preestablecido.

3.3 Optimización de la Grilla de turnos

Luego de haber concluido con el proceso de optimización lineal de la sección 3.2, para una jornada completa, la función objetivo Z de (6) se iguala al óptimo obtenido, y se convierte en una restricción de igualdad que se adiciona al conjunto de restricciones de desigualdades en (6). Puesto que se estudia un sistema exigente, se buscará para cada período de evaluación, cumplir al menos los requerimientos (nivel de servicio) para dicho período. Por lo tanto, se exigirá para toda la jornada cumplir al menos el nivel de servicio fijado por la administración, lo que se logra maximizando el promedio efectivo ponderado de los TSFs de cada período, o sea maximizando la función objetivo Z en (7):

$$NLP\begin{cases} Maximizar \quad Z = \dfrac{\sum_i \lambda_i \cdot P_i(W \le AWT)}{\sum_i \lambda_i} \\ sujeto\ a \begin{cases} \displaystyle\sum_{i=m_p}^{M_p} x_i \ \ge \ R_p\,, \qquad para\ p = 1..T \ \ y \ \ x_t \ \in \ \mathbb{Z}^+ \\ \displaystyle\sum_{p=1}^{T-L+1} x_p = CantAgNec \end{cases} \end{cases} \tag{7}$$

Donde

$$P_i(W \le AWT) = 1 - ErlangC(k,r) \cdot e^{-k\mu(1-\rho)AWT}\,;$$
$$AWT > 0, \quad r = \frac{\lambda_i}{\mu_i}\,, \quad \rho = \frac{\lambda_i}{k_i\mu_i}$$

Puesto que la función ErlangC(k,r), y por ende, $P_i(W \le AWT)$ es no lineal, se plantea así un problema de Programación Matemática No Lineal con restricciones lineales, cuya solución se obtiene a partir de la aplicación del método de Combinaciones Lineales o Algoritmo de Frank-Wolfe, tal como se describe en [28].

4 Pruebas y Resultados

La herramienta de simulación se puso a prueba con un cierto Call Center de envergadura media, con niveles de rendimientos muy exigentes. Dicho Call Center abre 9 horas al día de 8 a 5 pm, y las tasas de llamadas estimadas por la gerencia en cada hora se pueden ver en la tabla 1 al final de éste párrafo. Se ha observado que la duración de las llamadas se puede ajustar con una distribución exponencial de media 3 minutos; y que existen clientes impacientes, cuya impaciencia se distribuye exponencial con media 1.5 minutos. Se quiere tener un nivel de servicio que garantice que el 95% de los clientes esperen menos de 20 segundos para ser atendidos. Se busca determinar el mínimo número de agentes que se deben programar en cada hora para satisfacer el requerimiento de servicio. Se sabe que los agentes (operarios) trabajan turnos de 4 horas contiguas. Se requiere determinar el número (óptimo) de agentes que el Call Center debe contratar para una jornada, como así también la asignación más adecuada de turnos.

Tabla 1. Cantidad de llamadas estimativas que se prevé recibir período por período para una jornada típica.

Horas	λ
8:00 - 9:00	40
9:00 - 10:00	50
10:00 - 11:00	70
11:00 - 12:00	110
12:00 - 1:00	120
1:00 - 2:00	30
2:00 - 3:00	20
3:00 - 4:00	10
4:00 - 5:00	10

Tabla 2. Primera aproximación del resultado

Horas	λ	μ	$r = \lambda/\mu$	k	ρ	$C(k,r)$	$P(W \leq 20\text{s})$ %
8:00 - 9:00	40	20	2.00	5	0.400	5.97015%	95.72%
9:00 - 10:00	50	20	2.50	6	0.417	4.74448%	96.78%
10:00 -11:00	70	20	3.50	8	0.438	2.98857%	98.19%
11:00 - 12:00	110	20	5.50	10	0.550	6.27879%	96.19%
12:00 - 1:00	120	20	6.00	11	0.545	4.92220%	97.18%
1:00 - 2:00	30	20	1.50	5	0.300	2.01392%	98.63%
2:00 - 3:00	20	20	1.00	4	0.250	2.04082%	98.54%
3:00 - 4:00	10	20	0.50	3	0.167	1.51515%	98.85%
4:00 - 5:00	10	20	0.50	3	0.167	1.51515%	98.85%
						Promedio Efectivo	97.15%

Luego, la grilla de cálculo obtenida manualmente se muestra en tabla 2. La columna *k* de la tabla 2 indica la cantidad de agentes requeridos en cada período de una hora para lograr al menos un nivel de servicio del 95%. Un valor inferior al mostrado en ésta columna, implicaría un nivel de servicio por debajo del 95%. Un valor mayor, implicaría más cantidad de agentes, por encima del necesario. Puesto que se busca minimizar costos, éste valor debería ser el menor valor tal que cumpla el nivel de servicio impuesto por la gerencia. Por lo tanto, los valores de dicha columna se obtienen de una función que recibe como parámetro, entre otros, el nivel de servicio requerido (NS). Esto es, k_i = *Agentes Necesarios*(λ, μ, *NS*, *AWT*), que el simulador obtiene automáticamente. La última columna expresa el nivel de servicio teórico en términos del TSF ($TSF_p = P(W \leq AWT)$) alcanzado para la cantidad de agentes encontrados en la columna *k*. Los valores de las columnas *r* y ρ son inmediatos:

$$r = \frac{\lambda_i}{\mu_i}, \quad \rho = \frac{\lambda_i}{k_i \mu_i}, \qquad para\ i = 1..9$$

La columna C(*r*, *k*) es $P(W>0)$ = Erlang-C(*r*, *k*) de la expresión (1). El promedio efectivo se obtiene como:

$$Prom.Efect = \frac{\sum_i \lambda_i \cdot P_i(W \leq 20s)}{\sum_i \lambda_i} \qquad (7)$$

donde AWT = 20 segundos. Este promedio del 97.15% indica que con los agentes especificados en la columna *k*, se logra un nivel de servicio superior al requerido como mínimo. Pero este valor es muy hipotético, ya que no se sabe cómo se distribuirían los agentes para lograr esa especificación, ni tampoco, el número de agentes necesarios para esa especificación. Teniendo en cuenta el caso de estudio planteado en la tabla 2, se tiene:

Tabla 3. Problema a Optimizar (LP)

	Períodos	**1** 8:00 - 9:00	**2** 9:00 - 10:00	**3** 10:00 -11:00	**4** 11:00 - 12:00	**5** 12:00 - 1:00	**6** 1:00 - 2:00		**Necesarios (Restricciones)**	**Rendimiento Necesario**
	Z →	X_1	$+X_2$	$+X_3$	$+X_4$	$+X_5$	$+X_6$			
1	8:00 - 9:00	X_1						≥	5	95.72%
2	9:00 - 10:00	X_1	$+X_2$					≥	6	96.78%
3	10:00 -11:00	X_1	$+X_2$	$+X_3$				≥	8	98.19%
4	11:00 - 12:00	X_1	$+X_2$	$+X_3$	$+X_4$			≥	10	96.19%
5	12:00 - 1:00		X_2	$+X_3$	$+X_4$	$+X_5$		≥	11	97.18%
6	1:00 - 2:00			X_3	$+X_4$	$+X_5$	$+X_6$	≥	5	98.63%
7	2:00 - 3:00				X_4	$+X_5$	$+X_6$	≥	4	98.54%
8	3:00 - 4:00					X_5	$+X_6$	≥	3	98.85%
9	4:00 - 5:00						$+X_6$	≥	3	98.85%
Promedio efectivo										97.15%

En la tabla 3, se expresa el problema a optimizar en la primera fase. La fila Z, constituye la función objetivo que se debe minimizar, sujeto a las restricciones dadas por los valores de la columna *k* en la tabla 2. Luego de la optimización se obtiene:

Tabla 4. Resultado de la Optimización de Agentes (LP)

Cantidad Óptima de agentes	Asignación de Turnos		Nivel de Servicio Logrado
	Turno	Ingresan	
19	1	**5**	**96.39 %**
	2	**2**	
	3	**1**	
	4	**1**	
	5	**7**	
	6	**3**	

El esquema del problema resuelto se muestra en la tabla 5.

Tabla 5. Resultado de la Optimización de Agentes (LP)

	Períodos	**1** 8:00 - 9:00	**2** 9:00 - 10:00	**3** 10:00 -11:00	**4** 11:00 - 12:00	**5** 12:00 - 1:00	**6** 1:00 - 2:00		**Totales Obtenidos**		**Restricciones**	**Rendimiento Necesario**	**Rendimiento Calculado**
	Z →	5	2	1	1	7	3	=	**19**				
1	8:00 - 9:00	5						=	5	≥	5	95.72%	95.72%
2	9:00 - 10:00	5	2					=	7	≥	6	96.78%	99.06%
3	10:00 -11:00	5	2	1				=	8	≥	8	98.19%	98.19%
4	11:00 - 12:00	5	2	1	1			=	9	≥	10	96.19%	91.20%
5	12:00 - 1:00		2	1	1	7		=	11	≥	11	97.18%	97.18%
6	1:00 - 2:00			1	1	7	3	=	12	≥	5	98.63%	100 %
7	2:00 - 3:00				1	7	3	=	11	≥	4	98.54%	100 %
8	3:00 - 4:00					7	3	=	10	≥	3	98.85%	100 %
9	4:00 - 5:00						3	=	3	≥	3	98.85%	98.85%
Promedio efectivo												97.15%	**96.39%**

Luego de la optimización de los agentes, se logra una cantidad mínima que cumple el objetivo de al menos un 95%, al obtener un 96.39% de nivel de servicio, aunque un 0.76% por debajo del que se esperaba.

Si configuramos el simulador con los datos del caso de estudio, obtendremos el siguiente resultado:

Tabla 6. Resultado de la Optimización de Agentes (LP)

Cantidad Óptima de agentes	Asignación de Turnos		Nivel de Servicio Logrado
	Turno	Ingresan	
19	1	**5**	**98.86 %**
	2	**6**	
	3	**2**	
	4	**2**	
	5	**1**	
	6	**3**	

El esquema del problema resuelto se muestra en la tabla 7.

Tabla 7. Resultado de la Optimización de Agentes (LP)

	Períodos	**1** 8:00 - 9:00	**2** 9:00 - 10:00	**3** 10:00 -11:00	**4** 11:00 - 12:00	**5** 12:00 - 1:00	**6** 1:00 - 2:00		**Totales Obtenidos**		**Restricciones**	**Rendimiento Necesario**	**Rendimiento Calculado**
	Z →	5	6	2	2	1	3	=	**19**				
1	8:00 - 9:00	5						=	5	≥	5	95.72%	95.72%
2	9:00 - 10:00	5	6					=	11	≥	6	96.78%	99.99%
3	10:00 -11:00	5	6	2				=	13	≥	8	98.19%	99.99%
4	11:00 - 12:00	5	6	2	2			=	15	≥	10	96.19%	99.98%
5	12:00 - 1:00		6	2	2	1		=	11	≥	11	97.18%	97.18%
6	1:00 - 2:00			2	2	1	3	=	8	≥	5	98.63%	99.99 %
7	2:00 - 3:00				2	1	3	=	6	≥	4	98.54%	99.96 %
8	3:00 - 4:00					1	3	=	4	≥	3	98.85%	99.88 %
9	4:00 - 5:00						3	=	3	≥	3	98.85%	98.85%
Promedio efectivo												97.15%	**98.86%**

Puede observarse en la tabla 7, que el simulador logra predecir la misma cantidad de agentes óptimos (19 agentes), con una maximización en la programación de turnos, al determinar una política de (5, 6, 2, 2, 1, 3) con la que se logra un nivel de servicio del 98.86%, superior al que se esperaba.

5 Conclusión

La simulación es una herramienta de análisis muy poderosa para Diseñar, Evaluar y Predecir el comportamiento de procesos y sistemas que poseen alto grado de complejidad como son los Centros de Llamadas o Contacto Telefónico. Es por ello que se desarrolló un simulador para poder describir el comportamiento de estos

Centros de Comunicaciones, construir teorías o hipótesis referentes a dichos sistemas, y poder crear modelos que predicen el comportamiento futuro del sistema.

El diseño de Calls Centers modernos constituye un gran desafío, y a la vez, un área de investigación de mucho interés. Por lo que, contar con una herramienta que permita la experimentación con modelos en diversos escenarios, facilita el estudio y la comprensión del comportamiento de estos sistemas.

Es por ello que se ha dedicado un gran esfuerzo en el diseño e implementación de un simulador flexible, y de fácil operación para el usuario. Este software sigue las características de la Simulación de Eventos Discretos, en cuyo kernel las operaciones son impulsadas por eventos.

La innovación introducida en éste simulador permite hacer una evaluación, no solo de modelos característicos de la Teoría de Cola, sino también de las principales medidas de rendimiento de un Centro de Llamadas Telefónicas, con la inclusión de módulos de optimización lineal y no lineal que operan dinámicamente, para inferir una predicción final.

4 Referencias

1. Atlason J., Epelman M. A., Henderson S. G.; "Optimizing Call Center Staffing using Simulation and Analytic Center Cutting Plane Methods". Management Science. Vol. 54, No. 2, pp. 295-309. DOI: 10.1287/mnsc.1070.0774. (2008)
2. Bhulai S., Koole G., Pot A.; "Simple Methods for Shift Scheduling in Multiskill Call Centers". Manufacturing & Service Operations Management. Vol. 10, No. 3, pp. 411-420. DOI: 10.1287/msom.1070.0172. (2008)
3. Pot A., Koole G., Bhulai S.; "A Simple Staffing Method for Multiskill Call Centers". Manufacturing & Service Operations Management. Vol. 10, No. 3, pp. 421–428. DOI: 10.1287/msom.1070.0173. (2008)
4. Buffa E. S., Cosgrove M. J., Luce B. J., "An integrated work shift scheduling system" Decision Sciences, Vol. 7, pp. 620-630. (1976)
5. Cezik M. T., L'Ecuyer P.; "Staffing Multiskill Call Centers via Linear Programming and Simulation". Management Science. Vol. 54. Nro. 2, pp. 310-323. (2008)
6. Koole G., van der Sluis E.; "Optimal shift scheduling with a global service level constraint". IIE Transactions Vol. 35, pp. 1049-1055. (2003)
7. Caprara A., Monaci M., Toth P.; "Models and algorithms for a staff scheduling problem". Mathematical Programming; Vol. 98(1-3), pp. 445-476. (2003)
8. L'Ecuyer P.; "Modeling and Optimization Problems in Contact Centers". QEST, Third International Conference on the Quantitative Evaluation of Systems, pp. 145-156. (2006)
9. Fukunaga A., Hamilton E., Fama J., Andre D., Matan O., Nourbakhsh; "Staff Scheduling for Inbound Call Centers and Customer Contact Centers". AI Magazine. Special issue on selected papers from Innovative Applications of AI 2002, Vol. 23(4) , pp.30-40, Winter 2002. (2002)
10. Brown L., Gans N., Mandelbaum A., Sakov A., Shen H., Zeltyn S., Zhao L.; "Statistical Analysis of a Telephone Call Center: A queueing-science perspective". Journal of the American Statistical Association. Vol. 100. No. 469. (2005)
11. Mandelbaum A., Zeltyn S.; "The Impact of Customers Patience on Delay and Abandonment: Some Empirically-Driven Experiments with the M/M/n+G Queue". OR Spectrum, Vol. 26, pp. 377-411. (2004).

12. Mandelbaum A., Zeltyn S.; “Service Engineering in Action: The Palm/Erlang-A Queue, with Applications to Call Centers”. Advances in Services Innovations, Spath D., Fähnrich, K.-P. Springer-Verlag, pp. 17-48. (2005).
13. Koole G., Mandelbaum A.; “Queueing models of call centers: An introduction”. Annals of Operations Research. Vol. 113, No. 1-4, pp. 41–59. (2002).
14. Erdem A. S., Gedikoglu B.; “A DSS for Shift Design and Workforce Allocation in a Call Center”. Technology Management for the Global Future. PICMET 2006. Vol. 3, pp. 1279-1289. (2006).
15. Avramidis A. N., Chan W., Gendreau M., L’Ecuyer P., Pisacane O.; “Optimizing daily agent scheduling in a multiskill call center”. European Journal of Operational Research. Vol. 200, No. 3, pp. 822-832. (2010).
16. Buist E., Chan W., L'Ecuyer P.; “Speeding up call center simulation and optimization by Markov chain uniformization”. Winter Simulation Conference. Proceedings of the 40th Conference on Winter Simulation, pp. 1652-1660. (2008).
17. Deslauriers A., L'Ecuyer P., Pichitlamken J., Ingolfsson A., Avramidis, A. N.; “Markov chain models of a telephone call center with call blending”. Computers and Operations Research, Vol. 34, No. 6, pp. 1616-1645. (2007).
18. Robbins T. R., Harrison T. P.; “A simulation based scheduling model for call centers with uncertain arrival rates”. Proceedings of the 40th Conference on Winter Simulation. pp. 2884-2890. (2008).
19. Hishinuma C., Kanakubo M., Goto T., "An Agent Scheduling Optimization for Call Centers" APSCC, The 2nd IEEE Asia-Pacific Service Computing Conference (APSCC 2007), pp. 423-430. (2007).
20. Peyravi F., Keshavarzi A., "Agent Based Model for Call Centers Using Knowledge Management", AMS, 2009 Third Asia International Conference on Modeling & Simulation, pp.51-56. (2009).
21. Avramidis A. N., L'Ecuyer P., “Modeling and simulation of call centers”, Proceedings of the 2005 Winter Simulation Conference, pp. 144-152. (2005).
22. L’Ecuyer P., Buist E., “Variance reduction in the simulation of call centers”. Proceedings of the 38th conference on Winter simulation, pp. 604-613. (2006).
23. Kleijnen J. P. C., Beers van W. C. M., Nieuwenhuyse van I., “Constrained optimization in expensive simulation: Novel approach”. European Journal of Operational Research. Vol. 202, No. 1, pp. 164-174. (2010).
24. Garnett O., Mandelbaum A., “An Introduction to Skills-Based Routing and its Operational Complexities”. Manuscrito no publicado, Technion, Haifa, Israel. http://ie.technion.ac.il/serveng/Lectures/SBR.pdf
25. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: Tutorial, review, and research prospects. Manufacturing and Service Operations Management (M&SOM), 5 (2), 79–141. (2003).
26. Koole, G.: Call Center Mathematics: A scientific method for understanding and improving contact centers. eBook. Enero, 2007. http://www.math.vu.nl/~koole/ccmath/book.pdf
27. Mandelbaum A., Zeltyn S.: Service Engineering in Action: The Palm/Erlang-A Queue, with Applications to Call Centers. Advances in Services Innovations, pp. 17-48, Spath D., Fähnrich, K.-P. (Eds.), Springer-Verlag. (2007).
28. Barberis A., Veiga R.: Programación óptima de turnos en un Call Center usando el método de Combinaciones Lineales. AADECA 2010. XXII Congreso Argentino de Control Automático. (2010).
29. Cook W., Cunningham W., Pulleyblank Schrijver A.: Combinatorial Optimization. Cap. 6. John Wiley &.Sons (2000).
30. Schrijver A.: Theory of Linear and Integer Programming. Cap. 18 – 23. John Wiley (1999).
31. Taha, H.: Investigación de Operaciones. Cap. 21. Pp. 761-763. Edición 7ma. Pearson Educación (2004).