Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Modelización, simulación y optimización de Centros de Contactos Telefónicos

Tesis presentada para optar al título de Doctor de la Universidad de Buenos Aires en el Área de Ciencias de la Computación

Lic. Ángel Rubén Barberis

Director de Tesis: Dr. Hugo Scolnik
Consejero de Estudios: Dr. Hugo Scolnik

Lugar de Trabajo:
Dpto. de Informática. Facultad de Ciencias Exactas. UNSa.
Dpto. de Computación. Facultad de Cs. Exactas y Naturales. UBA.

Buenos Aires, 2018

Modelización, simulación y optimización de Centros de Contactos Telefónicos

Resumen

El diseño y dimensionamiento de los Centros de Llamadas/Contactos Telefónicos constituyen un gran desafío para la administración, ya que debe identificar y obtener los parámetros adecuados de rendimientos que logren un equilibrio entre la eficiencia operacional (minimización de costos de personal) y la calidad del servicio (accesibilidad a los agentes). La combinación de los factores de costo, calidad y satisfacción no es trivial, por lo que, muchos investigadores siguen estudiando diversos modelos matemáticos y probabilísticos para optimizar los recursos de personal, determinar proyecciones de desempeño, y poder conocer así, aspectos cuantitativos de los niveles operacionales de la organización.

El presente trabajo, expone la optimización de recursos de personal y la determinación óptima de la grilla de turnos bajo la restricción de lograr el mayor nivel de servicio en la estructura de los Centros de Llamados/Contactos Telefónicos. Esta optimización se lleva a cabo mediante un proceso de dos fases: 1) Se calcula la cantidad óptima de agentes telefónicos sujeta a un nivel de servicio predeterminado, mediante un programa de optimización lineal. 2) Conocido la cantidad óptima de agentes, se realiza un proceso de optimización no lineal entera, con el objeto de determinar la grilla de distribución óptima de turnos, de modo tal que alcance el mayor nivel de servicio por encima del preestablecido. También se realiza una caracterización del problema de optimización, de la que se derivan tres algoritmos novedosos que alcanzan soluciones óptimas capaces de lograr gran precisión y velocidad. Por otro lado, se describe el software de simulación de Centros de Llamadas Telefónicas, que permite el estudio de diferentes escenarios para la optimización con características particulares, que hacen de ésta una excelente herramienta de soporte para la toma de decisiones.

Palabras Claves: Modelización; Simulación; Optimización de Call Center.

Modeling, simulation and optimization of Telephone Contacts Centers

Abstract

The design and dimensioning of Telephone Call/Contacts Centers is a major challenge for the organizational management, as it must identify and obtain appropriate performance parameters that achieve a balance between operational efficiency (minimizing staff costs) and quality service (accessibility to agents).

The combination of the factors of cost, quality and satisfaction is not trivial, so that many researchers continue to explore various mathematical and probabilistic models to optimize staff resources, determine performance projections, and to estimate the organization´s operational levels. The models that support operational management are typically analytical, and are originated in the areas of Operational Research and Queue Theory.

This thesis presents new algorithms for optimizing staff resources and for the optimal determination of the grid shift under the constraint of achieving the highest level of service within the structure of Telephone Call/Contacts Centers. This optimization is performed through a two-step process: 1) the optimal determination of telephone operators is made subject to a predetermined service level management. This is achieved through a linear integer optimization program 2) after the optimal number of agents is computed; an integer nonlinear optimization is performed in order to determine the optimal shifts allocation grid that achieves the highest level of service, above the level pursued by the management. A characterization of the optimization problem is also made through which, three novel algorithms that achieve optimal solutions are presented. On the other hand, simulation software of call centers implementing the new algorithms for optimizing different scenarios has been developed, providing an excellent support tool for decision making.

Agradecimientos

En primer lugar quiero agradecer la gran ayuda recibida de diversas personas de la Universidad Nacional de Salta que, de forma directa e indirectamente, han contribuido a la realización de los estudios de doctorado. De manera especial quiero destacar la gran labor, desinteresada y comprometida del Dr. Hugo Scolnik, cuya disposición a cualquier hora del día, permitió sortear diversos obstáculos hasta la concreción final de la tesis. Por último, y no por ello menos importante, quiero agradecer a mi esposa, Lorena, quién con su paciencia perseverante y consejos me ayudó a superar las dificultades, obrando como un gran soporte anímico, al brindarme opciones fáciles para alcanzar cada meta propuesta hasta la finalización de la tesis. –

Mi sabiduría viene de esta tierra …

Salta

Se dedica esta tesis a…

…mi hija Antonella Geanel

Contenido

Lista de Figuras

Lista de Tablas

Lista de Acrónimos

ACD: Automatic Call Distributor
AHT: Average Handling Time
ANI: Automatic Number Identification
ARIMA: Autoregressive Integrated Moving Average
ASA: Average Speed of Answer
AWT: Acceptable Waiting Time
BLD: Búsqueda Local Directa
C1P: Consecutive Ones Property
CLT: Centros de Llamadas Telefónicas
CCT: Centros de Contactos Telefónicos
CL/CT: Centros de Llamadas/Contactos Telefónicos
COAg: Cantidad Optima de Agentes
CTI: Computer Telephony Integration
DNIS: Dialed Number Identification Service
FCR: First Call Resolution
FIFO: First In, First Out
FSA: Fast Simulated Annealing
GSA: Generalized Simulated Annealing
ILP: Integer Linear Programming
INLP: Integer Non Linear Programming
IVR: Interactive Voice Response
LIFO: Last In, First Out
LP: Linear Programming
NLP: Non Linear Programming
NS: Nivel de Servicio
PABX: Private Automatic Branch eXchange
PASTA: Poisson Arrivals See Time Average
PBX: Private Branch eXchange
PCMO: Problema Combinatorio Multi Objetivo
PLE: Programa Lineal Entero
PNLE: Programa No Lineal Entero
PSTN: Public Switched Telephone Network
SA: Simulated Annealing
TSF: Telephone Service Factor
TU: Totalmente Unimodular
VRU: Voice Response Units

1 Introducción

1.1 Conceptos Iniciales y Presentación del Problema

Los avances tecnológicos de los últimos años y la disminución de los costos en las comunicaciones han permitido a las organizaciones mejorar la calidad de los servicios en la atención al usuario, como así también poder ofrecer una cartera de servicios adicionales que antes simplemente no se podían hacer. La estructura tradicional de las organizaciones que brindan servicios con atención presencial de sus usuarios constituye una gran pérdida de oportunidades que impacta fuertemente en los costos. La falta de espacios adecuados para albergar grandes cantidades de personas y la generación de largas colas con tiempos de espera muy significativos convierten al usuario en una persona con alto grado de insatisfacción, induciéndolo potencialmente, a buscar una entidad alternativa que le brinde un servicio con una mejor y mayor interacción con la organización. Esta pérdida de competitividad agrava las economías de las organizaciones de servicios, ya que se enfrentan a una disminución progresiva de la cartera de usuarios. En consecuencia, las organizaciones recurren a una alternativa tecnológica que les permita optimizar los recursos de manera adecuada, captar mayor cantidad de usuarios sin que éstos requieran la atención presencial (atención virtual), y en definitiva, mejorar la calidad de atención a un bajo costo operativo. A estas alternativas tecnológicas se las conoce como Centros de Llamadas Telefónicas (Telephone Call Centers)[1] o Centros de Contactos Telefónicos (Telephone Contact Centers)[2]. Los Centros de Contactos Telefónicos (CCT) son una evolución tecnológica natural de los Centros de Llamadas Telefónicas (CLT), que permiten el desarrollo de la atención al usuario de manera no presencial. Esto último se da, básicamente, vía telefónica y comunicaciones online a través de internet, lo que permite reducir los costos por no tener que acondicionar grandes espacios para la atención presencial.

Durante más de 20 años la compañía de telecomunicaciones estadounidense AT&T ha dedicado un gran esfuerzo en la investigación para el diseño y administración de los CLT, con una gran producción basada en modelos estocásticos o probabilísticos.

Los primeros CLT se iniciaron con las consultas telefónicas de personas que necesitaban obtener información sobre determinados productos o servicios. Los usuarios dependían básicamente de la disponibilidad de los empleados para atender el teléfono. La evolución de los servicios hizo que esta forma de atención se transformara en equipos con

[1] El **Telephone Call Center** es un centro de llamadas telefónicas; un sistema de administración y gestión que se realiza a través de un solo canal, el telefónico, y cuya principal actividad es la recepción o emisión de información, que se realiza de manera rápida y concisa con una atención exclusiva entre el usuario y el teleoperador. Fuente: http://unitel-tc.com/diferencia-entre-call-center-y-contact-center/.

[2] El **Telephone Contact Center**, además de ser un Call Center, es un centro de comunicación unificada que permite la emisión y recepción de información a través de un sistema multi-canal: llamadas telefónicas, correos electrónicos, faxs y comunicaciones online, incluyendo la mensajería instantánea a través de las redes sociales. Básicamente es un Centro de Administración de las Relaciones con los usuarios, es decir, un medio tecnológico que permite integrar las diferentes áreas de negocio de una organización para recibir y emitir información a todos los usuarios, con el fin de poder categorizarlos, ubicarlos, convencerlos, investigarlos, retenerlos, venderles y fidelizarlos.

dedicación exclusiva para responder llamadas telefónicas, con capacitación provista de manera específica.

El desarrollo de la tecnología y el acceso masivo a las telecomunicaciones generaron en las organizaciones la necesidad de elegir a la comunicación telefónica como una de las principales formas de interacción con sus usuarios. Es así como surgieron los Centros de Llamadas/Contactos Telefónicos (CL/CT).

El desarrollo del servicio y la adopción en forma masiva por parte de las organizaciones impulsaron la adecuación a la nueva forma de gestionar sus servicios, con estructuras especialmente dedicadas a planear y operar CL/CT, actividades que resultan vitales para la gestión. En tal sentido, las exigencias originadas en la ampliación de estos servicios, los requerimientos de inversión y el "know how" necesario para sus operaciones, han determinado que muchas organizaciones se especialicen en proveer servicios basados en la comunicación telefónica. Esto demanda personal altamente especializado, una administración profesionalizada y de una organización estructural de bajo costo que provea servicios de excelencia para un consumidor (usuario) exigente, que pretende una interacción eficiente con su prestadora del servicio. La combinación de los factores costo, calidad y satisfacción no es trivial, ya que muchos científicos siguen estudiando diversos modelos matemáticos y probabilísticos para optimizar recursos de personal en distintas bandas horarias; como así también, el empleo de la simulación como una herramienta para determinar proyecciones de funcionamiento y desempeño, y conocer así aspectos cuantitativos de los niveles operacionales de la organización.

Los Modelos que dan soporte a la toma de decisiones operacionales son típicamente analíticos y complejos, y sobre éstos se centra un subconjunto de modelos de estudios que, en general, se originan en las áreas de la Investigación Operativa y en la Teoría de Colas en particular.

En resumen, para que una organización dedicada al servicio se inserte competitivamente en el medio necesita prestar mucha atención a la calidad del servicio que brinda a sus usuarios, pues este servicio se ha convertido en un factor determinante para las personas al momento de decidirse por un producto o servicio específico. Uno de los aspectos de la calidad del servicio es el tiempo: nadie quiere esperar demasiado para ser atendido por la entidad que le brinda un servicio, y sobre todo si es una emergencia. Por esta razón, las tareas de dimensionamiento de recursos en actividades relacionadas con el servicio al usuario son de vital importancia, pues no sólo exigen cuantificar la mínima cantidad de recursos necesarios para la atención, sino también considerar un determinado nivel de servicio para ello. En este sentido, la presente investigación se orienta hacia la formulación de algoritmos que garanticen tanto el dimensionamiento mínimo de los recursos de personal, como la obtención del máximo nivel de servicio plasmado como el mayor nivel de satisfacción de los usuarios por la atención de los agentes[3] que son distribuidos óptimamente según una grilla de turnos.

[3] Los **agentes** son los operarios telefónicos que atienden las llamadas entrantes originadas por los usuarios del servicio.

1.2 Motivación

La estimación del recurso humano necesario, el diseño de los turnos y la distribución de personal son los principales problemas que enfrenta la administración de los CL/CT. Los objetivos del administrador de estas organizaciones son distribuir y actualizar dinámicamente el personal para que las llamadas entrantes sean contestadas en el menor tiempo posible, bajo la restricción de cumplir con un nivel de servicio fijado por la administración. Dichos objetivos buscan disminuir el impacto de la sobrecarga de llamadas entrantes en un intervalo de observación cualquiera, y ajustar dinámicamente el ritmo de atención de los intervalos siguientes a los niveles de servicio deseados. Es por ello que el diseño y dimensionamiento de estos tipos de organizaciones constituyen un área de gran interés para una organización proveedora de productos y servicios, ya que debe identificar parámetros adecuados de rendimientos que logren un equilibrio entre la eficiencia operacional (minimización de costos de personal) y la calidad del servicio (accesibilidad de los agentes). Con tal motivo, se recurre permanentemente a investigaciones de la literatura en busca de nuevas estrategias de optimización y planificación de los recursos de personal que garanticen el mayor nivel de servicio a un bajo costo operacional.

La investigación realizada propone un enfoque alternativo para dar solución a la optimización y planificación de los recursos de personal. Se encamina sobre un área no explorada por las líneas de investigaciones actuales al incursionar en el mundo de la programación matemática no lineal entera (PNLE), con la posibilidad de encontrar nuevos algoritmos que se adecúen eficientemente para la optimización de funciones objetivos derivadas de la familia de ecuaciones Erlang usadas para los CL/CT. Por otro lado, con el uso de la simulación como herramienta de estudio, se provee: un análisis comparativo de los resultados algorítmicos y el estudio de diversos escenarios basados en colas LIFO (Last In, First Out) y FIFO (First In, First Out) con el objeto de brindar una perspectiva diferente a lo que se viene estudiando en la literatura actual.

1.3 Objetivos Propuestos

Los objetivos propuestos en la presente tesis son: Abrir una nueva línea de investigación orientada a la optimización matemática no lineal entera, desarrollando algoritmos eficientes y robustos que tengan en cuenta la característica de unimodularidad de las restricciones del modelo de CL/CT, en la optimización y programación de turnos bajo el régimen de la calidad y la eficiencia en el nivel de servicio. Como así también, evaluar la performance de esas estructuras organizativas mediante la modelización y

simulación de Sistemas de Colas con disciplinas FIFO y LIFO en ambientes de Single Skill[4] para comparar los resultados entre sí y obtener conclusiones.

1.4 Alcances de la Investigación

La complejidad del problema de planificación del personal depende, por una parte, del tamaño y de la estructura del recurso humano; y, por la otra, de la naturaleza heterogénea de las comunicaciones que se llevan a cabo con el usuario del servicio. En consecuencia, la comunidad científica aborda el problema de optimización de los CL/CT desde diversas perspectivas, proponiendo y desarrollando modelos y soluciones algorítmicas para resolver problemas específicos, ya que no se registran en la literatura modelos matemáticos para los problemas de planificación del personal en general. Por otro lado, no hay soluciones algorítmicas que hayan obtenido resultados óptimos claramente, ni mucho menos hay un consenso en cuanto a la mejor estrategia de implementación. En consecuencia, los resultados de la tesis permiten abrir una nueva línea de investigación orientada al desarrollo de algoritmos basados en la programación matemática no lineal entera, debido a la alta no linealidad de la función objetivo propuesta en el trabajo. La caracterización realizada del modelo de estudio permite el desarrollo de algoritmos altamente precisos (para problemas de mediana envergadura a chico se obtiene soluciones óptimas globales, y en problemas grandes, soluciones muy cercanas a los globales), y muy veloces. La disponibilidad de algoritmos precisos y veloces facilita el desarrollo de nuevos métodos de optimización en ambientes de multi-habilidades y multi-canales, ya que son propicios para tales entornos. Las hipótesis de que los escenarios de CL/CT en las que se implementan colas con disciplina LIFO mejoran las medidas de rendimientos pueden resultar interesantes en algunas implementaciones estructurales, debido al uso de colas virtuales invisibles al usuario. Una de las características de los Centros de Contactos Telefónicos consiste en poder anticipar el requerimiento de la llamada entrante que se encuentra encolada. Entonces es posible agilizar el flujo de llamadas por contestar, dando prioridad a aquellas que impliquen la menor intervención del operador. Obviamente esto penalizaría severamente a aquellas llamadas que necesitan mayor interlocución con el operador telefónico. Bajo la suposición de que la cantidad de este tipo de llamadas es menor, se podría aumentar significativamente el rendimiento de los CL/CT.

[4] Los ambientes de **Single Skill** son aquellos en los que se usa un solo canal de comunicación, el telefónico.

1.5 Contribución de la Investigación

La tesis se encamina sobre un área temática no abordada por las investigaciones desarrolladas hasta el momento, ya que no se registran publicaciones en la literatura actual. Se ha incursionado en el mundo de la programación matemática no lineal entera con el objetivo de encontrar nuevos algoritmos que se adecuen eficientemente a la optimización de modelos de CL/CT, en la optimización del recurso humano operativo y la distribución óptima de agentes en los turnos, buscando siempre el menor costo operativo y la mayor satisfacción del usuario. El proceso de optimización abordado en la tesis, se desglosa en dos fases:

1) Determinación de la cantidad mínima de agentes a programar. Se trata básicamente de un programa de optimización lineal entera (LP).

2) Determinación de la grilla de turnos óptima que maximiza el nivel de servicio. Se trata de un programa de optimización no lineal entera (NLP).

Para la optimización LP de fase 1, la función objetivo es lineal y está limitada por un conjunto de restricciones lineales convexas de desigualdades. La matriz de coeficientes de las restricciones goza de la propiedad de **1**s consecutivos (C1P), y se demuestra que ésta es una matriz Totalmente Unimodular (TU). Esta caracterización hace que el conjunto de las soluciones factibles sea finito y entero, por lo que la aplicación de cualquier algoritmo de optimización lineal dará como resultado un óptimo global.

En el caso de la optimización NLP de fase 2, el modelo de CL/CT describe a una función objetivo altamente no lineal, seudo-cóncava, no suave y sin posibilidad de calcular sus derivadas por su complejidad. Está limitada por un conjunto de restricciones lineales convexas de desigualdades (las mismas que en LP), más una restricción de igualdad. Si se descartara la restricción de igualdad, la matriz de coeficientes de las restantes restricciones gozaría de la propiedad de C1P, por lo que sería una matriz TU. La incorporación de la restricción de igualdad hace que se pierda la propiedad C1P, por lo que podría decirse que el resultado es "*aproximadamente unimodular*". Al mismo tiempo reduce el conjunto de soluciones factibles, y si se considerara adecuadamente dicha restricción, se lograrían algoritmos de gran velocidad en la convergencia y alta precisión en la solución. Con base en lo descripto, se propone tres nuevos algoritmos que resuelven los problemas de optimización no lineal entera sin cálculo de derivadas. Éstos son el algoritmo de Búsqueda Local Directa (BLD), la Adaptación del Método de Powell para la optimización no lineal entera y la optimización basada en la estrategia de Fast Simulated Annealing. Estos algoritmos abordan la optimización como si se tratara de un modelo con restricciones TU y obtienen resultados sorprendentes, sobre todo, cuando la cantidad de variables y restricciones superan las 300. Por otro lado, se desarrolló un software de simulación, con diseño visual y detalles simplificados, que permite realizar optimización on-line para inferir la carga operativa óptima del CL/CT, sin perder de vista el nivel de servicio para el usuario. El simulador incorpora las rutinas de optimización mencionadas y otras características de procesos no considerados en software actuales de optimización en el rubro de los CL/CT. Dichas características hacen atractivo el uso del simulador para

investigaciones y aplicaciones reales en el tema. Los software que simulan modelos de CL/CT suelen utilizar una distribución de Poisson para generar las llamadas esperadas con el propósito de simular un escenario posible y hacer una optimización off-line de los recursos. Actualmente los intervalos de observaciones sólo se tienen en cuenta para el cálculo de las medidas de rendimientos. En consecuencia, no se considera el funcionamiento ni el rendimiento del CL/CT durante el transcurso del período de observación, ya que en algunos casos podría ocurrir una sobrecarga de agentes en un intervalo (personal ocioso) y saturación de tareas por faltante de personal en otros. Si los intervalos de observación son de 30 o 60 minutos, entonces las medidas de rendimientos mostrarán resultados no deseados y poco reales. El simulador desarrollado realiza una simulación de las llamadas producidas dentro del intervalo de observación, siguiendo una función de distribución que puede seleccionarse entre una exponencial, exponencial ajustada, uniforme y permite el diseño personalizado de una función de distribución bimodal, de modo tal que al final del intervalo se logra la cantidad de llamadas esperadas como si se hubiera generado a través de una distribución de Poisson. De esta manera, las medidas de rendimientos se independizan de la longitud (en unidades de tiempo) del intervalo de observación, haciéndolas más reales y creíbles. Por otro lado, los diversos escenarios que se pueden simular incluyen ambientes de simple y multi-canal, con y sin abandonos, y reintentos de llamadas, entre otras complejidades.

1.6 Sumario de los Capítulos Subsiguientes

Esta presentación está organizada en capítulos. El Mundo de los CL/CT del Capítulo 2, describe el universo que concierne a los Centros de Llamadas o Contactos Telefónicos. Dicha descripción incluye antecedentes y estado del arte, el funcionamiento general de los CL/CT y el proceso de optimización básica de sus modelos. El Capítulo 3, expone los modelos de dimensionamiento basados en fórmulas de Erlang y las correspondientes derivaciones de las medidas de rendimientos. El Capítulo 4 expone la modelización y optimización de Call Centers desde la perspectiva de la optimización multi-objetivos y presenta una caracterización del problema. El Capítulo 5 presenta la descripción detallada de las tres propuestas algorítmicas y algunos resultados computacionales. El Capítulo 6 muestra los detalles de implementación del Simulador y algunos resultados experimentales de relevancia. El Capítulo 7 contiene las conclusiones respecto de la investigación llevada a cabo.

2 El Mundo de los Call/Contact Centers

2.1 Antecedentes y Estado del Arte

Los Centros de Llamadas Telefónicas o sus contemporáneos, los Centros de Contactos Telefónicos, son los medios más idóneos para relacionar a las organizaciones que ofrecen productos y/o brindan servicios con sus usuarios. Es por ello, que los CL/CT son un componente importante en cualquier estructura organizacional que involucre a una gran masa de personas.

En materia académica se ha notado un creciente desarrollo científico que, directa o indirectamente ha contribuido con aportes muy significativos en lo que respecta a la teoría de los CL/CT.

Una buena descripción general de la gestión y administración de los CL/CT se puede analizar en Cleveland & Mayben (1997) [1]. Este material describe las características primordiales a tener en cuenta cuando se administra un Centro de Llamadas Telefónicas (CLT). Explica los inconvenientes relacionados con las estimaciones del volumen de llamadas a recibir, las medidas de performance a través del nivel de servicio, y también varios otros aspectos vinculados con la administración de los CLT.

Otro material destacado en la literatura es Stolletz (2003) [2]. Allí, utilizando modelos de colas Markovianas se realiza un análisis de resultados, tanto técnicos como económicos de diferentes diseños CLT. El autor ha demostrado que el rendimiento de un CLT depende de varios parámetros que están relacionados con el perfil del usuario, las características de los agentes y las políticas de enrutamiento de llamadas particulares entre otros aspectos. Por otro lado, realiza una clasificación de los rasgos relevantes de los modelos de colas acordes a los CLT.

El libro de Ger Koole (2013) [3], trata todos los aspectos genéricos de los CL/CT, desde las fórmulas básicas de Erlang hasta temas avanzados, como el ruteo de llamadas basadas en habilidades y ambiente multi-canal.

Por su parte, Mehrotra (1997) [4], expone una introducción a los conceptos básicos relacionados con los CLT, brinda una breve descripción del proceso de planificación del personal y el rol de la simulación en el estudio de los modelos para estas organizaciones.

El tutorial de Gans et al. (2003) [5], ofrece un estudio detallado de esta literatura, aunque se recomienda la lectura de uno más reciente en Aksin et al. (2007) [6] y Mehrotra et al. (2013) [7]. Mandelbaum (2006) [8], realiza una recopilación de trabajos hasta 2004 sobre CLT.

Si bien hay numerosas investigaciones en torno a la teoría de los CL/CT, todavía se trata de una corriente muy joven. Es por ello que aún no hay mucho consenso en cómo abordar la problemática de la optimización de los CL/CTs en general. No obstante ello, existe una coincidencia científica en lo que respecta a la estructura del *proceso de planificación del personal*. El enfoque paso a paso presentado por Buffa et al. (1976) [9], constituye la base para muchos estudios relacionados con la administración y gestión del personal. Estas ideas fueron reforzadas en un artículo presentado por Mehrotra (1997) [4]. A su vez, Mason et al. (1998) [10], describen de manera general el proceso y se centran específicamente en la determinación de los requerimientos de personal y la planificación

de turnos. Más tarde, Grossman (2001) [11] y Stolletz (2003) [2] presentan un esquema clarificado en el que se describe cada etapa que conforma el proceso. A partir de allí, surgieron varios trabajos de investigación que siguen la misma línea de pensamiento, entre los que se puede destacar [7, 12-16], entre otros más recientes.

La planificación del personal y la asignación de turnos de trabajos en general es el proceso mediante el cual se definen horarios de trabajos optimizados, se planifica los requerimientos de personal adecuado, para luego ser asignados a dichos horarios de modo que se puedan satisfacer objetivos o políticas impuestas por la administración [17]. El proceso general se encuentra bien estudiado en la literatura científica actual, por ejemplo se destacan los trabajos [17-22]. Una clasificación completa de las investigaciones relacionadas con el tema se describe en [23].

Hay una considerable literatura en optimización de recursos de personal y planificación de turnos de trabajos que contribuyen significativamente a la elaboración de la teoría general relacionada con los CL/CT. Cabe mencionar a continuación los trabajos más destacados: Alex Fukunaga et al. (2002) [24] quienes hacen uso de técnicas de búsqueda con algoritmos de inteligencia artificial para optimizar la planificación de personal. Caprara et al. (2003) [25] proponen la utilización de Programación Matemática Entera, Programación Dinámica y Heurística, para determinar la mejor política de turnos, teniendo en cuenta los días de descanso en un Call Center de Emergencia. El trabajo de Cezik et al. (2008) [26], aplican la Programación Lineal Entera para optimizar el costo de personal y combinan sus resultados con simulación para evaluar los niveles de servicios. Hishinuma et al. (2007) [27], proponen un método práctico y computación evolucionaria para determinar la planificación óptima. Ingolfsson et al. (2010) [16], combinan Programación Matemática Entera con búsqueda aleatoria como un método para determinar un esquema de turnos de bajo costo considerando las variaciones del nivel de servicio en el tiempo.

Por otra parte, gran variabilidad del volumen de llamadas entrantes y la incertidumbre inherente a sus estimaciones, provocan que los niveles de servicios también varíen en una jornada de observación. Por lo tanto, algunos científicos ponen mayor énfasis en dicha variabilidad a la hora de realizar sus investigaciones. Tal es el caso de Robbins et al. (2010) [28], que proponen formular el problema como un modelo de Programación Estocástica Entera Mixta para determinar la planificación óptima de un CLT, con un nivel de servicio global aceptable. Un estudio más reciente de Mattia et al. (2014) [29], exponen un modelo de optimización robusta que reacciona ante una desviación del nivel de personal planificado y el número de empleados actuales, buscando garantizar el nivel de servicio deseado.

En resumen, las líneas de investigaciones en materia de optimización de CLT están orientadas principalmente a la modelización con herramientas de la Teoría de Colas. Algunos trabajos relevantes se pueden encontrar en [30-36]. Por otro lado, se utilizan técnicas de Programación Matemática Lineal Entera [13, 16, 25, 26, 37-39] que se combinan con otras estrategias, como Cadenas de Markov [40-43], Heurística [14, 27, 44-48], Modelos basados en Agentes con Gestión del Conocimiento [49], el Método de Optimización Variacional para determinar el número de agentes y la cantidad de líneas telefónicas [50], y la simulación en su diversas formas [10, 51-55].

Por último, la literatura de los CL/CT se divide en dos grandes grupos de estudios:
1) Las investigaciones relacionadas con los Call Centers que son ambientes de tipo *Single–Skill*. En estos ambientes, tanto las llamadas entrantes como los servicios de atención son homogéneos y utilizan un solo canal de comunicación, el telefónico. Casi toda la literatura desarrollada inicialmente está basada en este enfoque. [5, 31, 56-60]
2) Las investigaciones relacionadas con los Contact Centers, que son ambientes de tipo *Multi–Skill*. En estos ambientes, las llamadas entrantes son heterogéneas y ruteadas a través de múltiples canales hacia agentes con habilidades múltiples y especializadas. En los últimos años, las investigaciones basadas en Multi-Skill han crecido considerablemente. Buenas referencias sobre el tema se puede encontrar en [13, 14, 26, 45, 47, 61, 62].

Una buena descripción de cada enfoque se lee en Avramidis et al. (2010) [39].

2.2 Funcionamiento de los Centros de Llamadas/Contactos Telefónicos

Los primeros Centros de Llamadas Telefónicas brindaban servicios únicamente a través del teléfono y en forma personalizada. Actualmente, los Centros de Llamadas han evolucionado a complejos sistemas interrelacionados donde se conjugan la atención personalizada con la atención automática, en contactos entrantes (receptivos) y salientes (emisores), y no sólo a través de llamadas telefónicas, sino también a través de otros medios tales como el correo electrónico, la mensajería instantánea, la navegación por Internet y las redes sociales.

2.2.1 Estructura Tecnológica

Un Centro de Llamadas o Contactos Telefónicos es un conjunto formado básicamente por personal, computadoras y equipos de telecomunicaciones, que permiten la prestación de servicios a través del teléfono o la computadora. El entorno de trabajo de un gran centro de llamadas (ver Figura 1) puede ser concebido como una sala sin fin, con numerosos cubículos de espacios abiertos, en los que las personas con audífonos se sientan frente a las terminales del computador, proporcionando a los usuarios teleservicios para personas invisibles.

Los servicios que ofrecen los centros de llamadas son muy variados: Atención al usuario, mesa de ayuda y respuestas de emergencias, entre otros. La organización del trabajo en CL/CT puede variar radicalmente. En algunos casos, no se requieren grandes habilidades para la atención del usuario, en cuyo caso la llamada entrante se transfiere al primer agente disponible. En los entornos que requieren trabajos altamente calificados, cada agente puede estar capacitado para manejar sólo un subconjunto de los tipos de llamadas, por lo tanto se emplean "enrutamientos basados en habilidades" para redireccionar las llamadas a los agentes apropiados.

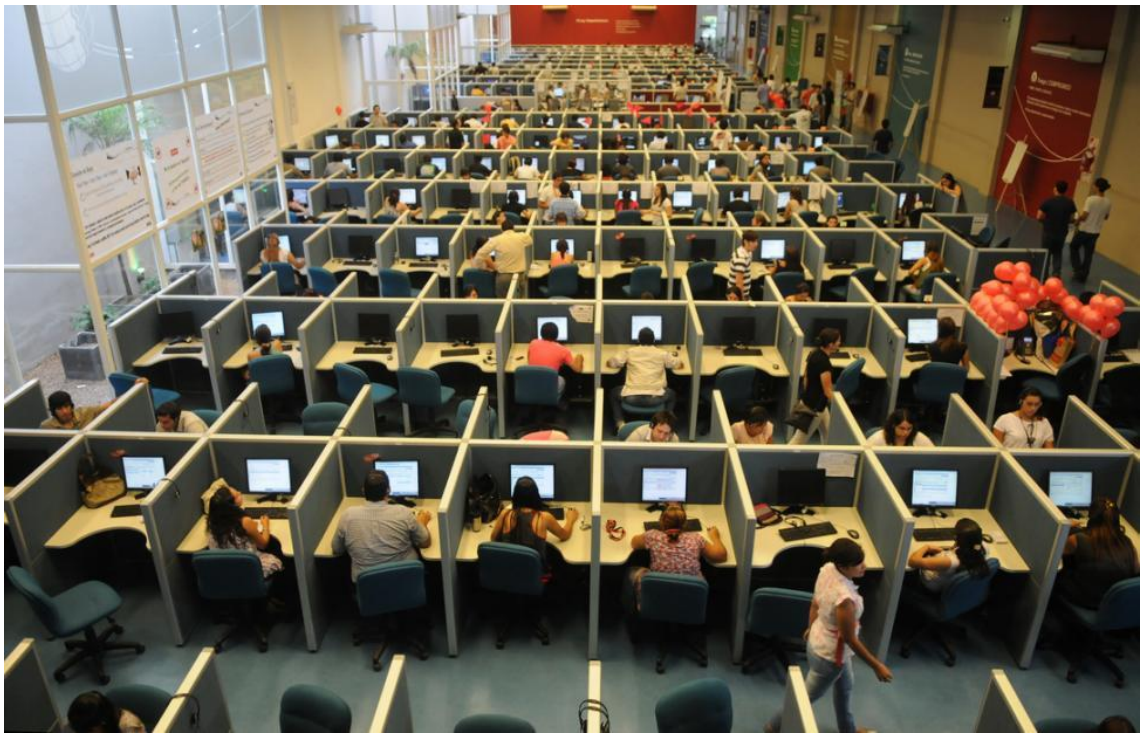

Figura 1: Ambiente de Trabajo en un Call Center[5].

Una característica esencial de los CL/CT es el manejo del tráfico de llamadas entrantes (inbound) o salientes (outbound) [7, 11]. Los CL/CT que administran llamadas entrantes (también llamados Centros Receptivos) son aquellos que enlazan una comunicación desde fuera hacia dentro de los CLT, conectando la llamada de una persona externa con un agente telefónico. Los Centros que administran llamadas salientes (también conocidos como Centros Emisores), son aquellos en donde el agente telefónico realiza la comunicación desde adentro hacia afuera del CLT. Estos tipos de operaciones están tradicionalmente asociados con el marketing y las oportunidades de negocios, por ejemplo, un Centro Receptivo puede iniciar una llamada saliente a un usuario de alto valor empresarial que ha abandonado su llamada antes de ser atendido.

Básicamente (Figura 2), los usuarios acceden a los servicios de los CL/CT a través de la red de telefonía pública denominada PSTN (Public Switched Telephone Network). La labor principal del PSTN consiste en identificar el número de la línea en la que se está produciendo la llamada ANI (Automatic Number Identification) y el número que se está marcando (número destino) conocido como DNIS (Dialed Number Identification Service) de la llamada. Con esta información la compañía telefónica establece una comunicación entre el dispositivo de llamada origen y el destino. Cuando el destino es un CL/CT las llamadas son dirigidas a una unidad PABX (Private Automatic Branch eXchange) o PBX por simplificación de la sigla. El PABX es una central telefónica automática perteneciente a la organización que brinda el servicio. Este dispositivo maneja el tráfico telefónico entrante y saliente de la organización, está conectada directamente a la red de telefonía pública (PSTN), y tiene autonomía sobre cualquier otra central telefónica interna. Hay organizaciones corporativas como Telecom Personal en Argentina que poseen una red de

[5] Foto obtenida de La Gaceta Tucumán. Diario Digital para Internet.

CL/CT ubicados en distintas regiones o provincias del interior del país. Cuando un abonado inicia una llamada al servicio de atención al usuario, el PSTN identifica la ubicación de la llamada y combina el ANI con el DNIS para direccionar la llamada al PABX del CL/CT más cercana a su ubicación. Si el CL/CT está saturado entonces se direccionará al siguiente PBX más cercano disponible.

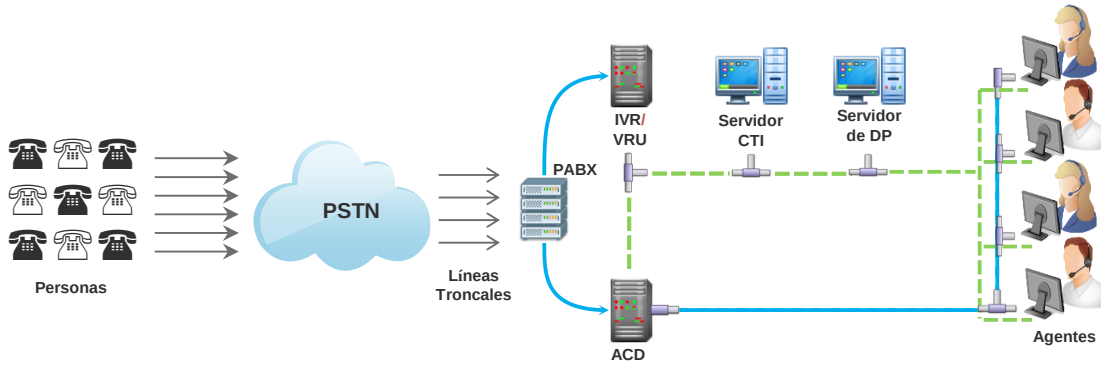

Figura 2: Diagrama del Esquema Tecnológico de un CL/CT.

El enlace entre el PSTN y el PABX se materializa a través de varias líneas telefónicas denominadas *líneas troncales*. La cantidad de líneas troncales disponibles depende de la cantidad de líneas telefónicas contratadas por la entidad prestadora del servicio. Luego de obtenido el DNIS de una llamada en curso, si hay una o más líneas troncales libres entonces el PSTN establece un enlace de comunicación con el PABX. De otro modo, si no hay líneas troncales libres el PSTN dará señal de ocupado al llamador.

Cuando se establece un enlace de comunicación con un PABX, éste puede direccionar la llamada a un dispositivo IVR (Interactive Voice Response), también conocido como VRU (Voice Response Units), o a un distribuidor de llamadas ACD (Automatic Call Distributor). El sistema ACD implementa algoritmos de distribución que aseguran que todos los agentes reciban en promedio el mismo número de llamadas. Los algoritmos de distribución pueden variar según sea el fabricante.

Algunos CLT no poseen dispositivos IVR/VRU por lo que el PABX se conecta directamente al ACD. En la mayoría de los CL/CT, el PABX se conecta directamente al dispositivo IVR/VRU, permitiendo que la persona que inicia la llamada desarrolle su consulta según sus necesidades. Por ejemplo, una persona conectada al sistema podría escuchar: "presione 1" si desea escuchar su saldo; "presione 2" si desea cambiar sus opciones personales, "presione 3" si desea ser atendido por un representante. De esta manera, una persona que realiza una llamada tendrá, básicamente, dos opciones de atención: **1**) interactuar permanentemente con el IVR y completar el servicio sin necesidad de hablar con un agente; o **2**) optar por hablar con un agente telefónico, en cuyo caso, el IVR direccionará la llamada al dispositivo ACD.

El kit tecnológico también incluye a los servidores de integración de telefonía informática CTI (Computer Telephony Integration). Esta tecnología permite sincronizar llamadas telefónicas con aplicaciones informáticas, posibilitando la aparición automática de los datos del usuario en los monitores de las computadoras de los agentes. Los sistemas CTI disponen de funciones adicionales, tales como la identificación de llamadas que permite identificar al usuario iniciador de la comunicación, o el enrutamiento de llamadas,

por ejemplo, en función de criterios geográficos o de negocio, redirigir la llamada a un agente u otro. Ambas funciones facilitan la trazabilidad de los contactos realizados con los usuarios y la automatización de tareas. Finalmente, el combo tecnológico se completa con el uso de *servidores de datos personales*. Muchas organizaciones suelen tener clasificados a sus usuarios según su grado de importancia, por ejemplo los bancos. Cuando se produce una llamada de un usuario preferencial, el agente telefónico o el CTI accede a sistemas de informaciones para resolver de la mejor manera posible la interacción con el usuario. Un detalle más generalizado se puede consultar en [5, 34, 54].

2.2.2 Dinámica de una Llamada Telefónica

En el proceso de llamadas, el universo de usuarios que potencialmente se pondrían en contacto con el CL/CT es mayor a la cantidad de líneas troncales disponibles. Por lo tanto, una llamada entrante podría recibir señal de ocupado si no hubiera una línea troncal disponible para el enlace telefónico y quien intenta comunicarse puede desistir de la llamada o reintentar nuevamente. Cuando una persona desiste de una llamada por tono de ocupado, se considera una llamada perdida [33].

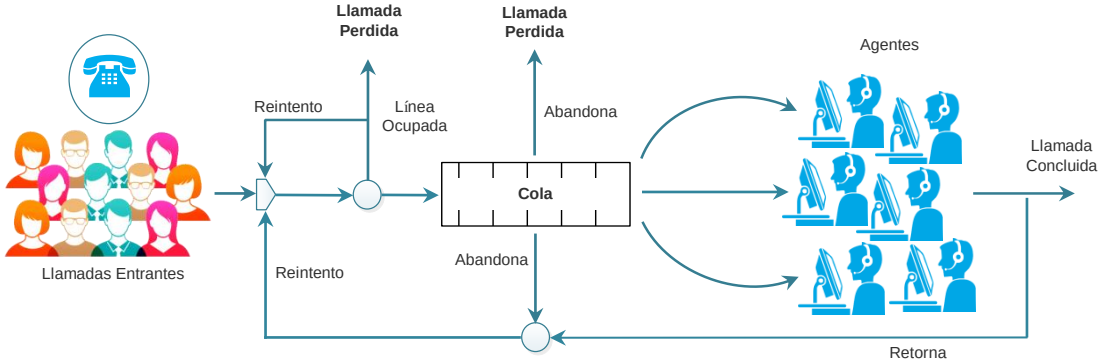

Figura 3: Dinámica del proceso de una llamada telefónica.

Cuando una llamada entrante logra un enlace telefónico, se dice que ha ingresado al sistema, y se la encola según una disciplina FIFO a la espera de ser atendido por un agente. En el momento en que un canal esté disponible para atender un nuevo usuario, la llamada se presenta directamente al agente. Cuando todos los agentes están ocupados, a la persona cuya llamada se encuentra en la cola de espera se le proporciona mensajes de interés o simplemente un audio musical. Si la persona decide esperar, finalmente la llamada será presentada a un agente disponible. Si el tiempo de espera llegara a ser superior al de la paciencia entonces la persona cortará la llamada y abandonará la cola de espera, en cuyo caso se considera una llamada *abandonada*. En muchos casos, una llamada abandonada puede reingresar al sistema, fenómeno que se conoce como reintento de llamada. En otros casos, la llamada que está siendo atendida por un agente pierde el enlace telefónico (se corta la llamada) antes de concluir el servicio de atención. En estos casos, la persona puede dar por concluida la llamada o retornar nuevamente al sistema a través de un reintento de llamada. La dinámica descripta se esquematiza en la Figura 3.

2.2.3 El Interés por el Estudio de los Centros de Llamadas Telefónicas

Los Centros de Llamadas o su evolución, los Centros de Contactos Telefónicos constituyen un componente importante en las organizaciones de servicios, y por lo tanto, muy creciente en América Latina, EE.UU. y en la economía mundial [6]. El principal objetivo de estas organizaciones es brindar un servicio adecuado, efectivo y eficiente a sus usuarios, utilizando la mínima cantidad de recursos posibles. Aún con las modernas tecnologías, el principal costo de un CLT se atribuye al Recurso Humano, lo que representa típicamente más del 50% de los gastos (Figura 4). El personal es el recurso más caro que se gestiona en términos económicos [63], ya que la mano de obra directa representa un costo de entre 60–80% del presupuesto total del funcionamiento de los CL/CT [6, 7]. Por lo tanto la optimización de los recursos y el dimensionamiento adecuado de los CL/CT es de vital importancia para la administración, ya que traen aparejados beneficios, no sólo en término de reducción de costos (eficiencia), sino también en término de mejoras en el nivel de atención al usuario (eficacia). Por lo tanto, la reducción de los costos obtenidos con buenos algoritmos de optimización puede ser sustancial en la gestión de los CL/CT [14].

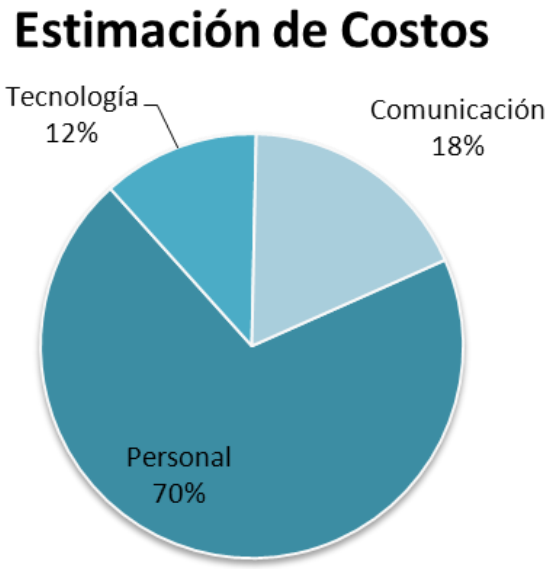

Figura 4: Estimación de Costos Básicos en Centro de Contactos Telefónicos.

El problema de optimización de recursos en los CL/CT con la restricción de alcanzar el mayor nivel de servicio para no perder el usuario que confía en la entidad es muy complejo. El dimensionamiento de los recursos de personal para una jornada laboral depende del volumen de llamadas entrantes que se espera recibir. No hay modelos analíticos que puedan predecir el volumen de llamadas a recibir, por lo tanto, se hacen inferencias basadas en datos históricos con los que se obtienen resultados aproximados. Con estos datos se procede al dimensionamiento del personal y a la determinación de la política de asignación de turnos que maximice la satisfacción del usuario. Esto último tampoco es una tarea sencilla, debido a las grandes y permanentes fluctuaciones de los volúmenes de llamadas durante el día.

En general, la complejidad del problema de planificación del personal depende del tamaño y de la estructura del recurso humano, por lo que la programación y asignación de turnos (horarios de trabajo) es una tarea compleja [64]. La planificación o programación del personal está considerada como un problema NP-Completo, ya que no se puede resolver en un plazo razonable de tiempo con métodos de fuerza bruta [65], mientras que, el proceso de asignación de turnos es considerado como un problema NP-Duro [64, 66]. Por lo tanto, cualquier variación en tamaño y combinación de ambos procesos conlleva a resolver un problema aún mucho más complejo, ya sea como NP-Duro o NP-Completo [66, 67]. Si bien, se han desarrollado modelos y soluciones algorítmicas para resolver problemas específicos, no existen modelos matemáticos en la literatura para los problemas de planificación del personal en general, ni mucho menos para la determinación óptima de la asignación de turnos, en lo que respecta a la teoría de los CL/CT.

2.3 Esquema General del Estudio de Call Centers

El proceso de dimensionamiento y planificación no es una tarea sencilla, sobre todo por las grandes fluctuaciones del volumen de llamadas entrantes y el alto grado de incertidumbre inherente a su predicción. En consecuencia, el proceso general se divide en varias etapas; cada una de ellas conforma toda un área de investigación en sí misma [11]. Varios autores en la literatura desglosan el esquema en 5 actividades [7, 10, 11], mientras que otros lo resumen en 4 etapas [2, 4, 9, 12-16]. A continuación se describe el proceso básico del esquema de 4 etapas:

1. Estimación del volumen de llamadas a recibir (Forecasting).

Esta etapa se compone, básicamente, de dos tareas:

***a*)** *Recopilación de datos históricos*. Es el primer paso del proceso de administración de personal. El objetivo es disponer de una muestra representativa de la información histórica que puede ser analizada para predecir los volúmenes de llamadas a recibir y patrones futuros. Estos datos normalmente son obtenidos del distribuidor automático de llamadas (ACD) y representan el número de llamadas recibidas y la información de los tiempos asociados a éstas para un período representativo. Los datos deben ser cuidadosamente revisados para asegurarse de que no existan errores y evitar que sean no representativos, en cuyo caso, son descartados antes de la previsión.

***b*)** *Estimación del volumen de trabajo*. Es el segundo paso en la planificación de recursos y consiste en la aplicación de modelos de pronósticos sobre la información histórica con el fin de predecir la carga de trabajo futuro. Para cumplir el objetivo se utilizan modelos estadísticos y técnicas de estimación, tales como modelos de regresión, exponenciales o cualquiera de sus variantes y, modelos de análisis de series de tiempos basados en

ARIMA[6]. La jornada laboral se divide en intervalos de tiempos representativos, generalmente entre 15 minutos y una hora. Para cada intervalo de tiempo se predice el volumen de llamadas a recibir, y se realiza una estimación de la performance. Se determinan los niveles de servicios en función del personal estimado para cada intervalo de observación. La cantidad de personal estimado en función del nivel de servicio requerido constituirá la restricción para las etapas siguientes.

En la literatura de los CLT se pueden encontrar varios trabajos relevantes relacionados con esta etapa del proceso [68-75].

2. Planificación del personal necesario.

En el análisis de resultados se evalúan diversas alternativas para determinar el impacto de la dotación de personal en el servicio, los niveles de productividad y costos. Básicamente, el objetivo de esta etapa es calcular, para cada intervalo de observación, la dotación de personal necesaria para atender un volumen de trabajo determinado y alcanzar un nivel de servicio objetivo. Los cálculos se efectúan en función de la carga de trabajo estimado en la etapa 1. Se tienen en cuenta el tiempo de espera de las personas en la cola de llamadas y el porcentaje de abandonos entre otras medidas de rendimientos. Una vez que se tienen las restricciones de personal para el nivel de servicio objetivo, se aplican diversas técnicas de Optimización Lineal Entera para determinar la cantidad de personal a disponer en una jornada, buscando una erogación mínima del costo para la organización.

Hay un número significativo de investigadores que proponen trabajos relacionados con las etapas 2 y 3 al mismo tiempo [2, 10, 13-16, 39, 55].

3. Planificación de turnos de trabajos.

En términos generales, el proceso trata de crear un conjunto de programas de mano de obra que mejor responda a la carga laboral esperada. Las necesidades básicas de personal son calculadas y adaptadas a las restricciones de capacidad de la organización. Se imponen reglas de asignación de turnos y restricciones para el diseño de una grilla horaria.

La determinación de la cantidad mínima de agentes que alcancen el nivel de servicio objetivo no es suficiente. Un turno de trabajo es un conjunto contiguo de intervalos de tiempos representativos en el que se incluye el período de descanso del agente telefónico. Los turnos de trabajos no se inician todos al mismo tiempo, sino que se distribuyen a lo largo de la jornada de atención al usuario. El gran desafío de los administradores es determinar cuántos agentes asignar a cada turno. Una mala distribución de los agentes en los turnos de trabajo puede conllevar al cumplimiento estricto del nivel

[6] En estadística, más precisamente en análisis de series de tiempo, un modelo de media móvil auto-regresivo integrado ARIMA (Autoregressive Integrated Moving Average) es una generalización del modelo ARMA (Autoregressive Moving Average). Estos modelos propuestos por Box y Jenkins en 1968, son aplicados a datos de series de tiempo para interpretar mejor las muestras o para predecir aspectos futuros en la serie (forecasting). Se aplican en algunos casos en donde los datos muestran evidencia de ser no estacionarios, para reducirlos y convertirlos a datos aproximadamente estacionarios. Más detalles en http://www.statsref.com/HTML/index.html?arima.html

de servicio objetivo, y en algunos casos, hasta un nivel inferior a éste. Una distribución adecuada de los agentes puede hacer que se satisfaga un nivel de servicio superior al objetivo. Por lo tanto, la determinación de la mínima cantidad de agentes (menor costo) que logren un nivel de servicio superior al propuesto (eficacia) es el objetivo de los CL/CT. Es por ello, que en esta etapa se aplican diversas técnicas de optimización para la determinación de la mejor política de asignación de turnos. En la literatura hay varios trabajos que abordan esta etapa en conjunción con la anterior. Al solo efecto de mencionar algunos, se recomienda la lectura de las bibliografías [7, 14, 28, 39, 76, 77].

4. Asignación de los agentes a los turnos (rostering).

Una vez que se ha determinado la grilla de turnos y la cantidad óptima de operarios que deben asignarse a cada uno, se procede a la asignación individual de los agentes a los mismos. En términos generales, el paso final en el proceso de administración de personal implica el seguimiento de los resultados reales que se obtienen a partir de la planificación del personal preestablecido y del nivel de servicio requerido. El volumen actual de llamadas, la administración de los tiempos y el personal disponible, se comparan con la previsión realizada en las etapas anteriores y se realizan los ajustes necesarios para cumplir con los niveles de servicio establecidos.

Una administración ineficiente del personal en los CL/CT provoca un mal servicio, usuarios insatisfechos, empleados desequilibrados, y altos costos. El exceso de personal resulta para la organización baja productividad, alta contracción, y agentes sin tareas. Por otra parte, la falta de personal lleva a largos tiempos de espera, un nivel de servicio inconsistente, agentes con exceso de trabajo y una experiencia negativa de los usuarios.

El dimensionamiento del personal en función de los niveles de servicios esperados es una tarea difícil, debido principalmente a la variabilidad y aleatoriedad de las llamadas entrantes, la invisibilidad de las colas de llamadas y la alta expectativa del servicio por parte de los usuarios. Para aproximar una solución al problema, se debe proponer un enfoque sistemático que aborde, de la mejor manera posible, cada una de las etapas del proceso, es decir, obtener datos históricos en calidad y cantidad adecuada, desarrollar y adecuar buenos modelos de pronósticos de la carga laboral, disponer y aplicar algoritmos eficientes y robustos para el cálculo óptimo del personal requerido, imponer buenas políticas de planificación y asignación de turnos, y por último, monitorear regularmente la administración del personal y los niveles de servicios que se están produciendo. Un resumen simplificado del esquema general de proceso se muestra en la Figura 5.

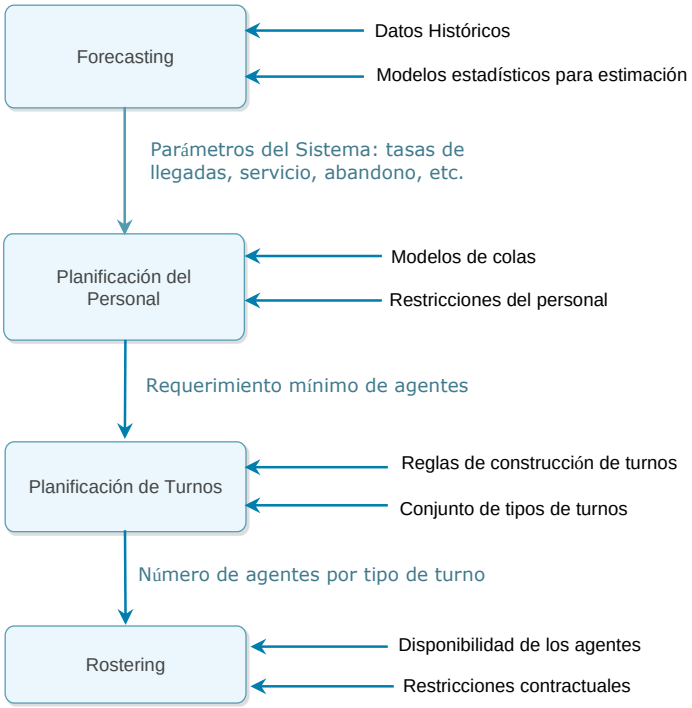

Figura 5: Esquema General del proceso de optimización de Call Centers.

3 Modelos de Call/Contact Centers

La Teoría de Colas, como parte de la *teoría de probabilidades* ha surgido a partir de investigaciones en el campo de la ingeniería de teletráfico clásico en los años 20. El Ingeniero danés, Agner Krarup Erlang (1878-1929)[7], ayudó a la Copenhagen Telephone Exchange Company —CTEC—, entonces subsidiaria de la compañía telefónica Bell, a resolver problemas avanzados de comunicaciones. Erlang dio inicio a la Teoría de Colas, aplicando teorías de probabilidades a los problemas de tráfico telefónico y publicó el primer artículo sobre la nueva área de investigación en 1909 denominado "*La teoría de las probabilidades y las conversaciones telefónicas*"[8] en el cual demostró que la Distribución de Poisson es aplicable al tráfico telefónico aleatorio. Específicamente se preocupó del estudio del problema de dimensionamiento de líneas y centrales de conmutación telefónica para el servicio de llamadas. En 1917 publicó el artículo "*Solución de algunos problemas en la teoría de probabilidades de importancia en centrales telefónicas automáticas*"[9], que contiene su fórmula clásica para el cálculo de pérdidas de llamadas y tiempos de espera. Una de sus preocupaciones consistió en mejorar los tiempos de espera para concretar una llamada telefónica. Él había observado que algunas personas empleaban demasiado tiempo en sus conversaciones (más de una hora), mientras que otras, algunos minutos. Luego de varios años de investigaciones Erlang publicó sus resultados en un compendio titulado "*Telephone Waiting Times*" (Las esperas para hablar por teléfono) en danés [78], trabajo que fue muy utilizado por otras compañías telefónicas, dando así continuidad a la nueva área de investigación llamada "Teoría de Colas".

3.1 Modelos de Dimensionamientos

Un modelo de dimensionamiento es aquel que permite determinar la cantidad de recursos (líneas telefónicas, agentes, etc.) necesarios para cumplir un nivel de servicio deseado [30]. Se cuenta básicamente con tres modelos:

Erlang–B: permite determinar la cantidad de líneas telefónicas necesarias para mantener un grado de servicio objetivo.

Erlang–C: permite determinar la cantidad de agentes necesarios para mantener un nivel de servicio objetivo sin considerar el abandono y re-intento de llamadas. Supone una cola de espera y paciencia infinita.

Erlang–A: permite determinar la cantidad de agentes necesarios para mantener un nivel de servicio objetivo, considerando las llamadas que abandonan la cola por impaciencia. Supone una cola de espera infinita mientras haya paciencia, y no considera los re-intentos de llamadas.

[7] https://plus.maths.org/content/os/issue2/erlang/index

[8] "**The Theory of Probabilities and Telephone Conversations**", Nyt Tidsskrift for Matematik B, vol 20, 1909.

[9] "**Solution of some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges**", Elektrotkeknikeren, vol 13, 1917.

Los parámetros comunes que conforman los modelos son:

Tasa de llegada – λ: Es la cantidad media de llegadas por unidad de tiempo. Si quinientas llamadas llegan en una hora, en promedio, entonces la tasa de llegada es de 500 llamadas / hora. El recíproco (1 / λ) es entonces la cantidad promedio de tiempo que separa una llamada entrante de otra.

Tasa de Servicio – μ: Es el número medio de llamadas atendidas por unidad de tiempo. Si se tarda 20 minutos para dar servicio a una llamada, entonces 3 llamadas pueden ser atendidas en una hora y la tasa de servicio es de 3 llamadas / hora. El recíproco (1 / μ) es el tiempo medio necesario para dar servicio a una llamada, expresada en unidades de tiempo por llamada; es decir, 0,333 horas / llamada o 20 minutos / llamada en el ejemplo actual.

Número de líneas – *n*: Es la cantidad de líneas telefónicas independientes. Determina el número de llamadas que se pueden servir al mismo tiempo.

Capacidad: Es el tamaño de la cola de llamadas. Representa el número total de llamadas que pueden ser atendidas en un momento dado, así como el número total de llamadas que se pueden mantener hasta que quede libre una línea. Nótese que en un sistema sin cola de espera, la capacidad es igual al número de líneas.

Hora Pico (Busy Hour): Es el período de una hora más activo del día. Durante este período las llamadas entrantes son más propensas a ser bloqueadas o rechazadas, por lo que para este momento se calculan las estadísticas.

Tráfico de la Hora Pico (BHT): Es la cantidad de tráfico en el sistema durante la hora pico. Se puede medir de varias maneras, incluyendo minutos de llamadas y horas de las mismas. En el ejemplo, donde 500 llamadas ingresan por hora y cada llamada tarda 20 minutos (0.333 horas) para el servicio, entonces el sistema tiene que manejar 500 × 0,333 = 166,5 horas de llamadas durante la hora pico del día. Básicamente su cálculo es
BHT = (Duración Promedio + Retardo Promedio) × Llamadas por hora / 3600.

El resultado exhibe la ocupación total de las líneas telefónicas en horas, incluyendo el periodo de Retardo Promedio durante el cual las llamadas están en espera en el Call Center y ocupando líneas. Es importante hacer notar que el Tráfico de una Hora de Ocupación debe representar la carga de tráfico más ocupada que un centro de llamadas puede soportar. El conjunto de líneas que se designe debe ser lo suficientemente grande como para satisfacer no sólo la carga de tráfico más pesada del día sino cada pico de tráfico.

Erlang ρ: Otra unidad común es la de Erlang, que representa la relación de la tasa de llegada sobre la tasa de servicio (λ / μ). Por ejemplo, si esperamos 500 llamadas / hora, la tasa de llegada es lambda = 500. Si la duración media de una llamada es de 20 minutos, o 0.333 horas, entonces la tasa de servicio es mu = 3. De esta manera, el BHT se calcula como λ / μ = 500 / 3 = 166,67. (BHT $\cong$ Erlang ρ)

Grado de Servicio en un sistema de pérdida – $\mathbf{P_B}$: Es el porcentaje de llamadas entrantes que son rechazadas durante la hora pico, debido a que todas las líneas están ocupadas en el momento de la llamada. Téngase en cuenta que esta estadística es una función del número de líneas y por lo tanto si aumentan las mismas se tendrá un menor porcentaje de llamadas perdidas, o sea un mejor grado de servicio. De esta manera el *grado de servicio* se relaciona directamente con la probabilidad de bloqueo. Un mayor grado de servicio para el usuario significa asegurar una baja probabilidad de bloqueo durante las horas pico. Si se proporciona un grado más alto de servicio entonces se requerirá aumentar el número de recursos en el sistema. A la inversa, se puede reducir el número de recursos para disminuir el costo, pero a expensas de grado de servicio.

Nivel de Servicio en un sistema con cola de espera – $\mathbf{P_W}$: Es el porcentaje de llamadas que son atendidas dentro de un tiempo aceptable de espera. Si la demora de la atención supera el tiempo aceptable de espera, la persona abandonará la cola y el sistema registrará una llamada perdida. Un porcentaje de las llamadas perdidas retornan al sistema inmediatamente después del abandono, y lo hace como una nueva llamada entrante. Este hecho distorsiona el cálculo de las medidas de rendimientos si no se las tiene en cuenta.

Tiempo de Espera: Es el período de tiempo en que la persona permanece en la cola de espera antes de ser atendido por un agente.

3.1.1 Modelo de Erlang-B (M/M/*n*/*n*)

La fórmula de Erlang-B, también conocida como fórmula de pérdida de Erlang, describe la probabilidad de pérdidas de llamadas telefónicas en un conjunto de líneas troncales activas. La fórmula deducida por Agner Krarup Erlang en 1917[10] [79] no se limita únicamente a la planificación de redes telefónicas puesto que describe una probabilidad en un sistema de colas; por ejemplo podría utilizarse en ciertos sistemas de inventario con pérdidas de ventas.

El modelo de Cola Erlang-B es utilizado por los diseñadores de sistemas telefónicos para determinar el número de líneas troncales necesarias con el objeto de administrar un volumen específico de llamadas en un intervalo de tiempo. Se trata de un modelo en el que el sistema cuenta con *n* servidores o líneas de comunicación sin cola de espera. A cada llamada que ingresa se le asigna directamente, sin demora de espera, un servidor hasta completar la totalidad disponible. Si los *n* servidores están ocupados y se produce una nueva llamada entrante, entonces ésta será bloqueada y se perderá definitivamente. El modelo no considera el reintento de llamadas razón por la cual este modelo también recibe el nombre de Sistema de pérdida de Erlang.

[10] "**Solution of some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges**", Elektrotkeknikeren, vol 13, 1917.

Dado que algunas personas reintentan la llamada luego de confirmar la falta de enlace telefónico, la fórmula Erlang-B puede calcular una cantidad insuficiente de líneas troncales. Sin embargo, por lo general resulta exacta en aquellas situaciones en que los tonos de ocupado son escasos.

La Teoría de Colas considera a éste modelo como una cola tipo M/M/*n*/*n* en donde la capacidad *n* es igual a la cantidad de servidores o líneas telefónicas. Se trata de un modelo en el que las llamadas que llegan al sistema lo hacen según un proceso de Poisson, con tasa λ, por lo que éstas son independientes. Cada llamada tiene una duración distribuida exponencialmente con media 1/μ. Por lo que fácilmente, mediante un modelo de nacimiento-muerte, pueden deducirse las expresiones de cálculo [80-82]. Así, la fórmula de Erlang-B es:

$$B(n,\rho) = \frac{\frac{\rho^n}{n!}}{\sum_{i=0}^{n}\frac{\rho^i}{i!}} \text{ , donde } \rho = \frac{\lambda}{\mu}, \quad \text{y} \quad n = \text{Cant. de servidores.} \qquad (1)$$

Esta expresión calcula la probabilidad de que una nueva llamada que se produce en el sistema sea bloqueada debido a que todos los servidores estén ocupados. Esto también es igual a la probabilidad de que ninguno de los *n* servidores estén libres. Así, la probabilidad de bloqueo puede expresarse como:

$P_B = B(n,\rho)$ donde $\boldsymbol{P_B}$ es la probabilidad de bloqueo

$\rho = \lambda/\mu$ = es la cantidad total de tráfico ofrecido en erlangs.

El *Erlang* es una unidad adimensional utilizada en telefonía como una medida estadística del volumen de tráfico. En general, si la tasa de llamadas entrantes es de λ por unidad de tiempo y la duración media de una llamada es *h*, entonces el tráfico ρ en Erlangs es: $\rho = \lambda \cdot h$

El tráfico medido en Erlangs se usa para calcular el nivel de servicio o grado de servicio de una red telefónica.

3.1.2 Modelo de Erlang-C (M/M/*n*)

El modelo de Erlang C incorpora una cola de espera infinita al modelo de Erlang B, de manera que una llamada telefónica entrante que recibe señal de ocupado porque todos los servidores están ocupados, no sea bloqueada.

El modelo de cola más simple y ampliamente utilizado en la administración de los Centros de Llamadas/Contactos Telefónicos es el sistema M/M/*n*, también llamado modelo de Erlang-C. Dado la tasa de arribo λ, la duración promedio del servicio μ^{-1} y *n* servidores trabajando en paralelo, la fórmula de Erlang-C, describe la probabilidad de que una llamada telefónica entrante tenga que esperar en la cola porque todos los servidores están ocupados o, dicho de otra manera, define la fracción del tiempo en que la persona que llama tenga que esperar en la cola antes de ser atendido por un agente. El modelo de

Erlang-C es muy restrictivo. Asume, entre otras cosas, recursos infinitos como cola de espera infinita y paciencia infinita del usuario; un ambiente en estado estacionario en el cual las llegadas se conforman según un proceso de Poisson; la duración de los servicios están exponencialmente distribuidos; los usuarios y los servidores son estadísticamente idénticos y actúan independientemente uno de otro. El modelo no reconoce los parámetros dependientes del tiempo, el comportamiento de los abandonos y la heterogeneidad de los usuarios, cuando el sistema real sí los soporta. La teoría de colas trata de determinar cuáles de estos factores son los más importantes. Utilizando un enfoque markoviano de nacimiento-muerte se puede deducir la expresión de la probabilidad de esperar en la cola cuando todos los canales de comunicación estén ocupados [80, 82]:

$$P_W = C(n,r) = 1 - \frac{\sum_{m=0}^{n-1} \frac{r^m}{m!}}{\sum_{m=0}^{n-1} \frac{r^m}{m!} + \frac{r^n}{n!(1-\rho)}} = \frac{\frac{r^n}{n!(1-\rho)}}{\sum_{m=0}^{n-1} \frac{r^m}{m!} + \frac{r^n}{n!(1-\rho)}} \quad \text{donde } r = \frac{\lambda}{\mu},\ \rho = \frac{\lambda}{n\mu} \qquad (2)$$

Donde:

r es la intensidad total del tráfico ofrecido en unidades de Erlangs.

n es la cantidad de servidores o número de líneas troncales.

$\boldsymbol{P_W}$ es la probabilidad de que un usuario tenga que esperar para ser atendido.

Se asume que las llamadas entrantes pueden ser modeladas usando una distribución de Poisson y que el tiempo de espera de las llamadas son descriptas por una distribución exponencial.

El modelo de bloqueo (1) puede utilizarse para calcular el grado de congestión en un sistema con modelo de Erlang–C, por lo tanto, la relación puede expresarse de la siguiente manera [80]:

$$C(n,r) = \frac{nB(n,r)}{n - r + rB(n,r)} \qquad (3)$$

3.1.3 Modelo de Erlang-A (M/M/*n*+M)

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 usuario ingresa al sistema se encuentra con un tiempo de espera ofrecido (AWT)[11], que se define como el tiempo que tendría que esperar la persona antes de ser atendido por un agente suponiendo que su paciencia es infinita. Si el tiempo ofrecido AWT excede el tiempo de paciencia del usuario, entonces la llamada se pierde al abandonar el sistema, o bien espera hasta recibir el servicio. El parámetro de paciencia θ no es más que la tasa de abandono individual. La Teoría de Colas identifica al modelo como una cola tipo

[11] AWT: **A**cceptable **W**aiting **T**ime

M/M/n+M para referirse al Modelo de Erlang-A (A en alusión al Abandono)[83], en donde el +M se refiere a la distribución exponencial del tiempo de paciencia [84]. Básicamente el Erlang-A combina las principales características de los Modelos de Erlang-C y Erlang-B. Dada la tasa de arribo λ, la duración promedio del servicio μ^{-1}, la tasa de paciencia θ, y n agentes trabajando en paralelo, la fórmula de Erlang-A mide, en forma teórica, la probabilidad de que una llamada entrante encuentre a todos los agentes ocupados y, por ende, tenga que esperar en la cola considerando la distribución del tiempo de paciencia de los usuarios. Así, la probabilidad de que una llamada entrante tenga que esperar en la cola un tiempo mayor a t es [33, 85]:

$$P(W>t)=\frac{\lambda\, e^{-\theta t}z(t)}{\varepsilon+\lambda z(0)} \quad \text{donde} \quad z(x)=\frac{\exp^{\frac{\lambda}{\theta}}}{\theta}\cdot\left(\frac{\theta}{\lambda}\right)^{\frac{n\mu}{\theta}}\cdot\gamma\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}e^{-\theta x}\right) \qquad (4)$$

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

$$\varepsilon\cong\int_0^{\infty}\left(1+\frac{t\mu}{\lambda}\right)^{n-1}dt$$

Por lo tanto, la probabilidad de que una llamada entrante tenga que esperar en la cola antes de ser atendida, es cuando $t = 0$ en la expresión anterior, esto es:

$$P(W>0)=ErlangA(\lambda,\mu,\theta,n)=\frac{\lambda z}{\varepsilon+\lambda z} \quad \text{donde} \quad z=\frac{\exp^{\frac{\lambda}{\theta}}}{\theta}\cdot\left(\frac{\theta}{\lambda}\right)^{\frac{n\mu}{\theta}}\cdot\gamma\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}\right) \qquad (5)$$

Ya sea que se usen modelos de Erlang-C o Erlang-A, se pueden derivar de ellos medidas de rendimientos capaces de posibilitar una administración eficiente del personal, con una eventual reducción de los costos, manteniendo al mismo tiempo un nivel de atención adecuado al usuario.

3.2 Medidas de Rendimientos

El proceso de dimensionamiento de los recursos consta básicamente de tres pasos:

1. **Dimensionamiento de recursos de comunicación**: Consiste en determinar la cantidad de líneas telefónicas troncales necesarias para administrar un grado de servicio estimado. Para cumplir este cometido se emplea el modelo de Erlang-B.

2. **Dimensionamiento de recursos de personal**: A cada línea telefónica troncal activa se le asigna un agente que tendrá por misión atender los requerimientos de servicios de las personas que se comunican con la empresa. Los agentes trabajan una cantidad de horas contiguas que conforman un turno de trabajo. El dimensionamiento del personal consiste en determinar la cantidad mínima de agentes que se planificará teniendo en cuenta el nivel

de servicio que se pretende alcanzar a lo largo de una jornada de trabajo. Para el dimensionamiento se puede usar ya sea el modelo de Erlang-C o un modelo de Erlang-A.

3. **Maximización de los niveles de servicios**: La determinación de la cantidad mínima de agentes a planificar para una jornada laboral sólo garantiza la reducción de los costos de personal para un nivel de servicio preestablecido. El objetivo de la maximización de los niveles de servicios consiste en establecer la mejor política de asignación de turnos de modo tal que se logre el mayor nivel de servicio en la atención al usuario, buscando valores por encima del nivel de servicio preestablecido. Para ello se utilizan algunos indicadores de rendimientos que dependen del modelo de Erlang utilizado.

3.2.1 Nivel de Servicio

Las actividades de los CL/CT son impulsadas por llamadas telefónicas que realizan los usuarios que, desde la visión del sistema, se producen aleatoriamente. Cuando una persona llama, tiene dos opciones: **1**) esperar hasta que un recurso agente esté disponible, o **2**) abandonar la llamada (colgar) y volver a intentarlo más tarde. Desde la perspectiva del Centro de Atención Telefónica, una larga cola de llamadas en espera puede resultar en muchas llamadas abandonadas, demasiados reintentos de comunicación y, por supuesto, una gran insatisfacción del usuario. Los administradores necesitan equilibrar el rendimiento de la actividad de respuesta al usuario con el de los costos, siendo éstos la cantidad de agentes planificados. El rendimiento de respuesta a los usuarios se mide normalmente por "*nivel de servicio*", llamándose así, al porcentaje de llamadas contestadas dentro de un tiempo objetivo [2], y está directamente relacionado con el nivel de satisfacción de las personas que llaman. Generalmente el nivel de servicio (NS) involucra diversos aspectos como los relacionados con la calidad de las respuestas, el tiempo de espera de las personas, la razón de abandono de las llamadas, la cantidad de llamadas resueltas con un solo contacto (FCR)[12], entre otros. El rendimiento de los costos es típicamente medido en términos de los niveles de utilización del personal y constituye el componente principal en los gastos de funcionamiento de un CL/CT [30]. Los modelos de pronósticos, la capacidad, la dotación de personal, el enrutamiento, la programación y la estimación de rendimiento son problemas importantes e interrelacionados que, en conjunto, constituyen un gran desafío para la optimización [11]. Por ello, la optimización en la planificación de los CL/CT resulta de gran interés para los investigadores que están en la permanente búsqueda de nuevas estrategias que optimicen el diseño y el dimensionamiento, como así también, establezcan las medidas de rendimientos más adecuadas para los objetivos de la organización.

[12] **First Call Resolution** (FCR) es el indicador que mide la cantidad de llamadas resueltas en el primer contacto del usuario. Básicamente, es el porcentaje de llamadas recibidas que pudieron ser resueltas sin necesidad de una segunda llamada.

En el proceso de optimización, el horizonte de planificación, generalmente una jornada laboral dentro del día, se divide en pequeños intervalos de tiempo de igual longitud, bloques de 15, 30 o 60 minutos [6] denominados períodos (también conocidos como "time blocks" o "time slots") para poder realizar una precisa estimación de las llamadas entrantes y una adecuada asignación de los agentes. La administración debe dotar de personal a cada bloque de tiempo (períodos) para asegurar que los usuarios sean atendidos en tiempo y forma, sin que esto implique incurrir en una mala utilización de los agentes. La cantidad de agentes que se asignará a cada período dependerá de la cantidad de llamadas que se espera obtener en cada intervalo de tiempo.

El administrador de los CL/CT intenta satisfacer los niveles de servicios fijados como política, teniendo en cuenta el presupuesto y otras restricciones, tales como el número de puestos de trabajo, la infraestructura y la disponibilidad de la mano de obra. Usando los modelos de pronósticos, la administración establece el nivel de servicio objetivo y un tiempo de espera aceptable antes de que un usuario sea atendido por un agente. Al tiempo de espera aceptable se lo conoce como AWT por su sigla en inglés (Acceptable Waiting Time). En la jerga técnica es muy habitual hablar de la regla 80-20, esto quiere decir, que el 80% de las llamadas recibidas son atendidas dentro de los 20 segundos (AWT) de espera en la cola [3]. También, se utilizan otras reglas como las 90-20, 95-20, 90-15 que son un poco más exigentes.

El presente trabajo tiene como objetivo el cálculo del nivel óptimo de personal, que se define como el número mínimo de agentes necesarios para responder X% de las llamadas en, a lo sumo, AWT segundos. Para ello, se debe disponer de un modelo de pronóstico capaz de predecir el nivel de servicio para un número fijo de agentes y, luego, variando el número de los mismos, se encontrará el nivel de personal adecuado.

Las grandes fluctuaciones en las cantidades de llamadas entrantes que se producen en períodos grandes de observación, y aún en períodos equivalentes a una jornada diaria, provocan grandes variaciones en el NS que dificultan fuertemente la manera de estimar promedios estadísticos en el análisis de rendimiento [86]. Para analizar el NS se pueden usar varios indicadores de rendimientos como ser: la probabilidad de tener que esperar en la cola hasta ser atendido P(W>0), la probabilidad de abandonar antes de ser atendido P(Ab), la fracción de personas atendidas que tuvieron que esperar menos de un tiempo de espera aceptable P(W< t) y la fracción de las llamadas que esperan ser atendidas durante un lapso mayor al tiempo de espera aceptable conocido como *Tiempo Promedio de Espera* E[W]. Los indicadores más utilizados son P(W<*t*) conocido como el Factor de Servicios Telefónicos (TSF) y E[W] que es la velocidad media de respuesta (ASA - Average Speed of Answer). El tiempo de espera aceptable (AWT), generalmente, es deducido por la administración en base a estimaciones estadísticas. En la presente investigación se ha utilizado como indicador del NS a la métrica del Factor de Servicio Telefónico (TSF).

3.2.2 Medidas de Rendimientos según Erlang-C

Cuando el proceso de llegada Poisson es independiente del estado del sistema, la probabilidad de que la llegada arbitraria de un nuevo usuario deba ponerse en la cola de espera es igual a la proporción del segmento de tiempo en que todos los servidores están ocupados. Esto se conoce como la propiedad PASTA (Poisson Arrivals See Time Average) [87]. Esta propiedad dice que la distribución del número de personas en una cola de espera en el momento que llega una nueva persona de un proceso de llegada de Poisson es igual a la distribución del número de personas en estado estacionario del sistema.

Por lo tanto, la probabilidad de que un usuario tenga que esperar es la probabilidad de que cuando llegue al sistema no encuentre servidores disponibles para ser atendido. De acuerdo con la propiedad de PASTA esto es igual a la probabilidad de que en estado estable todos los servidores estén ocupados, es decir $C(n, r)$, por lo que considerando (2), se tiene:

$$P(W > 0) = C(n, r) = \text{ErlangC}(n, r).$$

Por otro lado, si es que hay personas esperando ser atendidos en la cola del sistema, el tiempo de espera se distribuye exponencialmente. Esta distribución se obtiene condicionando el número de personas que llegan a la cola, así:

$$P(W \le t \mid W > 0) = 1 - e^{-(n\mu - \lambda)t}, \quad \text{siendo } t \ge 0$$

Factor de Servicios Telefónicos (TSF): $P(W \le t)$

El nivel del servicio está usualmente caracterizado por el TSF y mide el porcentaje de llamadas contestadas antes de n segundos, siendo n la meta de servicio acordada por la administración, también conocida como Tiempo de Espera Aceptable (AWT). La cantidad de tiempo real que se especifica como meta, se elige basándose en una variedad de consideraciones.

De esta manera, la probabilidad de que una llamada se encuentre en la cola de espera un tiempo menor a un cierto valor t fijado por la administración será [5, 80]:

$$P(W \le t) = P(W \le t \mid W = 0) \cdot P(W = 0) + P(W \le t \mid W > 0) \cdot P(W > 0)$$

$$P(W \le t) = 1 \cdot \left[1 - C(n,r)\right] + \left[1 - e^{-n\mu\,(1-\rho)t}\right] \cdot C(n,r)$$

Donde $\rho = \dfrac{\lambda}{n\mu}$ y $r = \dfrac{\lambda}{\mu}$.

$$\therefore \quad TSF = P(W \le t) = 1 - e^{-n\mu\,(1-\rho)\,t} \cdot C(n,r) = 1 - C(n,r) \cdot e^{-\mu\,(n-r)\,t} \qquad (6)$$

3.2.3 Medidas de Rendimientos según Erlang-A

En el modelo de Erlang-A los usuarios llegan al sistema de colas de acuerdo a un proceso de Poisson (λ). A las personas se les asocia un tiempo de paciencia τ que se distribuye Exp(θ), i.i.d. entre ellas. Y los tiempos de servicios son Exp(μ), i.i.d. Por último, los procesos de llegadas, la paciencia y el servicio son independientes entre sí. Por lo tanto, teniendo en cuenta la expresión (4) se puede deducir el TSF como indicador del NS bajo el modelo de Erlang-A [85]:

$$TSF = P(W \le t) = \ 1 - P(W > t) = 1 - \frac{\lambda\, e^{-\theta t} z(t)}{\varepsilon + \lambda z(0)}$$

$$\text{donde} \quad z(x) = \frac{\exp^{\frac{\lambda}{\theta}}}{\theta} \cdot \left(\frac{\theta}{\lambda}\right)^{\frac{n\mu}{\theta}} \cdot \gamma\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta} e^{-\theta x}\right) \tag{7}$$

Las expresiones $\gamma(x, y)$ y ε son como las definidas en (4).

El nivel del servicio para un período determinado depende de una serie de variables como la cantidad de llamadas entrantes, el número de agentes necesarios y, el tiempo medio operativo (AHT – Average Handling Time)[13]. Todas estas variables conforman los parámetros necesarios para el modelo de optimización, cuyos objetivos son calcular el nivel de personal adecuado y planificar los turnos óptimos de tal manera que maximicen los niveles de servicios en la atención al usuario.

[13] En la industria de los Call/Contact Centers, el Average Handling Time (AHT) es el Tiempo Medio Operativo (TMO), que incluye el tiempo de llamada en el que se está hablando con el usuario (talking time), el tiempo durante el cual el usuario es puesto en espera (holding time) y, luego de la llamada, el tiempo que se usa para dejar notas/registros de lo ocurrido durante el contacto (after call work mode).

4 Optimización

4.1 Introducción

Las etapas 2 y 3 del esquema general del estudio de Call Centers (sección 2.3), conforman básicamente el proceso de optimización. Estas pueden ser abordadas en forma conjunta o de manera independiente, al igual que todas las demás etapas. La etapa 2 se divide en dos sub-procesos: *a*) *Requerimientos*, que posibilitan determinar las restricciones básicas del problema de optimización en función de las demandas estimadas en la etapa 1, y b) *Dimensionamiento*, en el que se estima la dotación de personal con el mínimo costo para la contratación. En la etapa 3, se planifica una grilla de turnos óptimos que maximizan los niveles de servicios en función de la cantidad predeterminada de agentes (calculada en el dimensionamiento). Algunas investigaciones [13, 16] consideran a la modelización de Call Centers conformada por cuatro pasos secuenciales en el que no se contempla las tareas relacionadas a la etapa 3 y desglosan los procesos de la etapa 2 en dos pasos sucesivos.

La optimización de los recursos, tradicionalmente, se ha llevado a cabo con diferentes enfoques y complejidades algorítmicas.

El proceso de optimización de la etapa 2 ha recibido la mayor atención de los investigadores, ya que, se le ha dedicado gran esfuerzo algorítmico y se propusieron diversas estrategias para lograr la reducción de los costos de personal [13, 16, 25, 28]. Mientras que el abordaje de la etapa 3 se procesa en segunda instancia, y se orienta hacia la obtención aproximada de una grilla óptima de turnos. En este contexto, se describen diversas estrategias que incluyen simulación [13], algoritmos heurísticos [25], entre otros. En la literatura se encuentran trabajos en el que se realiza un abordaje conjunto de las etapas 2 y 3, por ejemplo [76] cuyos autores proponen un algoritmo simple de búsqueda local para obtener una solución óptima para ambos problemas. El algoritmo de fácil implementación sólo puede ser utilizado en problemas de pequeña envergadura debido a la lentitud del proceso de búsqueda sobre un espacio factible acotado. En [88], los investigadores generan diversos escenarios posibles buscando aquél cuyo esquema de turnos represente un bajo costo de personal y alto nivel de servicio, para el que usan un algoritmo genético. Por otro lado en [28], descomponen el problema original en varios programas de optimizaciones para los que se usa un algoritmo basado en la técnica del plano de corte llamado método de L-Shaped en programación estocástica. El algoritmo resuelve un programa maestro en el que se decide el dimensionamiento del personal; con esta información, se resuelven una serie de sub-problemas en el que se estiman los niveles de servicios para cada escenario posible. Basado en las soluciones de los sub-problemas, el algoritmo en cada iteración adiciona una nueva restricción al problema maestro y así hasta obtener una solución satisfactoria.

Los trabajos de investigación descriptos ponen en evidencia un nuevo enfoque no abordado explícitamente en la literatura de los Call Centers basado en la toma de decisiones multi-objetivos. Así, el proceso de optimización de Call Centers busca, por un lado, minimizar los costos de contratación de operarios (*dimensionamiento*), y por el otro, maximizar los niveles de servicios teniendo en cuenta la dotación de personal disponible (etapa 3). Se trata básicamente de una optimización entera bi-criterios que contrapone los

objetivos de tener el menor costo para la administración del personal, que implica una disminución drástica de los niveles de servicios, y la máxima satisfacción del usuario (medidos en términos de Niveles de Servicios), que implica la disponibilidad de una mayor cantidad posibles de agentes.

La tesis expone un nuevo enfoque de la optimización de Call Centers sustentado en la toma de decisiones multi-objetivos, y basada en la optimización lexicográfica. Así, el proceso de optimización global puede ser desglosado en dos sub-procesos mono-objetivo y de resolución secuencial según el orden de importancia, para explotar las características particulares de las restricciones del modelo. Desde el punto de vista lexicográfico, se le asigna la mayor preferencia al objetivo de dimensionamiento, para el que se resuelve un problema de Programación Lineal Entera (PLE) a los efectos de fijar la dotación de personal de costo mínimo. Teniendo en cuenta la cantidad mínima necesaria de agentes para alcanzar el nivel de satisfacción deseado, se resuelve un problema de Programación No Lineal Entera (PNLE) para especificar la mejor política de planificación de turnos que maximice el nivel de servicio (NS). Para esta última tarea se propone nuevos algoritmos que resuelven el problema sin uso de derivadas.

4.2 Optimización Multi-Objetivo

Muchos problemas de optimización del mundo real son formulados, naturalmente, en términos de programación no lineal con múltiples objetivos. Debido a la falta de estrategias de soluciones adecuadas, tales problemas son convenientemente convertidos a problemas de simples objetivos para su resolución. Particularmente, los problemas combinatorios multi-objetivos (PCMO) poseen un espacio de decisión integral, lo que hace difícil diseñar algoritmos eficientes para resolverlos adecuadamente. En general, la literatura considera a los PCMO con complejidad computacional NP-completos [89, 90] y NP-duros [91, 92], aunque hay evidencias que demuestran que los problemas de flujos de redes son intratables [93-95]. Un resumen detallado se puede leer en [96].

En la teoría de decisión con objetivos múltiples, varias funciones deben ser optimizadas sistemática y simultáneamente [97] sobre un conjunto factible de soluciones. En la mayoría de las veces, los objetivos se encuentran en conflictos, lo que imposibilita la búsqueda de una solución donde todos los criterios alcancen su valor óptimo. Esto quiere decir, que ninguno de los valores objetivos pueden ser mejorados sin empeorar al menos uno de los restantes [98]. En general, la optimalidad se refiere a encontrar la mejor solución posible o una buena aproximación de ésta, dado un conjunto de limitaciones o restricciones. Sin embargo, en el contexto multi-criterios, es necesario precisar el concepto. Un PCMO se formula matemáticamente como

$$\text{Optimizar } F(x) = [f_1(x),\ldots,f_p(x)] \qquad (1)$$
$$\text{Sujeto a:} \quad x \in D \subset \mathbb{Z}^n$$

Donde p es el número de funciones objetivos; x es el vector de variables de decisión y D las restricciones del problema que forma el espacio de decisión factible. $F(x)$ es el vector de funciones objetivos denominado espacio de decisión objetivo. El propósito de (1) es encontrar una solución que optimice a $F(x)$, donde cada una de las funciones componentes pueden ser maximizadas o minimizadas según sea el problema.

Al no poder lograrse optimalidad sin conflicto a cada objetivo individual, se torna necesario establecer mecanismos matemáticos que posibiliten la comparación de soluciones alternativas hasta lograr la deseada. En consecuencia, es necesario definir un esquema de relaciones binarias que permitan establecer jerarquías entre dos soluciones cualesquiera del espacio de decisión factible. En tal sentido se tiene:

Definición 1 [98]: Preferencia estricta e Indiferencia.
Sea R_1 una relación binaria en un conjunto W. R_1 es una preferencia estricta sobre W si y sólo si R_1 sirve para introducir una jerarquía entre los elementos de W. R_1 es entonces denotado por el símbolo $<_o$.
Sea R_2 una relación binaria en un conjunto W. R_2 es una indiferencia sobre W si R_2 sirve para introducir una noción de igualdad entre los elementos de W. R_2 es denotado por el símbolo $=_o$.

Definición 2 [98]: Preferencia. Sea R una relación binaria en un conjunto W. R es una preferencia en W si y sólo si $R = R_1 \cup R_2$ es una unión disjunta de los conjuntos de preferencia estricta y uno de indiferencia. R es denotado por el símbolo $\leq_o$.

Teniendo en cuenta las definiciones anteriores, y siendo $x, y \in W$, se tiene que:

$x <_o y$ si y sólo si, $x \leq_o y \ \wedge \ y \nleq_o x$. (relación asimétrica)
$x =_o y$ si y sólo si, $x \leq_o y \ \wedge \ y \leq_o x$. (relación simétrica)

Definición 3 [98]: Elementos Optimales. Sean W un conjunto arbitrario, $\leq_o$ una relación de preferencia sobre W y sea $x_0, x_1 \in W$.
x_0 es el mayor elemento de W con respecto a $\leq_o$ si y sólo si, $x \leq_o x_0$ para cada $x \in W$.
x_1 es un elemento maximal de W con respecto a $\leq_o$ si y sólo si, $x_1 \leq_o x$ implica que $x_1 =_o x$ para cada $x \in W$.

Definición análoga se aplica para el caso del elemento menor y minimal de W.

En el contexto de la optimización multi-criterio la relación binaria $\leq_o$ de la definición 2 se la conoce como relación de *dominancia de Pareto*. A partir de la definición 1 se derivan los conceptos: Sean las soluciones factibles $x_1, x_2 \in D$ entonces se dice que x_1 es *equivalente* a x_2 si y sólo si, $F(x_1) = F(x_2)$. Por otro lado, la relación binaria $<_o$ expresa la idea de dominancia estricta.

Definición 4: Una solución factible $x^* \in D$ se llama *eficiente* (u óptimo de Pareto) bajo una relación $\leq_o$–minimal, si no existe otra solución factible $x \in D$ tal que $f_k(x) \leq_o f_k(x^*)\ \forall k = 1\ldots p$ y $f_k(x) <_o f_k(x^*)$ en al menos un k, $1 \leq k \leq p$. El conjunto de las soluciones eficientes es denotado por D_E.

La definición 4 expresa que ninguna solución es al menos tan buena como x^* para todos los objetivos y estrictamente mejor para al menos uno de ellos. Este concepto divide dicotómicamente el espacio de decisión factible, por lo que $D_E \subset D$. Luego sea $x_1, x_2 \in D$ tal que $F(x_1) \leq_o F(x_2)$ y $F(x_1) \neq F(x_2)$ respecto de $\leq_o$–minimal entonces en el espacio de decisión objetivo se dice que $F(x_1)$ *domina a* $F(x_2)$. Sea $x^* \in D_E$ entonces la imagen $y = F(x^*)$ se denomina vector (punto) objetivo *no dominado*. El conjunto de los puntos no dominados se denota por Y_N. Al conjunto $Y_N \subset Y$ que es imagen de $D_E \subset D$ se lo denomina *frontera* de puntos no dominados (o frente óptimo de Pareto).

El frente óptimo de Pareto puede ser lineal, cóncavo, convexo, continuo o discontinuo dependiendo de las funciones objetivos integrantes del problema. Todas las soluciones pertenecientes a la frontera son igualmente buenas, y no se puede especificar si alguna de ellas es preferible sobre las otras, excepto en aquellos casos en que se haya definido una preferencia a priori.

La optimización clásica de Call Centers puede llevarse a cabo mediante un proceso bi-objetivo de dos fases. En la primera fase se busca el conjunto de soluciones eficientes que determina el frente óptimo de Pareto. En la segunda, se busca en el espacio objetivo reducido al frente óptimo de Pareto, la solución más adecuada para el criterio objetivo perseguido. Dado que la técnica de dos fases [99] es usado como un método general para resolver PCMO [100], en el ámbito de los Call Centers es propicio la utilización combinada con la estrategia lexicográfica [101], debido a que en la historia de la optimización de estas organizaciones se la ha adjudicado mayor importancia a la minimización de costos respecto al nivel de servicio.

Por lo tanto, la teoría lexicográfica combinada con la técnica de dos fases constituye un método general para la resolución de PCMO, que permite establecer un orden de preferencia a priori entre las funciones objetivos. De esta manera, se resuelve en secuencia una colección de problemas de optimización mono-objetivo, en el que salvo el primero, los restantes incorporan como restricciones los semiplanos soportes resultantes de las optimizaciones anteriores. Luego, el modelo matemático del problema (1) basado en el enfoque de optimización de dos fases con estrategia lexicográfica es:

$$
\begin{array}{lll}
\textbf{Fase 1}: & \text{Optimizar } y_1 = f_1(x) & \\
 & \text{Sujeto a: } x \in D \subset \mathbb{Z}^n & \\
\textbf{Fase 2}: & \text{Optimizar } y_i = f_i(x) \quad \forall i = 2\ldots p. & (2) \\
 & \text{Sujeto a: } x \in D \subset \mathbb{Z}^n & \\
 & \qquad f_j(x) = y_j \quad \forall j = 1\ldots i-1. &
\end{array}
$$

En (2), cada nueva restricción que se incorpora al problema mono-objetivo de la secuencia reduce la búsqueda en el espacio de decisión factible, de modo que el último resuelto proporciona un óptimo de Pareto que da solución al problema (1) [92].

4.3 El Problema a Optimizar

4.3.1 Definición del Problema como una Optimización Bi-Objetivos

El problema de optimizar los recursos de personal en los CL/CT responde a dos tipos de demandas: demandas con *restricciones duras* y aquellas que son con *restricciones blandas* [76]. Las demandas con *restricciones duras* especifican el número de agentes que deben programarse según el NS requerido. Las demandas con *restricciones blandas*, posibilitan que un mayor número de agentes programados en un período pueda compensar una escasez en otro período. La mayoría de los modelos de planificación para CL/CT, incluyendo los modelos implícitos de programación de descansos, aplican demandas con *restricciones duras* para el nivel de servicio en cada período en que se divide el horizonte de planificación, forzando implícitamente a que un período básico cumpla el NS fijado por la administración. La planificación en períodos individuales básicos que aseguran el NS requerido se basa en el enfoque Estacionario e Independiente Período por Período (SIPP)[14] [102]. El enfoque SIPP comienza dividiendo la jornada de trabajo en períodos de planificación que consisten en intervalos de 15, 30 o 60 minutos. A lo largo de un período, supone una tasa constante de llamadas telefónicas y la performance es considerada independiente de otros intervalos. Luego se construye una serie de modelos de colas estacionarias, generalmente con modelos Erlang, una para cada período de planificación [103]. Cada uno de estos modelos, específico para un período, se resuelve de manera independiente para establecer la cantidad de agentes necesarios bajo un régimen estacionario y cumplir así, con el nivel de servicio objetivo para ese período. Luego, se determina el nivel mínimo de agentes (*dimensionamiento*) para toda la jornada laboral, usando como restricción la cantidad requerida para mantener el modelo de programación estable. En un proceso posterior se calcula la distribución óptima del personal (*planificación de turnos*) para cada período de planificación usando como restricción la cantidad mínima de agentes programada y el número de agentes requerido para mantener el régimen estacionario. Así, queda diseñado el programa de optimización bi-objetivos conformado por los procesos básicos de optimización de personal (costo mínimo) y la determinación de las políticas de distribución que maximizan los NS. El *dimensionamiento* define el plantel mínimo que satisface los requerimientos para responder a variables exógenas (llamadas entrantes), que normalmente es resuelta mediante una programación lineal entera. Por otro lado, la etapa de *Planificación de turnos* tiene por misión distribuir adecuadamente el personal en los distintos turnos disponibles de manera que se logre el mayor NS global, por lo que la función objetivo será la composición aditiva y normalizada

[14] SIPP = Stationary Independent Period by Period.

de los rendimientos de servicios de cada periodo planificado. De esta manera, se configura un programa de optimización no lineal entero con los mismos requerimientos usados para el dimensionamiento. El modelo matemático bi-objetivo es:

$$\begin{aligned} &mín\ f_1(x);\ máx\ f_2(x) \\ &s.a. \quad x \in D \subset \mathbb{Z}^n;\ \ x \geq 0 \end{aligned} \tag{3}$$

Donde:

$f_1(x)$ es la función objetivo lineal que minimiza los costos de contratación de personal para una jornada diaria.

$f_2(x)$ es la función objetivo no lineal utilizada para determinar la política que maximiza los niveles de servicios en la atención telefónica de los usuarios.

D es el espacio de decisión factible.

4.3.2 Formulación de las Restricciones

Se considera el problema de dimensionamiento del personal y de la planificación de turnos en la administración de CL/CT, utilizando demandas con restricciones blandas. Según los modelos de pronósticos utilizados, la administración brinda el NS requerido, AWT y el AHT, las cuales serán constantes a lo largo de la jornada laboral. En este problema, el horizonte de planificación corresponde a una jornada laboral que no supera el día. La jornada se divide en m intervalos de tiempos de igual longitud denominados períodos. Las llamadas entrantes llegan a cada período según un proceso de Poisson con tasa λ_i, $i \in \{1,\ldots, m\}$ las cuales, se consideran constantes dentro del intervalo i. Los agentes trabajan en turnos de T intervalos consecutivos ($0 < T < m$). Cada agente puede iniciar su turno laboral en cualquiera de los primeros $n = m - T + 1$ intervalos. De esta manera, el turno k ($k \leq n$) comienza en el intervalo I_k y finaliza al comienzo del intervalo $I_k + T$. Con los datos brindados por la administración, se determina para cada intervalo $i \in \{1,\ldots, m\}$ el número de agentes necesarios r_i para preservar el régimen estable según el NS requerido (etapa de *requerimiento*). Los r_i constituirán las restricciones de cada período i de la planificación. Se define x_k, con $k \in \{1,\ldots, n\}$ como la cantidad de agentes asignada al turno k. Luego, el número de agentes que se encuentran trabajando en el intervalo i bajo la restricción r_i es:

$$\sum_{k=máx(1,\ i-T+1)}^{mín(i,\ m-T+1)} x_k \geq r_i, \quad i \in 1,\cdots,m \tag{4}$$

Esta expresión constituye la restricción que debe cumplirse en el intervalo i-ésimo.

Por lo tanto, se tiene n turnos posibles y m intervalos de observación independientes entre sí, cada uno con el requerimiento de r_i agentes para mantener el régimen estable bajo la exigencia de cumplir el NS global. La expresión en (4) genera una matriz de restricciones con elementos 1s y 0s, tal que $A \in \mathbb{Z}^{m \times n}$, por lo que los coeficientes se configuran de la siguiente manera:

$$a_{i,j} = \begin{cases} 1, & \text{si el periodo } i \text{ es cubierto por el turno } j \\ 0, & \text{en cualquier otro caso.} \end{cases}$$
$$i \in 1,\cdots,m \;\; ; \;\; j \in 1,\cdots,n$$

De esta manera, se tiene una matriz de intervalos con T unos (**1**s) consecutivos; esto es:

$$A^{m\,x\,n} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 0 & & \vdots & \vdots \\ 1 & 1 & 1 & \ddots & 0 & 0 \\ \vdots & 1 & 1 & \ddots & 0 & 0 \\ 1 & \vdots & 1 & & 1 & 0 \\ 0 & 1 & \vdots & & \vdots & 1 \\ 0 & 0 & 1 & & 1 & \vdots \\ 0 & 0 & 0 & \ddots & 1 & 1 \\ \vdots & \vdots & \vdots & \ddots & 1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \tag{5}$$

Así, el problema (3) queda conformado de la siguiente forma:

$$\begin{aligned} & \textit{mín}\; f_1(x);\; \textit{máx}\; f_2(x) \\ & s.a. \quad Ax \geq r \;\; ; \;\; x \geq 0;\; x \in \mathbb{Z}^n \wedge A \in \{0,1\}^{m \times n} \end{aligned} \tag{6}$$

4.3.3 Caracterización de las Restricciones

Antes de describir las características de la matriz en (5) es necesario tener en cuenta las siguientes definiciones.

Definición 5: Una matriz con elementos {0, 1} tiene la propiedad de **1**s consecutivos (C1P) si existe una permutación de sus filas tales que los elementos no nulos en cada columna estén en forma consecutiva. Esto es, para toda columna $c \in \{1,...,n\}$ de la matriz resultante A satisface la siguiente condición para todo $i_1, i_2 \in 1,\cdots,m$:

$$a_{c,i_1} = 1 \text{ y } a_{c,i_2} = 1 \;\Rightarrow\; a_{c,i} = 1 \quad \forall i_1 < i < i_2$$

Definición 6: Una matriz A($m \times n$) es Totalmente Unimodular (TU) si y sólo si, el valor del determinante de cada submatriz cuadrada de A es {0, ±1}.

Las Matrices Totalmente Unimodulares producen una clase privilegiada de problemas de programación lineal con soluciones óptimas enteras. En la literatura actual se pueden encontrar numerosos trabajos sobre el tema. Un resumen completo de la temática se encuentra en [104-107].

La matriz en (5) goza de la propiedad de tener 1s consecutivos (C1P). Las matrices así definidas son *Totalmente Unimodulares* [108]. Esta caracterización es muy importante para nuestro problema.

En definitiva, la matriz de coeficientes de las restricciones del problema (6) es TU si en el modelo no se contempla explícitamente los intervalos de descanso entre períodos de conversación. Si esto último llegara a considerarse o se contemplara un horizonte de observación que incluya varias jornadas de trabajo, entonces la matriz asociada a las restricciones puede no ser TU, y tampoco podría tener la propiedad C1P. Dado cualquiera de estos casos, el análisis del problema es aún más complejo. Modelar un problema de Call Center que induzca a tener una matriz de restricciones TU, resulta beneficioso para el diseño e implementación de algoritmos eficientes en la optimización multi-objetivo.

4.4 Optimización Bi-Objetivo del Problema de Call Centers

La propiedad de unimodularidad total en las restricciones de un PCMO no es suficiente para calcular el conjunto de soluciones de Pareto óptimas usando algoritmos eficientes de la optimización mono-objetivo [109]. La idea se enfatiza cuando se usa el enfoque lexicográfico, que incorpora a las restricciones el hiperplano soporte (restricción de igualdad) resultante de la optimización mono-objetivo anterior en la secuencia. Cada restricción de igualdad que se incorpora a las ya existentes, quita la propiedad de unimodularidad total a la matriz de las restricciones subyacentes. En definitiva el problema a optimizar es:

$$
\begin{array}{lll}
\textbf{Fase 1}: & \Omega = \text{Minimizar } f_1(x) & \\
& \text{sa:} \quad Ax \geq r \; ; \; x \geq 0 \; ; \; x \in \mathbb{Z}^n & \\
\textbf{Fase 2}: & \text{Maximizar } f_2(x) & (6^*) \\
& \text{sa:} \quad Ax \geq r \; ; \; x \geq 0 \; ; \; x \in \mathbb{Z}^n & \\
& \quad f_1(x) = \Omega &
\end{array}
$$

Dado que (6*) es un problema bi-objetivo lexicográfico, en el que $f_1(x)$ es un objetivo lineal entero con restricciones convexas y totalmente unimodulares, entonces es posible utilizar algoritmos eficientes de la literatura para calcular el conjunto de soluciones Pareto óptimas en la primera fase, que constituirá el espacio de búsqueda para la optimización no lineal entera del segundo objetivo en la segunda fase.

4.4.1 Dimensionamiento del Personal (fase 1)

El objetivo de esta etapa es determinar la cantidad de personal que se necesitará para una jornada diaria en función del volumen de llamadas telefónicas que se espera recibir, buscando siempre el costo mínimo en la contratación. Si $c \in \mathbb{R}^n$, es el vector de costo para los turnos a planificar, entonces $f_1(x) = c^T x$, donde $x \in \mathbb{Z}^n$ y representa el vector que expresa las cantidades de agentes que deben ponerse activos en cada turno. De esta manera el problema de Programación Lineal Entera (PLE) a resolver es:

$$\begin{aligned} &mín \;\; c^T x \\ &s.a. \quad Ax \geq r \;\; ; \;\; x \geq 0; \; x \in \mathbb{Z}^n \end{aligned} \tag{7}$$

Una manera rápida y eficiente de dar solución a (7) es relajar la restricción de integralidad y resolver como un Programa Lineal (LP). La técnica de relajación transforma un problema de optimización (PLE) NP-duro en un problema equivalente que es resoluble en tiempo polinomial, si se garantiza que el poliedro asociado al problema (7) posea todos sus vértices enteros.

Para determinar si el poliedro del problema (7) tiene soluciones integrales, se necesita recordar algunas propiedades.

Propiedad 4.4.1: propiedades importantes de matrices TU:
(***i***). La transpuesta de una matriz TU es también TU.
(***ii***). Cualquier submatriz de una TU es también TU.
(***iii***). Si la matriz A es TU, entonces [A I] es también TU.

Teorema 4.4.1a (Veinott & Danzig, 1968)[110] Sea $A \in \mathbb{Z}^{m \times n}$ una matriz de rango fila completo, entonces el poliedro $P = \{x: Ax = b, x \geq 0\}$ es un conjunto finito de vértices enteros para cualquier vector entero $b \in \mathbb{Z}^m$, si y sólo si, A es TU.

Teorema 4.4.1b (Hoffman & Kruskal, 1956)[111]

Sea $A \in \mathbb{Z}^{m \times n}$, entonces el poliedro $P = \{x: Ax \leq b, x \geq 0\}$ tiene todos sus vértices enteros para cualquier vector entero $b \in \mathbb{Z}^m$, si y sólo si, A es TU.

El teorema de Hoffman y Kruskal realiza un gran aporte para la resolución de problemas PLE. Muestra que si A es TU, entonces para cualquier vector entero $b \in \mathbb{Z}^m$, el poliedro $P = \{x: Ax \leq b, x \geq 0\}$ = cápsula.conv $(P \cap \mathbb{Z}^m)$. Por lo tanto, el poliedro P tiene puntos extremos enteros. De esta manera se garantiza que el poliedro asociado al problema (7) tiene vértices integrales. En consecuencia, se podría relajar las restricciones de integralidad y resolver el PLE como un problema LP de variables continuas y obtener así una solución en los enteros. Esta conjetura se demuestra con la siguiente proposición.

Proposición 1: Si $A \in \mathbb{Z}^{m \times n}$ es TU, entonces para cualquier vector $b \in \mathbb{Z}^m$ y $c \in \mathbb{R}^n$, el problema LP: $\{max\ c^T x \quad s.a \quad Ax \leq b, \quad x \geq 0\}$ tiene solución óptima entera, si existe ésta.

Demostración: Para resolver el problema LP con restricciones de desigualdades se agregan variables superfluas *s*, convirtiendo el problema LP a uno con restricciones de igualdades, tal que, para $z = \begin{pmatrix} x \\ s \end{pmatrix}$ se tiene [A I] $z = b$, $z \geq 0$. Puesto que A es TU, por la propiedad 4.4.1 (*iii*), la matriz [A I] es TU. Luego, por aplicación del teorema 4.4.1a se concluye que el problema LP tiene solución óptima entera, y éste es global si es que existe un óptimo. ▪

De esta manera, se puede relajar la restricción de integralidad en (7) y aplicar el método Simplex para encontrar la solución óptima, debido a que la matriz de las restricciones del modelo es TU, por lo tanto, los puntos extremos del poliedro asociado al problema son enteros. En consecuencia, la solución óptima del problema relajado LP lo es también del problema original PLE.

Ahora, el desafío es probar que el hiperplano soporte resultante de resolver el problema (7) constituye el conjunto de soluciones eficientes (Pareto óptimas) para el problema bi-objetivo (6). Para ello es necesario expresar algunos conceptos.

Sea el polítopo $P = \{x \in \mathbb{Z}^n: Ax \geq b, x \geq 0\}$ que expresa las restricciones del problema (7) y sea $\Omega = \min \{c^T x : x \in P\}$ con $c \in \mathbb{R}^n$. Como se busca conocer la cantidad mínima de agentes a contratar entonces $c = \mathbf{1}^{m \times 1}$. Sea además $F = \{x \in P: \mathbf{1}^T x = \Omega\}$ el hiperplano soporte de (7). Supóngase que $x^* \in F$, luego una estrategia simple de pasar de una posición x^* a una variante $x' \in F$ es descontar cierta cantidad de una de sus componentes y sumarla a otra de una posición diferente. El efecto resultante sería por ejemplo $x' = x^* + \lambda\,[-1, 0, 1, 0, \cdots, 0]$, con $\lambda \in \mathbb{Z}$. Usando esta regla y con $\lambda = 1$ se podrían conocer todos los vectores vecinos a x^* simplemente haciendo variar las posiciones de las componentes no nulas para lograr todas las combinaciones posibles. Cada combinación determina un vector direccional que permite pasar de un vector a otro vecino con las mismas propiedades. Si se tiene cuidado de no volver a un punto ya recorrido entonces se estaría transitando vecindades diferentes, que con una estrategia adecuada, tal vez se pueda conocer completamente el espacio de decisión F. Véase Figura 6.

Para dar mayor claridad a la conjetura, se define a Δ como el conjunto de vectores direccionales que tienen permutados las posiciones de sus elementos, y que estos, se construyen solamente con dos componentes unitarias de signos opuestos, y las restantes iguales a cero. Esto es:

$$\begin{aligned} d_1 &= [-1, 1, 0, \quad \cdots, \quad 0] \\ d_2 &= [-1, 0, 1, 0, \cdots, \quad 0] \\ d_3 &= [-1, 0, 0, 1, 0, \cdots, 0] \\ &\cdots \\ d_{n\cdot(n-2)} &= [0, \cdots, \ 1, 0, -1] \\ d_{n\cdot(n-1)} &= [0, \cdots, \ 0, 1, -1]. \end{aligned} \qquad (8)$$

Así, $\Delta = \{d_1, d_2, d_3, \cdots, d_{n(n-1)}\}$ está formado por $n! / (n-2)! = n \cdot (n-1)$ vectores.

Propiedad 4.4.1c: Sea el conjunto de coordenadas direccionales Δ formado con los vectores *n*-dimensionales de (8) entonces los siguientes postulados son válidos:

1. $\sum_{j=1}^{n} d_i = \mathbf{1}^{\mathrm{T}} d = 0, \ \forall d \in \Delta$
2. $\sum_{i=1}^{n(n-1)} d^{(i)} = 0$ esto es, la suma de todos los elementos de Δ resulta en un vector nulo.
3. Sea $\lambda = \sum_{i}^{K} \alpha_i \ d^{(i)}$ para $0 < K \leq n\,(n-1)$, $\alpha_i \in \mathbb{Z}-\{0\}$ y algunos $d^{(i)} \in \Delta$. Entonces $\mathbf{1}^{\mathrm{T}} \lambda = 0$.

Esto quiere decir que la combinación lineal de un subconjunto de vectores de Δ resulta un vector integral tal que la suma en valor absoluto de las componentes negativas es igual a la suma de las positivas.

Definición 7: Sea $x \in P$. Se llama vecindad de x al conjunto $V(x) = \{v \in \mathbb{Z}^n : \boldsymbol{v} = x + d_i\}$ donde $d_i \in \Delta$, $i = 1 \ldots n\,(n-1)$.

La definición 7 establece que para un $x \in P$ se tiene $n\,(n-1)$ vértices vecinos tal que $\mathbf{1}^{\mathrm{T}} v = \mathbf{1}^{\mathrm{T}} x$, $\forall v \in V(x)$. También si $x, x' \in P$ entonces $V(x)$ y $V(x')$ poseen $n \cdot (n-1)/2$ vértices comunes, por lo que, $V(x) \cap V(x') \neq \varnothing$ y $V(x) \cup V(x')$ incorpora $n \cdot (n-1)/2$ vértices nuevos a $V(x)$, esto se puede ver en el ejemplo de la Figura 6.

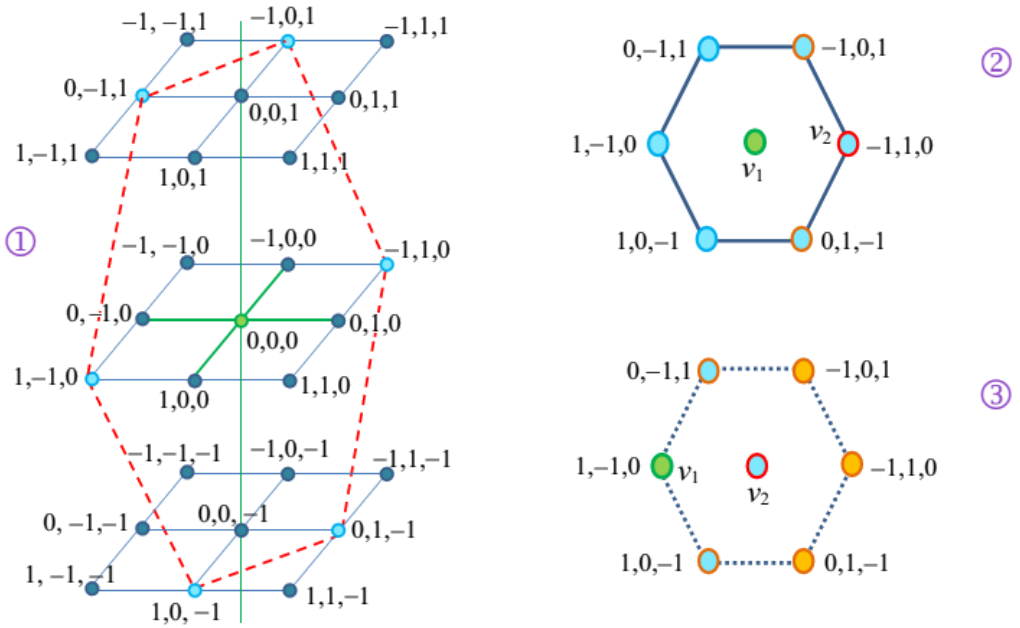

Figura 6: ① Cubo unitario. Estructura de vecindad para un problema de dimensión 3. ② Vecindad de v_1. ③ Vecindad de v_2. Los vértices en color naranja son los no comunes.

La estructura de vecindad como la descripta con la figura 6 es la que se tiene en cuenta durante la aplicación algorítmica de la optimización para problemas de dimensiones superiores.

La proposición siguiente prueba que el hiperplano soporte resultante de (7) conforma un conjunto de soluciones eficientes (espacio de decisión) para el problema de la segunda fase de (6*).

Proposición 2: Sea el polítopo $P = \{x \in \mathbb{Z}^n: Ax \geq b, x \geq 0\}$ de un problema de Call Center sin modelado explícito del descanso entre períodos de conversación, donde $A \in \mathbb{Z}^{m\times n}$ es una matriz TU y $b \in \mathbb{Z}^m$ y sea $F = \{x \in P : c^T x = \Omega\}$ donde $\Omega = \min\{c^T x : x \in P\}$ con $c \in \mathbb{R}^n$. Entonces F es el conjunto no vacío de soluciones integrales eficientes para la optimización bi-objetivo lexicográfica (6*), con cardinalidad finita $1 \leq |F| < \infty$.

Demostración: Sin pérdida de generalidad se adoptará $c = \mathbf{1}^{n\times 1}$. Por construcción del problema (sección 4.3.2), la cantidad de turnos que se evaluarán dentro de una jornada de observación con longitud de m períodos es $n = m - T + 1$, con $m, T \in \mathbb{Z}$ y $0 < T < m$. Primero se probará que $F \neq \varnothing$, luego, que forma un conjunto de soluciones eficientes para (6*), y finalmente $1 \leq |F| < \infty$.

1) $F \neq \varnothing$. Dado $0 < T < m$, el peor de los escenarios puede ocurrir cuando $T = 1$ en el que los turnos observados no están solapados, lo que genera una matriz identidad $\boldsymbol{I} \in \mathbb{Z}^{m\times m}$, cuyo sistema de desigualdad $\boldsymbol{I}x \geq b$ admite como solución única a $b \in \mathbb{Z}^m$, tal que $\Omega = \mathbf{1}^T b$. Cualquier otro vector $v \in \mathbb{Z}^m$, $v \neq b$ que sea $\mathbf{1}^T v = \Omega$ violaría las restricciones en P, por lo que $v \notin P$. En consecuencia la cardinalidad $|F| = 1$ y por lo tanto, $F \neq \varnothing$.

2) F es el conjunto de soluciones eficientes para (6*). En cualquier caso diferente de 1), en el horizonte de observación (una jornada de trabajo diario) se evidencia turnos solapados que genera una matriz de restricciones con la propiedad C1P, con $m > n$ y con rango columna completo; por lo que, F no es una cara (face) minimal de P [106]. Dado que A tiene la propiedad C1P entonces es TU, lo que garantiza que P es un poliedro con puntos extremos integrales. Dado que los elementos v de F son los de P que cumplen la propiedad de $\mathbf{1}^T v = \Omega$, entonces $F \subset P$, por lo que también F está formado por vectores integrales. Sea π una permutación de los índices $\{1,\ldots,n\}$ y $d_\pi = [0_{\pi(1)},\ldots,-\mathbf{1}_{\pi(k)},0_{\pi(k+1)}\ldots, \mathbf{1}_{\pi(s)}, \ldots 0_{\pi(n)}] \in \Delta$ con $0 < k,s \leq n$ y $k \neq s$. Sea $E(\gamma) = \{v: \mathbf{1}^T v = \gamma \wedge v = x + \lambda d_\pi, x \in P\}$ un conjunto convexo con $\lambda,\gamma \in \mathbb{Z}$, tal que $\{E(\gamma) \cap P\} \neq \varnothing$; entonces $\{E(\gamma) \cap P\}$ conforma el espacio de decisión factible del problema bi-objetivo lexicográfico (6*).
Para demostrarlo, sea un $v \in \{E(\gamma) \cap P\}$, entonces $v \in E(\gamma)$ por lo que $\mathbf{1}^T v = \gamma$ para algún $\gamma,\lambda \in \mathbb{Z}$ y, $v = x + \lambda d_\pi$ cumple las restricciones en P, $\forall x \in P$. Esto quiere decir que:

$$a_i^T v \geq b_i \;\Leftrightarrow\; a_i^T (x + \lambda d_\pi) \geq b_i$$
$$\Rightarrow \quad a_i^T (x + [0, \cdots, -\lambda_{\pi(k)}, 0\cdots, \lambda_{\pi(s)},\cdots 0]) \geq b_i \text{ con } i \in \{1,\cdots, n\} \qquad (9)$$

Teniendo en cuenta la formación de la matriz A como en (5), y sean los índices de una permutación $\pi(k) \neq \pi(s)$ en (9), se tiene los siguientes casos:

a) $\pi(k) = i = 1$ entonces $a_{11}\ (x_1 - \lambda) \geq b_1$. Esta desigualdad es válida aún en el peor de los casos. Puede ocurrir que $a_{11}\ (x_1 - \lambda) = b_1$, o $a_{11}\ (x_1 - \lambda) > b_1$, dado que $v \in P$.

b) $\pi(s) = i = m$ entonces $a_{mn}\ (x_n + \lambda) \geq b_m$. Si $a_{mn}\ x_n = b_m$, $\Rightarrow a_{mn}\ (x_n + \lambda) > b_m$, o bien que $a_{mn}\ (x_n + \lambda) >> b_m$.

c) $\pi(k) = k$, $\pi(s) = s$ y $a_{ik} = a_{is} \in \{0, 1\}$ entonces $a_i^T\ v \geq b_i$ preserva la desigualdad de $a_i^T\ x \geq b_i$ para $i \in \{1, \cdots, m\}$.

d) $\pi(k) = k$, $\pi(s) = s$ y $a_{ik} \neq a_{is} \in \{0, 1\}$ entonces el análisis es similar ya sea a 1) o 2).

Como Ω es solución de $f_1(x)$ en (6*), se concluye que $\{E(\Omega) \cap P\}$ es el conjunto de soluciones factibles del problema de maximización de $f_2(x)$, por lo tanto, es el conjunto de soluciones factible del problema bi-objetivo lexicográfico (6*).

Ahora se prueba que $\{E(\Omega) \cap P\}$ forma un conjunto de soluciones eficientes. Sea v el vector que da solución al problema de minimización de $f_1(x)$ en (6*), por lo tanto, $v \in \{E(\Omega) \cap P\}$. Luego, $f_1(v) \leq \Omega$, y no existe otro valor $v^* \in P \neq v$ tal que $f_1(v^*) < f_1(v)$, dado que Ω es solución óptima (global) del problema de minimización. Por otro lado, supóngase que $f_1(v^*) < \Omega$, entonces $Av^* < r$ por teorema 2 de (Koole et al., 2003)[76]. Como $v^* \in P$ se produce una contradicción. Por consiguiente, no existe otra solución factible $v^* \in P$ tal que $f_1(v^*) \leq f_1(v)$.

Luego, supóngase que el vector $v \in P$ da solución a $f_1(x)$, es decir $f_1(v) = \Omega$, y es usado en el problema de maximización, de forma tal que $|NS - f_2(v)|$ sea mínima. Si v^* es un elemento cualquiera de P y $v^* \neq v$, entonces $f_2(v) < f_2(v^*)$ ya que $|NS - f_2(v)| < |NS - f_2(v^*)|$ $\forall v^* \in P$, por lo que $f_2(v)$ constituye la base a mejorar en el problema de maximización. En consecuencia, no existe otra solución factible $v^* \in P$ tal que $f_k(v^*) \leq f_k(v)$ $\forall k = 1, 2$.

Por otro lado, dado que $f_2(x)$ es monótona creciente a partir del NS fijado como objetivo entonces $NS \leq f_2(v) < f_2(v^*)$ $\forall v^* \in P$, por ser $v \neq v^*$. Por lo tanto, existe un k tal que $1 \leq k \leq 2$ para el cual se cumple la relación de desigualdad estricta.

De esta manera, v verifica las condiciones de la definición 4, por lo que es un vector eficiente u óptimo de Pareto, consecuentemente, $\{E(\Omega) \cap P\}$ constituye el conjunto de soluciones eficientes para el problema (6*).

Ahora se necesita probar que $\{E(\Omega) \cap P\} = F$ con lo que se demostraría que $F \subset P$. Sea $x \in P$ el vector que da solución al problema de la fase 1, es decir que $\mathbf{1}^T x = \Omega$ entonces $x \in E(\Omega)$, por lo que, $x \in \{E(\Omega) \cap P\}$, y por definición de F, $x \in F$, se concluye entonces que $\{E(\Omega) \cap P\} \subseteq F$.

Por otro lado, sea un vector $u \in F$ entonces $u \in P$, luego se tiene:

$$\mathbf{1}^T u = \Omega \Leftrightarrow \mathbf{1}^T u + \mathbf{1}^T \mathbf{0} = \Omega \Leftrightarrow \mathbf{1}^T u + \mathbf{1}^T \lambda \mathbf{0} = \Omega \Leftrightarrow \mathbf{1}^T u + \mathbf{1}^T \lambda d_\pi = \Omega \Leftrightarrow$$
$$\Leftrightarrow \mathbf{1}^T (u + \lambda d_\pi) = \Omega \Rightarrow \text{si } v = u + \lambda d_\pi$$

entonces $v \in E(\Omega)$; y si v cumple las restricciones de P se tiene que $v \in \{E(\Omega) \cap P\}$. En consecuencia $F \subseteq \{E(\Omega) \cap P\}$, con el que se concluye que $\{E(\Omega) \cap P\} = F$.

3) $1 \leq |F| < \infty$. Sea $V(x)$ como en la definición 7, el conjunto de vértices vecinos de $x \in P$. Luego, la cardinalidad del conjunto de vecinos de x es $|V(x)| = n\,(n-1) = |\Delta|$. Sean $x, x' \in F$ tal que $x' \in V(x)$ entonces la cardinalidad de $|V(x) \cup V(x')| = 3/2 \cdot n\,(n-1)$, ya que V(x) y V($x'$) tienen en común la mitad de sus elementos. Luego, sean r vectores distintos $v_1, v_2, v_3, \ldots, v_r \in F$ tal que $v_2 \in V(v_1)$, $v_3 \in V(v_2)$, …, $v_r \in V(v_{r-1})$ entonces se generan r conjuntos distintos entre sí, ya que para un $i \in \{2, \cdots, r\}$ el elemento $v_i \in V(v_{i-1})$ y $V(v_{i-1})$ tiene la mitad de sus elementos comunes con $V(v_i)$, esto es $V(v_1) \cap V(v_2) \neq \varnothing$, $V(v_2) \cap V(v_3) \neq \varnothing$, $\cdots$, $V(v_{r-1}) \cap V(v_r) \neq \varnothing$. Sea $F' = V(v_1) \cup V(v_2) \cup \cdots \cup V(v_r)$, $\forall v_r \in F$, luego, una vecindad cualquiera $V(v_i)$ $\forall i = \{1,\ldots, |F|\}$ en el que $v_i \in F$ puede tener elementos que no pertenezcan a P, y dado que F' se forma a partir de los nodos de F entonces $F \subset F'$ y la cardinalidad de F' es $|F'| = (|F|+1)/2 \cdot n\,(n-1)$. Por otro lado, generalizando la definición de $E(\gamma)$ como $E(\gamma) = \{v \in \mathbb{Z}^n : \mathbf{1}^T v = \gamma,\ v \geq 0,\ \gamma \in \mathbb{Z}\}$ se tiene que $|E(\gamma)| = \binom{\gamma+n-1}{\gamma} = \frac{(\gamma+n-1)!}{\gamma!\,(n-1)!}$ por lo que, $F' \subset E(\Omega)$. Como el espacio de recorrido de $E(\Omega)$ es finito entonces el espacio de decisión F también es finito dado que $F \subset F' \subset E(\Omega)$. Esto es:

$$1 \leq |F| < \frac{|F|+1}{2} \cdot n\,(n-1) < \frac{(\Omega+n-1)!}{\Omega!\,(n-1)!} < \infty\,. \blacksquare$$

El análisis de situaciones desfavorables se realiza del punto de vista matemático. En este contexto se tendrá cardinalidad $|F|=1$ cada vez que $T=1$, o en algunos casos aislados cuando $n = 2$, situaciones que no ocurren en la realidad de los centros de comunicaciones. En general, en el modelado de los Call Centers se contempla un $m \geq 8$ y $n \geq 3$, restricciones que se corresponde con una estructura de comunicación muy pequeña. La finitud del espacio de decisión está dada por la cardinalidad de $|F'|$ y $E(\Omega)$. Esto quiere decir, que el algoritmo en el mejor de los casos recorrerá un espacio de decisión equivalente a F', y en el peor de los escenarios, deberá decidir la solución en un espacio equivalente a $E(\Omega)$. Así, el conjunto F conforma el espacio de decisión factible sobre el que se desarrollará la fase 2 del proceso de optimización.

4.4.2 Planificación de Turnos (fase 2)

En esta fase, el problema a resolver es aún más complejo ya que la función objetivo $f_2(x)$ es una composición de expresiones derivadas de las fórmulas de Erlang-A para el NS. La métrica utilizada para medir el nivel de satisfacción de los usuarios es el Factor de Servicios Telefónicos (TSF) (ver sección 3.2.3). Si se tienen m intervalos independientes de observación entonces la función objetivo $f_2(x)$ como composición de las expresiones de cada período y normalizada respecto de las tasas de llegada será:

$$f_2(x) = \frac{\sum_{i=1}^{m}\left[\lambda_i \cdot TSF_i(\lambda_i, \mu_i, \theta, t, a_i^T x)\right]}{\sum_{i=1}^{m}\lambda_i} \qquad i \in 1,...,m \qquad (10)$$

Donde a_i es una fila de la matriz A de las restricciones de (6) y λ_i, μ_i, θ y t son datos conocidos. Así, el problema que encuentra la política de distribución óptima de agentes que maximizan los niveles de servicios en la atención del usuario está dado por:

$$\begin{aligned} &máx \;\; f_2(x) \\ &s.a. \quad Ax \geq r \;; \\ &\quad 1^T x = \Omega; \quad x \geq 0; \; x \in \mathbb{Z}^n \end{aligned} \qquad (11)$$

La ecuación de igualdad es el hiperplano soporte que optimiza el problema (7), al mismo tiempo, que genera un conjunto de vectores Pareto óptimos en el que se debe localizar una solución Pareto dominante maximal que optimice a (11) y constituya la solución al problema bi-objetivo (6*).

La función objetivo $f_2(x)$ es no lineal y no derivable, lo que hace difícil la optimización en un Programa No Lineal Entero (PNLE). No obstante, es una función acotada en el intervalo $(0, 1] \subset \mathbb{R}$, y dado que el espacio de decisión que genera (7) es no vacío, entonces siempre se podrá estimar una solución al problema.

Uno de los aspectos que dificulta el estudio de los problemas multicriterios de Call Centers es el hecho de que el frente de Pareto puede contener un número significativamente creciente de puntos respecto de la cantidad de variables de estudios (dimensión del problema). Dado que la función objetivo en (11) es acotada entonces el sistema de desigualdades tiene una cota superior en el que $f_2(x^s) = 1$ para algún x^s que cumple las restricciones básica (6). Por construcción del problema, el vector $r \in \mathbb{Z}^m$ se fija ante la imposición de un NS deseado menor al 100%. De igual forma, si se buscara un escenario ideal en el que NS = 100% entonces se determinaría la cota superior de las restricciones de desigualdad. Por otro lado, la restricción de igualdad implica una combinación de valores enteros de las componentes de $x \in \mathbb{Z}^n$, tal que la suma de estos sea igual a Ω, por lo que dicha cantidad es finita. Si se considera todos los escenarios posibles se tendría un polítopo de la forma:

$$P = \{x \in \mathbb{Z}^n : r \leq Ax \leq s, \quad \mathbf{1}^T x = \Omega \; ; \quad x \geq 0; \; r,s \in \mathbb{Z}^m \} \qquad (12)$$

El vector $s \in \mathbb{Z}^m$ se obtiene de la misma forma que r, es decir, determinando la dotación de personal que alcanza el 100% de NS en cada período de observación. Por lo tanto, el conjunto de restricciones en (11) genera un espacio de decisión acotado. Por otro lado, la restricción de igualdad constituye el hiperplano soporte que es solución a (7). Por lo que el espacio de decisión discreto de (11) está soportado completamente en la restricción de igualdad.

Dado que se tiene un espacio de búsqueda discreto y finito es posible encontrar un algoritmo que calcule todos los puntos n-dimensionales que forman a P en (12). En esta

línea argumental, la propiedad 4.4.1c fija las bases para establecer cuando dos puntos (vectores) cualesquiera del espacio de decisión (12) son conexos.

Proposición 3: Sean $x, x' \in P$ en el que $V(x) \cap V(x') = \varnothing$, entonces existe al menos un λ tal que $x' = x + \lambda$, donde $\lambda \in \mathbb{Z}^n : \lambda = \sum_i^K \alpha_i \; d_i$; con $0 < K \leq n\,(n-1)$, $\alpha_i \in \mathbb{Z}$, $d_i \in \Delta$.

Demostración: Sea $\Omega = \min \{\mathbf{1}^T x : x \in P\}$ el hiperplano soporte que da solución al problema (7). Por hipótesis $x, x' \in P$ entonces $\mathbf{1}^T x = \mathbf{1}^T x' = \Omega \Rightarrow \mathbf{1}^T x = \mathbf{1}^T x' \Rightarrow \mathbf{1}^T x' - \mathbf{1}^T x = 0 \Leftrightarrow \mathbf{1}^T (x' - x) = 0$. Como $x \neq x'$ entonces necesariamente existe un $\lambda \in \mathbb{Z}^n : \lambda = \sum_i^K \alpha_i \; d_i$; con algún $0 < K \leq n\,(n-1)$, $\alpha_i \in \mathbb{Z}$, $d_i \in \Delta$; que por el postulado 3 de la propiedad 4.4.1c, se tiene $\mathbf{1}^T \lambda = 0$. Luego, $\mathbf{1}^T (x' - x) = 0 \Leftrightarrow \mathbf{1}^T (x' - x) = \mathbf{1}^T \lambda \Leftrightarrow x' - x = \lambda \Leftrightarrow x' = x + \lambda$. ▪

La proposición 3 es fácil de comprobar. Por ejemplo, sean $P = \{x \in \mathbb{Z}^3 / \mathbf{1}^T x = 21\}$; para el cual se tiene $\Delta = \{(-1, 1, 0); (-1, 0, 1); (1, -1, 0); (0, -1, 1); (1, 0, -1); (0, 1, -1)\}$. Luego, si $v_1 = (7, 5, 9)$ y $v_2 = (5, 10, 6)$ dos vectores de P con $V(v_1) \cap V(v_2) = \varnothing$. Entonces

$$v_1 - v_2 = (2, -5, 3) = (2, -2, 0) + (0, -3, 3) = 2\,(1, -1, 0) + 3\,(0, -1, 1).$$

Luego $v_1 = [[[[v_2 + (1, -1, 0)] + (1, -1, 0)] + (0, -1, 1)] + (0, -1, 1)] + (0, -1, 1)$.

En general, fijado un $\Omega \in \mathbb{Z}$ y para cualquier par de vectores $u, v \in \mathbb{Z}^n$ tal que $\mathbf{1}^T u = \mathbf{1}^T v = \Omega$, siempre se podrá encontrar una secuencia de vectores de Δ que aplicados a algunos de $\{u, v\}$ se puede alcanzar el otro. La proposición 3 constituye la base de la búsqueda local que realizan los algoritmos que se proponen en la siguiente sección.

4.5 Propuestas Algorítmicas

Las propiedades descriptas en las secciones anteriores posibilitan el desarrollo de algoritmos específicos y eficientes que optimizan cada fase del problema (6*).

En el capítulo siguiente se describen tres algoritmos novedosos para la optimización exclusiva de la fase 2: El algoritmo BLD que realiza una búsqueda directa sin uso de los valores de las derivadas de la función objetivo; una adaptación del algoritmo de Powell de 1964, en el que se cambia el conjunto de direcciones de búsqueda y responde con un vector integral; y finalmente, un método metaheurístico basado en Simulated Annealing. Todos ellos alcanzan soluciones óptimas con gran velocidad de convergencia y precisión.

Páginas 56 a 146 eliminadas a pedido del autor.

9 Referencias Bibliográficas

[**1**] Cleveland B. and Mayben J. (**1997**). *Call center management on fast forward : succeeding in today's dynamic inbound environment*. 1 Ed. Call Center Press, Annapolis – Md, USA. http://trove.nla.gov.au/work/24749636

[**2**] Stolletz R. (**2003**). *Performance Analysis and Optimization of Inbound Call Centers*. Lecture Notes in Economics and Mathematical Systems. 1 Ed. (Vol. 528. np. X, 229) Springer-Verlag, Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-55506-0

[**3**] Koole G. (**2013**). *Call Center Optimization*. 1 Ed. MG books, Amsterdam. http://www.gerkoole.com/CCO/

[**4**] Mehrotra V. (**1997**). Ringing Up Big Business. *Operations Research/Management Science Today*. Vol. 24 (4):18–25. http://www.orms-today.org/orms-8-97/CallCenter.html

[**5**] Gans N., Koole G. and Mandelbaum A. (**2003**). Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management*. Vol. 5 (2):79–141. http://dx.doi.org/10.1287/msom.5.2.79.16071

[**6**] Aksin Z., Armony M. and Mehrotra V. (**2007**). The Modern Call Center: A Multi-Disciplinary Perspective on Operations Management Research. *Production and Operations Management*. Vol. 16 (6):665–688. http://dx.doi.org/10.1111/j.1937-5956.2007.tb00288.x

[**7**] Mehrotra V., Grossman T. A. and Samuelson D. A. (**2013**). Call and Contact Centers. *Encyclopedia of Operations Research and Management Science*. pp. 144-150. Springer US. Boston, MA. http://dx.doi.org/10.1007/978-1-4419-1153-7_95

[**8**] Mandelbaum A. (**2006**). *Call centers: research bibliography with abstracts*. Version 7.Technical Report.Technion–Israel Institute of Technology. Disponible en http://iew3.technion.ac.il/serveng/References/US7_CC_avi.pdf. Accedido: 12/03/2016

[**9**] Buffa E. S., Cosgrove M. J. and Luce B. J. (**1976**). An Integrated Work Shift Scheduling System. *Decision Sciences*. Vol. 7 (4):620-630. http://dx.doi.org/10.1111/j.1540-5915.1976.tb00706.x

[**10**] Mason A. J., Ryan D. M. and Panton D. M. (**1998**). Integrated Simulation, Heuristic and Optimisation Approaches to Staff Scheduling. *Operations Research*. Vol. 46 (2):161–175. http://dx.doi.org/10.1287/opre.46.2.161

[**11**] Grossman T. A., Oh S. L., Rohleder T. R. and Samuelson D. A. (**2001**). Call centers. *Encyclopedia of Operations Research and Management Science*. (2º Ed.). pp. 73-76. Springer US. Boston, MA - Kluwer Academic Publishers. http://dx.doi.org/10.1007/1-4020-0611-x_95

[**12**] Ertogral K. and Bamuqabel B. (**2008**). Developing staff schedules for a bilingual telecommunication call center with flexible workers. *Computers & Industrial Engineering*. Vol. 54 (1):118–127. http://dx.doi.org/10.1016/j.cie.2007.06.040

[**13**] Atlason J., Epelman M. A. and Henderson S. G. (**2008**). Optimizing Call Center Staffing Using Simulation and Analytic Center Cutting-Plane Methods. *Management Science*. Vol. 54 (2):295-309. http://dx.doi.org/10.1287/mnsc.1070.0774

[**14**] Bhulai S., Koole G. and Pot A. (**2008**). Simple Methods for Shift Scheduling in Multiskill Call Centers. *Manufacturing & Service Operations Management*. Vol. 10 (3):411-420. http://dx.doi.org/10.1287/msom.1070.0172

[**15**] Castillo I., Joro T. and Li Y. Y. (**2009**). Workforce scheduling with multiple objectives. *European Journal of Operational Research*. Vol. 196 (1):162–170. http://dx.doi.org/10.1016/j.ejor.2008.02.038

[**16**] Ingolfsson A., Campello F., Wu X. and Cabral E. (**2010**). Combining integer programming and the randomization method to schedule employees. *European Journal of Operational Research*. Vol. 202 (1):153-163. http://dx.doi.org/10.1016/j.ejor.2009.04.026

[**17**] Ernst A. T., Jiang H., Krishnamoorthy M. and Sier D. (**2004**). Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research*. Vol. 153 (1):3-27. http://dx.doi.org/10.1016/S0377-2217(03)00095-X

[**18**] Eveborn P. and Rönnqvist M. (**2004**). Scheduler – A System for Staff Planning. *Annals of Operations Research*. Vol. 128 (1):21-45. http://dx.doi.org/10.1023/b:anor.0000019097.93634.07

[**19**] Glover F. W. and McMillan C. (**1986**). The general employee scheduling problem. An integration of MS and AI. (Applications of Integer Programming). *Computers & Operations Research*. Vol. 13 (5):563-573. http://dx.doi.org/10.1016/0305-0548(86)90050-X

[**20**] Brucker P. and Qu R. (**2014**). Network flow models for intraday personnel scheduling problems. *Annals of Operations Research*. Vol. 218 (1):107-114. http://dx.doi.org/10.1007/s10479-012-1234-y

[**21**] Pinedo M. L. (**2005**). *Planning and Scheduling in Manufacturing and Services*. Springer Series in Operations Research. 1 Ed. (np. XVI, 506) Springer-Verlag, New York. http://dx.doi.org/10.1007/b139030

[**22**] Musliu N., Schaerf A. and Slany W. (**2004**). Local search for shift design. *European Journal of Operational Research*. Vol. 153 (1):51-64. http://dx.doi.org/10.1016/S0377-2217(03)00098-5

[**23**] Ernst A. T., Jiang H., Krishnamoorthy M., Owens B. and Sier D. (**2004**). An Annotated Bibliography of Personnel Scheduling and Rostering. *Annals of Operations Research*. Vol. 127 (1):21-144. http://dx.doi.org/10.1023/B:ANOR.0000019087.46656.e2

[**24**] Fukunaga A., Hamilton E., Fama J., Andre D., Matan O. and Nourbakhsh I. (**2002**). Staff scheduling for inbound call centers and customer contact centers. *AI Magazine*. Vol. 23 (4):30-40. http://www.aaai.org/ojs/index.php/aimagazine/article/view/1667

[**25**] Caprara A., Monaci M. and Toth P. (**2003**). Models and algorithms for a staff scheduling problem. *Mathematical Programming*. Vol. 98 (1):445-476. http://dx.doi.org/10.1007/s10107-003-0413-7

[**26**] Cezik M. T. and L'Ecuyer P. (**2008**). Staffing Multiskill Call Centers via Linear Programming and Simulation. *Management Science*. Vol. 54 (2):310-323. http://dx.doi.org/10.1287/mnsc.1070.0824

[**27**] Hishinuma C., Kanakubo M. and Goto T. (**2007**). An Agent Scheduling Optimization for Call Centers. *Proceedings of Asia-Pacific Service Computing Conference, The 2nd IEEE* pp. 423-430. IEEE Computing Society. http://dx.doi.org/10.1109/APSCC.2007.27

[**28**] Robbins T. R. and Harrison T. P. (**2010**). A stochastic programming model for scheduling call centers with global Service Level Agreements. *European Journal of Operational Research*. Vol. 207 (3):1608-1619. http://dx.doi.org/10.1016/j.ejor.2010.06.013

[**29**] Mattia S., Rossi F., Servilio M. and Smriglio S. (**2014**). Robust Shift Scheduling in Call Centers. *Combinatorial Optimization: Third International Symposium, ISCO 2014. Revised Selected Papers*. Vol. 8596 of the series Lecture Notes in Computer Science. pp. 336-346. Springer International Publishing. Cham. http://dx.doi.org/10.1007/978-3-319-09174-7_29

[**30**] Borst S., Mandelbaum A. and Reiman M. I. (**2004**). Dimensioning Large Call Centers. *Journal Operations Research*. Vol. 52 (1):17–34. http://dx.doi.org/10.1287/opre.1030.0081

[**31**] Brown L., Gans N., Mandelbaum A., Sakov A., Shen H., Zeltyn S. and Zhao L. (**2005**). Statistical Analysis of a Telephone Call Center. *Journal of the American Statistical Association*. Vol. 100 (469):36–50. http://dx.doi.org/10.1198/016214504000001808

[**32**] Mandelbaum A. and Zeltyn S. (**2004**). The impact of customers' patience on delay and abandonment: some empirically-driven experiments with the M/M/n + G queue. *OR Spectrum*. Vol. 26 (3):377-411. http://dx.doi.org/10.1007/s00291-004-0164-8

[**33**] Mandelbaum A. and Zeltyn S. (**2007**). Service Engineering in Action: The Palm/Erlang-A Queue, with Applications to Call Centers. *Advances in Services Innovations*. pp. 17-45. Springer-Verlag. Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-540-29860-1_2

[**34**] Koole G. and Mandelbaum A. (**2002**). Queueing Models of Call Centers: An Introduction. *Annals of Operations Research*. Vol. 113 (1-4):41-59. http://dx.doi.org/10.1023/a:1020949626017

[**35**] Jouini O., Pot A., Koole G. and Dallery Y. (**2010**). Online scheduling policies for multiclass call centers with impatient customers. *European Journal of Operational Research*. Vol. 207 (1):258-268. http://dx.doi.org/10.1016/j.ejor.2010.02.036

[**36**] Zohar E., Mandelbaum A. and Shimkin N. (**2002**). Adaptive Behavior of Impatient Customers in Tele-Queues: Theory and Empirical Support. *Management Science*. Vol. 48 (4):566-583. http://dx.doi.org/10.1287/mnsc.48.4.566.211

[**37**] L'Ecuyer P. (**2006**). Modeling and Optimization Problems in Contact Centers. *Proceedings of Third International Conference on Quantitative Evaluation of Systems, 2006. QEST 2006.* pp. 145–156. IEEE Computing Society. Riverside, CA. http://dx.doi.org/10.1109/QEST.2006.34

[**38**] Erdem A. S. and Gedikoglu B. (**2006**). A DSS for Shift Design and Workforce Allocation in a Call Center. *Proceedings of Technology Management for the Global Future, 2006. PICMET 2006* pp. 1279-1289. IEEE Computing Society. Istanbul. http://dx.doi.org/10.1109/PICMET.2006.296696

[**39**] Avramidis A. N., Chan W., Gendreau M., L'Ecuyer P. and Pisacane O. (**2010**). Optimizing daily agent scheduling in a multiskill call center. *European Journal of Operational Research*. Vol. 200 (3):822-832. http://dx.doi.org/10.1016/j.ejor.2009.01.042

[**40**] Buist E., Chan W. and L'Ecuyer P. (**2008**). Speeding up call center simulation and optimization by Markov chain uniformization. *Proceedings of Winter Simulation Conference, 2008. WSC 2008.* pp. 1652-1660. IEEE Computing Society. Austin, TX. http://dx.doi.org/10.1109/WSC.2008.4736250

[**41**] Deslauriers A., L'Ecuyer P., Pichitlamken J., Ingolfsson A. and Avramidis A. N. (**2007**). Markov chain models of a telephone call center with call blending. *Computers & Operations Research*. Vol. 34 (6):1616-1645. http://dx.doi.org/10.1016/j.cor.2005.06.019

[**42**] Bylina J., Bylina B., Zoła A. and Skaraczyński T. (**2009**). A Markovian Model of a Call Center with Time Varying Arrival Rate and Skill Based Routing. *Computer Networks: 16th Conference, CN 2009.* Vol. 39 of the series Communications in Computer and Information Science. pp. 26-33. Springer. Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-02671-3_4

[**43**] Schneps-Schneppe M. and Sedols J. (**2012**). Markov Models for Multi-Skill Call Centers. *International Journal of Networks and Communications*. Vol. 2 (4):55-61. http://dx.doi.org/10.5923/j.ijnc.20120204.03

[**44**] Henderson W. B. and Berry W. L. (**1976**). Heuristic Methods for Telephone Operator Shift Scheduling: An Experimental Analysis. *Management Science*. Vol. 22 (12):1372-1380. http://dx.doi.org/10.1287/mnsc.22.12.1372

[**45**] Pot A., Bhulai S. and Koole G. (**2008**). A Simple Staffing Method for Multiskill Call Centers. *Manufacturing & Service Operations Management*. Vol. 10 (3):421-428. http://dx.doi.org/doi:10.1287/msom.1070.0173

[**46**] Robbins T. R. and Harrison T. P. (**2008**). A simulation based scheduling model for call centers with uncertain arrival rates. *Proceedings of Winter Simulation Conference, 2008. WSC 2008.* pp. 2884-2890. IEEE Computing Society. Austin, TX. http://dx.doi.org/10.1109/WSC.2008.4736410

[**47**] Avramidis A. N., Chan W. and L'Ecuyer P. (**2009**). Staffing multi-skill call centers via search methods and a performance approximation. *IIE Transactions*. Vol. 41 (6):483-497. http://dx.doi.org/10.1080/07408170802322986

[**48**] Mendes F., Lucet C. and Moukrim A. (**2006**). Tabu Search to Plan Schedules in a Multiskill Customer Contact Center. *Proceedings of International Conference on Service Systems and Service Management, 2006* pp. 1126-1131. IEEE Computing Society. Troyes. http://dx.doi.org/10.1109/ICSSSM.2006.320666

[**49**] Peyravi F. and Keshavarzi A. (**2009**). Agent Based Model for Call Centers Using Knowledge Management. *Proceedings of Third Asia International Conference on Modelling & Simulation, 2009. AMS '09.* pp. 51-56. IEEE Computing Society. Bali. http://dx.doi.org/10.1109/AMS.2009.147

[**50**] Hampshire R. C. and Massey W. A. (**2005**). Variational optimization for call center staffing. *Proceedings of Richard Tapia Celebration of Diversity in Computing Conference, 2005* pp. 4-6. IEEE Computing Society. Albuquerque, New Mexico - USA. http://dx.doi.org/10.1109/RTCDC.2005.201631

[**51**] Pichitlamken J., Deslauriers A., L'Ecuyer P. and Avramidis A. N. (**2003**). Modelling and simulation of a telephone call center. *Proceedings of Winter Simulation Conference, 2003.* pp. 1805-1812 vol.2. IEEE Computing Society. http://dx.doi.org/10.1109/WSC.2003.1261636

[**52**] Avramidis A. N. and L'Ecuyer P. (**2005**). Modeling and simulation of call centers. *Proceedings of Winter Simulation Conference, 2005* pp. 144-152. IEEE Computing Society. Piscataway, New Jersey. http://dx.doi.org/10.1109/WSC.2005.1574247

[**53**] L'Ecuyer P. and Buist E. (**2006**). Variance Reduction in the Simulation of Call Centers. *Proceedings of Winter Simulation Conference, 2006. WSC 06.* pp. 604-613. IEEE Computing Society. Monterey, CA. http://dx.doi.org/10.1109/WSC.2006.323136

[**54**] Mehrotra V. and Fama J. (**2003**). Call center simulation modeling: methods, challenges, and opportunities. *Proceedings of Winter Simulation Conference, 2003.* pp. 135-143 Vol.1. IEEE Computing Society. http://dx.doi.org/10.1109/WSC.2003.1261416

[**55**] Atlason J., Epelman M. A. and Henderson S. G. (**2004**). Call Center Staffing with Simulation and Cutting Plane Methods. *Annals of Operations Research*. Vol. 127 (1):333–358. http://dx.doi.org/10.1023/B:ANOR.0000019095.91642.bb

[**56**] Avramidis A. N., Deslauriers A. and L'Ecuyer P. (**2004**). Modeling Daily Arrivals to a Telephone Call Center. *Management Science*. Vol. 50 (7):896-908. http://dx.doi.org/10.1287/mnsc.1040.0236

[**57**] Whitt W. (**2005**). Engineering Solution of a Basic Call-Center Model. *Management Science*. Vol. 51 (2):221-235. http://dx.doi.org/10.1287/mnsc.1040.0302

[**58**] Whitt W. (**2006**). Staffing a Call Center with Uncertain Arrival Rate and Absenteeism. *Production and Operations Management*. Vol. 15 (1):88-102. http://www.columbia.edu/~ww2040/POM2006.pdf

[**59**] Jouini O., Koole G. and Roubos A. (**2013**). Performance indicators for call centers with impatient customers. *IIE Transactions*. Vol. 45 (3):341-354. http://dx.doi.org/10.1080/0740817X.2012.712241

[**60**] Yu M., Gong J., Tang J. and Zhu H. (**2013**). The method of staffing a call center with delay information considering the customers' behavior. *Proceedings of Control and Decision Conference (CCDC), 2013 25th Chinese* pp. 4723-4727. IEEE Computing Society. Guiyang. http://dx.doi.org/10.1109/CCDC.2013.6561788

[**61**] Wallace R. B. and Whitt W. (**2005**). A Staffing Algorithm for Call Centers with Skill-Based Routing. *Manufacturing & Service Operations Management*. Vol. 7 (4):276-294. http://dx.doi.org/10.1287/msom.1050.0086

[**62**] Jaoua A., L'Ecuyer P. and Delorme L. (**2013**). Call-type dependence in multiskill call centers. *SIMULATION*. Vol. 89 (6):722-734. http://dx.doi.org/10.1177/0037549713479405

[**63**] Alfares H. K. (**2004**). Survey, Categorization, and Comparison of Recent Tour Scheduling Literature. *Annals of Operations Research*. Vol. 127 (1-4):145-175. http://dx.doi.org/10.1023/B:ANOR.0000019088.98647.e2

[**64**] Chuin Lau H. (**1996**). On the complexity of manpower shift scheduling. *Computers & Operations Research*. Vol. 23 (1):93-102. http://dx.doi.org/10.1016/0305-0548(94)00094-O

[**65**] Dean J. S. (**2008**). Staff Scheduling by a Genetic Algorithm with a Two-Dimensional Chromosome Structure. *Proceedings of The 7th International Conference on the Practice and Theory of Automated Timetabling*. 18-22 August. Montréal, Canada. http://www.patatconference.org/patat2008/proceedings/Dean-WA3c.pdf

[**66**] Brucker P., Qu R. and Burke E. (**2011**). Personnel scheduling: Models and complexity. *European Journal of Operational Research*. Vol. 210 (3):467-473. http://dx.doi.org/10.1016/j.ejor.2010.11.017

[**67**] Garey M. R. and Johnson D. S. (**1979**). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Series of Books in the Mathematical Sciences. 1 Ed. (np. 340) W. H. Freeman and Company, New York, USA. http://dl.acm.org/citation.cfm?id=578533

[**68**] Shen H. and Huang J. Z.; (**2005**). *Forecasting Arrivals to a Telephone Call Center*. **Technical Report** UNC/STOR/05/05. Department of Statistics and Operations Research. University of North Carolina, Chapel Hill. http://stat-or.unc.edu/research/techpdf/svdfore.pdf

[**69**] Shen H. and Huang J. Z. (**2005**). Analysis of call centre arrival data using singular value decomposition. *Applied Stochastic Models in Business and Industry*. Vol. 21 (3):251-263. http://dx.doi.org/10.1002/asmb.598

[**70**] Shen H. and Huang J. Z. (**2008**). Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. Vol. 2 (2):601-623. http://dx.doi.org/10.1214/08-AOAS164

[**71**] Soyer R. and Tarimcilar M. M. (**2008**). Modeling and Analysis of Call Center Arrival Data: A Bayesian Approach. *Management Science*. Vol. 54 (2):266-278. http://dx.doi.org/10.1287/mnsc.1070.0776

[**72**] Aldor-Noiman S., Feigin P. D. and Mandelbaum A. (**2009**). Workload forecasting for a call center: Methodology and a case study. *The Annals of Applied Statistics*. Vol. 3 (4):1403-1447. http://dx.doi.org/10.1214/09-AOAS255

[**73**] Taylor J. W. (**2012**). Density Forecasting of Intraday Call Center Arrivals Using Models Based on Exponential Smoothing. *Management Science*. Vol. 58 (3):534-549. http://dx.doi.org/10.1287/mnsc.1110.1434

[**74**] Ibrahim R., Regnard N., L'Ecuyer P. and Shen H. (**2012**). On the modeling and forecasting of call center arrivals. *Proceedings of Winter Simulation Conference (WSC)* pp. 1-12. IEEE Computing Society. Berlin. http://dx.doi.org/10.1109/WSC.2012.6465292

[**75**] Ibrahim R. and L'Ecuyer P. (**2013**). Forecasting Call Center Arrivals: Fixed-Effects, Mixed-Effects, and Bivariate Models. *Manufacturing & Service Operations Management*. Vol. 15 (1):72-85. http://dx.doi.org/10.1287/msom.1120.0405

[**76**] Koole G. and van der Sluis E. (**2003**). Optimal Shift Scheduling with a Global Service Level Constraint. *IIE Transactions*. Vol. 35 (11):1049-1055. http://dx.doi.org/10.1080/07408170304398

[**77**] Mehrotra V., Ozlük O. and Saltzman R. (**2010**). Intelligent Procedures for Intra-Day Updating of Call Center Agent Schedules. *Production and Operations Management*. Vol. 19 (3):353-367. http://dx.doi.org/10.1111/j.1937-5956.2009.01097.x

[**78**] Erlang A. K. (**1920**). Telefon-Ventetider. Et Stykke Sandsynlighedsregning. *Matematisk Tidsskrift*. B. pp. 25-42. Matematisk Forening I Kobenhavn. AARGANG 1920.

[**79**] Kelly F. P. (**1991**). Loss Networks. *The Annals of Applied Probability*. Vol. 1 (3):319-378. http://www.jstor.org/stable/2959742

[**80**] Gross D., Shortie J. F., Thompson J. M. and Harris C. M. (**2008**). *Fundamentals of Queueing Theory*. Wiley Series in Probability and Statistics. 4 Ed. (np. 528) John Wiley & Sons, Inc., Hoboken, New Jersey. http://www.wiley.com/WileyCDA/WileyTitle/productCd-047179127X.html

[**81**] Iversen V. B. (**2015**). *Teletraffic engineering and network planning*. Handbook. DTU Fotonik. http://orbit.dtu.dk/en/publications/teletraffic-engineering-and-network-planning(1770eb79-c4d7-4eff-8bd8-55ac72898d7b).html

[**82**] Lakatos L., Szeidl L. and Telek M. (**2013**). *Introduction to Queueing Systems with Telecommunication Applications*. Mathematics and Statistics. (np. 396) Springer Publishing Company, US. http://dx.doi.org/10.1007/978-1-4614-5317-8

[**83**] Garnett O. N., Mandelbaum A. and Reiman M. I. (**2002**). Designing a Call Center with Impatient Customers. *Manufacturing & Service Operations Management*. Vol. 4 (3):208-227. http://dx.doi.org/10.1287/msom.4.3.208.7753

[**84**] Baccelli F. and Hebuterne G. (**1981**). On Queues with Impatient Customers. *Paper of PERFORMANCE 81*. pp. 159-180. North Holland Publishing Company. Amsterdam. https://hal.archives-ouvertes.fr/file/index/docid/76467/filename/RR-0094.pdf

[**85**] Zeltyn S. and Mandelbaum A. (**2005**). Call Centers with Impatient Customers: Many-Server Asymptotics of the M/M/n + G Queue. *Queueing Systems*. Vol. 51 (3-4):361-402. http://dx.doi.org/10.1007/s11134-005-3699-8

[**86**] Roubos A., Koole G. and Stolletz R. (**2012**). Service-Level Variability of Inbound Call Centers. *Manufacturing & Service Operations Management*. Vol. 14 (3):402-413. http://dx.doi.org/10.1287/msom.1120.0382

[**87**] Wolff R. W. (**1982**). Poisson Arrivals See Time Averages. *Operations Research*. Vol. 30 (2):223-231. http://dx.doi.org/10.1287/opre.30.2.223

[**88**] Ingolfsson A., Amanul Haque M. and Umnikov A. (**2002**). Accounting for time-varying queueing effects in workforce scheduling. *European Journal of Operational Research*. Vol. 139 (3):585-597. http://dx.doi.org/10.1016/S0377-2217(01)00169-2

[**89**] Ehrgott M. (**1996**). On matroids with multiple objectives. *Optimization*. Vol. 38 (1):73-84. http://dx.doi.org/10.1080/02331939608844238

[**90**] Ehrgott M. (**2000**). Approximation algorithms for combinatorial multicriteria optimization problems. *International Transactions in Operational Research*. Vol. 7 (1):5-31. http://dx.doi.org/10.1016/S0969-6016(99)00024-6

[**91**] Serafini P. (**1987**). Some Considerations about Computational Complexity for Multi Objective Combinatorial Problems. *Recent Advances and Historical Development of Vector Optimization: Proceedings of an International Conference on Vector Optimization Held at the Technical University of Darmstadt, FRG, August 4–7, 1986*. pp. 222-232. Springer. Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-642-46618-2_15

[**92**] Ehrgott M. (**2005**). *Multicriteria Optimization*. 2nd Ed. (np. XIII, 323) Springer-Verlag, Berlin, Heidelberg. http://www.springer.com/us/book/9783540213987

[**93**] Ruhe G. (**1988**). Complexity results for multicriterial and parametric network flows using a pathological graph of Zadeh. *Zeitschrift für Operations Research*. Vol. 32 (1):9-27. http://dx.doi.org/10.1007/bf01920568

[**94**] Ruzika S. and Hamacher H. W. (**2009**). A Survey on Multiple Objective Minimum Spanning Tree Problems. *Algorithmics of Large and Complex Networks: Design, Analysis, and Simulation*. pp. 104-116. Springer Berlin Heidelberg. Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-642-02094-0_6

[**95**] Hamacher H. W. and Ruhe G. (**1994**). On spanning tree problems with multiple objectives. *Annals of Operations Research*. Vol. 52 (4):209-230. http://dx.doi.org/10.1007/bf02032304

[**96**] Ehrgott M. and Gandibleux X. (**2003**). Multiple Objective Combinatorial Optimization — A Tutorial. *Multi-Objective Programming and Goal Programming: Theory and Applications*. pp. 3-18. Springer Berlin Heidelberg. Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-540-36510-5_1

[**97**] Marler R. T. and Arora J. S. (**2004**). Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization*. Vol. 26 (6):369-395. http://dx.doi.org/10.1007/s00158-003-0368-6

[**98**] Stadler W. (**1979**). A survey of multicriteria optimization or the vector maximum problem, part I: 1776–1960. *Journal of Optimization Theory and Applications*. Vol. 29 (1):1-52. http://dx.doi.org/10.1007/bf00932634

[**99**] Ulungu E. L. and Teghem J. (**1995**). The two phases method: An efficient procedure to solve bi-objective combinatorial optimization problems. *Foundations of Computing and Decision Sciences*. Vol. 20 (2):149-165. http://fcds.cs.put.poznan.pl/FCDS2/Old/1995.htm

[**100**] Przybylski A., Gandibleux X. and Ehrgott M. (**2010**). A two phase method for multi-objective integer programming and its application to the assignment problem with three objectives. *Discrete Optimization*. Vol. 7 (3):149-165. http://dx.doi.org/10.1016/j.disopt.2010.03.005

[**101**] Fishburn P. C. (**1974**). Exceptional Paper—Lexicographic Orders, Utilities and Decision Rules: A Survey. *Management Science*. Vol. 20 (11):1442-1471. http://dx.doi.org/10.1287/mnsc.20.11.1442

[**102**] Green L. V., Kolesar P. J. and Soares J. (**2001**). Improving the Sipp Approach for Staffing Service Systems That Have Cyclic Demands. *Operations Research*. Vol. 49 (4):549-564. http://dx.doi.org/10.1287/opre.49.4.549.11228

[**103**] Green L. V., Kolesar P. J. and Soares J. (**2003**). An Improved Heuristic for Staffing Telephone Call Centers with Limited Operating Hours. *Production and Operations Management*. Vol. 12 (1):46-61. http://dx.doi.org/10.1111/j.1937-5956.2003.tb00197.x

[**104**] Schrijver A. (**1986**). *Theory of Linear and Integer Programming*. SERIES IN DISCRETE MATHEMATICS. (np. 485) John Wiley & Sons, Great Britain. http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471982326.html

[**105**] Papadimitriou C. H. and Steiglitz K. (**1998**). *Combinatorial Optimization*. Algorithms and Complexity. 2 Ed. (np. 513) Dover Publications, Mineola, New York.

[**106**] Cook W. J., Cunningham w. H., Pulleyblank W. R. and Schrijver A. (**1998**). *Combinatorial Optimization*. Wiley-Interscience series in discrete mathematics and optimization. (np. 368) John Wiley & Sons, New York. http://www.wiley.com/WileyCDA/WileyTitle/productCd-047155894X,subjectCd-MA40.html

[**107**] Chen D.-S., Batson R. G. and Dang Y. (**2010**). *Applied Integer Programming*. Modeling and Solution. (np. 488) John Wiley & Sons, USA. http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470373067.html

[**108**] Veinott (Jr.) A. F. and Wagner H. M. (**1962**). Optimal Capacity Scheduling. *Operations Research*. Vol. 10 (4):Part-I: 518-532, Part-II: 533-546. http://dx.doi.org/10.1287/opre.10.4.518 doi part-II: 10.1287/opre.10.4.533

[**109**] Kouvelis P. and Carlson R. C. (**1992**). Total unimodularity applications in bi-objective discrete optimization. *Operations Research Letters*. Vol. 11 (1):61-65. http://dx.doi.org/10.1016/0167-6377(92)90064-A

[**110**] Veinott (Jr.) A. F. and Dantzig G. B. (**1968**). Integral Extreme Points. *SIAM Review*. Vol. 10 (3):371-372. http://dx.doi.org/doi:10.1137/1010063

[**111**] Hoffman A. J. and Kruskal J. B. (**2010**). Integral Boundary Points of Convex Polyhedra. *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*. pp. 49-76. Springer-Verlag. Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-540-68279-0_3

[**112**] Frank M. and Wolfe P. (**1956**). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*. Vol. 3 (1-2):95-110. http://dx.doi.org/10.1002/nav.3800030109

[**113**] Clarkson K. L. (**2010**). Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Trans. Algorithms*. Vol. 6 (4):1-30. http://dx.doi.org/10.1145/1824777.1824783

[**114**] Bertsekas D. P. (**1999**). *Nonlinear Programming*. 2nd Ed. (np. 776) Athena Scientific, USA. http://www.athenasc.com/nonlinbook.html

[**115**] Ponnusamy S. (**2012**). Sequences: Convergence and Divergence. *Foundations of Mathematical Analysis*. pp. 23-70. Birkhäuser (Springer Science + Business Media). Boston, MA. http://dx.doi.org/10.1007/978-0-8176-8292-7_2

[**116**] Powell M. J. D. (**1964**). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*. Vol. 7 (2):155-162. http://dx.doi.org/10.1093/comjnl/7.2.155

[**117**] Fletcher R. (**1965**). Function Minimization Without Evaluating Derivatives -- a Review. *The Computer Journal*. Vol. 8 (1):33-41. http://dx.doi.org/10.1093/comjnl/8.1.33

[**118**] Smith C. S.; (**1962**). *The automatic computation of maximum likelihood estimates*. **Technical Report** SC 846/MR/40. Scientific Department. Pneumoconiosis Field Research. National Coal Board. London. http://www.iom-world.org/pubs/IOM_R6901.pdf

[**119**] Fletcher R. (**1987**). *Practical Methods of Optimization*. 2 Ed. (np. 450) A Wiley - Interscience Publication, Great Britain. http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471494631.html

[**120**] Lewis R. M., Torczon V. and Trosset M. W. (**2000**). Direct search methods: then and now. *Journal of Computational and Applied Mathematics*. Vol. 124 (1–2):191-207. http://dx.doi.org/10.1016/S0377-0427(00)00423-4

[**121**] Powell M. J. D. (**1972**). Quadratic Termination Properties of Minimization Algorithms I. Statement and Discussion of Results. *IMA Journal of Applied Mathematics*. Vol. 10 (3):333-342. http://dx.doi.org/10.1093/imamat/10.3.333

[**122**] Powell M. J. D. (**1972**). Quadratic Termination Properties of Minimization Algorithms II. Proofs of Theorems. *IMA Journal of Applied Mathematics*. Vol. 10 (3):343-357. http://dx.doi.org/10.1093/imamat/10.3.343

[**123**] Zangwill W. I. (**1967**). Minimizing a function without calculating derivatives. *The Computer Journal*. Vol. 10 (3):293-296. http://dx.doi.org/10.1093/comjnl/10.3.293

[**124**] Adby P. R. and Dempster M. A. H. (**1974**). *Introduction to Optimization Methods*. Chapman and Hall Mathematics Series. Springer, Netherlands. http://link.springer.com/book/10.1007/978-94-009-5705-3

[**125**] Brent R. P. (**1973**). *Algorithms for Minimization without Derivatives*. Prentice-Hall Series in Automatic Computation. (np. 204) Prentice-Hall Inc., Englewood Cliffs, New Jersey.

[**126**] Buhmann M. D. and Fletcher R.; (**1996**). *M.J.D. Powell's work in univariate and multivariate approximation theory and his contribution to optimization*. **Research Report** 96-16. Seminar für Angewandte Mathematik, ETH. University of Dundee, Switzerland. http://e-collection.library.ethz.ch/eserv/eth:24735/eth-24735-01.pdf

[**127**] Powell M. J. D. (**1971**). Recent advances in unconstrained optimization. *Mathematical Programming*. Vol. 1 (1):26-57. http://dx.doi.org/10.1007/bf01584071

[**128**] Powell M. J. D.; (**1998**). *Direct search algorithms for optimization calculations*. **Acta Numerica.** Vol. 7, pp. 287-336. Cambridge Journals. Cambridge University Press, United Kingdom. http://dx.doi.org/10.1017/S0962492900002841

[**129**] Acton F. S. (**1990**). *Numerical Methods That (Usually) Work*. (revision of 1970 edition published by Harper & Rowe, New York). (np. 569) The Mathematical Association of America, Washington D.C. http://trove.nla.gov.au/work/20271372?selectedversion=NBD8110389

[**130**] Press W. H., Teukolsky S. A., Vetterling W. T. and Flannery B. P. (**2007**). *Numerical Recipes*. The Art of Scientific Computing. 3 Ed. Cambridge University Press, New York. www.cambridge.org/9780521880688

[**131**] Glover F. W. (**1986**). Future paths for integer programming and links to artificial intelligence. (Applications of Integer Programming). *Computers & Operations Research*. Vol. 13 (5):533-549. http://dx.doi.org/10.1016/0305-0548(86)90048-1

[**132**] Glover F. W. and Laguna M. (**1997**). *Tabu Search*. Kluwer Academic Publishers, Springer. US. http://dx.doi.org/10.1007/978-1-4615-6089-0

[**133**] Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H. and Teller E. (**1953**). Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*. Vol. 21 (6):1087-1092. http://dx.doi.org/10.1063/1.1699114

[**134**] Kirkpatrick S., Gelatt C. D. and Vecchi M. P. (**1983**). Optimization by Simulated Annealing. *Science*. Vol. 220 (4598):671-680. http://dx.doi.org/10.1126/science.220.4598.671

[**135**] Černý V. (**1985**). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*. Vol. 45 (1):41-51. http://dx.doi.org/10.1007/bf00940812

[**136**] Talbi E.-G. (**2009**). *Metaheuristics*. From Design to Implementation. (np. 624) John Wiley & Sons, Inc., Hoboken, New Jersey. http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470278587.html

[**137**] Rose J., Klebsch W. and Wolf J. (**1990**). Temperature measurement and equilibrium dynamics of simulated annealing placements. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*. Vol. 9 (3):253-259. http://dx.doi.org/10.1109/43.46801

[**138**] Miki M., Hiroyasu T. and Jitta T. (**2003**). Adaptive Simulated Annealing for maximum temperature. *Proceedings of IEEE International Conference on Systems, Man and Cybernetics, 2003* pp. 20-25. http://dx.doi.org/10.1109/ICSMC.2003.1243786

[**139**] Hajek B. (**1988**). Cooling Schedules for Optimal Annealing. *Mathematics of Operations Research*. Vol. 13 (2):311-329. http://dx.doi.org/10.1287/moor.13.2.311

[**140**] Azizi N. and Zolfaghari S. (**2004**). Adaptive temperature control for simulated annealing: a comparative study. *Computers & Operations Research*. Vol. 31 (14):2439-2451. http://dx.doi.org/10.1016/S0305-0548(03)00197-7

[**141**] Atiqullah M. M. (**2004**). An Efficient Simple Cooling Schedule for Simulated Annealing. *Computational Science and Its Applications – ICCSA 2004*. Vol. 3045 of the series Lecture Notes in Computer Science. pp. 396-404. Springer-Verlag. Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-540-24767-8_41

[**142**] Koulamas C., Antony S. R. and Jaen R. (**1994**). A survey of simulated annealing applications to operations research problems. *Omega*. Vol. 22 (1):41-56. http://dx.doi.org/10.1016/0305-0483(94)90006-X

[**143**] Eglese R. W. (**1990**). Simulated annealing: A tool for operational research. *European Journal of Operational Research*. Vol. 46 (3):271-281. http://dx.doi.org/10.1016/0377-2217(90)90001-R

[**144**] Dekkers A. and Aarts E. (**1991**). Global optimization and simulated annealing. *Mathematical Programming*. Vol. 50 (1-3):367-393. http://dx.doi.org/10.1007/bf01594945

[**145**] van Laarhoven P. J. M. and Aarts E. H. L. (**1987**). *Simulated Annealing: Theory and Applications*. Mathematics and Its Applications. (Vol. 37. np. 196) Springer Science+Business Media B.V, Netherlands. http://link.springer.com/book/10.1007%2F978-94-015-7744-1

[**146**] Corana A., Marchesi M., Martini C. and Ridella S. (**1987**). Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm. *Journal of ACM Transactions on Mathematical Software (TOMS)*. Vol. 13 (3):262-280. http://dx.doi.org/10.1145/29380.29864

[**147**] Lundy M. and Mees A. (**1986**). Convergence of an annealing algorithm. *Mathematical Programming*. Vol. 34 (1):111-124. http://dx.doi.org/10.1007/bf01582166

[**148**] Aarts E. H. L. and Korst J. H. M. (**1989**). *Simulated Annealing and Boltzmann Machines*. A Stochastic Approach to Combinatorial Optimization and Neural Computing. (np. 274) John Wiley & Sons Ltd., Great Britain.
http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471921467.html

[**149**] Granville V., Krivanek M. and Rasson J. P. (**1994**). Simulated annealing: a proof of convergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 16 (6):652-656. http://dx.doi.org/10.1109/34.295910

[**150**] Saleh Elmohamed M. A., Coddington P. and Fox G. (**1998**). A comparison of annealing techniques for academic course scheduling. *Practice and Theory of Automated Timetabling II: Second International Conference, PATAT'97*. Vol. 1408 of the series Lecture Notes in Computer Science. pp. 92-112. Springer-Verlag. Berlin, Heidelberg. http://dx.doi.org/10.1007/BFb0055883

[**151**] Ingber L. (**1993**). Simulated annealing: Practice versus theory. *Mathematical and Computer Modelling*. Vol. 18 (11):29-57. http://dx.doi.org/10.1016/0895-7177(93)90204-C

[**152**] Szu H. and Hartley R. (**1987**). Fast simulated annealing. *Physics Letters A*. Vol. 122 (3-4):157-162. http://dx.doi.org/10.1016/0375-9601(87)90796-1

[**153**] Tsallis C. and Stariolo D. A. (**1996**). Generalized simulated annealing. *Physica A: Statistical Mechanics and its Applications*. Vol. 233 (1-2):395-406.
http://dx.doi.org/10.1016/S0378-4371(96)00271-3

[**154**] Xiang Y. and Gong X. G. (**2000**). Efficiency of generalized simulated annealing. *Physical Review E*. Vol. 62 (3):4473-4476. http://dx.doi.org/10.1103/PhysRevE.62.4473

[**155**] Ingber L. (**1989**). Very fast simulated re-annealing. *Mathematical and Computer Modelling*. Vol. 12 (8):967-973. http://dx.doi.org/10.1016/0895-7177(89)90202-1

[**156**] Ingber L. (**1996**). Adaptive Simulated Annealing (ASA): lessons learned. *Control and Cybernetics*. Vol. 25 (1):33-54. http://arxiv.org/abs/cs/0001018

[**157**] Cheh K. M., Goldberg J. B. and Askin R. G. (**1991**). A note on the effect of neighborhood structure in simulated annealing. *Computers & Operations Research*. Vol. 18 (6):537-547. http://dx.doi.org/10.1016/0305-0548(91)90059-Z

[**158**] Yao X., Liu Y. and Lin G. (**1999**). Evolutionary programming made faster. *IEEE Transactions on Evolutionary Computation*. Vol. 3 (2):82-102. http://dx.doi.org/10.1109/4235.771163

[**159**] Courrieu P. (**1997**). The Hyperbell Algorithm for Global Optimization: A Random Walk Using Cauchy Densities. *Journal of Global Optimization*. Vol. 10 (1):37-55. http://dx.doi.org/10.1023/a:1008230212303

[**160**] Courat J. P., Raynaud G., Mrad I. and Siarry P. (**1994**). Electronic component model minimization based on log simulated annealing. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*. Vol. 41 (12):790-795. http://dx.doi.org/10.1109/81.340841

[**161**] Bilbro G. L. and Snyder W. E. (**1991**). Optimization of functions with many minima. *IEEE Transactions on Systems, Man, and Cybernetics*. Vol. 21 (4):840-849. http://dx.doi.org/10.1109/21.108301

[**162**] Han Y.-S., Snyder W. E. and Bilbro G. L. (**1990**). Pose determination using tree annealing. *Proceedings of IEEE International Conference on Robotics and Automation* pp. 427-432 vol.1. http://dx.doi.org/10.1109/ROBOT.1990.126014

[**163**] Romeo F. and Sangiovanni-Vincentelli A. (**1991**). A theoretical framework for simulated annealing. *Algorithmica*. Vol. 6 (1):302-345. http://dx.doi.org/10.1007/bf01759049

[**164**] Aarts E. H. L. and van Laarhoven P. J. M. (**1989**). Simulated annealing: An introduction. *Statistica Neerlandica*. Vol. 43 (1):31-52. http://dx.doi.org/10.1111/j.1467-9574.1989.tb01245.x

[**165**] Strenski P. N. and Kirkpatrick S. (**1991**). Analysis of finite length annealing schedules. *Algorithmica*. Vol. 6 (1):346-366. http://dx.doi.org/10.1007/bf01759050

[**166**] Potts C. N. and Van Wassenhove L. N. (**1991**). Single Machine Tardiness Sequencing Heuristics. *IIE Transactions*. Vol. 23 (4):346-354. http://dx.doi.org/10.1080/07408179108963868

[**167**] Yao X. (**1995**). A new simulated annealing algorithm. *International Journal of Computer Mathematics*. Vol. 56 (3-4):161-168. http://dx.doi.org/10.1080/00207169508804397

[**168**] Aarts E. H. L. and van Laarhoven P. J. M. (**1985**). A New Polynomial-Time Cooling Schedule. *Proceedings of IEEE International Conference on Computer-Aided Design* pp. 206-208. IEEE. Santa Clara.

[**169**] Youhua W., Weili Y. and Guansheng Z. (**1996**). Adaptive simulated annealing for the optimal design of electromagnetic devices. *IEEE Transactions on Magnetics*. Vol. 32 (3):1214-1217. http://dx.doi.org/10.1109/20.497462

[**170**] Cardoso M. F., Salcedo R. L. and de Azevedo S. F. (**1994**). Nonequilibrium Simulated Annealing: A Faster Approach to Combinatorial Minimization. *Industrial & Engineering Chemistry Research*. Vol. 33 (8):1908-1918. http://dx.doi.org/10.1021/ie00032a005

[**171**] Ortner M., Descombes X. and Zerubia J.; (**2007**). *An adaptive simulated annealing cooling schedule for object detection in images*. **Research Report** RR-6336. INRIA. inria-00181764, version 5. https://hal.inria.fr/inria-00181764/

[**172**] Sanvicente-Sánchez H. and Frausto-Solís J. (**2004**). A Method to Establish the Cooling Scheme in Simulated Annealing Like Algorithms. *Computational Science and Its Applications – ICCSA 2004: International Conference, Assisi, Italy, May 14-17, 2004, Proceedings, Part III*. pp. 755-763. Springer. Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-540-24767-8_80

[**173**] Monticelli A. J., Romero R. and Asada E. N. (**2008**). Fundamentals of Simulated Annealing. *Modern Heuristic Optimization Techniques:Theory and Applications to Power Systems*. pp. 123-146. Wiley-IEEE Press http://dx.doi.org/10.1002/9780470225868.ch7

[**174**] Frausto-Solis J., Román E. F., Romero D., Soberon X. and Liñán-García E. (**2007**). Analytically Tuned Simulated Annealing Applied to the Protein Folding Problem. *Computational Science – ICCS 2007: 7th International Conference, Beijing, China, May 27 - 30, 2007, Proceedings, Part II*. pp. 370-377. Springer. Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-540-72586-2_53

[**175**] Wah B. W. and Wang T. (**1999**). Simulated Annealing with Asymptotic Convergence for Nonlinear Constrained Global Optimization. *Principles and Practice of Constraint Programming – CP'99: 5th International Conference, Alexandria, VA, USA, October 11-14, 1999. Proceedings*. pp. 461-475. Springer. Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-540-48085-3_33

[**176**] Johnson D. S., Aragon C. R., McGeoch L. A. and Schevon C. (**1991**). Optimization by Simulated Annealing: An Experimental Evaluation; Part II, Graph Coloring and Number Partitioning. *Operations Research*. Vol. 39 (3):378-406. http://dx.doi.org/10.1287/opre.39.3.378

[**177**] Aarts E. H. L., Korst J. H. M. and van Laarhoven P. J. M. (**1997**). Simulated Annealing. *Local Search in Conbinatorial Optimization*. pp. 91-120. John Wiley and Sons. New York.

[**178**] Jacobson S. H. and Yücesan E. (**2004**). Analyzing the Performance of Generalized Hill Climbing Algorithms. *Journal of Heuristics*. Vol. 10 (4):387-405. http://dx.doi.org/10.1023/B:HEUR.0000034712.48917.a9

[**179**] Loper M. L. (**2015**). *Modeling and Simulation in the Systems Engineering Life Cycle: Core Concepts and Accompanying Lectures*. Simulation Foundations, Methods and Applications. 1 Ed. (np. XIX, 410) Springer-Verlag, London. http://dx.doi.org/10.1007/978-1-4471-5634-5

[**180**] Yilmaz L. (**2015**). *Concepts and Methodologies for Modeling and Simulation: A Tribute to Tuncer Ören*. Simulation Foundations, Methods and Applications. 1 Ed. (np. XV, 352) Springer International Publishing, Switzerland. http://dx.doi.org/10.1007/978-3-319-15096-3

[**181**] Sencer A. and Basarir-Ozel B. (**2013**). A simulation-based decision support system for workforce management in call centers. *SIMULATION*. Vol. 89 (4):481-497. http://dx.doi.org/10.1177/0037549712470169

[**182**] Chokshi R. (**1999**). Decision support for call center management using simulation. *Proceedings of Winter Simulation Conference, 1999* pp. 1634-1639 vol.2. IEEE. Phoenix, AZ. http://dx.doi.org/10.1109/WSC.1999.816903

[**183**] Gulati S. and Malcolm S. A. (**2001**). Call center scheduling technology evaluation using simulation. *Proceedings of Winter Simulation Conference, 2001* pp. 1438-1442 vol.2. IEEE. Arlington, VA. http://dx.doi.org/10.1109/WSC.2001.977467

[**184**] Bouzada M. A. C. (**2009**). Scenario Analysis within a Call Center Using Simulation. *Journal of Operations and Supply Chain Management*. Vol. 2 (1):89-103. http://bibliotecadigital.fgv.br/ojs/index.php/joscm/article/view/13936

[**185**] Wallace R. B. and Saltzman R. M. (**2005**). Comparing skill-based routing call center simulations using C programming and Arena models. *Proceedings of Winter Simulation Conference, 2005 (WSC 2005)* pp. 2636-2644. IEEE. http://dx.doi.org/10.1109/WSC.2005.1574563

[**186**] Mazzuchi T. A. and Wallace R. B. (**2004**). Analyzing skill-based routing call centers using discrete-event simulation and design experiment. *Proceedings of Winter Simulation Conference, 2004 (WSC 2004)* pp. 1812-1820 vol.2. IEEE. http://dx.doi.org/10.1109/WSC.2004.1371534

[**187**] Steinmann G. and De Freitas Filho P. J. (**2013**). Using simulation to evaluate call forecasting algorithms for inbound call center. *Proceedings of Winter Simulations Conference, 2013 (WSC 2013)* pp. 1132-1139. IEEE. Washington, DC. http://dx.doi.org/10.1109/WSC.2013.6721502

[**188**] Kim S.-M., Nah J.-E. and Kim S.-M. (**2011**). The Staffing Problem at the Call Center by Optimization and Simulation. *IE interfaces*. Vol. 24 (1):40-50. http://dx.doi.org/10.7232/IEIF.2011.24.1.040

[**189**] Bapat V. and Pruitte E. B. (**1998**). Using simulation in call centers. *Proceedings of Winter Simulation Conference, 1998 (WSC'98)* pp. 1395-1399. IEEE. http://dx.doi.org/10.1109/WSC.1998.746007

[**190**] Robinson S. (**2004**). *Simulation*. The Practice of Model Development and Use. (np. 339) John Wiley & Sons Ltd, England. http://eu.wiley.com//legacy/wileychi/robinson/

[**191**] Mathew B. and Nambiar M. K. (**2013**). A Tutorial on Modelling Call Centres using Discrete Event Simulation. *Paper of 27th European Conference on Modelling and Simulation (ECMS'13).* May 27-30, 2013. Aalesund University College. Alesund, Norway. http://www.scs-europe.net/dlib/2013/ecms13papers/ibs_ECMS2013_0072.pdf

[**192**] Akhtar S. and Latif M. (**2010**). Exploiting Simulation for Call Centre Optimization. *Proceedings of The World Congress on Engineering 2010 (WCE'10).* pp. 2112-2117. Vol. 3. International Association of Engineers (IAENG). London, U.K. http://www.iaeng.org/publication/WCE2010/WCE2010_pp2112-2117.pdf

[**193**] Tanir O. and Booth R. J. (**1999**). Call center simulation in Bell Canada. *Proceedings of Winter Simulation Conference, 1999.* pp. 1640-1647 vol.2. IEEE. Phoenix, AZ. http://dx.doi.org/10.1109/WSC.1999.816904

[**194**] Lewis B. G., Herbert R. D., Summons P. F. and Chivers W. J. (**2007**). Agent-based simulation of a multi-queue emergency services call centre to evaluate resource allocation. *Paper of MODSIM 2007*. . December 2007. pp. 11-17. International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand. http://www.mssanz.org.au/MODSIM07/papers/1_s5/Agent-basedSims5_Lewis.pdf

[**195**] Lewis B. G., Herbert R. D. and Chivers W. J. (**2010**). Modelling service levels in a call centre with an agent-based model. *World Review of Science, Technology and Sustainable Development*. Vol. 7 (1-2):2-12. http://dx.doi.org/10.1504/WRSTSD.2010.032339

[**196**] Park S. K. and Miller K. W. (**1988**). Random number generators: good ones are hard to find. *Magazine Communications of the ACM*. Vol. 31 (10):1192-1201. http://dx.doi.org/10.1145/63039.63042

[**197**] Hull T. E. and Dobell A. R. (**1962**). Random Number Generators. *SIAM Review*. Vol. 4 (3):230-254. http://dx.doi.org/10.1137/1004061

[**198**] Allard J. L., Dobell A. R. and Hull T. E. (**1963**). Mixed Congruential Random Number Generators for Decimal Machines. *Journal of the ACM (JACM)*. Vol. 10 (2):131-141. http://dx.doi.org/10.1145/321160.321163

[**199**] Lehmer D. H. (**1949**). Mathematical methods in large-scale computing units. *Proceedings of 2nd Symposium on Large-Scale Digital Calculating Machinery* pp. 141-146. Harvard University Press. https://archive.org/details/proceedings_of_a_second_symposium_on_large-scale_

[**200**] L'Ecuyer P. (**1999**). Tables of linear congruential generators of different sizes and good lattice structure. *Mathematics of Computation*. Vol. 68 (225):249-260. http://dx.doi.org/10.1090/S0025-5718-99-00996-5

[**201**] Steckley S. G., Henderson S. G. and Mehrotra V.; (**2004**). *Service System Planning in the presence of a Random Arrival Rate*. **Technical Report** 1416. Cornell University Operations Research and Industrial Engineering. NY. https://ecommons.cornell.edu/handle/1813/9291

[**202**] Jongbloed G. and Koole G. (**2001**). Managing uncertainty in call centres using Poisson mixtures. *Applied Stochastic Models in Business and Industry*. Vol. 17 (4):307-318. http://dx.doi.org/10.1002/asmb.444

[**203**] Robbins T. R. (**2007**). Addressing Arrival Rate Uncertainty in Call Center Workforce Management. *Proceedings of Service Operations and Logistics, and Informatics (SOLI 2007)* pp. 1-6. IEEE International. Philadelphia, USA. http://dx.doi.org/10.1109/SOLI.2007.4383934

[**204**] Liao S. Q. (**2011**). Staffing and shift-scheduling of call centers under call arrival rate uncertainty.Thesis of Docteur. 2011-07-01. Génie Industrielle. Ecole Centrale Paris.Francia. https://tel.archives-ouvertes.fr/tel-00635534

[**205**] Weinberg J., Brown L. D. and Stroud J. R. (**2007**). Bayesian Forecasting of an Inhomogeneous Poisson Process With Applications to Call Center Data. *Journal of the American Statistical Association*. Vol. 102 (480):1185-1198. http://dx.doi.org/10.1198/016214506000001455

[**206**] Whitt W. (**2006**). Fluid Models for Multiserver Queues with Abandonments. *Operations Research*. Vol. 54 (1):37-54. http://dx.doi.org/10.1287/opre.1050.0227

[**207**] Aguir S. M., Karaesmen F., Akşin O. Z. and Chauvet F. (**2004**). The impact of retrials on call center performance. *Operations Research Spectrum*. Vol. 26 (3):353-376. http://dx.doi.org/10.1007/s00291-004-0165-7

[**208**] Hoffman K. L. and Harris C. M. (**1986**). Estimation of a caller retrial rate for a telephone information system. *European Journal of Operational Research*. Vol. 27 (2):207-214. http://dx.doi.org/10.1016/0377-2217(86)90062-7

[**209**] Aguir M. S., Akşin O. Z., Karaesmen F. and Dallery Y. (**2008**). On the interaction between retrials and sizing of call centers. *European Journal of Operational Research*. Vol. 191 (2):398-408. http://dx.doi.org/10.1016/j.ejor.2007.06.051

[**210**] Ding S., Koole G. and van der Mei R. D. (**2015**). On the estimation of the true demand in call centers with redials and reconnects. *European Journal of Operational Research*. Vol. 246 (1):250-262. http://dx.doi.org/10.1016/j.ejor.2015.04.018

[**211**] Buist E. and L'Ecuyer P. (**2005**). A Java library for simulating contact centers. *Proceedings of the Winter Simulation Conference, 2005.* pp. 10. http://dx.doi.org/10.1109/WSC.2005.1574295

[**212**] Garnett O. and Mandelbaum A. (**2000**). *An Introduction to Skills-Based Routing and its Operational Complexities*.Teaching-Note.Service Engineering. Technion, Israel. Disponible en http://iew3.technion.ac.il/serveng/Lectures/SBR.pdf. Accedido: 01/07/2016.

[**213**] Stidham Jr. S. (**2002**). Analysis, Design, and Control of Queueing Systems. *Operations Research*. Vol. 50 (1):197-216. http://dx.doi.org/10.1287/opre.50.1.197.17783

[**214**] Larson R. C. (**1987**). OR Forum—Perspectives on Queues: Social Justice and the Psychology of Queueing. *Operations Research*. Vol. 35 (6):895-905. http://dx.doi.org/10.1287/opre.35.6.895

[**215**] Law A. M. and Kelton W. D. (**2000**). *Simulation Modeling and Analysis*. 3 Ed. McGraw-Hill, USA. http://www.mhhe.com/engcs/industrial/lawkelton/

[**216**] Sohrab H. H. (**2014**). *Basic Real Analysis*. Springer Science+Business Media. 2nd Ed. (np. XI, 683) Birkhäuser Basel, New York. http://www.springer.com/in/book/9781493918409