ESCALONAMENTO DE AGENTES EM *CALL CENTERS* RECEPTIVOS MULTILÍNGUES UTILIZANDO PROGRAMAÇÃO INTEIRA

Luiz Henrique Santanna Barbosa
Universidade Federal Fluminense, UFF
Departamento de Engenharia de Produção
luizhsb@gmail.com

Carlos Francisco Simões Gomes
Universidade Federal Fluminense, UFF
Departamento de Engenharia de Produção
cfsg1@bol.com.br

RESUMO

Call centers têm se tornado cada vez mais presentes no dia-a-dia das empresas e seus clientes. Aproximadamente 70% dos custos de um *call center* estão relacionados a custos com pessoal. Uma gestão eficiente permite com que os serviços por telefone sejam oferecidos com maior qualidade e menor custo. O presente trabalho tem como objetivo apresentar um modelo de programação inteira para resolver o problema do escalonamento de agentes em *call centers* com roteamento de chamadas baseado em habilidades, comparando-o com abordagens existentes. O programa determina a quantidade e combinação ótima de agentes e suas habilidades, levando em consideração custos diferenciados de acordo com as habilidades e as escalas dos agentes. De acordo com osexperimentos demonstrados neste estudo, foi possível observar economia significativa de custos com pessoal, sem prejuízo à qualidade do atendimento.

PALAVRAS CHAVE.Callcenter,Escalonamento,Programação inteira.

AD&GP - PO na Administração & Gestão da Produção, OA - Outras aplicações em PO, PM - Programação Matemática.

ABSTRACT

Call centers are an increasingly important part of today's business world. Approximately 70% of call center costs are personnel related. Efficient management helps centers provide high-quality services at low costs. This paper aims to present an integer programming model that solves shift scheduling problems in inbound call centers with skills-based routing. The program determines the optimal combination of skills and agents, taking into account varying costs according to the agents' skills and work shifts. Numerical experiments are run using a general purpose integer programming solver. As a result from these experiments, it is possible to observe significant reductions in personnel-related operational costs without negatively affecting the overall service level.

KEYWORDS.Call center,Scheduling,Integer programming.

AD & GP - OR in Administration & Production Management, OA - Other applications in OR, PM - Mathematical Programming.

1 Introdução

C*all centers* podem ser definidos como um conjunto de recursos, principalmente pessoas, computadores e equipamentos de telecomunicações, que permite o fornecimento de serviços através do telefone (Gans*et al.*, 2003). Os *call centers* normalmente são classificados pela origem da chamada. *Call centers* receptivos (*inbound*) são aqueles cuja ligação é originada pelo cliente final;*Call centers* ativos (*outbound*), as chamadas são originadas pela própria central, como em serviços de telemarketing (Reynolds, 2003).

Nas últimas décadas, os *call centers* têm se proliferado e crescido em importância para as empresas. Com o crescimento econômico brasileiro, cresce também a atenção das empresas ao relacionamento com seus clientes, e o *call center* é uma das principais ferramentas para isto (Reynolds, 2003). Estima-se que o setor tenha movimentado aproximadamente 29 bilhões de reais em 2011, com um crescimento de 9,5% em relação a 2010, ritmo que se mantém por três anos consecutivos (E-Consulting, 2012). Além disto, custos com pessoal em *call centers* chegam a representar 70% do custo total (Gans*et al.*, 2003), de forma que quaisquer reduções de custos operacionais podem ser extremamente relevantes. Através de uma gestão eficiente e da utilização de técnicas objetivas para o planejamento operacional do *call center*, é possível alcançar reduções significativas de custo, sem deixar de atender os requisitos de qualidade.

1.1 Propósito deste Artigo

O planejamento operacional de *call centers* é um processo composto por diversas atividades que são executadas de forma encadeada. Primeiro é realizada a previsão de chamadas para cada intervalo de tempo do dia, normalmente utilizando modelos de séries temporais com componentes de tendência e sazonalidade. Em seguida é feito o dimensionamento, que consiste em identificar, para cada um dos intervalos de tempo, a quantidade de agentes necessária para atender o volume de chamadas previsto dentro de um determinado nível de serviço. É comum a utilização de técnicas de simulação e teoria de filas, notavelmente Erlang C, nesta etapa. Na terceira etapa, denominada de escalonamento, é determinada a combinação ótima das escalas de trabalho de forma a cobrir a demanda esperada. Por fim, na quarta e última atividade, as escalas são atribuídas aos agentes específicos que estão disponíveis para o *callcenter*, podendo levar em consideração questões como preferências de turno dos agentes, intervalos, pausas, entre outras coisas (Koole, 2007; Stolletz, 2003).

Este artigo concentra-se na terceira etapa epropõe uma modelagem alternativa para o problema do escalonamento em *call centers* com roteamento de chamadas por habilidades. Ao final, é feita uma análise dos resultados em um estudo de caso.

2 Roteamento de Chamadas por Habilidades

A quantidade de serviços oferecidos por telefone tem crescido de maneira acentuada, de forma que, para a maioria dos *call centers*, não é possível que todos os agentes atendam qualquer tipo de ligação. Normalmente, o sistema de telefonia captura informações sobre a chamada para poder determinar quais agentes podem atender a ligação e direcioná-la corretamente.

Em *call centers* tradicionais, os agentes são separados em departamentos, onde cada departamento é responsável por um ou mais tipos de ligação. Os departamentos são totalmente independentes, e cada tipo de ligação pode ser resolvido por um único departamento. Os agentes pertencem a um único departamento (Reynolds, 2003). Nestes *call centers*, o planejamento operacional é relativamente simples.No entanto, esta estrutura pode se demonstrar ineficiente. Muitas vezes, os horários de maior volume de chamadas são diferentes para cada tipo de ligação, podendo haver uma taxa de ociosidade alta de agentes. Além disto, como os departamentos são independentes e dedicados a um tipo específico de ligações, o tamanho de cada departamento é relativamente pequeno, como se cada departamento fosse um *callcenter* independente. Se estes

departamentos fossem combinados, o número de agentes por departamento seria maior e, com isto, haveria redução de custo por economia de escala (Gans*et al*, 2003; Koole, 2007).

Call centers modernos fazem uso do roteamento de chamadas baseado em habilidades, ou *skill-basedrouting* (SBR). Nestas centrais, os agentes são classificados quanto às habilidades que eles possuem e as chamadas são roteadas para qualquer agente que possua a habilidade necessária para atendê-la (Wallace e Whitt, 2005). Isto permite maior flexibilidade para lidar com variações no volume de chamadas, nivelar a qualidade do serviço entre os tipos de ligação, além de aumentar a quantidade de agentes disponíveis, resultando em reduções de custo. Porém, a complexidade do processo de planejamentoaumenta consideravelmente, pois, em geral, quanto maior o número de habilidades de um agente, maior também será o seu custo. Encontrar a composição ótima de agentes e suas habilidades é o resultado esperado do processo de escalonamento em *call centers* com SBR (Koole, 2007).

3 Escalonamento de Agentes em Call Centers

Diversos trabalhos tratam a questão do escalonamento, porém ainda há relativamente poucos que tratam múltiplas habilidades. Dantzig (1954) fornece um modelo de escalonamento através de programação inteira que pode ser aplicado em *call centers*simples. Pinedo (2005) estende este modelo para incorporar penalidades para o não cumprimento de demandas e para o excesso de agentes. Thompson (1997), Atlason*et al.* (2004) e Koole (2010) apresentam modelos onde, além de um nível de serviço mínimo, há também um nível de serviço agregado. No entanto, não levam em consideração as habilidades dos agentes.

Cordone*et al.* (2011) apresentam uma heurística onde é feita a relaxação de um programa inteiro e as variáveis são manipuladas iterativamente até que se atinja um resultado inteiro. O algoritmo leva em consideração as habilidades dos agentes e incorpora pausas para descanso, além de balancear os tipos de contrato de trabalho (período integral ou parcial). No entanto, o algoritmo não leva em consideração custos diferenciados de acordo com as habilidades dos agentes. Além disto, o programa recebe como entrada a fração com que um agente de cada perfil contribui para o suprimento da demanda de cada tipo de chamada, em cada intervalo de tempo. Entendemos que este dado deveria ser resultado da otimização, e não um dado de entrada.

Cezik e L'Ecuyer (2008) apresentam algoritmo para dimensionamento de *call centers* com múltiplas habilidade através de simulação e programação linear, sem, porém, resolver o problema do escalonamento. Bhulai*et al.*(2008) propõem algoritmo para escalonamento com múltiplas habilidades em duas fases: na primeira, utilizam o modelo de Cezik e L'Ecuyer (2008) para determinar as quantidades ótimas de agentes para cada perfil; na segunda, utilizam um programa inteiro para resolver o problema do escalonamento.O modelo proposto é resultado de oportunidades de melhoria identificadas no trabalho de Bhulai*et al*. (2008).

3.1 Escalonamento paraCall Centers Simples

O modelo de Dantzig(1954) básico para escalonamento de*call centers*simples não leva em consideração segmentação por habilidades. Como dados de entrada, são fornecidos: o conjunto$K$ possíveis escalas, o conjunto$T$de intervalos de tempo,$c_k$com o custo de um agente com a escala k es$_t$com a quantidade de agentes necessários no intervalo de tempo t. Por fim, a matriz $a_{k,t}$ é definida da seguinte forma:

$$a_{k,t} = \begin{cases} 1, \text{se na escala k estiver disponível no tempo t} \\ 0, \text{caso contrário} \end{cases} \quad (1)$$

O conjunto de variáveis de decisão x_k indica a quantidade de agentes alocados à escala k. O programa utiliza restrições rígidas para garantir que em cada intervalo de tempo, a quantidade de agentes disponíveis, isto é, alocados em escalas que estejam em serviço naquele intervalo de

tempo, seja igual ou superior à quantidade de agentes requeridos. O objetivo é minimizar o custo total das escalas. A formulação completa é dada a seguir:

$$\min \sum_{k \in K} c_k x_k \quad (2)$$

sujeito a

$$\sum_{k \in K} a_{k,t} x_k \geq s_t \forall t \in T \quad (3)$$

$$x_k \geq 0 \text{ e inteiro}, \quad \forall k \in K \quad (4)$$

3.2 Escalonamento para Call Centers com Múltiplas Habilidades

Nas abordagens analisadas, o modelo de Dantzig (1954) é expandido para incorporar múltiplas habilidades. Para fins de simplicidade, considera-se que, para cada tipo de ligação há apenas uma habilidade que um agente deverá possuir para que seja capaz de atender chamadas daquele tipo. Os perfis são as combinações de habilidades que sejam técnica e economicamente viáveis. Desta forma, os perfis é que são atribuídos aos agentes e não as habilidades, garantindo que agentes possuam apenas as combinações de habilidades viáveis.

No modelo de escalonamento multi-habilidades, mantém-se os conjuntos *K* e *T* e a matriz $a_{k,t}$, de forma semelhante ao modelo acima. Acrescenta-se o conjunto *H* com as habilidades existentes no *callcenter* e o conjunto *P* com os perfis de agentes.

O custo deve variar não apenas em função da escala, mas também do perfil do agente, e é dado pela matriz $c_{k,p}$. Isto permite, por exemplo, que um agente com mais habilidades, ou uma especialidade mais restrita, possua um custo diferenciado dos demais agentes. A variável de decisão também deve ser modificada: $x_{k,p}$ informa a quantidade de agentes com o perfil *p*alocados à escala *k*. Com isto, chega-se à função objetivo (5). Esta mesma função objetivo é utilizada nas duas formulações descritas a seguir. As restrições, porém, diferem em cada formulação e serão descritas de forma completa nas respectivas seções.

$$\min \sum_{k \in K} \sum_{p \in P} c_{k,p} x_{k,p} \quad (5)$$

A principal modificação no modelo está na restrição para o atendimento da demanda. O programa não deve permitir que um agente seja contabilizado para suprir a demanda de tipos de ligação para os quais não possua habilidade e nem que, no mesmo intervalo de tempo, um agente com múltiplas habilidades seja contabilizado para suprir a demanda de mais de um tipo de ligação. Porém, deve permitir que osagentes supram a demanda de tipos de ligações diferentes, desde que em intervalos de tempo diferentes e que possuam a habilidade necessária. O resultado deve garantir que em tempo de execução, seja possível realizar a atribuição ilustrada pelaFigura 1em cada intervalo de tempo:

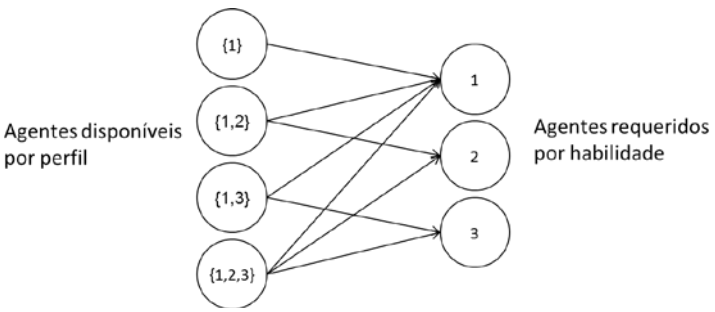

Figura 1: Atribuição de agentes por perfil às demandas, por tipo de ligação.

3.2.1 Formulação de Bhulai*et al.*(2008)

No trabalho de Bhulai*et al.* (2008) é apresentado um modelo de programação inteira para tratar o escalonamento em *call centers* com múltiplas habilidades. A primeira parte do algoritmo onde é feito o dimensionamento não será analisada, pois não faz parte do escopo do presente trabalho. Além disto, forma de apresentação das variáveis foi modificada, para manter a consistência com os conceitos utilizados neste trabalho.

Neste modelo, em cada intervalo de tempo é feita uma atribuição para determinar a quantidade de agentes que estarão alocados a cada tipo de ligação. Esta atribuição é feita através das variáveis de decisão $y_{t,p,h}$, que indica a quantidade de agentes com perfil *p* que no intervalo de tempo *t* suprem a demanda por agentes com habilidade *h*. Além das variáveis de decisão $x_{k,p}$ e dos custos dos agentes, $c_{k,p}$, já mencionados na seção anterior, é introduzida ainda a matriz $b_{p,h}$, que indica com 1 caso agentes com perfil *p* possuam a habilidade *h*, ou 0 caso contrário. Também introduzidos os elementos:

- *P:* conjunto de perfis de agentes existentes;
- *T:* intervalos de tempo considerados no horizonte de planejamento;
- *H:* conjunto com as habilidades existentes;
- $s_{t,h}$: Quantidade de agentes com a habilidade *h*requeridos no intervalo de tempo *t*.

Por fim, é mantida a matriz $a_{k,t}$ do modelo de Dantzig (1954), que indica a disponibilidade de um agente com a escala *k* no tempo *t*, conforme descrito na seção anterior.

A seguir, a formulação adaptada do trabalho de Bhulai*et al* (2008). A função objetivo (5) é a mesma apresentada anteriormente, acrescentada das restrições (6)a(9).

$$\min \sum_{k\in K}\sum_{p\in P} c_{k,p}x_{k,p} \quad (5)$$

sujeito a

$$\sum_{p\in P} b_{p,h}y_{t,p,h} \geq s_{t,h} \forall t \in T, \forall h \in H \quad (6)$$

$$\sum_{k\in K} a_{k,t}x_{k,p} = \sum_{h\in H} b_{p,h}y_{t,p,h} \; \forall t \in T, \forall p \in P \quad (7)$$

$$x_{k,p} \geq 0 \text{ e inteiro}, \quad \forall k \in K, \forall p \in P \quad (8)$$

$$y_{t,p,h} \geq 0 \text{ e inteiro}, \quad \forall t \in T, \forall p \in P, \forall h \in H \quad (9)$$

A restrição (6) garante que a demanda por agentes para cada habilidade é suprida em cada intervalo de tempo. A restrição (7)é responsável por alocar os agentes disponíveis em cada perfil a uma habilidade específica, de forma que, em cada intervalo de tempo, um agente seja contabilizado para suprir a demanda de apenas uma habilidade.

Assim, em cada intervalo de tempo, é realizada uma atribuição de agentes de um perfil ao tipo de chamada que estará atendendo. Esta atribuição tem como único propósito assegurar a consistência no suprimento da demanda, garantindo que um agente com mais de uma habilidade seja contabilizado para suprir a demanda de apenas um tipo de ligação em cada intervalo de tempo. Para o planejamento operacional, esta informação é irrelevante, tendo em vista que o roteamento das chamadas éfeito em tempo de execução pelo sistema de telefonia (Koole e Pot, 2006). Desta forma, as variáveis de decisão $y_{t,p,h}$ não possuem utilidade prática, a não ser assegurar a consistência dos dados. Apesar de ter apresentado bons tempos de execução (Bhulai*et al.*, 2008), o número variáveis cresce de forma acelerada ao acrescentar mais tipos de chamadas, perfis de agentes ou ao utilizar um horizonte maior de planejamento.

3.2.2 Formulação Proposta

Com isto em vista, buscou-se aprimorar este modelo reduzindo o número de variáveis e, consequentemente, o tempo de execução. O modelo proposto utiliza a mesma função objetivo (5), apresentada anteriormente. No entanto, utiliza outra abordagem para garantir com que a demanda seja atendida pelos agentes disponíveis, de acordo com as habilidades que possuem. A formulação proposta baseia-se na premissa de que não há necessidade de realizar a atribuição de agentes de um perfil à habilidade cuja demanda estará suprindo. Uma vez que esta é uma decisão dinâmica tomada em tempo execução pelo sistema de telefonia, basta garantir que esta atribuição seja possível. Para isto, introduzimos a seguinte função:

$$f(L,p) = \begin{cases} 1, \text{ se o perfil } p \text{ possui pelo menos uma habilidade do conjunto } L \\ 0, \text{ caso contrário} \end{cases} \quad (10)$$

Em seguida, complementa-se o modelo com as restrições que garantem que a demanda de cada tipo de ligação seja atendida.

$$\sum_{k\in K}\sum_{p\in P} f(L,p)\,a_{k,t}\,x_{k,s} \geq \sum_{h\in L} s_{t,h} \;\forall t \in T, \forall L \subseteq H \quad (11)$$

$$x_{k,p} \geq 0 \text{ e inteiro}, \quad \forall k \in K, \forall p \in P \quad (12)$$

A matriz $a_{k,t}$, já descrita no modelo de Dantzig (1954) e na formulação baseada em Bhulai*et al* (2008), é utilizada também nesta formulação, bem como a matriz $s_{t,h}$, que indica a demanda de agentes com habilidade *h* no intervalo de tempo *t*. Os custos, por escala e perfil de agente, são dados pela matriz $c_{k,p}$ e a mesma variável de decisão, $x_{k,p}$, é utilizada, indicando a quantidade de agentes com perfil *p* que devem ser alocados à escala *k*.

A restrição (11) baseia-se no teorema de Gale (1957) para garantir que a atribuição seja possível. Nela, garantimos que para cada subconjunto *L*do conjunto de habilidades *H*, a quantidade de agentes disponíveis que possuam ao menos uma das habilidades de *L* seja igual ou superior à soma das demandas das habilidades contidas em *L*. A seção seguinte descreve em maior detalhe o teorema de Gale (1957) e a sua aplicabilidade no problema em questão.

A formulação completa é dada pela função objetivo (5) e as restrições (11) e (12):

$$\min \sum_{k\in K}\sum_{p\in P} c_{k,p}\,x_{k,p} \quad (5)$$

sujeito a

$$\sum_{k\in K}\sum_{p\in P} f(L,p)\,a_{k,t}\,x_{k,s} \geq \sum_{h\in L} s_{t,h} \;\forall t \in T, \forall L \subseteq H \quad (11)$$

$$x_{k,p} \geq 0 \text{ e inteiro}, \quad \forall k \in K, \forall p \in P \quad (12)$$

Assim, obtém-se um modelo relativamente simples, capaz de resolver o problema do escalonamento multi-habilidades com um tempo computacional bem pequeno (segundos ou frações de segundos), de acordo com a quantidade de escalas, de habilidades e de perfis.

3.2.3 Prova de Corretude do Modelo Proposto

Conforme visto anteriormente, Bhulai*et al* (2008) modelam o problema de escalonamento de agentes em *call centers* com múltiplas habilidades como problemas menores de fluxo de redes, ou mais especificamente de atribuição, em cada intervalo de tempo. Em cada intervalo de tempo é feita a atribuição dos agentes disponíveis em cada perfil à demanda por cada

habilidade. Todavia, também observamos que, para fins de planejamento, esta atribuição não tem valor prático; basta garantir que seja possível realiza-la em tempo de execução.

Gale (1957) demonstra o seguinte teorema para verificar a viabilidade de um problema de fluxo de redes:

Teorema. Um problema de fluxo de redes terá solução se, e somente se, para todo subconjunto *S* de *N*, a soma das demandas dos nós pertencentes a *S* seja inferior à soma das capacidades dos arcos com direção aos elementos de *S*.

Para ilustrar este teorema e sua aplicação neste problema, será utilizado um exemplo simples em um *callcenter* com duashabilidades e três perfis diferentes, conforme a Figura 2. Na notação adotada, cada perfil indica entre chaves as habilidades que ele possui. Existe uma demanda de 30 agentes com a habilidade 1 e 50 agentes com a habilidade 2, representada pelos arcos com origem nos nós de habilidade. Os arcos possuem restrição quanto à sua capacidade, indicada acima de cada arco. Os arcos com origem nos nós de perfil e destino nos nós de habilidades têm capacidade infinita. Já os arcos com destino nos nós de perfil possuem restrição quanto à capacidade máxima de 20, 30 e 30 para os perfis {1}, {1, 2} e {2}, respectivamente.

Aplicando o teorema de Gale (1957), é possível verificar a viabilidade de resolução deste problema. Para isto, selecionam-se os subconjuntos de *N* a fim de verificar que a soma das demandas de seus elementos seja menor ou igual à capacidade dos arcos entrantes, destacados em vermelho na Figura 2. Os subconjuntos contendo apenas nós de habilidade serão ignorados, tendo em vista que seus arcos entrantes possuem capacidade infinita. O leitor poderá observar neste exemplo que, para todos os subconjuntos *S*, a soma dos arcos que saem do subconjunto, representando a demanda, é menor ou igual à soma das capacidades dos arcos entrantes, representando os agentes disponíveis. Logo, de acordo com este teorema, o problema é viável.

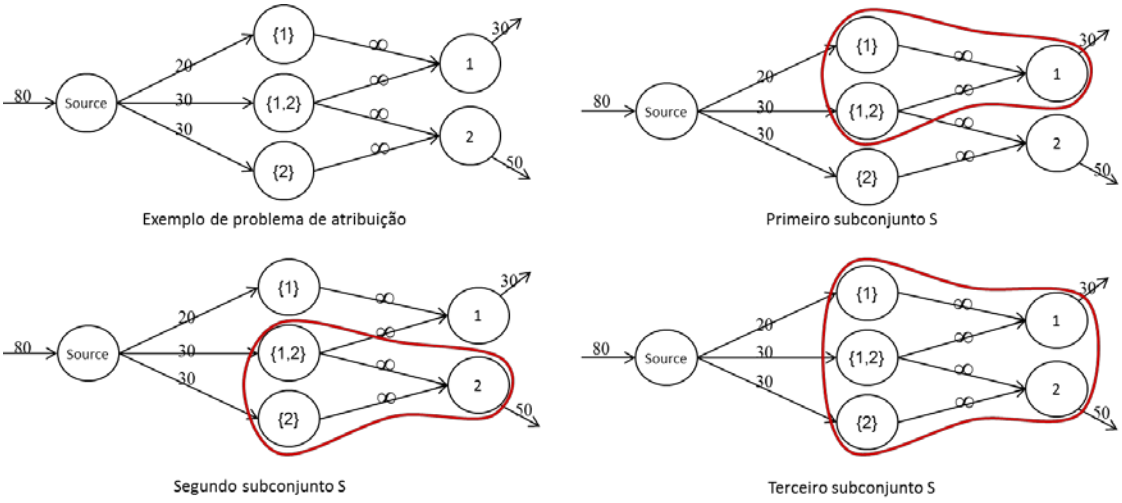

Figura 2: Exemplo de fluxo de redes de *callcenter* com duas habilidades e três perfis de agentes.

De modo mais geral, este raciocínio mostra que para que um problema com duas habilidades com demandas d_1 e d_2, para que exista solução as disponibilidades de agentes $c_{\{1\}}$, $c_{\{2\}}$ e $c_{\{1,2\}}$ devem satisfazer as seguintes restrições:

$$c_{s,\{1\}} + c_{s,\{1,2\}} \geq d_1 \quad (13)$$

$$c_{s,\{1,2\}} + c_{s,\{2\}} \geq d_2 \quad (14)$$

$$c_{s,\{1\}} + c_{s,\{1,2\}} + c_{s,\{2\}} \geq d_1 + d_2 \quad (155)$$

Generalizando, temos que é possível realizar uma atribuição de agentes para habilidades se, para cada subconjunto *L* do conjunto de habilidades *H*, a soma dos agentes disponíveis com

perfis compatíveis (isto é, com pelo menos uma habilidade do conjunto*L*) deve ser superior ou igual à soma das demandas das habilidades de *L*. A formulação (5, 11, 12) proposta neste trabalho baseia-se neste princípio para garantir que,em cada intervalo de tempo, a atribuição de perfis de agentes às habilidades que os agentes estarão alocados seja possível.

4 Estudo de Caso

O estudo foi conduzido em *callcenter* brasileiro de médio porte, com serviço de atendimento a clientes que recebe chamadas nos idiomas português, inglês e espanhol. Estes, portanto, são os possíveis tipos de chamadas existentes e, para que estas chamadas possam atendidas, é necessário que o agente possua habilidade de conversação no respectivo idioma.O *callcenter* estudado possui os seguintes perfis de agentes:

- Monolíngue: Apenas chamadas em português. Custo médio: R$800,00.
- Bilíngue-inglês: Chamadas em português e inglês. Custo médio: R$1.200,00.
- Bilíngue-espanhol:Chamadas em português e espanhol. Custo médio: R$1.200,00.
- Trilíngue: Atende chamadas em todos os idiomas. Custo médio: R$1.600,00.

Neste *callcenter*, a previsão através de sistema de informação e o volume de chamadas previsto é exportado para planilha eletrônica. A previsão é feita em intervalo intra-diário de 30 minutos. Uma vez feita a previsão, o dimensionamento é feito em planilha eletrônica com implementação da fórmula Erlang C. O horário de serviço do *call center* é de 8h00min às 20h00min, havendo dois turnos de 6 horas de trabalho, um começando às 8h00min, e outro começando às 14h00min. Após o dimensionamento, a gestão da empresa observa o intervalo de tempo com maior demanda de agentes e utiliza esta informação para determinar a quantidade de agentes que deverá existir em cada turno. A demanda média por intervalo de tempo foi fornecida pelo *callcenter* e pode ser observada na Tabela 3.

Existem algumas limitações importantes que puderam ser observadas neste modelo de gestão. A primeira decorre de que há poucos turnos de trabalho, e é necessário um número maior de agentes para que seja possível cobrir os horários de pico de demanda durante a operação do *callcenter*. Caso a empresa utilizasse turnos começando a cada hora, seria possível organizar a escala dos agentes para atender a demanda dos horários de pico com menos agentes no quadro geral de funcionários. A segunda limitação é o fato de o planejamento ser feito de forma independente para cada tipo de ligação. Ou seja, na prática, os benefícios de se ter um *callcenter* segmentado por habilidades não são aproveitados, de forma que agentes bilíngues com habilidades de conversação em inglês e português são contabilizados apenas para suprir a demanda de agentes com habilidades em inglês. Como regra geral, agentes trilíngues não são utilizados pelo fato de terem o salário maior do que os demais. No entanto, em determinadas situações, é possível que uma solução ótima inclua agentes trilíngues.

ATabela 1 resume a quantidade de agentes por perfil e por escala necessários para suprir a demanda prevista pelo modelo de planejamento original utilizado pelo *callcenter*.

Perfil / Escala	Monolíngues	Bilíngues (Inglês)	Bilíngues (Espanhol)	Trilíngues
08:00 – 14:00	482	106	33	0
14:00 – 20:00	430	110	23	0
Total:	912	216	56	0

Tabela 1: Quantidade de agentes por escala e conjunto de habilidades antes da otimização

O custo total com salários de agentes, com base no salário médio por perfil de agente, é de R$1.056.000,00. No entanto, após a execução do programa, é possível uma redução

considerável do total de agentes e do custo final, sem deixar de atender a demanda, conforme será descrito nas seções seguintes.

O programa foi implementado e resolvido pelo pacote computacional IBM ILOG CPLEX Optimizer (IBM, 2010), utilizando sempre um *thread*de execução. Também foi utilizada a biblioteca UFFLP (UFFLP, 2012) como interface deimplementação dos programas. A UFFLP é uma biblioteca que fornece uma interface para implementação de modelos de programação inteira e mista, simplificando a integração com resolvedores (Pessoa e Uchoa, 2011). O programa foi executado em computador com 3.1GHz de processamento, 4GB de memória RAM e sistema operacional Microsoft Windows 7 com 64 bits (Microsoft, 2012).

4.1 Aplicação do Modelo

Na primeira instância testada, criaram-se sete escalas, começando de hora em hora, a partir das 8h00min até às 14h00min e desprezando-se as pausas e intervalos. Foram utilizados os mesmos dados de demanda do modelo original, com intervalos de tempo intra-diários de 30 minutos. Em ambas as formulações, os resultados, em termos numéricos, foram equivalentes e o problema foi resolvido em 0,01 segundos, gerando resultado conforme Tabela 2.

Escala \ Perfil	Monolíngues	Bilíngues (Inglês)	Bilíngues (Espanhol)	Trilíngues
08:00 – 14:00	437	83	28	0
09:00 – 15:00	9	13	3	0
10:00 – 16:00	8	0	0	0
11:00 – 17:00	14	9	2	0
12:00 – 18:00	166	41	7	0
13:00 – 19:00	39	17	5	0
14:00 – 20:00	193	30	6	0
Total:	866	193	52	0

Tabela 2: Quantidade de agentes por escala e conjunto de habilidades após otimização

O custo total da solução encontrada pelo pacote de otimização foi de R$985.600,00. Ao comparar com o plano de escalas utilizado pela empresa, observa-se uma economia de aproximadamente 7% com o pessoal. Os motivos para a economia de escala podem ser atribuídos em primeiro lugar ao maior número de possibilidades de escalas, uma vez que os agentes podem iniciar o serviço a cada hora, e não apenas em dois turnos distintos. Além disto, conforme a análise a seguir, o modelo também foi capaz de incorporar a flexibilidade da segmentação por habilidades no planejamento, reduzindo a quantidade de agentes. A Tabela 3 compara a demanda intra-diária com a quantidade de agentes disponíveis em cada intervalo de tempo.

Intervalo de Tempo	Demanda			Agentes Disponíveis - Modelo Original			Agentes Disponíveis - Após Otimização		
	Português	Inglês	Espanhol	Mono-língue	Bilíngue (Inglês)	Bilíngue (Espanhol)	Mono-língue	Bilíngue (Inglês)	Bilíngue (Espanhol)
08:00 - 08:30	268	22	3	482	106	33	437	83	28
08:30 - 09:00	332	23	5	482	106	33	437	83	28
09:00 - 09:30	389	32	9	482	106	33	446	96	31
09:30 - 10:00	450	39	10	482	106	33	446	96	31
10:00 - 10:30	466	96	14	482	106	33	454	96	31
10:30 - 11:00	474	92	15	482	106	33	454	96	31
11:00 - 11:30	482	100	24	482	106	33	468	105	33
11:30 - 12:00	459	105	33	482	106	33	468	105	33
12:00 - 12:30	375	96	25	482	106	33	634	146	40
12:30 - 13:00	364	105	25	482	106	33	634	146	40

13:00 - 13:30	390	106	23	482	106	33	673	163	45
13:30 - 14:00	418	100	20	482	106	33	673	163	45
14:00 - 14:30	430	110	22	430	110	23	429	110	23
14:30 - 15:00	419	106	23	430	110	23	429	110	23
15:00 - 15:30	403	96	20	430	110	23	420	97	20
15:30 - 16:00	428	91	18	430	110	23	420	97	20
16:00 - 16:30	412	97	20	430	110	23	412	97	20
16:30 - 17:00	369	84	15	430	110	23	412	97	20
17:00 - 17:30	338	83	18	430	110	23	398	88	18
17:30 - 18:00	286	76	17	430	110	23	398	88	18
18:00 - 18:30	232	47	11	430	110	23	232	47	11
18:30 - 19:00	204	37	10	430	110	23	232	47	11
19:00 - 19:30	193	30	6	430	110	23	193	30	6
19:30 - 20:00	189	29	4	430	110	23	193	30	6

Tabela 3: Demanda x Agentes Disponíveis por Intervalo de Tempo após Otimização

Na Tabela 3, é possível observar que existe um nível considerável de ociosidade no plano de escalas original, principalmente nos intervalos de tempo próximos ao início e ao término do serviço. No plano de escalas gerado pelo programa, esta ociosidade é diminuída consideravelmente. Merece destaque tambémo fato do modelo incorporar a flexibilidade das múltiplas habilidades no escalonamento,gerando resultados melhores do que a abordagem tradicional. Pode-se tomar como exemplo o intervalo entre 15h30min e 16h00min: a demanda para agentes com capacidade de atendimento em português é de 428 agentes. Porém, há apenas 420 agentes monolíngues disponíveis. O programa faz uso do excedente de bilíngues neste intervalo—6 agentes bilíngues-inglês e 2 agentes bilíngues-espanhol—para atender a demanda das chamadas em português.

Em seguida, foi realizado novo teste do modelo, a fim de incorporar restrições trabalhistas para obter resultados ainda mais precisos. A legislação para o setor determina que em turnos de 6 horas diárias haja duas pausas de 10 minutos, a primeira após a primeira hora de trabalho e a segunda antes da última. Ambas as pausas contam como tempo de trabalho. Deve haver ainda um intervalo de 20 minutos entre as duas pausas, que não conta como tempo de trabalho (Brasil, 2007).

Para o segundo teste, a demanda intra-diária foi então dividida em intervalos de 15 minutos. As pausas de 10 minutos foram arredondadas para 15 minutos e as pausas de 20 minutos para 30. Este arredondamento é razoável, pois, na prática, há o tempo de preparação para o atendimento que não é contado como parte da pausa. Desta forma, as escalas não são mais consideradas como períodos ininterruptos de trabalho. Assim, as escalas possuem intervalos de tempo dentro do período de expediente em que o agente não estará de serviço. Salienta-se que uma escala com início às 8:00 terá seu término às 14:20, e não as 14:00, como era considerado anteriormente, uma vez que o intervalo de 20 minutos entre as pausas não é contado como parte do expediente.Foram geradas 275 combinações diferentes de pausas e intervalos seguindo as regras da legislação e, a partir destas combinações, foram geradas 1.925 escalas diferentes, iniciando de hora em hora, das 8h00min às 14h00min.

Os resultados foram bastante semelhantes ao do experimento anterior, em termos de ordem de grandeza quanto ao número de agentes necessários e à consistência da solução, garantindo com que toda a demanda fosse atendida por um agente com a habilidade necessária. A tabela com o detalhamento dos resultados é omitida a fim de manter a brevidade do trabalho.

Para esta instância, o pacote de otimização conseguiu encontrar a solução ótima após 4,21 segundos na primeira formulação, baseada no trabalho de Bhulai*et al* (2008), e 0,89 segundos para a formulação proposta neste trabalho. Isto sugere que o modelo proposto apresenta um tempo de execução melhor, mesmo para uma instância consideravelmente maior. Como o número de restrições da formulação proposta cresce de forma exponencial, de acordo com o número de habilidades existentes, o modelo proposto tende a ter desempenho melhor principalmente em instâncias com grande número de intervalos de tempo e relativamente poucas habilidades.

O custo total da solução foi de R$962.000,00, ainda menor do que o do primeiro experimento, representando uma economia de quase 9% em relação ao custo atual. Isto ocorre pois, apesar da existência de pausas e intervalos ao longo do expediente do agente, a duração é maior, devido ao fato do intervalo não contar como tempo de trabalho. O programa conseguiu acomodar as pausas e intervalos dos agentes, dentro do universo de combinações existentes, nos momentos de menor demanda, de forma que não foi necessário alocar mais agentes para cobrir estas pausas. A Tabela 4 apresenta o resumo das instâncias testadas os resultados obtidos.

Instância	Qtd. Escalas	Qtd. Intervalos de Tempo	Tempo de Execução (segundos)		Custo total
			Formulação Bhulai*et al.* (2008)	Formulação Proposta	
Modelo Original	2	12	-	-	R$1.056.000,00
Instância 1	7	24	0,01	0,01	R$985.600,00
Instância 2	1.925	48	4,21	0,89	R$962.000,00

Tabela 4: Resumo dos experimentos.

5 Considerações Finais e Conclusões

Como resultado do presente trabalho, verificou-se que o modelo proposto foi capaz de resolver de forma eficiente o problema do escalonamento de agentes de *call centers* com múltiplas habilidades para as instâncias testadas. Sendo assim, a abordagem proposta apresenta uma alternativa a ser considerada, pois é capaz de incorporar a flexibilidade de *call centers* com SBR em um modelo simples de ser implementado.

Algumas ressalvas devem ser feitas. O modelo presume que um agente generalista consegue realizar um atendimento tão bem, isto é, com o mesmo tempo médio de atendimento, que um agente especialista para um dado tipo de chamada. No caso do *callcenter*estudado, esta premissa se demonstrou verdadeira, pois os únicos agentes especialistas da central de atendimento são monolíngues com habilidade para atendimento em português. Os agentes generalistas, bilíngues e trilíngues, são de nacionalidade brasileira, de forma que o nível de habilidade deles para atendimento em português é equivalente aos monolíngues.

Além disto, é necessária a execução de testes em instâncias maiores, com mais combinações de escalas, para confirmar a eficiência do modelo. Contudo, optamos pela metodologia de estudo de caso devido ao trabalho apresentar elementos de caráter inédito, o que limita o número de instâncias testadas, porém permite uma análise mais detalhada dos resultados. Não obstante, o trabalho de Wallace e Whitt (2005) sugere que não é eficiente a criação de perfis de agentes com mais de duas habilidades, de forma que, ainda que sejam necessários testes em instâncias maiores, o número de combinações de habilidades continua sendo relativamente baixo.

O modelo proposto, bem como o apresentado por Bhulai*et al*. (2008), utiliza uma abordagem baseada em cobertura de conjuntos e existe ampla literatura atestando que estes problemas são bem resolvidos pelos pacotes de otimização existentes no mercado. Sendo assim, o resultado da pesquisa apresenta uma nova alternativa na utilização da pesquisa operacional para a resolução de problemas de gestão de *call centers*.

ReferênciasBibliográficas

Atlason, J.; Epelman, M.; Henerson, S.(2004), Call center staffing with simulation and cutting plane methods. *Annals of Operations Research*, vol. 127, p. 333–358.

Bhulai, S; Koole, G.; Pot, A.(2008), Simple Methods for Shift Scheduling in Multi-Skill Call Centers.*Manufacturing & Services Operations Management* 10, p. 411-420.

Brasil. Ministério do Trabalho e Emprego. Portaria SIT n.º 13, de 21 de junho de 2007 - NR 07. Altera Norma Regulamentadora 17 - Ergonomia. *Diário Oficial da União*, Brasília, 02 de abr de 2007.

Cezik, M.; L'Ecuyer, P.(2008), Staffingmultiskill call centers via linear programming and simulation.*Management Science*, v. 54, n. 2, p. 310-323.
Cordone, R.; Piselli, A.; Ravizza, P.; Righini, G.(2011), Optimization of Multi-Skill Call Centers Contracts and Work-shifts, *Service Science*, vol. 3, n° 1: 67-81.
Dantzig, G. B.(1954), A comment on Edie's "traffic delays at toll booths", *Operations Research*, v. 2, n. 3: 339-341.
E-Consulting (2012), Anuário Brasileiro de Relacionamento com o Cliente 2011/2012. Disponível em: <http://www.portaldocallcenter.com.br>. Acesso em: 02 mar. 2011.
Gale, D.(1957) A Theorem on Flows in Networks.*Pacific Journal of Mathematics*, v. 7, n. 2, p. 1073-1082.
Gans, N.; Koole, G.; Mandelbaum, A.(2003), Telephone call centers: Tutorial, review and research prospects. *Manufacturing Service Operations Management*, v. 5, p. 79–141.
Koole, G. (2007). Call Center Mathematics.*VrijeUniversiteit Amsterdam*. Disponível em: <http://www.cs.vu.nl/~koole/ccmath>. Acessoem: 17 jan. 2012.
_____.(2010). Optimization of Business Processes.*VrijeUniversiteit Amsterdam*.Disponível em: <http://www.gerkoole.com>. Acesso em: 23 fev. 2012.
Koole, G.; Pot, A. (2006).An Overview of Routing and Staffing Algorithms in Multi-Skill Customer Contact Center.Submetido para publicação. Disponível em: <http://www.aukepot.com>. Acesso em: 03 fev. 2012.
IBM(2010), IBM ILOG CPLEX Optimizer. Disponível em: <http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/>. Acesso em: 28/04/2012.
Microsoft(2012), Microsoft Windows 7. Disponível em: <http://windows.microsoft.com/pt-BR/windows7/products/home>. Acesso em 05 mai. 2012.
Pessoa, A.; Uchoa, E.(2011), UFFLP: Integrando Programação Inteira e Mista e Planilhas de Cálculo. Mini-curso apresentado no *XLIII Simpósio Brasileiro de Pesquisa Operacional*. Disponível em: <http://www.logis.uff.br/~artur/UFFLP/>. Acesso em: 02 mar. 2012.
Pinedo, M.*Scheduling*: Theory, Algorithms, and Systems. New York: Springer, 2008.
Reynolds, P.*Call Center Staffing*: The Complete, Practical Guide to Workforce Management. The Call Center School, Lebanon, TN, EUA, 2003.
Stolletz, R.*Performance Analysis and Optimization of Inbound Call Centers*.Springer-Verlag. Berlin, Germany, 2003.
Thompson, G.(1997), Labor staffing and scheduling models for controlling service levels. *Naval Research Logistics*. vol. 44, n° 8: 719–740.
UFFLP(2012). UFFLP: An easy API for Mixed, Integer and Linear Programming. Departamento de Engenharia de Produção, Universidade Federal Fluminense. Disponível em: <http://www.logis.uff.br/~artur/UFFLP/>. Acesso em: 02/05/2012.
Wallace, R.; Whitt, W.(2005), A Staffing Algorithm for Call Centers with Skill-Based Routing. *Manufacturing & Service Operations Management*, v. 7, n. 4: 276-294.