Algorithme de recherche locale pour la résolution d'un problème réel de tournées d'inventaires

Thierry Benoist **Frédéric Gardi** **Antoine Jeanjean**
Bouygues e-lab, Paris
{ tbenoist, fgardi, ajeanjean }@bouygues.com

Bertrand Estellon
Laboratoire d'Informatique Fondamentale,
Faculté des Sciences de Luminy, Marseille
bertrand.estellon@lif.univ-mrs.fr

Inventory Routing Problem (IRP)

Vendor Managed Inventory : stocks des clients gérés par le fournisseur → minimiser les coûts de réapprovisionnement à long terme

Mais à court terme : Qui livrer ? Quand ? Combien ? Comment ?

IRP réel : posé en 2007 au e-lab par un grand groupe industriel français.

IRP déterministe : consommations des clients et productions des usines à court terme sont fournies par un logiciel de prévision. Les seuils de sécurité doivent couvrir le risque relatif à ces prévisions.

Présentation du problème

Ressources (chauffeurs, tracteurs, remorques) : non homogènes, situées à des dépôts, fenêtres de disponibilité dans le temps, incompatibilité entre ressources.

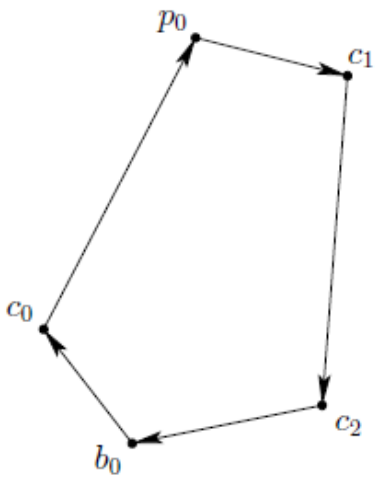
Sites (clients, usines) : capacité de stockage finie, fenêtres d'ouverture dans le temps, accessibilité restreinte aux ressources, consommation ou production non linéaire (discrète sur un ensemble de pas de temps).

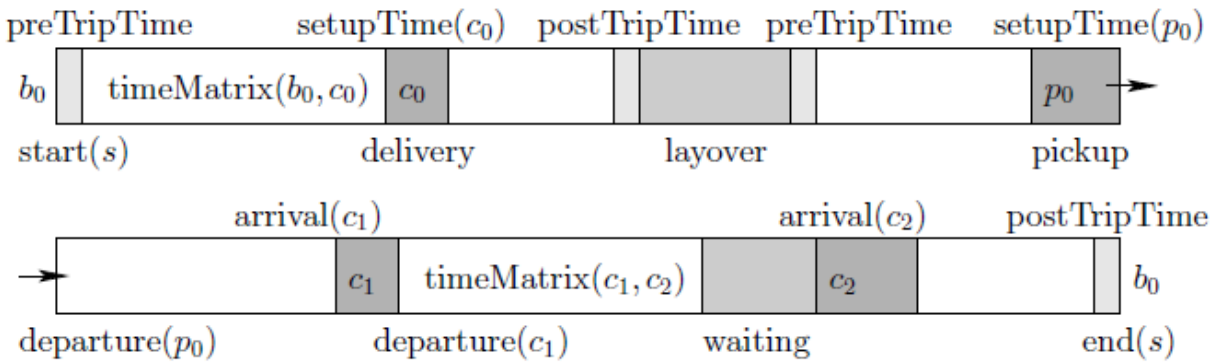
Tournée : un triplet de ressources, la date de départ du dépôt, les dates d'arrivée et quantités livrées (resp. chargées) des opérations réalisées chez les clients (resp. usines) de la tournée, avant retour au dépôt.

Présentation du problème

Contraintes réglementaires (RH) : temps de conduite et de travail des chauffeurs bornés : dès qu'un maximum est atteint (conduite ou travail), le chauffeur doit prendre une pause d'une durée minimale.

Quelques contraintes métier : livrer une quantité minimale lors de la visite d'un client, client à livrer immédiatement après chargement à l'usine, …

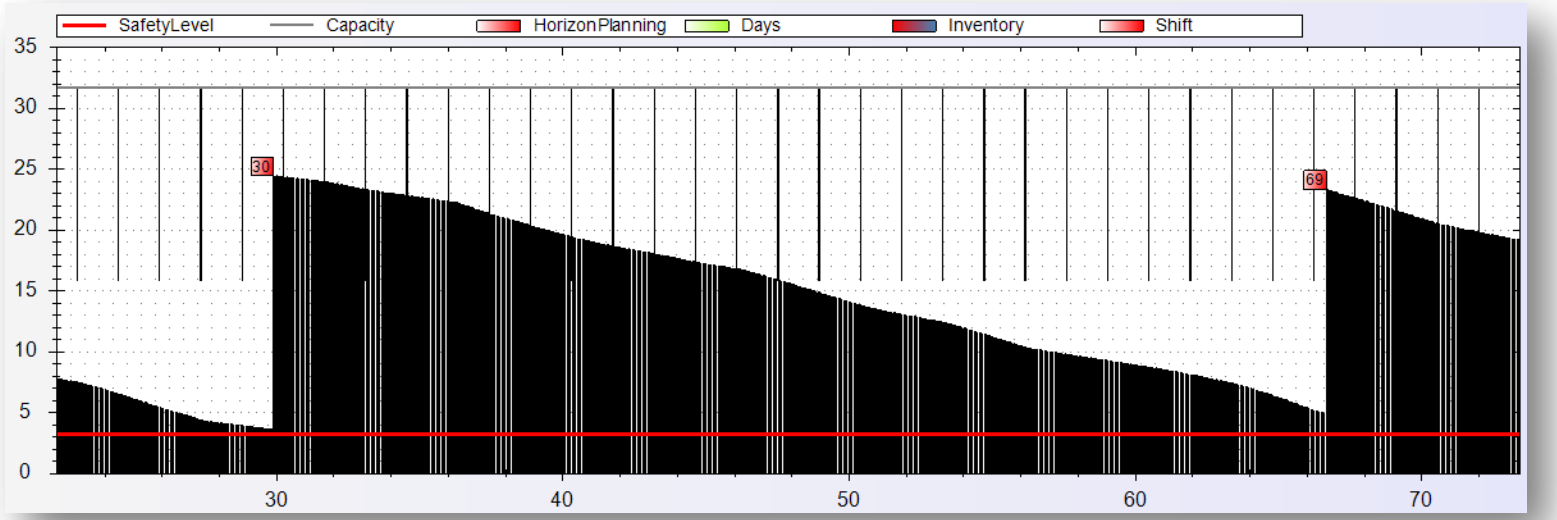




Présentation du problème

Dynamique des inventaires : les quantités livrées (aux clients) ou chargés (aux usines) respectent, pour tous les inventaires (clients, usines, remorques), les contraintes :

- de capacité
- de conservation des flux



Présentation du problème

Objectifs à long terme :

1) Minimiser le nombre *MO* de commandes insatisfaites (commande = quantité ou niveau demandé dans une fenêtre de temps).

2) Minimiser le nombre *SO* de pas de temps passé en-dessous les seuils de sécurité.

3) Minimiser le ratio logistique $LR = SC / DQ$, avec *SC* la somme des coûts des tournées et *DQ* la somme des quantités livrées.

En pratique, les objectifs *MO* et *SO* doivent être (facilement) atteints. En effet, toute commande insatisfaite ou pénurie n'est pas acceptable pour un client.

Présentation du problème

Processus de planification à court terme :

Le jour J, on planifie la distribution jusqu'à J+15.

On fixe les tournées dont la date de début se situe le jour J :

- les ressources participant à ces tournées sont immobilisées jusqu'à la fin de celles-ci

- les niveaux (futurs) des inventaires des clients/usines impactés par ces tournées sont actualisés

Le processus est itéré jour après jour.

Travaux antérieurs

Travaux nombreux dans la littérature RO, mais :

- travaux théoriques (ex : politique de distribution optimale)
- travaux sur des problèmes purs (ex : pas de gestion du chargement)
- peu de références sur des applications concrètes (logiciels exploités ?)

Deux références sur la résolution pratique d'IRP :

Bell, Dalberto, Fisher, Greenfield, Jaikumar, Kedia, Mack, Prutzman (1983).
Improving the distribution of industrial gases with an on-line computerized routing and scheduling optimizer.
Interfaces 13(6), pp. 4-23.

Campbell, Savelsbergh (2004).
A decomposition approach for the inventory-routing problem.
Transportation Science 38(4), pp. 488-502.

Complexité du problème

NP-difficile : IRP → VRP → TSP

Échelle des instances en pratique :

- horizon de planification de 15 jours
- des centaines de clients (jusqu'à 1500)
- des dizaines d'usines (jusqu'à 50)
- des dizaines de dépôts (jusqu'à 50)
- des dizaines de ressources (jusqu'à 50 par type)
- temps continue (précision à la minute : 21600 minutes)
- consommations/productions à l'heure (360 pas de temps)

Temps de résolution limité à 5 minutes sur un ordinateur standard

Gain minimum attendu sur solutions métier : 8 % en moyenne

Contribution à la modélisation

Propriété fondamentale de l'IRP : effectuer un aller-retour chez un client pour y vider une remorque chargée à plein (*full drop*) est *LR*-optimal.

Preuve : aller-retour de coût minimum (inégalité triangulaire) et volume livré maximum.

Une difficulté de l'IRP : les décisions minimisant *LR* à court terme peuvent être sous optimales à long terme.

Exemple : Un client beaucoup plus éloigné du dépôt par rapport aux autres clients ne sera pas livré s'il n'a pas un risque de pénurie à court terme. Or, si un *full drop* est possible, pourquoi ne pas le faire ?

Contribution à la modélisation

En résumé, on souhaite : « ne pas remettre à demain ce que l'on peut faire de façon optimale aujourd'hui ».

Définition d'un objectif de remplacement pour la planification court terme allant dans le sens de l'objectif à long terme :

→ minimiser le surcoût global *LR'* par unité de produit livré, étant donné pour chaque client son ratio logistique optimal LR^* (= *full drop*).

$$SC^*(s) = \sum_{\substack{customer\ p \\ delivered\ over\ s}} LR^*(p) \times quantity(p)$$

$$LR' = \frac{\sum_s (SC(s) - SC^*(s))}{DQ}$$

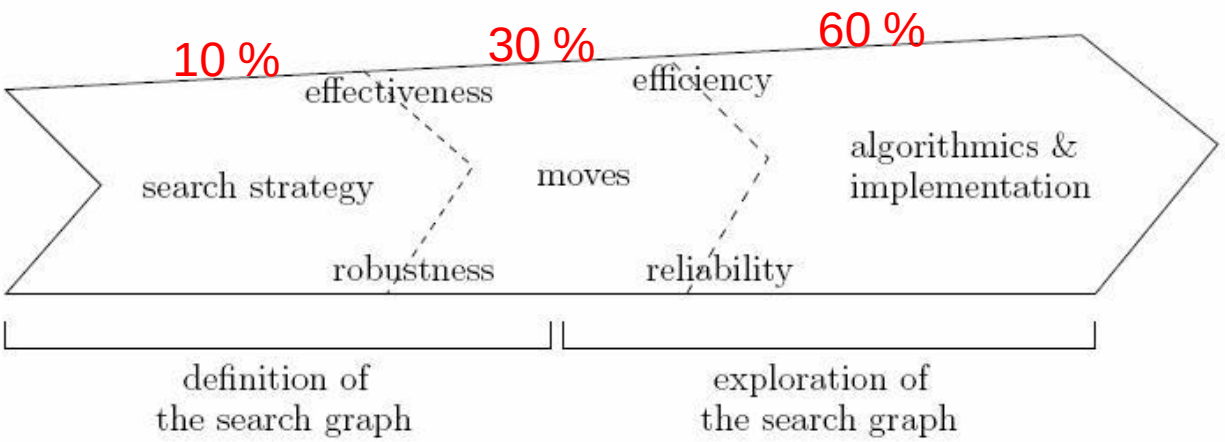
Contribution à la résolution

Algorithme de recherche locale pour la planification court terme :
- pur : pas de métaheuristique, pas d'hybridation
- direct : le problème n'est pas décomposé pour être résolu

→ espace de recherche = espace des solutions

Conception et implémentation suivent la méthodologie à 3 niveaux pour une « recherche locale haute performance » décrite dans :

B. Estellon, F. Gardi, K. Nouioua (2009).
High-performance local search for task scheduling with human resource allocation.
Proceedings of SLS 2009, *LNCS* 5752, pp. 1-15.



Contribution à la résolution

Heuristique générale :

- solution initiale construite par glouton (basé sur l'urgence des clients)
- *first-improvement descent* en 3 phases *MO*, *SO*, *LR'*
- choix stochastique (mais pas uniforme) des mouvements

Mouvements dérivant de transformations canoniques :

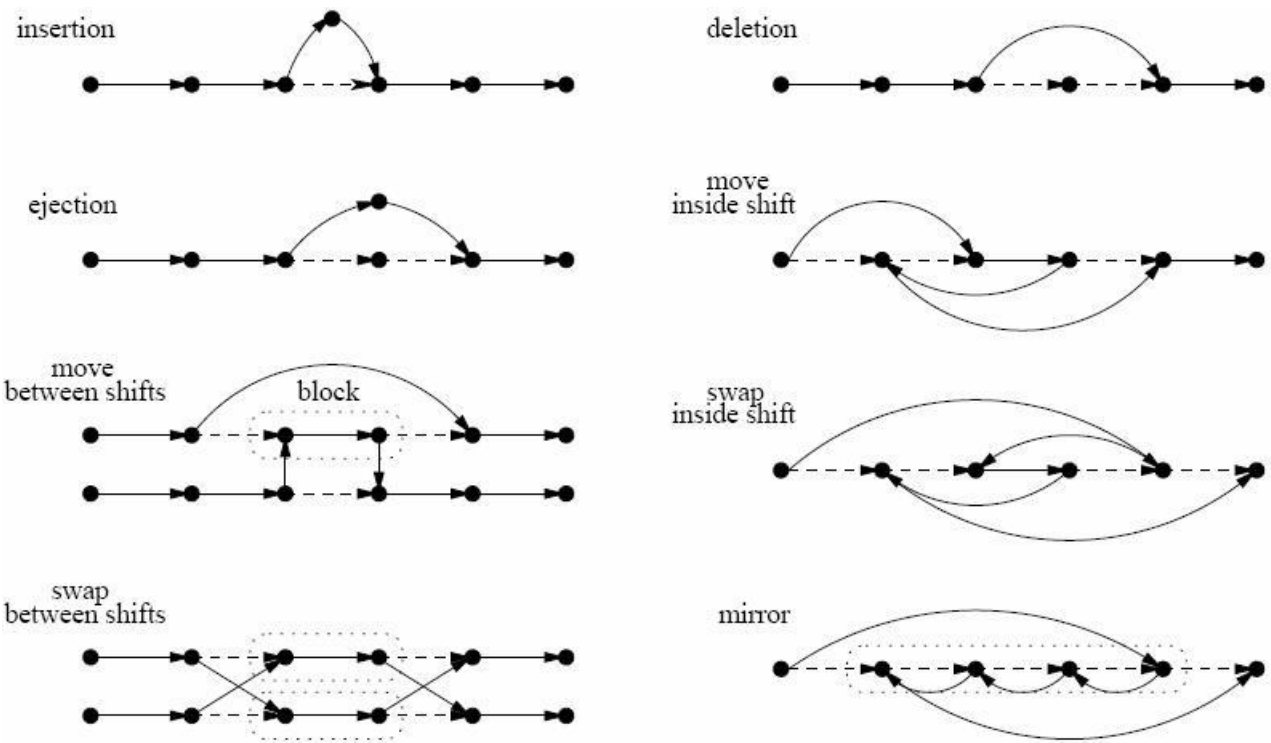
- sur opérations : insertion, suppression, éjection, déplacement intra/inter tournées, échange intra/inter tournées, inversion

- sur tournées : insertion, suppression, décalage temporel, réaffectation, échange, fusion, séparation

→ voisinage de taille $O(n^2)$ avec n le nombre d'opérations dans la solution courante. Mais la constante cachée par le O est large !

Contribution à la résolution

Transformations sur les opérations :



Contribution à la résolution

***** Rappel méthodologique important [SLS 2009] *****

Local search = incomplete search technique: its performance depends strongly on the number of solutions explored within the time limit.

algorithms = engine of local search

3 crucial routines for each transformation : evaluate, commit, rollback

1) incremental algorithms relying on special data structures, exploiting invariants of transformations → high-level efficiency

2) careful implementation (cache-aware programming, CPU & RAM profiling) → low-level efficiency

3) programming with assertions, data structures checked at each iteration in debug mode (checkers) → correctness & reliability

Contribution à la résolution

1) (Ré)ordonnancement d'une tournée de k opérations : minimiser le temps improductif sur la tournée → glouton chronologique

- prendre la pause le plus tard possible et convertir les temps d'attente devant les sites fermés en pause

- en temps $O(k)$ si les pauses ne sont pas stockés explicitement

- ordonnancement en avant ou en arrière rendu symétrique (tournée = liste d'opérations doublement chaînées)

- optimal si aucun temps d'attente n'apparaît sur la tournée (cas fréquent en pratique)

Contribution à la résolution

2) (Ré)affectation des volumes sur n opérations : maximiser le volume total livré → critique pour la performance

- Résolution exacte par flot maximum → temps $O(n^3)$

TROP LENT !

- Glouton poussant un maximum de flot à chaque nœud du DAG suivant un ordre topologique (= chronologique ici) → temps $O(n \log n)$

Propriété (intéressante en pratique) : optimal si chaque client est livré une seule fois sur l'horizon de planification

HUM…

- Même glouton mais avec réaffectation partielle (= locale) des volumes dans réseau → temps $O(n' \log n')$ avec $n' << n$

Contribution à la résolution

Difficulté : réaffecter la quantité livrée à un client à l'instant *t* alors que d'autres opérations ont lieu après *t* → risque de pénurie ou de dépassement de capacité après *t*.

→ mise en œuvre incrémentale très délicate : les réseaux avant et après transformation sont « superposés » dans la même structure de donnée.

Récompense : le glouton partiel s'exécute en temps quasiment constant en pratique.

T(glouton partiel) = T(glouton complet) / 100
T(glouton partiel) = T(flot maximum) / 2000

Écart moyen à la réaffectation optimale : 2 % (si *MO* = 0 et *SO* = 0)

BANCO !

Chiffres clés

- Programmé en C# 2.0 (pour machine virtuelle .NET 2.0)
- Environ 30 000 lignes de code dont 6 000 (20 %) de *checkers*
- Projet global : 300 jours-homme

- Plus de 10 000 mouvements par seconde
- Près de 10 millions de solutions visitées en 5 minutes
- 30 Mo de mémoire alloué, 300 Mo pour les très grandes instances

- Diversification naturelle à coût constant pour les 3 objectifs
- Taux d'acceptation des mouvements entre 1 et 10 %
- Un millier de mouvements strictement améliorant

- Gain moyen de 21 % sur notre algorithme glouton
- Gain moyen de 25 % sur les experts logistiques

Résultats numériques

Testé sur Intel Xeon X5365 : CPU 3 GHz, L1 64 Ko, L2 4 Mo, RAM 8 Go.
Pour chaque instance, moyenne sur 5 exécutions de 5 minutes.

instances court terme

data	customers	plants	bases	drivers	tractors	trailers	attempt	accept	improve	gain
A01	80	2	2	20	10	20	10.928 M	472221 (4.3 %)	494 (0.5 h%)	20.8 %
A02	108	1	1	35	18	35	5.450 M	355980 (6.5 %)	1030 (1.9 h%)	31.2 %
A03	132	1	1	20	17	15	5.454 M	293142 (5.4 %)	909 (1.7 h%)	25.6 %
A04	130	2	1	17	10	20	6.443 M	233130 (3.6 %)	964 (1.5 h%)	26.7 %
A05	125	2	1	20	18	20	12.477 M	423865 (3.4 %)	999 (0.8 h%)	34.5 %
A06	46	1	2	50	50	50	15.384 M	601032 (3.9 %)	779 (0.5 h%)	28.4 %
A07	80	2	2	20	10	20	7.646 M	480636 (6.3 %)	702 (0.9 h%)	25.4 %
A08	75	1	1	10	10	10	7.772 M	390098 (5.0 %)	593 (0.8 h%)	39.1 %
A09	150	2	1	20	20	20	8.582 M	289688 (3.4 %)	840 (1.0 h%)	28.6 %
A10	250	5	1	30	30	30	7.496 M	192871 (2.6 %)	1449 (1.9 h%)	30.9 %
A11	500	4	2	50	20	50	4.608 M	133799 (2.9 %)	2029 (4.4 h%)	26.7 %
A12	108	1	1	35	18	32	4.702 M	314355 (6.7 %)	1066 (2.3 h%)	26.4 %
A13	100	1	1	35	35	35	8.643 M	365894 (4.2 %)	818 (0.9 h%)	22.6 %
A14	70	1	1	50	5	10	10.515 M	693110 (6.6 %)	516 (0.5 h%)	32.6 %
A15	132	1	1	20	17	15	4.863 M	319638 (6.6 %)	697 (1.4 h%)	32.4 %
A16	130	2	1	17	10	20	10.391 M	335002 (3.2 %)	915 (0.9 h%)	30.1 %
A17	135	3	1	20	18	20	8.933 M	252043 (2.8 %)	1002 (1.1 h%)	32.6 %
average							8.252 M	361559 (4.5 %)	929 (1.3 h%)	29.0 %

instances long terme

data	customers	plants	bases	drivers	tractors	trailers	orders	wst 1 mn	avg 1 mn	avg 5 mn	avg 1 h
L1	75	6	1	35	21	5	56	23.8 %	24.6 %	26.3 %	26.5 %
L2	75	6	1	35	21	5	55	22.3 %	23.5 %	24.9 %	25.2 %
L3	175	8	1	35	21	12	189	5.2 %	5.8 %	8.3 %	11.2 %
L4	165	4	1	24	11	7	167	9.9 %	11.2 %	14.0 %	18.9 %
L5	198	8	7	12	12	12	40	32.5 %	34.2 %	35.7 %	35.9 %
average	138	6	2	28	17	8	101	18.7 %	19.9 %	21.8 %	23.5 %