2
Problema de Roteamento de Veículos com Restrição de Capacidade

2.1
Definição

O Problema de Roteamento de Veículos (*Vehicle Routing Problem* - VRP) foi originalmente estudado por Dantzig e Ramser [15] no contexto de distribuição de gasolina para postos de venda de combustíveis. A partir de um depósito central, o objetivo consistia na obtenção de um conjunto de rotas (isto é, caminhos fechados) de tamanho mínimo, cada qual realizada por um determinado veículo da frota, para abastecer cada um dos postos pré-definidos.

Desde então, o VRP tem sido alvo de diversas pesquisas, por sua relevância tanto teórica quanto prática, permitindo a aplicação e o desenvolvimento de diversas abordagens, seja entre pesquisadores do âmbito exato, seja entre aqueles do meio heurístico. Sua importância prática se justifica na medida em que o planejamento com processos de transporte permeia todos os estágios de produção e distribuição de uma organização, representando cerca de 10% a 20% dos custos totais dos bens de consumo [50]. Configura-se, portanto, como um dos problemas centrais da área de pesquisa operacional.

Muitas variantes do VRP foram estudadas ao longo das últimas décadas. Pode-se ter, por exemplo, restrições de tempo, em que uma rota não pode ultrapassar um limite de tempo L. Pode-se trabalhar com janelas de tempo, em que um cliente i é especificado para ser obrigatoriamente visitado no intervalo $[a_i, b_i]$. Pode ainda haver relações de precedência entre clientes, em que um dado cliente i deve ser atendido antes de um cliente j [33].

Neste trabalho, aborda-se a variante denominada de Problema de Roteamento de Veículos com Restrição de Capacidade (*Capacitated Vehicle Routing Problem* - CVRP), em que todos os veículos possuem uma capacidade única C.

Formalmente, tem-se um grafo não–orientado completo $G = (V, E)$, com vértices $V = \{v_0, v_1, \ldots, v_n\}$ e arestas $E = \{(v_i, v_j) \mid v_i, v_j \in V,\ i < j\}$. O vértice v_0 representa o depósito e os demais correspondem aos clientes, cada

qual possuindo uma demanda positiva d_i. Cada aresta $(v_i, v_j) \in E$ possui um custo (ou comprimento) associado c_{ij}, correspondendo ao caminho mais curto entre os vértices v_i e v_j. Desse modo, detalhes como aquele em que o caminho de v_i a v_j passa por um vértice v_k não são levados em consideração na formulação do problema.

Dados inteiros positivos C e K, deseja-se obter rotas para os K veículos de capacidade C da frota, satisfazendo as seguintes restrições:

i) Cada rota deve começar e terminar no depósito;

ii) Cada cliente deve ser atendido por exatamente um veículo; e

iii) A soma das demandas dos clientes em uma rota deve corresponder a, no máximo, C.

Por fim, o objetivo consiste em minimizar o custo total das rotas, sabendo-se que o custo de uma rota corresponde à soma dos custos das arestas que a compõem.

A Figura 2.1 ilustra uma possível solução para uma instância do CVRP, com 3 rotas partindo do depósito (representado pelo vértice 0):

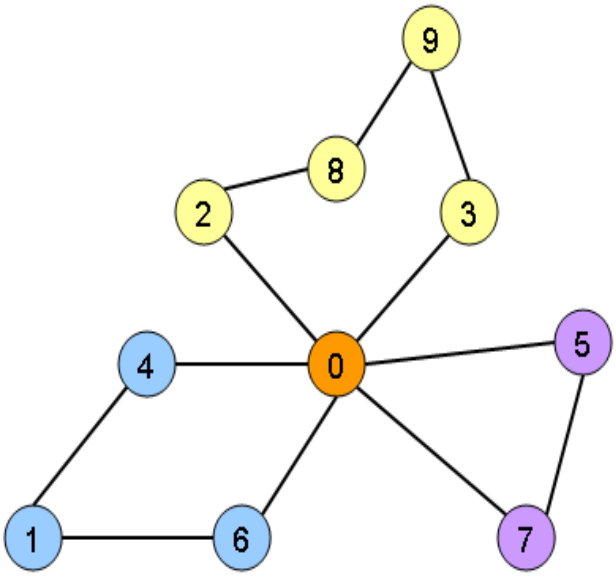

Figura 2.1: Exemplo de solução para uma instância do problema.

De fato, é fácil perceber que o CVRP corresponde a uma generalização do Problema do Caixeiro Viajante (*Travelling Salesman Problem* - TSP). Além disso, também engloba o Problema de Empacotamento (*Bin Packing Problem* - BPP) em sua estrutura. Tem-se, pois, que o CVRP pertence à classe $\mathcal{NP}$-Difícil, visto que os dois problemas supracitados também pertencem a essa [23].

2.2 Revisão de Abordagens Propostas

Há uma vasta literatura disponível para o VRP e também para o CVRP. Cite-se, a título de exemplo, a obra de Toth e Vigo de 2002 [50], destinada à descrição e análise de diversas abordagens ao CVRP e outras variantes, além de alguns exemplos de aplicações práticas dessas. Assim sendo, nesta seção, serão descritos alguns dos principais trabalhos disponíveis na literatura para o problema.

2.2.1 Abordagens Exatas

Os métodos exatos disponíveis até a década de 1980 estão reunidos e descritos no *survey* de Laporte e Norbert [35]. Laporte [33] também fornece uma boa revisão de métodos propostos até 1992. Ainda, alguns capítulos do livro de Toth e Vigo [50] são dedicados à descrição das principais abordagens exatas propostas até 2002.

O primeiro trabalho, de fato, de destaque foi o de Christofides *et al.* [11], usando um limite lagrangeano obtido a partir da resolução do problema de obtenção das q–rotas mínimas. O algoritmo de *Branch-and-Bound* resultante desse limite conseguiu resolver instâncias contendo de 10 a 25 vértices, um feito considerável para a época.

Esse mesmo trabalho de Christofides *et al.* [11] também propõe o uso de uma relaxação lagrangeana, através de um *lower bound* baseado em *k–degree center trees*, correspondendo a árvores geradoras mínimas em que o depósito possui grau $K \leq k \leq 2K$ e usando $2K - k$ arestas de menor custo.

Alguns trabalhos tratam o CVRP através de uma formulação por Particionamento de Conjuntos (*Set Partitioning*). Balinski e Quandt [7] foram pioneiros nessa vertente, propondo uma formulação com um número exponencial de colunas que exige a resolução do TSP com coleta de prêmios no subproblema de *Pricing*, o que se mostrou inviável. Agarwal *et al.* [5] e Hadjiconstantinou *et al.* [26] propuseram versões modificadas a fim de evitar incidir nesse problema.

Durante alguns anos, a comunidade de métodos exatos voltou-se à tentativa de compreensão da estrutura poliedral do problema, almejando obter novos procedimentos separadores de cortes. Contudo, a diferença de melhoria nos *lower bounds* obtidos se tornaram cada vez mais marginais. Alguns trabalhos de destaque, seguindo essa linha, são os algoritmos de *Branch-and-Cut* de Lysgaard *et al.* [38] e Augerat *et al.* [4].

Esses trabalhos normalmente descartavam a possibilidade de combinar geração de colunas e cortes, visto que novas variáveis duais introduzidas por cortes separados poderiam ocasionar uma alteração na estrutura do subproblema de *pricing*, tornando-o intratável. A pesquisa de Fukasawa *et al.* [22] foi a pioneira em expressar cortes sobre uma formulação na qual tal problema não ocorre para o CVRP. Tal formulação, contendo um número exponencial de colunas e de restrições, também utiliza o conceito de q–rotas. Com isso, o algoritmo de *Branch-and-Cut-and-Price* implementado foi capaz de atingir o ótimo para diversas instâncias ainda em aberto na literatura.

O trabalho de Baldacci *et al.* [6] propõe uma abordagem através da formulação por *Set Partitioning*, com a adição de cortes de capacidade e de clique, melhorando alguns limites inferiores e tempos de execução em relação ao trabalho de Fukasawa *et al.* [22].

2.2.2 Abordagens Heurísticas

Primeiramente, cabe fazer a distinção de que, normalmente, abordagens heurísticas lidam com instâncias em que o número de veículos consiste em uma variável de decisão. Já abordagens exatas usam aquelas em que esse dado é tido como uma das entradas do problema.

O trabalho de Laporte *et al.* [36] fornece uma boa visão dos principais métodos heurísticos até os anos 2000, fazendo uma distinção entre duas classes principais: as *heurísticas clássicas*, muito desenvolvidas entre as décadas de 1960 a 1990, e as metaeurísticas, que passaram a ser exploradas a partir de 1990.

A heurística de Clark e Wright [12] é tida como a primeira heurística proposta para o CVRP, sendo também denominada de *Método das Economias*. Seu funcionamento é simples: inicia com cada rota sendo formada por apenas um único vértice e procede fazendo junções, desde que essa operação diminua o valor da solução.

Outras heurísticas clássicas dizem respeito às de *Cluster-First Route-Second*, em que primeiramente são agrupados os clientes (etapa de *clusterização*) e então são definidas suas ordens nas rotas (etapa de roteamento). O trabalho de Gillet e Miller [24] agrupa os clientes por meio da rotação de um raio centrado no depósito e, logo após, resolve um TSP para cada rota obtida. Fisher e Jaikumar [21] também seguem essa linha, não se baseando em conceitos geométricos para realizar o agrupamento dos clientes, mas sim na resolução de um Problema de Alocação Generalizada (*Generalized Assignment Problem* - GAP) formulado sobre esses. Por fim, resolve um TSP para cada

cluster correspondendo à solução do GAP.

Existem também as heurísticas de *Route-First Cluster-Second*, com poucos trabalhos que as usam. De início, relaxa-se a restrição de capacidade dos veículos, para se obter uma única rota gigante (denominada de *giant tour*), contemplando todos os vértices do grafo. De forma subsequente, essa única rota é dividida em um conjunto de rotas viáveis. Beasley [8] foi o primeiro a propor esse tipo de heurística. Ademais, ressalta-se que uma das abordagens mais bem sucedidas recentemente é a realizada por Prins [41], que propõe um algoritmo genético e utiliza a ideia de *giant tour* em sua estrutura de tratamento dos cromossomos.

A partir de então, a ênfase esteve sobre metaeurísticas baseadas em busca local. Um trabalho de destaque nesse sentido é o de Taillard [49], que realiza uma busca tabu sobre uma vizinhança t–OPT tradicional com reotimização individual das rotas, de acordo com o número de iterações. Contudo, o principal diferencial da técnica sugerida é a decomposição do problema principal em subproblemas menores, através do particionamento dos vértices em setores centrados no depósito (para casos planares e não-planares). Assim, implementações paralelas podem se beneficiar dessa decomposição, distribuindo os diversos subproblemas entre os processadores ou núcleos disponíveis.

Rochat e Taillard [47] desenvolveram o conceito de *Memória Adaptativa* (*Adaptive Memory*), que corresponde a um conjunto de soluções de qualidade atualizado dinamicamente ao longo da busca. Periodicamente, alguns elementos, isto é, partes, dessas soluções são extraídos do conjunto e combinados, a fim de formar soluções novas e que sejam promissoras. Nesse processo, as rotas que pertencem a soluções melhores recebem um peso maior. Dessa maneira, através de uma busca tabu usando essa estratégia, conseguiram obter novos limites superiores para duas instâncias da literatura.

Outra abordagem interessante é a denominada *Very Large Neighborhood Search* de Ahuja *et al.* [2], em que uma vizinhança cíclica (*cyclic exchange*) é definida e otimizada através da resolução de um problema combinatório. Essa vizinhança consiste na transferência de um cliente v_i da rota r_1 para r_2, a rota r_2 transferindo um cliente v_j para r_3 e assim sucessivamente, até que um cliente v_k seja retirado de r_n e colocando em r_1. É, pois, uma uma generalização de vizinhanças 2–*exchange*, em que um par de vértices é trocado entre duas rotas. Outros trabalhos nessa linha foram produzidos por Rego e Roucairol [46], e também por Xu e Kelly [52] sobre *Ejection Chains*, correspondendo ao mesmo conceito de vizinhança cíclica. Esses estudos são exemplos em que vizinhanças maiores do que as tradicionais são exploradas em tempos de computação razoáveis.

Por fim, cabe destaque ao trabalho de Vidal *et al.* [51], que é o estado-da-arte em abordagens metaeurísticas para o CVRP. É proposto um algoritmo genético com mecanismo de controle populacional adaptativo, investindo na avaliação de um indivíduo através do custo da solução associada, bem como da contribuição à diversidade da população. É mantido um conjunto de indivíduos de elite, a fim de resguardar as características presentes em soluções consideradas boas. Ademais, esse estudo também faz uso da ideia de *giant tours* na composição de cromossomos, conforme sugerido por Prins [41]. Ressalte-se que este trabalho foi originalmente destinado ao *Periodic Vehicle Routing Problem* (Problema de Roteamento de Veículos Periódicos), e posteriormente aplicado ao CVRP.

2.3 Formulações do Problema

Nesta seção, são apresentadas e descritas as formulações contidas na implementação do *Branch-and-Cut-and-Price* (BCP) de Fukasawa *et al.* [22], usado neste trabalho. Essa corresponde a uma das melhores abordagens exatas atualmente, conforme revisão feita na subseção 2.2.1.

2.3.1 Formulação Clássica

A formulação considerada clássica para o problema, denominada de *Formulação de Dois Índices* (*Two-Index Formulation*), foi originalmente proposta por Laporte e Norbert [34]. Nessa, as variáveis de decisão x_{ij} representam o número de vezes que uma aresta $(v_i, v_j) \in E$ é atravessada por um veículo. Há, portanto, uma variável para cada aresta do grafo.

Seja definido o conjunto $V_+ = \{v_1, \ldots, v_n\}$ de vértices clientes (isto é, excluindo o vértice v_0 referente ao depósito). Além disso, para um dado $S \subseteq V_+$, seja $d(S) = \sum_{v_i \in S} d_i$, correspondendo à soma das demandas dos vértices em S. Seja também $\delta(S)$ o conjunto de arestas cruzando o corte definido pelo conjunto S, ou seja, com uma extremidade em S e outra fora. Utiliza-se também $k(S)$ como um valor representativo do número de veículos necessários para atender aos clientes no subconjunto S. Assim sendo, a formulação segue abaixo:

$$
F_1 = \left\{
\begin{array}{lllll}
\min & \sum_{e \in E} c_e.x_e & & & \\
 & \sum_{e \in \delta(\{v_i\})} x_e & = 2 & \forall\, v_i \in V_+ & (1) \\
 & \sum_{e \in \delta(\{v_0\})} x_e & = 2.K & & (2) \\
 & \sum_{e \in \delta(S)} x_e & \geq 2.k(S) & \forall\, S \subseteq V_+ & (3) \\
 & x_e & \leq 1 & \forall\, e \in E \setminus \delta(\{v_0\}) & (4) \\
 & x_e & \geq 0 & \forall\, e \in E &
\end{array}
\right.
$$

As restrições (1) refletem o fato de que, para todo vértice cliente, uma aresta deve entrar e outra deve sair, forçando que seja visitado por apenas um veículo. A restrição (2) determina que K veículos devem sair e retornar ao depósito. Por sua vez, as restrições (4) delimitam que arestas não-adjacentes ao depósito só podem ser utilizadas uma vez. Já arestas ligadas ao depósito podem ser atravessadas duas vezes, em uma rota que atenda apenas um cliente.

Finalmente, as restrições (3) eliminam a formação de subciclos e garantem também que, para todo subconjunto S, uma quantidade suficiente de

veículos deve atendê-lo. Calcular $k(S)$ exatamente é tão difícil quanto o *Bin Packing Problem* e, assim, consiste em um problema *Fortemente $\mathcal{NP}$-Difícil* [38]. Todavia, a formulação continua sendo válida ao se adotar o seguinte limite trivial: $k(S) = \lceil d(S)/C \rceil$.

Os vetores inteiros x pertencentes ao politopo definido por F_1 definem todas as soluções viáveis para o CVRP [22]. Observa-se, porém, que o número de restrições do tipo (3) é exponencial, visto que existe uma para cada subconjunto $S \subseteq V_+$. Logo, o cálculo do limite inferior fornecido por F_1 deve ser realizado através de um algoritmo de *Planos de Corte*, isto é, que gere de forma iterativa esse conjunto de restrições.

2.3.2 Formulação por Geração de Colunas

A primeira formulação com um número exponencial de colunas foi proposta por Balinski e Quandt [7], usando o Problema de Particionamento de Conjuntos. Uma vez geradas todas as rotas viáveis do problema, representa-se por J o conjunto dessas rotas j e por a_{ij} um coeficiente binário que indica se o vértice $v_i \in V_+$ aparece em j. Sejam também c_j^* o custo ótimo da rota j e x_j uma variável binária igual a 1 caso j seja usada na solução ótima, e 0, caso contrário. A formulação, portanto, segue abaixo:

$$SPF = \begin{cases} \min & \sum_{j \in J} c_j^*.x_j & & & \\ & \sum_{j \in J} a_{ij}.x_j & = & 1 & \forall\, v_i \in V_+ \\ & x_j & \in & \{0,1\} & \forall\, j \in J \end{cases}$$

em outras palavras, deseja-se minimizar o custo das rotas, sujeito à restrição de que todo cliente seja atendido por uma, e somente uma, rota.

Evidentemente, haverá tantas variáveis x_j quantas forem as rotas viáveis para o problema. Assim, tem-se que a formulação SPF possui um número exponencial de variáveis x_j. Além disso, calcular os custos c_j^* implica na resolução de um TSP para cada rota j.

Uma outra formulação com um número exponencial de colunas também pode ser obtida aplicando-se o conceito de q-rotas. Uma q-rota corresponde a um caminho que começa no depósito, passa por uma sequência de clientes com demanda total, no máximo, C e retorna ao depósito. Como cada cliente pode aparecer mais de uma vez em uma q-rota, tem-se que um conjunto de rotas válidas para o problema é subconjunto próprio do conjunto de q-rotas [22]. Consideram-se apenas as q–rotas que não contenham ciclos pequenos (de tamanho até 4), seguindo a ideia de Irnich e Villeneuve [31].

Nesta fase, definem-se variáveis λ correspondentes a q-rotas respeitando a restrição de capacidade. Desse modo, cada λ_j está associada a uma das p possíveis q-rotas. Seja Q uma matriz $m \times p$, em que as colunas são os vetores de incidência dessas p possíveis q-rotas. Assim, q_j^e é o coeficiente associado à aresta e na j-ésima coluna de Q, representando o número de vezes que a aresta e é utilizada na q-rota j. Define-se, portanto, a seguinte formulação F_2:

$$
F_2 = \left\{ \begin{array}{lrcll}
\min & \displaystyle\sum_{e \in E} c_e.x_e & & & \\
\displaystyle\sum_{j=1}^{p} q_j^e.\lambda_j & - \quad x_e & = & 0 \quad \forall\, e \in E & (5) \\
\displaystyle\sum_{j=1}^{p} \lambda_j & & = & K & (6) \\
& \displaystyle\sum_{e \in \delta(\{v_i\})} x_e & = & 2 \quad \forall\, v_i \in V_+ & (1) \\
& x_e & \geq & 0 \quad \forall\, e \in E & \\
\lambda_j & & \geq & 0 \quad \forall\, j \in \{1, \ldots, p\} &
\end{array} \right.
$$

As restrições do tipo (5) definem as variáveis x em relação às variáveis λ. Por seu turno, a restrição (6) define o número de veículos a ser utilizado. Tem-se, portanto, que essa formulação especifica igualmente todas as soluções viáveis para o CVRP [22].

No entanto, o número exponencial de q–rotas existentes implica em um número exponencial de variáveis λ. Assim, o limite inferior dado por F_2 deve ser calculado usando-se as técnicas de *Geração de Colunas* ou de *Relaxação Lagrangeana*, na medida em que realizar o *pricing* explicitamente passa a se tornar inviável.

Desta maneira, o problema de geração de colunas corresponde a obter as q–rotas de custo reduzido mínimo, o qual pode ser resolvido em tempo pseudo-polinomial por um algoritmo baseado em Programação Dinâmica. O uso de rotas válidas, isto é, que não contenham ciclos de quaisquer tamanhos, implicaria na resolução de um problema Fortemente $\mathcal{NP}$–Difícil como subproblema de *pricing* e, por isso, não é adotado.

2.3.3 Formulação com Número Exponencial de Colunas e Restrições

Ao se fazer a interseção dos politopos definidos por F_1 e F_2, obtém-se uma formulação em relação às variáveis x e λ, correspondendo a uma combinação de ambas as anteriores, e denominada de *Mestre Explícita*:

$$
F_3 = \begin{cases}
\min & \sum_{e\in E} c_e.x_e & & & & (0)\\
& \sum_{e\in\delta(\{v_i\})} x_e & = & 2 & \forall\, v_i \in V_+ & (1)\\
& \sum_{e\in\delta(\{v_0\})} x_e & = & 2.K & & (2)\\
& \sum_{e\in\delta(S)} x_e & \geq & 2.k(S) & \forall\, S \subseteq V_+ & (3)\\
& x_e & \leq & 1 & \forall\, e \in E \setminus \delta(\{v_0\}) & (4)\\
\sum_{j=1}^{p} q_j^e.\lambda_j - & x_e & = & 0 & \forall\, e \in E & (5)\\
\sum_{j=1}^{p} \lambda_j & & = & K & & (6)\\
& x_e & \geq & 0 & \forall\, e \in E & \\
\lambda_j & & \geq & 0 & \forall\, j \in \{1,\ldots,p\} &
\end{cases}
$$

A restrição (6) é redundante e pode ser eliminada, na medida em que já está definida por (2) e (5). O cálculo do limite fornecido pela formulação F_3 envolve a resolução de um programa linear com um número exponencial tanto de variáveis quanto de restrições, exigindo, portanto, o uso de algoritmos de planos de corte e de geração de colunas.

Pode-se, ainda, obter uma formulação mais compacta, alterando-se toda ocorrência das variáveis x_e em (0) - (4) pela forma equivalente dada pela restrição (5). A formulação resultante, denominada de *Dantzig-Wolfe Master*, segue abaixo:

$$
DWM = \begin{cases}
\min & \sum_{j=1}^{p}\sum_{e\in E} c_e.q_j^e.\lambda_j & & & & \\
& \sum_{j=1}^{p}\sum_{e\in\delta(v_i)} q_j^e.\lambda_j & = & 2 & \forall\, v_i \in V_+ & (1)\\
& \sum_{j=1}^{p}\sum_{e\in\delta(\{v_0\})} q_j^e.\lambda_j & = & 2.K & & (2)\\
& \sum_{j=1}^{p}\sum_{e\in\delta(S)} q_j^e.\lambda_j & \geq & 2.k(S) & \forall\, S \subseteq V_+ & (3)\\
& \sum_{j=1}^{p} q_j^e.\lambda_j & \leq & 1 & \forall\, e \in E \setminus \delta(\{v_0\}) & (4)\\
& \lambda_j & \geq & 0 & \forall\, j \in \{1,\ldots,p\} &
\end{cases}
$$

Lembrando-se que $\sum_{j=1}^{p} q_j^e.\lambda_j = x_e$, pode-se incluir um corte genérico $\sum_{e\in E} a_e x_e \geq b$ na formulação da seguinte forma: $\sum_{j=1}^{p} \left(\sum_{e\in E} a_e q_j^e\right).\lambda_j \geq b$.

Cortes adicionados sobre as variáveis x_e podem ser adicionados à formulação sem alterar o subproblema de *pricing*, caracterítica que torna o *Branch-and-Cut-and-Price*, sobre a formulação DWM, robusto.