# Heuristique branch and bound pour la sous-allocation et la réallocation

Antoine Rebecq [1]

[1] *INSEE - DMCSI division Sondages, antoine.rebecq@insee.fr*

**Résumé.** Dans le cadre des enquêtes INSEE, les logements à enquêter sont attribués aux enquêteurs par zone géographique. Chaque enquêteur ne traite que des logements situés dans la zone qui lui est attribuée. Ainsi, quand certaines zones sont couvertes par un faible nombre d'enquêteurs, une absence longue d'un enquêteur peut conduire à un cas dit de "zone orpheline". On cherche absolument à éviter ce cas de figure, par crainte de création d'un biais de non-réponse non-ignorable.

Pour ce faire, on définit un algorithme permettant de choisir la meilleure répartition des logements à enquêter parmi les enquêteurs disponibles à proximité de la zone orpheline. Le choix se fait selon la minimisation d'une fonction de coût sous contraintes. Ceci constitue un problème $\mathcal{NP}$-difficile, dont la résolution pour les valeurs typiques demande la création d'une heuristique spécifique. Quand cet algorithme est mis en œuvre en amont de la collecte (cas d'une absence prévue), le programme minimise la dispersion des poids de sondage ainsi que le nombre de fiches adresses perdues. Une des contraintes est le respect de la quotité de travail de l'enquêteur. On parle alors de sous-allocation. Quand l'algorithme est mis en œuvre pendant la collecte (cas d'une absence non prévue), la fonction d'objectif minimise le temps de collecte restant ainsi que la distance à parcourir pour les enquêteurs sélectionnés. On parle alors de réallocation.

**Mots-clés.** Sous-allocation, réallocation, optimisation non linéaire, échantillonnage, collecte adaptative ...

# 1 Position du problème

## 1.1 Risque de zone orpheline

Au cours de la collecte de l'Enquête Nationale Logement 2013 par l'INSEE, la répartition inégale des logements non enquêtés faisait courir le risque de laisser certaines zones géographiques échantillonnées non couvertes par l'enquête. On désigne par **zone orpheline** de telles zones. Craignant de voir apparaître des phénomènes non MAR (non-réponse non-ignorable), on préfère se donner une stratégie pour réallouer ces logements à des enquêteurs en mesure d'effectuer la collecte. Dans [5], Schouten, Calinescu et Luiten (2013) évoquent des plans de collecte adaptatifs mettant en jeu des problèmes d'optimisation linéaire et non linéaire.

## 1.2 Un problème $\mathcal{NP}$-difficile

La résolution de ce problème d'optimisation est classiquement un problème $\mathcal{NP}$-difficile. Un algorithme "naïf" de parcours exhaustif posséderait une complexité de l'ordre de

$\mathcal{O}({n_{enqueteurs}}^{n_{logements}})$. Le problème pratique met en jeu des valeurs de l'ordre de : $n_{enqueteurs} \simeq 50$ et $n_{logements} \simeq 200$. Le nombre total de possibilités est donc de $50^{200} \simeq 6 \cdot 10^{339}$. En imaginant que l'on utilise à pleine capacité 100 processeurs cadencés à 3 GHz (donc capables d'effectuer $3 \cdot 10^9$ opérations par seconde), le parcours exhaustif est effectué en $6 \cdot 10^{320}$ années. Ainsi, une résolution par recherche exhaustive est inenvisageable.

# 2 Une heuristique spécifique

Pour parvenir à résoudre le problème, on utilise une méthode classique de résolution, le backtracking avec algorithme branch and bound (partie 2.1) et on en réduit les dimensions grâce à des hypothèses adaptées (partie 2.2). Ceci se fait au prix d'une perte théorique d'optimalité, dont on discute en 2.2.2.

## 2.1 Résolution par méthode branch and bound

Plutôt que de parcourir l'intégralité des solutions possibles, on les répartit dans un arbre, que l'on explore selon un **"parcours en profondeur d'abord"**. Cette méthode porte le nom usuel de backtracking. Le fait de se donner une fonction de coût vérifiant une propriété de croissance permet de se placer comme un cas particulier des méthodes classiques de branch and bound introduites par Land et Doig en 1960 [3].

### 2.1.1 Principe

On se donne deux ensembles : $\mathcal{E} = \{e_1, \cdots, e_N\}$ et $\mathcal{L} = \{l_1, \cdots, l_M\}$, ainsi que les $\mathcal{L}_k = \{l_1, \cdots, l_k\}$ pour $k \in [[1, M]]$ tels que $\mathcal{L} = \bigcup_{k \in [[1,M]]} \mathcal{L}_k$. Dans les problèmes que l'on traite ensuite, l'ensemble $\mathcal{E}$ contient les enquêteurs pouvant être affectés aux logements de $\mathcal{L}$. Le problème consiste à choisir une application $S : \mathcal{L} \to \mathcal{E}$ qui minimise une **fonction de coût**, notée $C(S) = f(X(S))$. Le vecteur X, noté $X(S) = \begin{pmatrix} x_1(S) \\ \vdots \\ x_n(S) \end{pmatrix}$ est un vecteur de paramètres, qui dépend de l'allocation S choisie. Ainsi, dans l'application 3.2, X sera composé du temps anticipé de l'enquête et de la distance supplémentaire pour chaque enquêteur. Pour chaque nœud situé à distance $k \in [[1, M]]$ de la racine de l'arbre (ces nœuds sont donc au nombre de $N^k$, et devraient être en toute rigueur désignés par $k_{(i), i \in [[1, N^k]]}$, mais on s'autorise par souci de clarté à ne pas faire figurer l'indice $(i)$ dans la suite de l'article), on désigne les fonctions candidates partielles associées par : $S_k : \mathcal{L}_k \to \mathcal{E}$, ainsi que leur coût partiel : $C_k = f(X(S_k))$.

La seule propriété que doit vérifier la fonction de coût est une **propriété de croissance** :

$$\forall (p, q) \in [[1, M]] \text{ tels que } p \geq q : C(S_p) \geq C(S_q) \tag{1}$$

Le problème s'écrit ainsi :

$$\underset{S \in \mathcal{E}^{\mathcal{L}}}{\operatorname{argmin}} \; C(S)$$

L'algorithme de résolution est défini par récursivité :

- Initialisation : Définir $\mu = +\infty$
- Algorithme :
  - Si le nœud courant possède des nœuds enfants et si $C(S_k) < \mu$ : lancer l'algorithme pour tous les enfants du nœud courant.
  - Sinon : Si $C(S_k) < \mu$, poser $\mu = C(S_k)$ et sauvegarder l'état courant comme l'état minimal.
- Lancer l'algorithme à la racine de l'arbre.

Une variante consiste à utiliser une **propriété de croissance faible**. On fixe $\tau \in [[2, M]]$ tel que :

$$\forall (p, q) \in [[1, \tau]] \text{ tels que } p \geq q : C(S_p) \geq C(S_q) \tag{2}$$

$$\forall k \in [[\tau, M]], C(S_k) \geq C(S_\tau) \tag{3}$$

L'algorithme est alors légèrement modifié :

- Initialisation : Définir $\mu = +\infty$
- Algorithme :
  - Si le nœud courant possède des nœuds enfants et si ( $C(S_k) < \mu$ ou $k \geq \tau$ ) : lancer l'algorithme pour tous les enfants du nœud courant.
  - Sinon : Si $C(S_k) < \mu$, poser $\mu = C(S_k)$ et sauvegarder l'état courant comme l'état minimal.
- Lancer l'algorithme à la racine de l'arbre.

La vitesse de résolution du programme diminue dès lors que $\tau$ s'éloigne de $M$ (dans le cas limite $\tau = 1$, on est ramené à la recherche exhaustive). Il est à noter qu'on peut également sauvegarder comme état minimum l'état $k_0 = \underset{k \in [[\tau, M]]}{\operatorname{argmin}} C(S_k)$ si l'on souhaite rendre admissible une solution constituée d'un des $\mathcal{L}_k$ uniquement, au lieu de $\mathcal{L}$.

### 2.1.2 Exemple

Prenons l'exemple d'un ensemble de logements à enquêter $\mathcal{L} = \{l_1, l_2, l_3, l_4\}$, à répartir entre les éléments de l'ensemble $\mathcal{E} = \{A, B\}$ des enquêteurs disponibles. L'arbre correspondant est dessiné en figure 1. L'arbre se lit comme suit : chaque "étage" de nœuds indique une possibilité d'affectation du k-ème logement entre l'enquêteur A ou l'enquêteur B, et en gras figure le coût du choix de cette affectation. Le parcours en profondeur d'abord donne ainsi l'examen, dans l'ordre, des configurations de la table 1. L'intérêt de cette méthode consiste donc à éliminer ("couper") les nœuds enfants d'un nœud dont le coût dépasse déjà le coût minimal observé jusque là. Le nombre de possibilités à examiner par la méthode naïve était donc de $2^4 = 16$, mais il n'a fallu en tester que 7 en utilisant l'arbre.

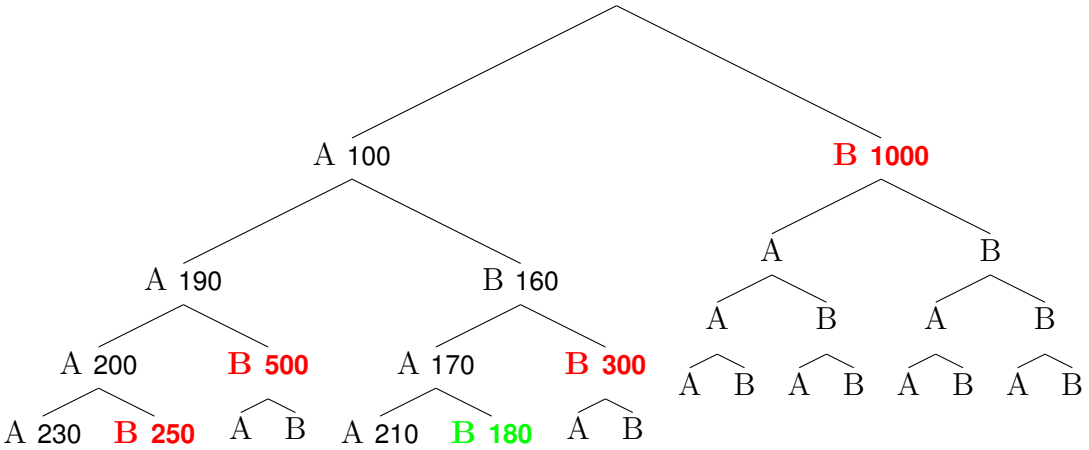


Figure 1: Arbre pour $\mathcal{L} = \{l_1, l_2, l_3, l_4\}$ et $\mathcal{E} = \{A, B\}$. Le coût de chaque nœud est inscrit en gras. En vert, le nœud contenant la solution de coût minimal. En rouge, les nœuds "coupés", c'est-à-dire dont les nœuds enfants ne sont pas explorés.

| $l_1$ | $l_2$ | $l_3$ | $l_4$ | coût |
|---|---|---|---|---|
| A | A | A | A | 230 |
| A | A | A | B | 250 |
| A | A | B | | 500 |
| A | B | A | A | 210 |
| A | B | A | B | 180 |
| A | B | B | | 300 |
| B | | | | 1000 |

Table 1: Parcours en profondeur de l'arbre de la figure 1

### 2.1.3 Contraintes

Des contraintes spécifiques au problème (par exemple, quotité de travail des enquêteurs dans 3.1) peuvent limiter le nombre de solutions acceptables. Dans le cadre de l'algorithme défini en 2.1.1, il suffit de fixer $C(S_k) = +\infty$. La fonction $C$ conserve sa propriété de croissance : la seule implication est de "couper" les nœuds enfants de k qui ne vérifient pas les contraintes, qui ne seront alors pas parcourus.

### 2.1.4 Amélioration par initiation

Reprenons l'exemple développé en figure 1. Si l'on avait interverti l'ordre de test de l'affectation $(A; B; A; B)$ et de l'affectation $(A; A; A; A)$, on aurait pu ne tester que 5 possibilités (au lieu de 7) avant de prouver que $(A; B; A; B)$ est la solution optimale. Bien entendu, en pratique, il est très peu probable de pouvoir tester la solution effectivement optimale en premier, mais définir une heuristique permettant d'examiner en premier une solution de coût faible peut permettre de gagner un temps de calcul appréciable. Comme le note Clausen [2], il convient toutefois de ne pas affecter des ressources trop importantes à l'élaboration d'une telle solution, le gain étant finalement relativement limité à l'échelle du problème. En pratique, lors de la résolution du problème de réallocation, les enquêteurs et les logements à enquêter étaient placés sur une carte de la région sélectionnée, ce qui

permettait de tracer à la main une allocation initiale s'approchant de façon satisfaisante de l'optimalité.

## 2.2 Réduction du problème

On se concentre ici sur l'application de l'algorithme général précédent à notre problème de réallocation de logements à enquêter.

### 2.2.1 Sélection d'enquêteurs

Au sein d'une même région de gestion, le nombre d'enquêteurs peut dépasser 50, ce qui est trop élevé pour permettre une intégration telle quelle dans l'algorithme de résolution. On peut proposer plusieurs méthodes de réduction du nombre d'éléments de $\mathcal{E}$ :

**Réduction géographique** Cette condition pose peu de problèmes pratiques : il s'agit de restreindre les contours de la réallocation à des zones géographiques suffisamment disjointes pour que la résolution séparée du problème dans chaque zone conduise à une perte d'optimalité la plus limitée possible. En France, le découpage naturel par région de gestion de l'INSEE est largement satisfaisant, même si la stratégie demande à être précisée dans les régions les plus denses (voir 3.2).

**Réduction aux enquêteurs les plus performants** Malgré la réduction géographique, le nombre d'enquêteurs du problème peut-être bien supérieur à 20, ce qui sera au-delà de nos capacités de calcul. Il faut alors opérer une sélection d'enquêteurs vis-à-vis d'un critère spécifique. Le choix d'un critère portant sur l'efficacité de l'enquêteur semble logique, car on cherche souvent à minimiser le temps de collecte global ou à maximiser l'efficacité de la collecte.

### 2.2.2 Groupes de logements enquêtés

Le nombre de logements du problème ne peut bien entendu pas être réduit. Dans les problèmes rencontrés en partie 3, ce nombre de logements est supérieur à 100, ce qui est bien trop grand en vertu de la formule donnée en 1.2.

Il est toutefois possible d'opérer des regroupements de logements. La méthode utilisée consiste à d'abord classer les logements à enquêter parmi des groupes (par exemple zone géographique $\times$ strate de tirage) de logements estimés équivalents en vue de la réduction du risque de zone orpheline. Au sein de chaque groupe, les logements sont intégrés au problème par groupes de $p$ logements, où $p$ est choisi de manière à intégrer un nombre total soluble en un temps raisonnable d'éléments au problème. Si ce nombre est de l'ordre de 20, le problème pourra être résolu en un temps de calcul raisonnable.

De cette façon, la solution donnée par notre programme de résolution est évidemment seulement sous-optimale. En effet, en notant $P = \{P_1, \cdots, P_K\}$ la partition de $\mathcal{L}_k$ avec les groupes constitués, cela revient à imposer la contrainte : $\forall (I, I') \in P_i, S(I) = S(I')$. Dans l'exemple de la sous-allocation (paragraphe 3.1), où l'objectif est à la fois de limiter la dispersion des poids et la présence de zone orpheline, cela se traduira par une dispersion légèrement inoptimale (et donc une légère augmentation de variance *in fine*). Notons

qu'en aucun cas la sous-optimalité ne peut se traduire par la non-affectation d'une partie des logements (ce qui reste le point le plus crucial).

# 3 Application

Les méthodes de sous-allocation et de réallocation ont été mises en oeuvre et testées dans le cadre de l'Enquête Nationale Logement 2013.

## 3.1 Avant la collecte : sous-allocation

Le problème et sa mise en place sont décrits dans Haag et Pendoli (2014) [4]. Le problème est le suivant : suite à une absence prévue d'un enquêteur auquel une zone de logements à enquêter a été affectée, on cherche à réallouer ces logements à d'autres enquêteurs situés dans les zones voisines. L'objectif est alors de minimiser conjointement le nombre de logements perdus et la dispersion des poids de tirage. Ceci permet de baisser la variance finale (augmentation du nombre de logements ainsi que dispersion plus faible en vue d'un calage, comme discuté par Ardilly (2006) [1]).

On peut utiliser le branch and bound pour résoudre le programme avec le choix suivant de fonction de coût.

**Fonction de coût** On écrit :

$$\mathcal{L} = \mathcal{L}_{deja\ affectees} \cup \mathcal{L}_{zone\ orpheline}$$

puis ensuite :

$$\mathcal{L}_k = \mathcal{L}_{deja\ affectees} \cup \mathcal{L}_{zone\ orpheline,k}$$

où $\mathcal{L}_{deja\ affectees}$ désigne les logements ayant un enquêteur qui leur est attribué au début du problème (le cardinal de l'ensemble de ces logements affectés se verra donc diminué par la sous-allocation), et $\mathcal{L}_{zone\ orpheline}$ désigne les logements de la zone orpheline. Cela revient en fait à parcourir l'arbre en ajoutant peu à peu dans les solutions candidates les fiches adresse de la zone initialement orpheline.

On utilise la fonction de coût suivante :

$$C = \alpha \cdot n^2_{logements\ affectes} + \beta \cdot \left( \mathrm{D} - \frac{1}{n_{region}} \cdot \sum_{logement \in region} (\omega_{logement} - \bar{\omega})^2 \right)^2$$

où $\omega_{logement}$ = poids des logements de la région, D = dispersion des poids des logements de la région dans le cas hypothétique où 1 seul logement serait enquêté dans la zone orpheline (voir [1]), et $\alpha$ et $\beta$ doivent être choisis empiriquement pour accentuer l'importance d'un des paramètres dans la fonction de coût.

Cette fonction de coût possède la propriété de croissance partielle, et il est judicieux d'utiliser la variante décrite en partie 2 où l'on fixe $\mu = \min_{k \in [[\tau, M]]} C(S_k)$. La preuve de ce résultat repose sur le plan de sondage utilisé, et fera l'objet d'un développement futur. Dans le cas du plan de sondage de l'Enquête Nationale Logement 2013, le tirage limite (borne supérieure et borne inférieure) les poids de chaque logement. Ainsi, à chaque

nœud k, l'insertion dans l'arbre d'un nouveau logement de l'ensemble $\mathcal{L}_{orphelines,k}$ fait nettement augmenter le facteur $\left( \mathrm{D} - \frac{1}{n_{region}} \cdot \sum_{logement \in region} (\omega_{logement} - \bar{\omega})^2 \right)^2$, ce qui assure la croissance jusqu'à un certain $\tau$.

Notons qu'expliciter la preuve permettra également de donner une formule analytique pour $\tau$ ainsi que des valeurs précises de $\alpha$ et $\beta$ permettant d'assurer la croissance partielle pour différents plans de sondages.

## 3.2 Pendant la collecte : réallocation

À ce stade, la collecte a déjà commencé, et on ne peut plus changer l'affectation des enquêteurs ou les poids de tirage correspondant à cette affectation. On s'aperçoit d'un trou de collecte, soit par absence non prévue d'un enquêteur, soit à cause de mauvaises performances de collecte. On cherche à réaffecter une partie des logements affectés à des enquêteurs peu performants à d'autres plus performants dans les zones voisines, ainsi qu'à prioriser la collecte de certains logements non réaffectés. Afin de ne pas affecter les poids de tirage, on ne peut respecter la contrainte de temps de travail pour les enquêteurs. Dans notre cas, une prolongation du délai de fin de collecte pour l'enquête Logement a permis d'augmenter la quotité totale disponible pour l'enquête.

**Fonction d'objectif** On cherche une réallocation avec les enquêteurs "les plus proches et les plus performants". On utilise la fonction de coût suivante :

$$C = \alpha \cdot \sum_{e_k \in \mathcal{E}} T_{e_k}^2 + \beta \cdot \sum_{e_k \in \mathcal{E}} d_{e_k}^4$$

où $d_{e_k}$ = temps supplémentaire de parcours pour chaque enquêteur, $T_{e_k}$ = temps nécessaire pour effectuer la collecte, estimé *via* une vitesse de collecte pour chaque enquêteur : $T_{e_k} = \frac{Tcollecte\ deja\ passe}{n_{logements\ deja\ enquetes}} \cdot n^{final}_{logements\ a\ enqueter}$, et $\alpha$ et $\beta$ doivent être choisis empiriquement pour accentuer l'importance de l'un ou l'autre des paramètres dans la fonction de coût. $C$ est bien croissante au sens défini en 2.1.1.

**Contraintes** On n'accepte pas les solutions imposant à un enquêteur un déplacement supérieur à soixante minutes depuis sa zone d'origine :

$$C(S_k) = +\infty \text{ si } d \geq 60 \text{ minutes}$$

**Application** La réaffectation se fait pour les enquêteurs de la région Île-de-France, densément peuplée et par conséquent dense en nombre de logements à enquêter. Cette région concentre également les problèmes de collecte pour l'Enquête Nationale Logement 2013. On groupe les logements enquêtés par (zone d'action enquêteur) × (strate de tirage), et on se concentre sur la réallocation des fiches adresse des groupes présentant des taux de collecte inférieurs à 25%. Ceci permet de sélectionner un peu plus de 200 logements à enquêter.

Le découpage du problème (75 enquêteurs) se fait en trois zones géographiques choisies à l'aide d'une carte géographique sur laquelle étaient placées les localisations de base des

enquêteurs et des zones de logements à enquêter. Il subsiste trois zones, dont la plus grande est constituée de 30 enquêteurs et 102 logements à réaffecter. On découpe ces logements selon 15 groupes de maximum 7 logements, comme expliqué en partie 2.2.2. On sélectionne les 10 meilleurs enquêteurs de la région géographique sélectionnée. Ces enquêteurs sont également distribués dans tout l'espace du problème, ce qui laisse supposer que l'existence d'une solution n'est pas une hypothèse absurde.

Le problème est finalement constitué de 10 enquêteurs et de 15 groupes de logements (dans lesquels figurent 102 logements à enquêter). On procède enfin à une pré-allocation "à la main" pour faciliter le calcul, comme discuté en 2.1.4. La résolution s'effectue en quelques dizaines de minutes sur un simple PC de bureau doté d'un Intel Centrino 2 cadencé à 2,4 GHz.

# 4 Conclusion

Ainsi, afin d'anticiper un éventuel biais de non-réponse non-ignorable, on préfère réallouer les zones d'enquête non couvertes à d'autres enquêteurs dont les zones d'action sont situées à proximité. Le problème de réallocation, bien que très coûteux en calculs, peut être traité informatiquement moyennant une simplification spécifique au problème et l'utilisation d'un algorithme d'optimisation fondé sur le principe du branch and bound. L'efficacité de l'application de cette méthode au problème de sous-allocation présenté en [1] demandera à évaluée de façon plus précise.

# Bibliographie

[1] P.Ardilly (2006), Les techniques de sondage, *Editions Technip.*
[2] J. Clausen (1999), Branch and Bound Algorithms - Principles and Examples. Department of Computer Science, University of Copenhagen, Universitetsparken 1, DK-2100 Copenhagen, Denmark.
[3] A. H. Land et A. G. Doig (1960), An automatic method of solving discrete programming problems. *Econometrica 28 (3)*, pp. 497-520
[4] P.A. Pendoli et O.Haag (2014), Méthodologie de la sous-allocation, *Preprint.*
[5] B. Schouten, M. Calinescu et A. Luiten (2013), Optimiser la qualité de la réponse au moyen de plans de collecte adaptatifs, *Techniques d'enquête, juin 2013, Vol. 39, N° 1, pp. 33-66*, Statistique Canada