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 os experimentos demonstrados neste estudo, foi possível observar economia significativa de custos com pessoal, sem prejuízo à qualidade do atendimento.

PALAVRAS CHAVE. Call center, Escalonamento, Programação inteira.

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.

1. INTRODUÇÃO

C*all centers* podem ser definidos como um conjunto de recursos, tais como pessoas, computadores e equipamentos de telecomunicações, que permitem o fornecimento de serviços através do telefone[1]. Os *call centers* normalmente são classificados pela origem da chamada:

- *Call centers* receptivos (*inbound*) são aqueles cuja ligação é originada pelo cliente final, como em casos de Serviço de Atendimento ao Cliente;
- *Call centers* ativos (*outbound*), as chamadas são originadas pela própria central, como em serviços de telemarketing[2].

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[2]. Prova disto que, em 2011, estima-se que o setor tenha movimentado aproximadamente 29 bilhões de reais, com um crescimento de 9,5% em relação a 2010, ritmo que se mantém por três anos consecutivos[3]. Estima-se ainda que os custos com pessoal em grandes *call centers* chegam a representar 70% do custo total[1], de forma que reduções de custos e economias de escalas alcançadas no processo de planejamento operacional tornam-se 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 previsto dentro dos requisitos de qualidade esperados. É comum a utilização de técnicas de simulação e teoria de filas nesta etapa. Na terceira etapa, denominada de escalonamento, é determinada a combinação ótima das escalas de trabalho de forma a cobrir a demanda esperada.

Este artigo concentra-se na terceira etapa, denominada de escalonamento. Nesta etapa, 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 *call center*, podendo levar em consideração questões como preferências de turno dos agentes, intervalos, pausas, entre outras coisas[4][5].

O presente trabalho irá propor 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 viável criar uma estrutura onde todos os agentes sejam capazes de atender qualquer tipo de ligação. Assim, 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. As chamadas são direcionadas para o departamento responsável. Os departamentos são totalmente independentes, e cada tipo de ligação pode ser resolvido por um único departamento. Os agentes pertencem a um único departamento e são especializados em resolver questões de ligações pelas quais são responsáveis[2]. Nestes *call centers*, o planejamento operacional é relativamente simples. No entanto, esta estratégia 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 *call center* 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[1][4].

Call centers modernos fazem uso do roteamento de chamadas baseado em habilidades, ou *skill-based routing* (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[6]. 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, contribuindo para reduções de custo por economia de escala. No entanto, aumenta consideravelmente a complexidade do processo de planejamento operacional, visto que normalmente, quanto maior o número de habilidades de um determinado 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[4].

2.1. ESCALONAMENTO DE AGENTES EM CALL CENTERS

O modelo básico para escalonamento em *call centers* não leva em consideração segmentação por habilidades e é dado pela modelagem de programação inteira a seguir[7]:

$$\min \sum_{k \in K} c_k x_k \tag{1}$$

sujeito a

$$\sum_{k \in K} a_{k,t} x_k \geq s_t \quad \forall t \in T \tag{2}$$

$$x_k \geq 0 \text{ e inteiro}, \quad \forall k \in K \tag{3}$$

Onde K é o conjunto de possíveis escalas, T são os intervalos de tempo para os quais o planejamento está sendo feito, c_k é o custo de um agente com a escala k e s_t é a quantidade de agentes necessários no intervalo de tempo t. A matriz $a_{k,t}$ é definida da seguinte forma:

$$a_{k,t} = \begin{cases} 1, \text{se na escala k estiver disponível no tempo t} \\ 0, \text{caso contrário} \end{cases} \tag{4}$$

As variáveis de decisão x_k indicam a quantidade de agentes que serão alocados com a escala k.

O modelo de Dantzig [7] é expandido para incorporar a possibilidade de escalonamento multi-habilidades. Introduzem-se os conceitos de habilidades e perfis (*skill sets*). Para fins de simplicidade, considera-se um relacionamento de um para um entre habilidades e tipos de ligação, de forma que, para cada tipo de ligação, há uma, e somente 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 e que sejam vantajosas para o *call center*, em termos de performance. A maneira como os perfis de agentes viáveis são determinados também foge ao escopo do presente trabalho. Desta forma, os perfis é que são atribuídos aos agentes e não as habilidades diretamente, garantindo que apenas as combinações viáveis sejam atribuídas.

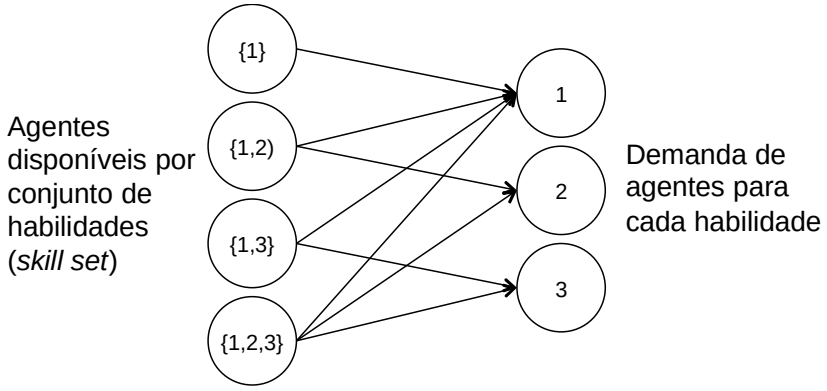
No modelo de escalonamento multi-habilidades, mantém-se os conjuntos K e T e a matriz $a_{k,t}$ no modelo, de forma semelhante ao modelo supracitado. Acrescenta-se o conjunto H com o universo de habilidades do *call center* e o conjunto P com os perfis de agentes.

O custo, agora, deve variar não apenas em função da escala, mas também em função do perfil atribuído ao agente, sendo representado pela matriz $c_{k,p}$. Isto permite, por exemplo, com que um agente que possua mais habilidades, ou uma especialidade mais restrita possua um custo diferenciado dos demais agentes do *call center*. A variável de decisão também deve ser modificada: $x_{k,p}$ informa a quantidade de agentes com o perfil p que possuem a escala k, chegando então 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.

$$\min \sum_{k \in K} \sum_{p \in P} c_{k,p} x_{k,p} \tag{5}$$

A principal modificação no modelo ocorre com a inclusão da restrição para o atendimento da demanda. A restrição de demanda deve garantir que, para cada intervalo de tempo, haja agentes suficientes para cada tipo de ligação para atender o volume previsto. A demanda agora não pode ser fornecida apenas em função do intervalo de tempo, mas deve ser segmentada para cada habilidade exigida, ou tipo de ligação.

A possibilidade de utilização de agentes com mais de uma habilidade faz com que seja necessário que o modelo não permita que, no mesmo intervalo de tempo, um agente com múltiplas habilidades seja contabilizado para suprir a demanda de mais de um tipo de chamada. Ao mesmo tempo, ele deve ser flexível o suficiente para permitir com que agentes com mais de uma habilidade supra a demanda de tipos de ligações diferentes, desde que em intervalos de tempo diferentes. Assim, o resultado da otimização deve garantir que, para cada intervalo de tempo, seja possível realizar a seguinte atribuição dos agentes disponíveis para a demanda por agentes de acordo com suas habilidades. A Figura 1 ilustra os perfis de agentes, os tipos de ligações e as possíveis atribuições entre eles:



Fonte: Elaborado pelo autor.

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

A diferença de cada formulação reside justamente na forma como cada abordagem garante a consistência dos agentes disponíveis e de suas habilidades ao suprir a demanda para cada tipo de ligação. A abordagem encontrada na literatura será descrita a seguir e, em seguida, será apresentada a abordagem proposta pelo presente trabalho.

2.1.1. Formulação de Bhulai *et al* [8]

No trabalho de Bhulai *et al* [8], é apresentado um algoritmo que, numa primeira etapa calcula as quantidades necessárias de agentes de cada perfil e, em seguida, cria as escalas destes agentes de forma a atender a demanda mínima calculada na etapa anterior. Considerando que o enfoque deste trabalho é no escalonamento, somente a segunda etapa do algoritmo será descrita e testada no presente trabalho, a fim de que se tenha uma base de comparação. Também não será considerada a heurística mencionada para reduzir o número de possíveis escalas.

A nomenclatura e a forma de apresentação das variáveis foram modificadas, para que fosse possível manter uma consistência com os conceitos utilizados ao longo deste trabalho e obter uma base de comparação para o desempenho do modelo proposto neste trabalho.

A demanda é calculada utilizando o algoritmo apresentado por Cezik e L'Ecuyer [9], que utiliza simulação para fornecer a quantidade de agentes necessários para cada intervalo de tempo, em cada perfil. No presente trabalho, a demanda é fornecida por tipo de chamada, e não por perfil de agentes, de forma que também foi necessário realizar esta adaptação na formulação de Bhulai *et al* [8]. No entanto, seria relativamente simples adaptar o modelo proposto pelo presente trabalho para atender demandas por perfil de agentes, em vez de por tipo de chamada.

No modelo de escalonamento de Bhulai *et al* [8], em cada intervalo de tempo é feita uma atribuição para determinar a quantidade de agentes que estará atendendo cada tipo de chamadas. Esta atribuição é feita através das variáveis de decisão $y_{t,p,h}$, onde t é o intervalo de tempo, p é o perfil do agente e h é o tipo de chamada (ou habilidade) que os agentes estarão atendendo. Além disto, é

introduzida a matriz $b_{p,h}$, que indica com 1 caso agentes com perfil *p* possuam a habilidade *h*, ou 0 caso contrário. São introduzidos ainda os elementos:

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

Por fim, é mantida a matriz $a_{k,t}$ do modelo de Dantzig [7], 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* [8]. A função objetivo (5) é a mesma apresentada anteriormente, acrescentada das restrições (6) e (9).

$$\min \sum_{k \in K} \sum_{p \in P} c_{k,p} x_{k,p} \tag{5}$$

sujeito a

$$\sum_{p \in P} b_{p,h} y_{t,p,h} \geq s_{t,h} \qquad \forall t \in T, \forall h \in H \tag{6}$$

$$\sum_{k \in K} a_{k,t} x_{k,p} = \sum_{h \in H} b_{p,h} y_{t,p,h} \qquad \forall t \in T, \forall p \in P \tag{7}$$

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

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

Desta forma, em cada intervalo de tempo, é realizada uma atribuição de quantidade de agentes com cada perfil com o tipo de chamada que estará atendendo durante um determinado intervalo de tempo. Esta atribuição tem como único propósito assegurar a consistência da quantidade de agentes utilizados no suprimento da demanda por cada tipo de chamada e a quantidade de agentes disponíveis, de acordo com as escalas e perfis determinados. Para o planejamento operacional, esta informação é irrelevante, tendo em vista que é uma decisão tomada em tempo de execução pelo sistema de telefonia. Desta forma, as variáveis de decisão $y_{t,p,h}$ não possuem utilidade, a não ser assegurar a consistência dos dados. Apesar de ter apresentado bons tempos de execução[8], o número variáveis cresce de forma acelerada ao acrescentar mais tipos de chamadas, perfis de agentes ou um horizonte maior de planejamento.

2.1.2. 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. Nesta formulação, garante-se que, para cada subconjunto do conjunto de habilidades *H*, a soma das demandas por estas habilidades seja inferior ou igual à soma dos agentes disponíveis, isto é, aqueles agentes cuja escala determine que estejam em serviço durante o intervalo de tempo, e que possuam capacidade para atender a demanda de pelo menos uma destas habilidades. Para isto, introduzimos a seguinte função:

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

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

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

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

A matriz $a_{k,t}$, já descrita no modelo de Dantzig [7] e na formulação baseada em Bhulai *et al* [8], 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*.

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

$$\min \sum_{k \in K} \sum_{p \in P} c_{k,p} x_{k,p} \tag{5}$$

sujeito a

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

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

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). O tempo de computação, obviamente, dependerá da quantidade de escalas, que varia em função do tempo de serviço, e da quantidade de habilidades e conjuntos de habilidades.

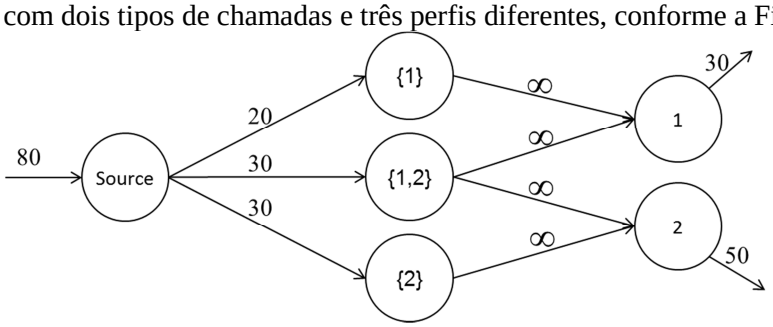
2.1.3. Prova de Corretude do Modelo Proposto

Conforme visto anteriormente, Bhulai *et al* [8] modelam o problema de escalonamento de agentes em *call centers* com múltiplas habilidades como problemas menores de fluxo de redes em cada intervalo de tempo. Desta forma, em cada intervalo de tempo é feita a atribuição dos agentes disponíveis em cada grupo à demanda por cada tipo de habilidade. Todavia, também pôde ser observado que, para fins de planejamento, a atribuição destes grupos de agentes aos tipos de chamada é irrelevante. Para o planejamento operacional, basta com que o resultado fornecido garanta com que esta atribuição seja possível em tempo de execução.

Gale[10] demonstra o seguinte teorema que permite 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 *call center* com dois tipos de chamadas e três perfis diferentes, conforme a Figura 2:



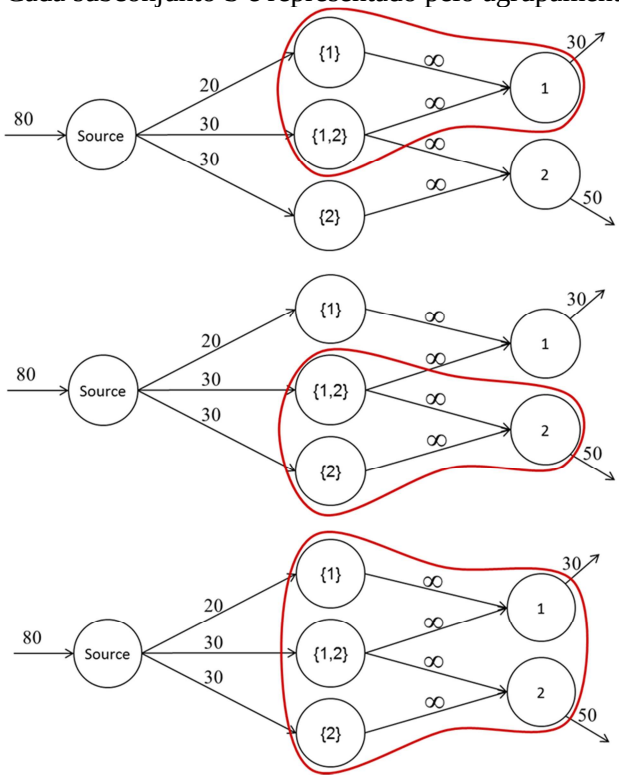
Fonte: Elaborado pelo autor.

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

Neste *call center*, o perfil {1} possui habilidade para atender apenas chamadas do tipo 1; o perfil {2} possui habilidade para atender apenas chamadas do tipo 2; o perfil {1, 2} possui habilidade para atender chamadas dos tipos 1 e 2. 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[10], 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. Os subconjuntos contendo apenas nós de habilidade serão ignorados, tendo em vista que seus arcos entrantes possuem capacidade infinita. A Figura 3 ilustra os subconjuntos S do conjunto N de nós que são relevantes para a análise de viabilidade do problema. Cada subconjunto S é representado pelo agrupamento de nós em vermelho.



Fonte: Elaborado pelo autor
Figura 3: Representação dos subconjuntos de N.

No primeiro subconjunto, a demanda total é de 30 agentes, e a soma das capacidades dos arcos entrantes é de 50. No segundo subconjunto, a demanda total é de 50 agentes e a soma das capacidades dos arcos entrantes é 60 agentes. No terceiro e último subconjunto S, a demanda total é de 80 agentes e a soma das capacidades dos arcos entrantes também é de 80 unidades. Portanto, como em todos os subconjuntos *S* a soma das demandas é menor ou igual à soma das capacidades dos arcos entrantes, de acordo com o teorema de Gale[10], o problema é viável.

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:

$$c_{s,\{1\}} + c_{s,\{1,2\}} \geq d_1 \tag{13}$$

$$c_{s,\{1,2\}} + c_{s,\{2\}} \geq d_2 \tag{14}$$

$$c_{s,\{1\}} + c_{s,\{1,2\}} + c_{s,\{2\}} \geq d_1 + d_2 \tag{155}$$

Generalizando para qualquer número de habilidades, temos que é possível realizar uma atribuição de agentes para habilidades se, e somente se, para cada subconjunto *L* do conjunto de habilidades *H*, a soma dos agentes disponíveis com perfis compatíveis com pelo menos uma habilidade em *L* deve ser superior ou igual à demanda total de *L*. A formulação (5), (11), (12) proposta

neste trabalho baseia-se neste princípio para garantir que a atribuição de perfis de agentes aos tipos de ligações possa ser realizada em cada intervalo de tempo.

3. ESTUDO DE CASO

Para o estudo de caso, procurou-se um *call center* que possuísse serviço receptivo e que o modelo de negócio permitisse a implementação da segmentação por habilidades, além de disponibilidade de dados relativos à volume de chamadas e custo operacional. O estudo foi conduzido em *call center* 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 *call center* estudado possui os seguintes perfis de agentes, ou conjuntos de habilidades, com seus respectivos custos mensais:

- Monolíngue: Possui habilidade apenas para chamadas em português. Custo médio mensal: R$800,00.
- Bilíngue-inglês: Possui habilidade para atender chamadas em português e inglês. Custo médio mensal: R$1.200,00.
- Bilíngue-espanhol. Possui habilidade para atender chamadas em português e espanhol. Custo médio mensal: R$1.200,00.
- Trilíngue: Possui habilidade para atender chamadas em português, espanhol e inglês. Custo médio mensal: R$1.600,00.

Neste *call center*, 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. Depois de feito 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.

Existem algumas limitações importantes que puderam ser observadas neste modelo de gestão. A primeira limitação 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 *call center*. Caso a empresa utilizasse turnos começando a cada hora a partir das 8:00, que é o horário de início de serviço, seria possível organizar a escala dos agentes de forma que seja possível atender a demanda dos horários de pico com menos agentes no quadro geral de funcionários. Isto não é feito atualmente pela empresa, pois o planejamento das escalas é realizado manualmente, o que dificulta consideravelmente este procedimento.

A segunda limitação é o fato do planejamento não levar em consideração a possibilidade de que em um intervalo de tempo de onde há sobra de agentes com mais de uma habilidade, esta sobra de agentes não é contabilizada para suprir a demanda por agentes com outra habilidade. Ou seja, na prática, os benefícios de se ter um *call center* 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. Apenas os agentes monolíngues são utilizados para suprir a demanda em português. Por fim, como regra geral, agentes trilíngues não são utilizados pelo fato de ter o salário maior do que os demais. No entanto, em determinadas situações, é possível que uma solução ótima inclua agentes trilíngues. Nos experimentos realizados e descritos abaixo, observa-se que a solução ótima de fato não incluía agentes trilíngues, porém no *call center* estudado, esta possibilidade não era sequer avaliada, conforme demonstra a Tabela 1, que exibe a demanda prevista fornecida pelo *call center*. O planejamento feito pelo *call center* seguindo o processo descrito acima em números de agentes.

Intervalo	Demanda			Agentes Disponíveis		
	Português	Inglês	Espanhol	Monolíngue	Bilíngue	Bilíngue

					(Inglês)	(Espanhol)
08:00 - 08:30	268	22	3	482	106	33
08:30 - 09:00	332	23	5	482	106	33
09:00 - 09:30	389	32	9	482	106	33
09:30 - 10:00	450	39	10	482	106	33
10:00 - 10:30	466	96	14	482	106	33
10:30 - 11:00	474	92	15	482	106	33
11:00 - 11:30	482	100	24	482	106	33
11:30 - 12:00	459	105	33	482	106	33
12:00 - 12:30	375	96	25	482	106	33
12:30 - 13:00	364	105	25	482	106	33
13:00 - 13:30	390	106	23	482	106	33
13:30 - 14:00	418	100	20	482	106	33
14:00 - 14:30	430	110	22	430	110	23
14:30 - 15:00	419	106	23	430	110	23
15:00 - 15:30	403	96	20	430	110	23
15:30 - 16:00	428	91	18	430	110	23
16:00 - 16:30	412	97	20	430	110	23
16:30 - 17:00	369	84	15	430	110	23
17:00 - 17:30	338	83	18	430	110	23
17:30 - 18:00	286	76	17	430	110	23
18:00 - 18:30	232	47	11	430	110	23
18:30 - 19:00	204	37	10	430	110	23
19:00 - 19:30	193	30	6	430	110	23
19:30 - 20:00	189	29	4	430	110	23

Tabela 1: Demanda vs. Planejamento do Call Center

É possível observar que, com esta forma de planejamento a ociosidade é grande, porém de forma desnecessária, pois há um grande número de agentes em serviço em períodos de baixa demanda. De acordo com este planejamento, é necessário manter 912 agentes monolíngues, 216 agentes bilíngues-inglês e 56 agentes bilíngues-espanhol. O custo total com salários de agentes, com base no salário médio por perfil de agente, foi de R$1.056.000,00.

O programa foi implementado e resolvido pelo pacote computacional IBM ILOG CPLEX Optimizer[11], utilizando sempre um *thread* de execução. Também foi utilizada a biblioteca UFFLP[12] como interface para a implementação dos programas. A UFFLP é uma biblioteca que fornece uma interface para implementação de modelos de programação inteira e mista, simplificando a integração com um resolvedor[13]. O programa foi executado em computador com 3.1GHz de processamento, 4GB de memória RAM, utilizando o sistema operacional Microsoft Windows 7 com 64 bits[14].

3.1. 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. Previsão foi realizada em intervalos de tempo intra-diários de 30 minutos. Manteve-se a mesma previsão de demanda e mesmo método para o dimensionamento para se chegar à quantidade de agentes por intervalo intra-diário utilizado na análise feita na seção anterior, a fim de que se pudesse ter uma base de comparação. Em ambas as formulações, os resultados, em termos numéricos, foram equivalentes e o problema foi resolvido em tempo menor do que 0,01 segundos, gerando resultado conforme Tabela 2.

Perfil / Escala	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 modelo considerou que não compensa o custo adicional para se trabalhar com agentes trilíngues. 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	Demanda			Agentes Disponíveis		
	Português	Inglês	Espanhol	Monolíngue	Bilíngue (Inglês)	Bilíngue (Espanhol)
08:00 - 08:30	268	22	3	437	83	28
08:30 - 09:00	332	23	5	437	83	28
09:00 - 09:30	389	32	9	446	96	31
09:30 - 10:00	450	39	10	446	96	31
10:00 - 10:30	466	96	14	454	96	31
10:30 - 11:00	474	92	15	454	96	31
11:00 - 11:30	482	100	24	468	105	33
11:30 - 12:00	459	105	33	468	105	33
12:00 - 12:30	375	96	25	634	146	40
12:30 - 13:00	364	105	25	634	146	40
13:00 - 13:30	390	106	23	673	163	45
13:30 - 14:00	418	100	20	673	163	45
14:00 - 14:30	430	110	22	429	110	23
14:30 - 15:00	419	106	23	429	110	23
15:00 - 15:30	403	96	20	420	97	20
15:30 - 16:00	428	91	18	420	97	20
16:00 - 16:30	412	97	20	412	97	20
16:30 - 17:00	369	84	15	412	97	20
17:00 - 17:30	338	83	18	398	88	18
17:30 - 18:00	286	76	17	398	88	18
18:00 - 18:30	232	47	11	232	47	11
18:30 - 19:00	204	37	10	232	47	11
19:00 - 19:30	193	30	6	193	30	6
19:30 - 20:00	189	29	4	193	30	6

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

Merece destaque 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. No entanto, 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. Desta forma, o modelo, além de robusto, é capaz de incorporar a flexibilidade do

roteamento por habilidades, para fornecer resultados ótimos—um avanço bastante significativo com relação aos modelos de escalonamento disponíveis atualmente.

Em seguida, foi realizado novo teste do modelo, buscando incorporar as restrições da legislação trabalhista para obter resultados ainda mais precisos. A legislação trabalhista para o setor determina que nas escalas 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. Além disto, deve haver um intervalo de 20 minutos entre as duas pausas, que não conta como tempo de trabalho[15].

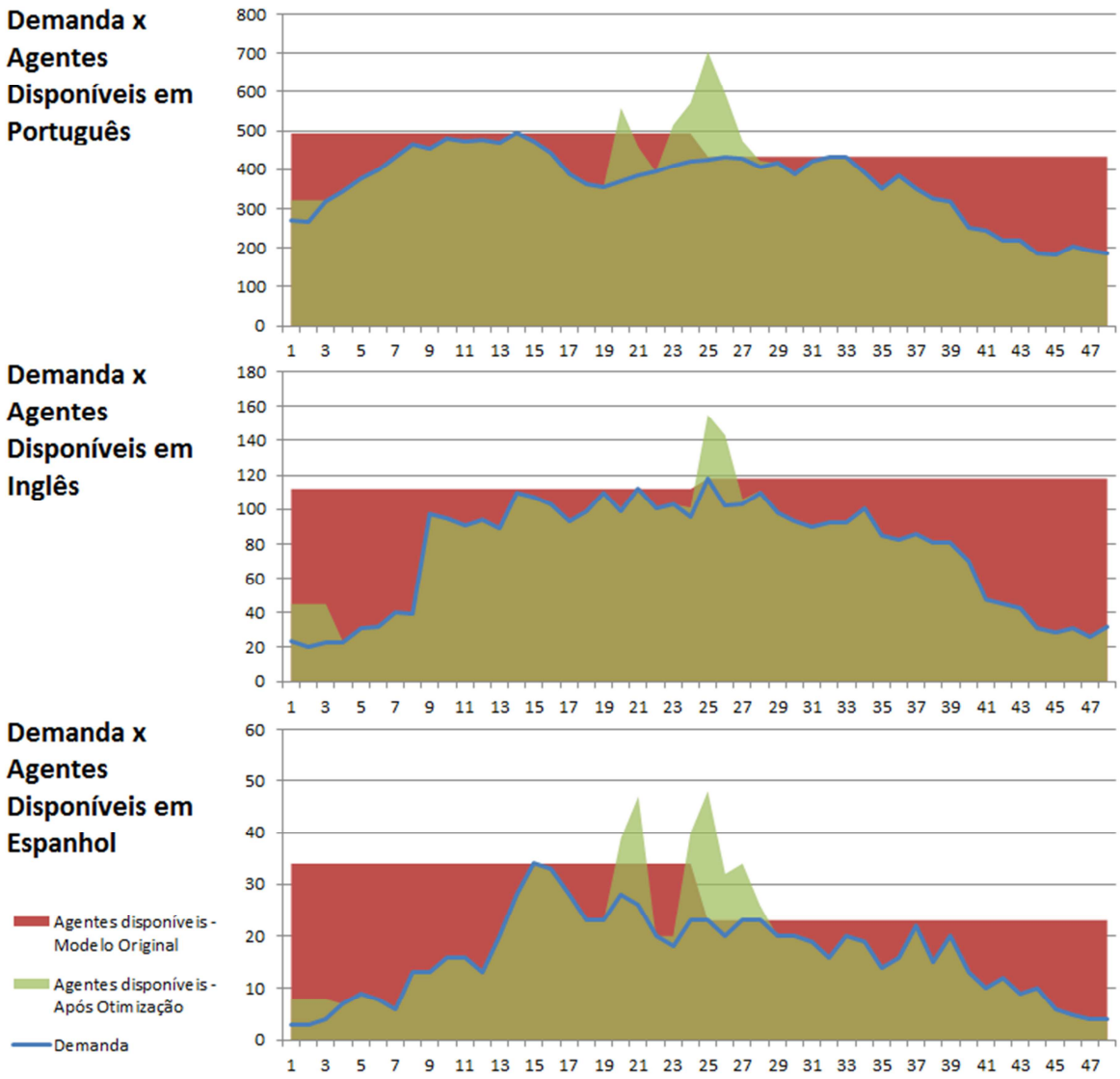
Para o segundo teste, a demanda intra-diária foi então dividida em intervalos de 15 minutos. As pausas de 10 minutos foram arredondadas para 15 minutos e as pausas de 20 minutos para 30. Este arredondamento é considerado razoável, pois, na prática, há o tempo de preparação para o atendimento que não 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* [8], 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.

Analisando o gráfico da Figura 4 é possível verificar que com as escalas geradas pelo modelo de otimização para múltiplas habilidades, é possível alcançar uma aderência à demanda muito maior, diminuindo consideravelmente a taxa de ociosidade dos agentes.



Fonte: Elaborado pelo autor

Figura 4: Demanda vs. Agentes disponíveis por idioma.

4. CONSIDERAÇÕES FINAIS E CONCLUSÕES

Como resultado do presente trabalho, foi possível verificar 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, com um tempo computacional relativamente baixo. Sendo assim, a abordagem proposta neste estudo apresenta uma alternativa a ser considerada, pois é capaz de incorporar a flexibilidade de *call centers* com roteamento de chamadas por habilidades em um modelo bastante simples de ser implementado, e que se demonstrou eficiente nos experimentos conduzidos.

Algumas ressalvas devem ser feitas. Conforme mencionado, o modelo considera 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 *call center* 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.

Outro ponto que deve ser considerado é o número de instâncias testadas. É necessária a execução de testes em instâncias maiores, com mais combinações de escalas para poder afirmar a eficiência do modelo. Contudo, a abordagem apresentada nesta pesquisa possui caráter que pode ser

considerado inédito e por este motivo carece de uma melhor análise. Por isto, torna-se necessária uma pesquisa mais detalhada, motivo pelo qual se adotou a metodologia de estudo de caso.

Como citado anteriormente, o trabalho de Wallace e Whitt [6] sugere que não é eficiente a criação de grupos 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. Por fim, o modelo do presente trabalho, bem como o apresentado por Bhulai, Koole e Pot [8], 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.

Não obstante, o resultado da pesquisa apresenta uma nova proposta na utilização da pesquisa operacional para a resolução de problemas de gestão de *call centers*.

5. REFERÊNCIAS BIBLIOGRÁFICAS

[1] GANNS, N.; KOOLE, Ger.; MANDELBAUM, A. Telephone call centers: Tutorial, review and research prospects. *Manufacturing Service Operations Management*, v. 5, p. 79–141, 2003.

[2] REYNOLDS, Penny. *Call Center Staffing*: The Complete, Practical Guide to Workforce Management. The Call Center School, Lebanon, TN, EUA, 2003.

[3] E-CONSULTING (2012). Anuário Brasileiro de Relacionamento com o Cliente 2011/2012. Disponível em: <http://www.portaldocallcenter.com.br>. Acesso em: 02 mar. 2011.

[4] KOOLE, Ger (2007). Call Center Mathematics. *Vrije Universiteit Amsterdam*. Disponível em: <http://www.cs.vu.nl/~koole/ccmath>. Acesso em: 17 jan. 2012.

[5] STOLLETZ, Raik. *Performance Analysis and Optimization of Inbound Call Centers*. Springer-Verlag. Berlin, Germany, 2003.

[6] WALLACE, Rodney; WHITT, Ward. A Staffing Algorithm for Call Centers with Skill-Based Routing. *Manufacturing & Service Operations Management*, v. 7, n. 4: 276-294, 2005.

[7] DANTZIG, G. B. A comment on Edie's "traffic delays at toll booths", *Operations Research*, v. 2, n. 3: 339-341, 1954.

[8] BHULAI, Sandjai; KOOLE, Ger; POT, Auke. Simple Methods for Shift Scheduling in Multi-Skill Call Centers. *Manufacturing & Services Operations Management* 10, p. 411-420: 2008.

[9] CEZIK, M.; L'ECUYER, P. Staffing multiskill call centers via linear programming and simulation. *Management Science*, v. 54, n. 2, p. 310-323, 2008.

[10] GALE, David. A Theorem on Flows in Networks. *Pacific Journal of Mathematics*, v. 7, n. 2, p. 1073-1082: 1957.

[11] IBM, 2010. IBM ILOG CPLEX Optimizer. Disponível em: <http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/>. Acesso em: 28/04/2012.

[12] 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.

[13] PESSOA, Artur; UCHOA, Eduardo. 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.

[14] MICROSOFT, 2012. Microsoft Windows 7. Disponível em: <http://windows.microsoft.com/pt-BR/windows7/products/home>. Acesso em 05 mai. 2012.

[15] 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.