RAIRO. RECHERCHE OPÉRATIONNELLE

VINCENT GIARD
CHRISTINE TRIOMPHE

Investissement et flexibilité organisationnelle

RAIRO. Recherche opérationnelle, tome 29, n° 3 (1995), p. 299-320

<http://www.numdam.org/item?id=RO_1995__29_3_299_0>

Recherche opérationnelle/Operations Research
(vol. 29, n° 3, 1995, p. 299 à 320)

INVESTISSEMENT ET FLEXIBILITÉ ORGANISATIONNELLE (*)

par Vincent GIARD ([1]) et Christine TRIOMPHE ([1])

Résumé. – *Tout dossier d'investissement repose sur une analyse des conséquences physiques induites par son introduction. Cette analyse est difficile lorsque la demande est à la fois fortement cyclique et aléatoire, lorsque la production s'effectue en partie à la commande et en partie sur stock et, enfin, lorsque l'organisation, relativement rigide jusqu'alors, doit être transformée en profondeur pour permettre une bonne utilisation du nouvel équipement. La démarche retenue pour éclairer la décision s'appuie sur une utilisation en cascade de deux modèles du fonctionnement du centre de tri. Le premier se place en univers certain, s'appuie sur une formulation de programmation linéaire mixte et est utilisé dans une approche de système d'aide à la décision. Le second est un modèle de simulation visant à tester simultanément la robustesse de la définition du système productif trouvée au premier modèle et un ensemble de règles adaptatives qui pilotent le système au cours de la simulation.*

Mots clés : Investissement, flexibilité, production.

Abstract. – *An investment file relies on an analysis of the physical consequences of its introduction. This analysis is difficult when demand is both cyclic and stochastic, when the production is partly processed on order and partly for stock and, lastly, when the organization, thus far relatively rigid, must be thoroughly revised in order to put the new equipment to good use. The procedure intended to support actual decision for the French Post Office rests on the succesive use of two models of the sorting center process. The first is deterministic, based on integer programming formulation and is used in a DSS way. The second is a simulation model aimed to test the robustness of both the system design found by the first model and of a set of adaptive rules which monitors the system during simulation.*

Keywords: Investment, flexibility, production.

Tout dossier d'investissement repose, dans un premier temps, sur une analyse des conséquences physiques induites par son introduction et, dans un second temps, sur un processus de valorisation de ces conséquences. L'analyse des conséquences physiques d'un investissement d'automatisation est relativement aisée lorsque celui-ci est destiné à satisfaire une demande qui est importante en volume et relativement stable dans le temps. Il n'en est pas de même lorsque, comme c'est le cas dans les centres de tri, la

(*) Reçu en août 1993.
([1]) IAE de Paris (Université Paris 1) et ENSPTT.

Recherche opérationnelle/Operations Research, 0399-0559/95/03/$ 4.00

demande est à la fois fortement cyclique (avec combinaison d'un cycle intra-journalier et de cycles plus longs) et aléatoire (la composante aléatoire étant relativement importante par rapport aux composantes certaines), lorsque la production s'effectue en partie à la commande (courrier urgent devant être traité en totalité avant l'heure de coupure laquelle correspond à la dernière remise possible de la production triée pour l'acheminement vers un autre centre de tri ou vers les bureaux de distribution) et en partie sur stock (courrier économique) et, enfin, lorsque l'organisation, relativement rigide jusqu'alors, doit être transformée en profondeur pour permettre une bonne utilisation d'un nouvel équipement. Le problème de la détermination des caractéristiques de l'équipement devient indissociable de celui de la réorganisation du processus de production et de la gestion des autres ressources (en particulier celle des opérateurs) ; il faut déterminer la flexibilité (obtenue par une combinaison de moyens physiques et organisationnels) de la nouvelle configuration productive ainsi que le niveau induit de qualité de service. C'est à un problème complexe de cette nature que *La Poste* est confrontée avec l'introduction d'une TOP (trieuse d'objets plats) dans un second centre de tri, avant de généraliser cette introduction dans les autres centres de tri. Ce travail est le fruit d'un important contrat de recherche avec *La Poste*.

La démarche retenue pour éclairer la décision s'appuie sur *une utilisation en cascade de deux modèles* du fonctionnement du centre de tri. Le premier se situe en univers certain et a pour objectif de définir les ressources permanentes du système productif et une organisation de la production satisfaisant les principales contraintes techniques et commerciales. On commencera (§ 1) par décrire les principes qui fondent ce modèle (renvoyant en annexe sa description exhaustive), avant de présenter le processus de modélisation suivi (§ 2) qui s'avère aussi important que le modèle lui-même dans la mesure où il peut ou non permettre de concilier la qualité d'une description précise et la possibilité d'une résolution numérique. Le second modèle (§ 3) s'appuie sur les ressources permanentes définies par le premier modèle et une description un peu plus fine des processus productifs ; il utilise une simulation sur plusieurs dizaines de jours des demandes en reprenant leurs caractéristiques aléatoires et cycliques. Il est évident que le programme de production trouvé en univers certain n'est plus pertinent, c'est pourquoi ce second modèle a pour premier objectif d'éprouver un ensemble de principes retenus pour piloter la production sans connaître à l'avance les demandes (ce qui le distingue des approches de simulation classique). Le second objectif de ce modèle est de tester la « robustesse » de la solution trouvée par le premier

modèle, c'est-à-dire la capacité des ressources permanentes à satisfaire des demandes aléatoires et cycliques, mais, dans la mesure où cette capacité dépend étroitement des principes de pilotage proposés, il ne s'agit pas d'une « pure » analyse de robustesse.

1. LA DESCRIPTION DU MODÈLE EN UNIVERS CERTAIN

Ce premier modèle (dont on peut trouver une présentation détaillée en annexe) se place en univers certain ; il retient un découpage temporel du quart d'heure sur une journée de fonctionnement dont les arrivées de courrier sont considérées comme caractéristiques du régime de croisière. Ce travail s'appuie sur une formulation par la *programmation linéaire mixte* dont la fonction objectif est de minimiser la somme des dépenses de personnel $\sum_h \Gamma_h \upsilon_h$, où υ_h est le nombre de personnes à qui le service h, caractérisé par un horaire de présence et un coût Γ_h, a été donné.

Comme tout système productif, un centre de tri est caractérisé par des équipements et des opérateurs. Dans le problème traité :

– les équipements en machines sont connus en nombre et en caractéristiques (possibilité de tri et/ou de lecture optique, mode d'alimentation, nombre de directions du tri des équipements automatisés, nombre d'antennes de la TOP, spécification du courrier à traiter, rendements, taux de rejet...) mais rien ne s'opposerait à ce que le nombre de certaines machines ou la définition de certaines caractéristiques soient des variables de commande du système ; c'est l'un des usages possibles du modèle, d'autant que l'analyse économique d'une transformation marginale du système productif ne peut en aucun cas s'effectuer sur la base de valeurs moyennes, du fait du caractère très fortement cyclique des arrivées et des contraintes de production à la commande qui pèse sur une partie de la production ;

– les opérateurs sont polyvalents, toutefois on pourra, sans perte de performance compte tenu de la taille des effectifs en jeu, spécialiser une partie d'entre eux sur du courrier « *Petit Format* » et le reste sur le « *Grand Format* », avec des avantages organisationnels (amélioration de la qualité due à une meilleure responsabilisation) ; le nombre d'opérateurs présents au cours d'une période est la somme des effectifs des services dont les horaires de présence incluent la période considérée (ces effectifs étant des variables de commande du système) ; l'affectation du personnel durant la période découle de la production retenue pour la période, sur chaque équipement (ces productions étant des variables de commande du système).

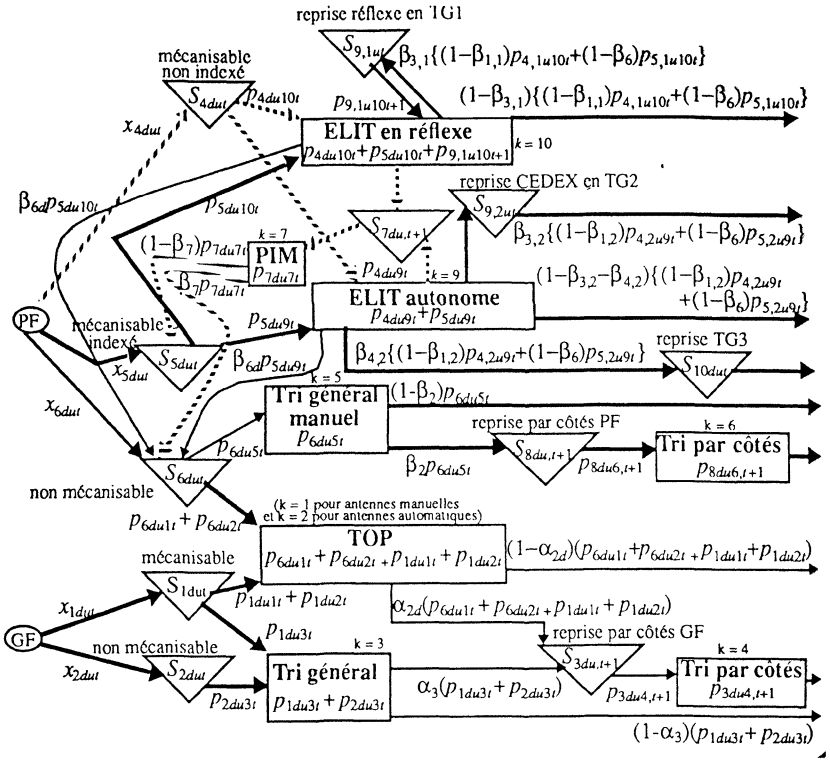

Figure 1. – Fonctionnement simplifié du système productif.

Le processus est décrit par la cartographie des flux de la figure 1 où l'on voit que la production s'effectue par l'intermédiaire de processeurs (machines avec leurs opérateurs ou postes de travail manuels) qui prélèvent des éléments à traiter dans des stocks et expédient les éléments traités dans des stocks:

– La dynamique du système est décrite à travers des équations de conservation des stocks (qui seront données, une fois introduites toutes les variables utilisées) qui stipulent que « le stock de début de période (et donc de fin de la période précédente) est égal au stock de début de la période précédente, augmenté des arrivées dans le stock au cours de la période précédente en provenance de l'extérieur ou de postes de travail, et diminué des prélèvements sur le stock, au cours de la période précédente, à destination d'autres postes de travail ou de l'extérieur ». Dans le système étudié, un stock est noté S_{idut} pour repérer sa localisation i (11 stocks retenus), la classe de destination du courrier traité d (2 occurrences: « trafic à destination du département » et « trafic à destination d'autres départements »), la catégorie d'urgence du courrier traité u (2 occurrences: « trafic urgent » et « trafic

non urgent »); il se définit au *début* de la période t ($24 \times 4 = 96$ périodes d'un quart d'heure), avant toute nouvelle arrivée de trafic. Les stocks sont initialisés en $t = 1$ à une valeur nulle pour les flux de courrier urgent et à une valeur « suffisante » pour les autres flux. La contrainte de niveau de service pesant sur le courrier urgent imposera d'avoir un stock nul, au moment de la coupure, pour ce type de courrier; pour les autres, une contrainte impose la reconstitution du stock initial, à la fin de la dernière période.

– L'utilisation du système est décrite par l'intermédiaire des productions p_{idukt} effectuées par un processeur k, au cours de la période t (mais réputée livrée en totalité au début de la période $t+1$), à partir d'éléments prélevés dans le stock i. La sommation par rapport à i, u et d (qui est conservé pour des raisons de généralité sachant que les plans de tri de ces flux sont incompatibles) permet de calculer la production du processeur sur la période, celle-ci devant être inférieure ou égale à la capacité disponible, laquelle est égale à $\pi_k \, n_{kt}$, où n_{kt} est l'effectif associé au processeur k (effectif limité techniquement) durant la période t et π_k est le taux de production de cette ressource; $\sum_i \sum_u \sum_d p_{idukt} \leq \pi_k \, n_{kt}$; techniquement, on autorise le mélange des flux de production quelle que soit la catégorie d'urgence (indice u) mais pour le seul trafic à destination du département ($d = 2$); les équations de conservation des stocks sont alors:

$$S_{idu,\,t+1} = S_{idut} + (x_{idut} + \sum_{i' \neq i} \sum_k \beta_{duk} \, P_{i'duk,\,t-1}) - \sum_k P_{idukt}$$

où x_{idut} correspond à des arrivées de flux venant de l'extérieur et où seulement la part β_{duk} des productions $P_{i'dukt}$ va dans le stock i.

– Les effectifs requis pendant une période s'obtiennent par sommation des effectifs utilisés par les différents processeurs; ils ne peuvent dépasser les effectifs disponibles $\sum_h \mu_{ht} \, v_h$, où μ_{ht} vaut 1 si le service h implique une présence pendant la période t, et 0, dans le cas contraire: $\sum_k n_{kt} \leq \sum_h \mu_{ht} v_h$.

– Un certain nombre de contraintes techniques additionnelles doivent être prises en compte:

• chacun des 2 groupes de destinations du courrier implique l'utilisation d'un plan de tri spécifique (qui joue le rôle d'une gamme de production); le mélange de production doit respecter cette contrainte d'unicité du plan de tri; pour ce faire on introduit autant de variables indicatrice δ_{dkt} qu'il y a

de plan de tri (avec $\delta_{dkt} = 1$, si la gamme d est utilisée et $\delta_{dkt} = 0$, dans le cas contraire), on oblige à ne pas utiliser plus d'un plan de tri en posant $\sum_d \delta_{dkt} \leq 1$ et l'on force la variable indicatrice δ_{dkt} à être égale à 1 en cas de production du flux d en posant les contraintes $\delta_{dkt} \geq [\sum_i \sum_u p_{idukt}]/H$, où H est une valeur quelconque supérieure à la production totale la plus élevée au cours d'une période ;

• entre deux changements de plan de tri il y a impossibilité de produire durant une période pour pouvoir vider les cases (ce qui s'apparente à un temps de réglage) ; cette interdiction sera respectée en posant $\delta_{2\,kt} + \delta_{1\,k,\,t+1} \leq 1$ et $\delta_{1\,kt} + \delta_{2\,k,\,t+1} \leq 1$ (solution difficilement généralisable à plus de 3 ou 4 catégories).

Les variables de commande du système sont :

– les effectifs associés à chaque service offert (*voir* fonction-objectif),

– la production, période par période, de chaque processeur ; celle-ci induit une utilisation des effectifs, doit être compatible avec les effectifs disponibles et doit respecter les contraintes de niveau de service et les principales contraintes techniques.

Les modèles élaborés, dans une perspective d'exploration de scénarios alternatifs, sont décrits dans le formalisme du logiciel GAMS (de Scientific Press, *voir* Brooke *et al.*) qui interprète une description analytique du problème par des relations algébriques utilisant directement des variables et paramètres indicées et permettant l'appel aux opérateurs Σ et Π et utilise l'ensemble des données élémentaires non redondantes, pour générer automatiquement les matrices de données qui seront soumises à un solveur (en l'occurrence, OSL d'IBM, l'un des plus puissants disponibles, *voir* Murphy *et al.*). Dans sa version complète, ce problème utilise 12 022 variables (dont 3 714 entières) et comporte 9 227 contraintes, ces valeurs étant obtenues après élimination des variables nécessairement nulles et des contraintes sans objet. Il est résolu sur un 486-DX cadencé à 50 MH, sous système d'exploitation DOS, avec gestion d'une mémoire étendue à 16 Mo.

2. LE PROCESSUS DE MODÉLISATION

Des considérations de performance militent en faveur d'une transformation du modèle, à condition que l'on ait la quasi-certitude que cette transformation ne dégrade pas la solution (§ 2.1). Par ailleurs, la formulation du problème

et en particulier la détermination de certains paramètres ou variables de commande (comme la définition des services à proposer) et la difficulté de décrire très finement une réalité complexe militent en faveur d'une utilisation des logiciels d'optimisation dans une optique d'aide à la décision où la solution trouvée rétroagit sur la formulation du problème à résoudre (§ 2.2).

2.1. La simplification du modèle

La taille du modèle incite à utiliser, dans la mesure du possible, la scission du problème initial en plusieurs problèmes indépendants ; deux possibilités ont été utilisées dans cette perspective.

L'élimination de certaines variables du modèle et la duplication de certaines contraintes avec restriction des variables utilisées à des ensembles disjoints peut transformer la matrice des contraintes en une matrice diagonale-bloc qui autorise la scission du problème initial en autant de problèmes indépendants que de blocs. Cette transformation n'est acceptable que si, pour des raisons évidentes, les variables éliminées ne jouent qu'un rôle mineur dans le problème, c'est-à-dire que leur retrait n'altère pas significativement la description du réel (en particulier, la solution doit rester physiquement acceptable) et qu'il ne dégrade pas la valeur de l'optimum. Il en est parfois ainsi pour des variables qui assurent la liaison physique entre deux sous-systèmes productifs relativement indépendants, différant notablement par la nature des ressources et des caractéristiques physiques de flux. Dans cette optique, on a pu remplacer le problème initial par 2 problèmes de taille voisine en :

– éliminant du modèle les flux $p_{6\,dult}$ et $p_{6\,du\,2\,t}$, qui correspondent à une possibilité d'utilisation de la TOP pour traiter du *Petit Format* mais dont l'intérêt pratique, est à l'évidence, très marginal,

– dupliquant les équations de contrainte d'effectifs utilisées ne dépassant pas les effectifs disponisbles après spécialisation des effectifs sur du *Grand Format* ou du *Petit Format*, ce qui est acceptable parce que ces effectifs restent importants dans chaque cas et que cette spécialisation améliore la responsabilisation.

On peut ensuite tirer parti de caractéristiques de solutions partielles qui sont induites par certaines contraintes. On a dit en introduction que dans le système productif étudié, il y a coexistence d'une production sur stock (tri du courrier non urgent) et d'une production à la commande (tri de courrier urgent) devant être traité en totalité avant l'heure de coupure. L'importance relative de la production à la commande et la brièveté du délai disponible pour exécuter des demandes étalées dans le temps font qu'il est certain que l'on

fera un appel massif aux centres les plus productifs. Il est alors possible de faible appel à un raisonnement de programmation dynamique pour déterminer un nombre limité de solutions optimales alternatives du problème partiel, que la solution optimale du problème global doit comporter. Le raisonnement proposé (de type programmation dynamique rétrograde) est assez simple.

Désignons par x_t, le trafic arrivé au début de la période t (avec t variant de 1, première période d'arrivée du trafic, à T, période à la fin de laquelle la production doit être achevée) et par X_t, le trafic arrivé jusqu'à la période t : $X_t = \sum_{\tau=1}^{t} x_\tau$. On calcule l'indicateur $K_t = \dfrac{X_T - X_{t-1}}{T - t + 1}$ qui s'interprète comme la production moyenne périodique à écouler de la période t à la période T pour satisfaire en totalité la demande exprimée entre ces deux périotes (avec $X_0 = 0$). Désignons par Q_t, la capacité constante de production définie de la période t à la période T. Il est évident :

– que si antérieurement à la période t, tout le trafic arrivé a été traité,

– et que si Q_t est supérieur aux différentes valeurs K_t calculées entre t et T, la contrainte de niveau de service sera respectée. La solution la plus économique, pour le problème des périodes t à T est obtenue pour la capacité la plus faible satisfaisant la demande, c'est-à-dire pour la plus forte des valeurs K_t. En reculant d'une période, la capacité nécessaire est la plus forte des deux valeurs K_{t-1} et Q_t : $Q_{t-1} = \text{Max}\,(Q_t,\ K_{t-1})$. En poussant le raisonnement jusqu'en $t = 1$, on obtient la capacité minimale à installer, permettant de produire *au plus tard et sur stock*, la production à la commande. Cet algorithme amène plusieurs remarques :

– désignons par θ, la période pour laquelle l'indicateur K_t est égal à la capacité installée ; cette capacité est en réalité la capacité minimale requise pour les seules périodes θ à T ; pour les périodes antérieures, une capacité plus faible peut être installée ; ce problème défini pour les périodes 1 à $\theta - 1$ est indépendant de celui des périodes θ à T (c'est-à-dire que l'on est donc en présence d'une propriété d'horizon de planification dans ce problème dynamique) ; la combinatoire des solutions alternatives reste en pratique très limitée et le choix de la plus intéressante d'entre elles dans le problème global s'effectue par le biais de variables indicatrices (valant 0 ou 1) dont la somme est égale à 1 (obligation de sélection de l'une des alternatives) et qui interviennent dans les contraintes de production et d'utilisation des ressources ;

– la capacité théorique ainsi calculée sera presque toujours inférieure à la capacité effectivement installée, pour des raisons d'indivisibilité des

ressources ; ceci conduit à pouvoir retarder l'heure de début de la production et, accessoirement, à limiter la combinatoire évoquée précédemment ;

– lorsque les ressources sont hétérogènes, il est normal, compte tenu des contraintes de niveau de service, d'utiliser en priorité les ressources les plus productives ; ceci conduit à définir l'augmentation de capacité par simple adjonction de la ressource la plus productive qui n'est pas encore utilisée.

2.2. Une approche orientée SIAD

Tout modèle implique un arbitrage entre la qualité de représentation du réel et la facilité (ou, parfois même, la possibilité) de manipulation de cette représentation à des coûts acceptables. Dans le cas de modèle d'optimisation, cette facilité est moindre et plus coûteuse en temps et capacité de traitements ; l'obligation de compromis conduit à considérer le modèle comme une étape d'un *processus itératif complexe où une solution trouvée rétroagit sur la formulation* du problème d'optimisation, jusqu'à ce que l'on ait obtenu une formulation et une solution jugées satisfaisantes. Appliqué à la prise de décisions complexes, le modèle d'optimisation ne s'inscrit pas dans une logique de décisions programmables (au sens de H. Simon) mais dans celle de décisions semi-programmables s'appuyant sur un système d'aide à la décision s'articulant sur le modèle d'optimisation. On illustrera cette conception par trois rétroactions importantes qui ont été utilisées.

a) Tout d'abord, les variables de commande retenues expriment un point de vue partiel du problème réel qui, généralement, dispose de davantage de degrés de liberté :

– de nombreux paramètres du problème sont en fait la traduction implicite soit de décisions anciennes que l'on se refuse à remettre en cause, soit de décisions nouvelles qui sont arbitrairement arrêtées pour limiter la complexité du problème à résoudre (ce sera le cas, par exemple, des spécifications retenues pour les équipements du système étudié)

– la définition des variables de commande, en particulier celles qui traduisent des alternatives organisationnelles, est également souvent arbitrairement limitée pour des raisons de calculs.

Un exemple intéressant de ce dernier cas de figure peut être trouvé avec la *définition des services offerts. A priori*, il est possile de définir plusieurs centaines de services différents mais il est totalement irréaliste de le faire car le problème consiste non seulement à déterminer des effectifs par service retenu mais encore à trouver un programme de production détaillé qui s'appuie sur les effectifs présents et respecte les nombreuses contraintes

productives. Après avoir résolu un problème reprenant l'offre actuelle de services, afin d'obtenir une solution optimisée de référence ne remettant pas en cause l'organisation actuelle, on a cherché à définir une offre de service optimale. La démarche suivie est en deux temps :

– un problème fictif, caractérisé par une offre de service comportant quelques services de nuit (s'arrêtant au moment de l'expédition du courrier trié aux bureaux de poste) et une dizaine de services différents de 2 heures allant de 6 heures à 24 heures, a été optimisé,

– les effectifs des services de nuit sont retenus tandis que ceux des services de 2 heures sont considérés comme une présence minimale qui est couverte, pour chaque tranche horaire, par le cumul des effectifs de services de 6 ou 8 heures comportant la tranche horaire considérée et que l'on définit pour minimiser l'excédent de l'offre par rapport à la demande

• soit empiriquement, avec des spécialistes de la gestion des ressources humaines pour réfléchir en commun à cette composante importante de la flexibilité organisationnelle qu'est l'offre de services, ce qu'illustre le tableau suivant de définition des services spécialisés dans le *Petit Format* (aucune présence n'est requise en matinée ; exemple de solution possible de services possibles quasi optimale, parce que n'induisant que 2 heures de travail perdues, mais d'autres solutions sont possibles, un peu moins bonnes économiquement mais préférables socialement) ;

Périodes	12H/14H	14H/16H	16H/18H	18H/20H	20H/22H	22H/24H	0H/5H	5H/6H30
Effectifs demandés	3	18	24	21	22	9	31	6
Services offerts + effectifs	3	3	3	0	0	0	0	0
	0	15	15	0	0	0	0	0
	0	0	0	16	16	0	0	0
	0	0	6	6	6	0	0	0
	0	0	0	0	0	9	9	0
	0	0	0	0	0	0	16	0
	0	0	0	0	0	0	6	6

• soit de manière « plus scientifique » (ce qui n'est pas nécessairement plus efficace du point de vue de l'optimum et sans doute pas du point de vue de la dynamique de groupe) en résolvant le programme linéaire

Min. $\left(\sum_{h'=1}^{H'} c_{h'}\, y_{h'}\right)$ où $y_{h'}$ est l'effectif retenu pour le service h' (le SGRH a défini les H' services possibles), de coût $c_{h'}$, sous contrainte d'une offre en personnel supérieure ou égale à la demande, c'est-à-dire : $\sum_{h'=1}^{H'} \tau_{h't}\, y_{h'} \geq n_t$, où $n_t = \sum_h \mu_{ht}\, v_h$ est l'effectif minimal présent durant la période t et $\tau_{h't} = 1$ si le service h' implique un travail possible sur la période t et $\tau_{h't} = 0$, dans le cas contraire.

b) La formalisation retenue relaxe nécessairement un certain nombre de contraintes (dont certaines sont du reste très difficiles à formaliser). Il faut donc toujours s'assurer de l'applicabilité de la solution trouvée. Dans la négative, on la transformera, par tâtonnement en cherchant à respecter les contraintes relaxées, sans dégrader l'optimum trouvé. Troix axes d'amélioration ont été suivis.

Tout d'abord, la dimension combinatoire du problème est telle qu'il existe un nombre extrêmement élevé de programmes alternatifs d'utilisation du personnel présent (et donc de solutions économiquement interchangeables) ; la majorité de ces programmes se caractérise par des changements d'affectation manifestement excessifs du personnel disponible qui se traduisent par un véritable « mouvement brownien » dans le système productif ; il a donc fallu chercher à lisser la présence du personnel sur chaque chantier :

– en introduisant, dans la fonction objectif du programme linéaire, une pénalité liée à la variation des effectifs sur le chantier (le coût du personnel restant dominant à l'optimum) ;

– en limitant la disponibilité de certains équipements faiblement utilisés, pour privilégier certains traitement « au fil de l'eau » (il est préférable, si les stocks sont suffisants, d'utiliser 2 personnes sur un chantier pendant 4 périodes d'un quart d'heure que d'en mobiliser 8 pendant 1 quart d'heure).

Ensuite, la possibilité de produire sur une même ressource, au cours d'une même période, simultanément à la commande (courrier urgent) et pour stock (courrier économique), une même catégorie de courrier (par exemple, *Grand Format* à destination extra-départementale, cette possibilité n'étant finalement pas exploitée dans le modèle retenu) ne doit pas se traduire par un panachage systématique ; une transformation de la solution en une solution économiquement équivalente doit être opérée pour limiter ce mélange, en particulier pour le courrier qui doit être envoyé à d'autres centres de tri

(le courrier économique mélangé au courrier urgent étant ultérieurement considéré comme du courrier urgent) ; pour ce faire, on utilise :

– des considérations de *flexibilité* qui font que l'on a pas intérêt à effectuer une programmation au plus tard de la production à la commande ; en effet, le volume de travail pour stock qui est programmé avant la coupure constitue un matelas permettant d'absorber des retards d'arrivée et/ou des augmentations de volume ; on peut donc déterminer la robustesse de la solution proposée à des perturbations exogènes ;

– des considérations de *continuité* qui inciteront à retarder le début de la production à la commande jusqu'au moment où l'on dispose d'une quantité de travail suffisante pour traiter en totalité sur stock la production à la commande, pour éviter tout mélange ; ce faisant, la date de fin de la production d'ensemble sera la plus précoce possible (ce qui va dans le sens du principe précédent) mais si cette date est également la plus tardive, il sera prudent d'accepter un certain mélange.

c) Enfin, l'existence d'une demande cyclique fait que les caractéristiques moyennes de l'univers certain ne permettent pas de définir de solution garantissant systématiquement que la production à la commande sera exécutée sans retard ; on peut alors décider de surdimensionner le système productif en se fondant sur le trafic urgent des journées les plus chargées (en utilisant l'algorithme proposé au § 2.1) ; on dispose alors pour les autres jours d'une marge de manœuvre supplémentaire pour absorber les aléas et le lissage dynamique de la charge s'effectue par un report positif ou négatif du courrier non urgent du stock de début d'une période, sur le stock du début de la période suivante.

3. LE MODÈLE DE SIMULATION

Le second modèle simule, en « temps continu » et sur plusieurs dizaines de jours, le fonctionnement du système productif calibré en univers certain mais devant faire face à des demandes aléatoires qui combinent des variations en volume et des variations sur les heures d'arrivées du courrier. Il est évident que toute simulation reprenant le programme d'affectation des ressources aux différentes productions ne peut qu'aboutir à un niveau de service dégradé pour la production à la commande (alors qu'il était nécessairement maximal en univers certain). Il est non moins évident que la recherche, par tâtonnement, d'une programmation « satisfaisante » pour ce jeu de données, qui s'appuierait implicitement sur une connaissance préalable de l'ensemble

des demandes à venir, ne fournit qu'une faible présomption de viabilité du calibrage du système productif proposé et ne fournit guère d'assistance aux opérationnels qui devront gérer un système productif de conception assez nouvelle, les laissant faire seuls leur apprentissage. C'est pourquoi il a été retenu, dans cette simulation, de tester la performance de la mise en œuvre de principes généraux de pilotage du système productif qui n'impliquent aucune connaisance des demandes à venir ou de ce qui se passe en temps réel dans le système productif. Si la démonstration est faite que l'application de principes faciles à mettre en œuvre sur le terrain assure au système productif une « bonne » capacité de réponse, il y a, de notre point de vue, une présomption suffisante de la viabilité du couple « système productif en univers certain – principes d'organisation de la production ». Examinons comment ce concept de flexibilité a été rendu opérationnel dans notre approche.

A capacité fixée, l'obtention de la flexibilité repose sur les potentialités des ressources et l'organisation du système de production. Il s'agit principalement d'exploiter la flexibilité qualitative des ressources définie lors de la conception du système et de développer la flexibilité organisationnelle. Pour faire face aux fluctuations de la demande, des solutions complémentaires, mais qui n'ont pas les mêmes impacts sur le système productif (on ne modifie pas les mêmes variables de commande), sont envisageables :

– des transferts de charge entre les processeurs ; qui sont limités par les possibilités de substitution (par exemple la production du courrier *Petit Format* sur le chantier TOP) ou des tranferts de charge d'un produit à un autre, par variation des plages de traitement allouées aux différents types de produits sur un même processeur ;

– la variation de la capacité des différents processeurs, par modification de l'affectation des opérateurs présents ;

– le lissage de la charge par des fusions de catégories de flux, selon des stratégies de différenciation définies *a priori* (par exemple, le mélange des flux de courrier urgent et non urgent, sous certaines conditions).

La mise en œuvre de ces principes nécessite une organisation de la production « flexible » articulée autour d'une **planification quotidienne** de la production intégrant le caractère aléatoire de la demande et d'un **pilotage en temps réel**, dont le but est de réviser l'organisation en place en fonction d'informations prévisionnelles et de l'état du système (ce qui se traduit par la possibilité de modification des variables de commande). Un certain nombre de règles sont des retombées directes du travail d'étude en

univers certain, qui a fourni des idées intéressantes de stratégie de réactivité. Ces règles définissent, en fonction de situations-types, les adaptations, cohérentes au niveau global, aux perturbations constatées ou anticipées. L'objectif de la simulation est alors de tester la capacité du système et de sa gestion à fournir un bon niveau de qualité de service, étant entendu que l'on teste simultanément une capacité installée et la manière de l'utiliser.

L'organisation de la production est réalisée sous contrainte de capacité (ressources permanentes déterminées en univers certain), mais l'utilisation du personnel disponible n'est pas définie une fois pour toute en début de simulation, contrairement à la démarche adoptée dans ce domaine (*voir* Cebry *et al.*, Oh, et Wert *et al.*). La simulation a été réalisée avec le logiciel Witness (d'AT&T ; *voir* Clark), plus particulièrement dédié aux systèmes industriels, et dont la terminologie utilisée pour nommer les blocs de construction fait référence aux différents éléments des ateliers classiques : *article, stock, machine, ressources, convoyeur*, etc. (qui seront en italique dans la suite du texte).

D'un point de vue opérationnel, un processeur est modélisé par une *machine* et une seule, mais le nombre d'opérateurs présents (*ressources*) varient, ce qui a pour effet de modifier le temps de cycle de la *machine*. Les différents types de courrier sont représentés par des *articles* de noms différents, un *article* étant équivalent à un lot d'objets car le nombre d'articles que le modèle peut manipuler est limité techniquement. Un article possède des *attributs* et véhicule ainsi des informations à l'intérieur du modèle. Comme pour le modèle utilisé en programmation mathématique, il n'y a pas de liaison directe d'une *machine* à une autre, un stock s'intercalant entre deux processeurs, ce qui permet de gérer les sorties de stock (avec des règles de sorties du type FIFO ou plus complexes, fondées sur des priorités) et de différer éventuellement la production du courrier. L'opération de transport d'un processeur à l'autre est explicitement prise en compte avec des règles de lotissement particulières (correspondant à un contenant) et de temporisation (« temps de séjour minimum » avant déplacement, plus temps de transport). Le routage des flux à la sortie d'un processeur ou d'un stock nécessite l'introduction de *machines* fictives, qui réalisent « l'éclatement » par un tirage aléatoire selon des pourcentages moyens (car il n'est pas possible de réaliser un routage différencié selon le type d'article avec une seule *machine*).

Le pilotage du système est effectué par l'intermédiaire d'un **superviseur**, modélisé par une *machine* fictive, dont le rôle est de scanner à intervalle

de temps fixé les indicateurs d'état (cette *machine* étant activée par l'entrée, toutes les 5 minutes, d'un *article* fictif), et de modifier en conséquence les variables de commande. Witness oblige à décrire les méta-règles utilisées sous forme procédurale (et non de manière déclarative). L'implémentation des règles dans le superviseur implique une structuration de la description avec des décisions de temporalité différente :

– une fois par jour, est réalisée une planification de la production qui détermine la programmation des productions à la commande, au plus tard, en fonction des quantités en stock et des volumes prévisionnels attendus, qui sont définis pour une probabilité choisie de défaillance du système ; il en résulte, pour chaque processeur et chaque type de flux, une heure de début de traitement et un nombre d'opérateurs requis ;

– à chaque arrivée, la charge et la capacité de production des processeurs en cours de tri de courrier « urgent » sont réévaluées et des décisions de modification des affectations des opérateurs sont prises lorsque l'organisation en place (déterminée ci-dessus) ne permet pas d'assurer l'intégralité de cette production dans le délai imparti, compte tenu des variations de volume, et si les conditions de déclenchement sont remplies (« surcharge » détectée et temps de travail minimum, la détection des surcharges s'effectuant par comparaison des quantités de courrier arrivées et des quantités de référence servant à la programmation des traitements) ;

– à tout moment mais relativement rarement en pratique, en fonction de seuils d'alerte ou de priorités (rupture de charge, temps de séjour dans le système), certaines productions ou des affectations sont également modifiées.

Concrètement, les règles de pilotages sont basées sur la connaissance des arrivées du courrier dans le système et sur quelques indicateurs locaux (principalement « rupture de charge » sur un processeur). Il s'agit donc de règles ne nécessitant pas un système d'information interne très sophistiqué et parfaitement utilisables sur le terrain.

Ce travail de simulation a permis de dégager un certain nombre de recommendations organisationnelle directement utilisables pour améliorer le pilotage du système réel. Il a permis également d'évaluer dans quelle mesure la réponse du système peut être améliorée lorsque les caractéristiques de certains flux entrants sont connues de façon prévisionnelle par rapport à la situation actuelle où ces informations ne sont connues qu'à la réception, le système d'information constituant un moyen additionnel d'améliorer la flexibilité organisationnelle.

4. ANNEXE : LE PROGRAMME LINÉAIRE

* Indices utilisés

– t est l'indice d'une **période**, variant de 1 à T (découpage temporel par quart d'heure, sur 24 heures) ; origine du temps fixé à 7 heures du matin ;

– d est l'indice d'une **classe de destination** : $d = 1$: TG1 (à destination des autres départements) ; $d = 2$: TG2 (à destination du département) ;

– u est l'indice d'une **catégorie de courrier** : $u = 1$: première catégorie (trafic urgent) ; $u = 2$: seconde catégorie (trafic non urgent) ;

– h est l'indice d'un service offert au personnel ;

– k est l'indice d'un **chantier** *ou* d'une partie d'un chantier (pour $k = 1$ ou 2) : $k = 1$ pour antennes manuelles du chantier TOP (supposé doté d'une seule machine) ; $k = 2$: antennes automatiques du chantier TOP ; $k = 3$: chantier *Grand Format manuel-Tri général* (comportant plusieurs casiers) ; $k = 4$: chantier *Grand Format manuel-Tri par côtés* (comportant plusieurs casiers) ; $k = 5$: chantier *Petit Format manuel-Tri général* (comportant plusieurs casiers) ; $k = 6$: chantier *Petit Format manuel-Tri par côtés* (comportant plusieurs casiers) ; $k = 7$: chantier PIM (comportant plusieurs postes d'indexation) ; $k = 8$: chantier *Petit Format manuel-Tri au pouce* ; $k = 9$: chantier *Petit Format-automatique* : machines (ELIT) non utilisées en tri réflexe, $k = 10$: chantier *Petit Format-automatique* : machines (ELIT) utilisées en tri réflexe ;

– i est l'indice du **stock de courrier** : les 11 stocks sont représentés à la figure 1 et décrits « littéralement » par les relations 1 à 12 de conservation des stocks ;

* Paramètres

– x_{idut} est le **trafic** alimentant le stock i, pour la classe de destination d et la catégorie u, arrivé au début de la période t (unité : pli ou paquet) ;

– μ_{ht} est un paramètre binaire valant 1 si les personnes rattachées au service h sont présentes durant la période t et 0, dans le cas contraire ;

– β_j et α_i (ou β_{jd} et α_{jd}) sont des fractions de courrier (éventuellement de la classe de destination d) correspondant à des taux de rejet, de reprise, etc. ;

– Γ_h est le coût induit par l'usage du service h ;

– π_k est le taux de production du chantier k.

*** Variables**

– v_h, est le nombre de **personnes** à qui le service h a été donné (l'effectif présent durant la période t étant alors $\sum_h \mu_{ht} \cdot v_h$);

– n_{kt} est le nombre de **personnes affectées au chantier** k durant la période t, sauf pour les chantiers $k = 9$ et 10 (ELIT), où l'automatisation conduisant à imposer une équipe complète, l'effectif présent se déduit par une simple pondération des variables binaires n_{kt} (respectivement par 3 et 6);

– S_{idut} est le niveau du **stock** i, pour la classe de destination d et la catégorie u, au **début** de la période t (avant toute arrivée de trafic);

– p_{idukt} est la **production** (tri ou indexation) d'objets prélevés sur le stock i, pour la classe de destination d et la catégorie u, traité durant la période t, par le chantier k; cette variable est la seule variable continue du programme linéaire;

– δ_{dkt} est une variable binaire valant 1 si la ressource k traite la destination d durant la période t, et 0 dans le cas contraire;

– η^+_{kt} est une variable qui vaut 0 si $(n_{kt} - n_{k,t-1}) \leq 0$ (non augmentation des effectifs) et $(n_{kt} - n_{k,t-1})$, dans le cas contraire;

– η^-_{kt} est une variable qui vaut 0 si $(n_{kt} - n_{k,t-1}) \geq 0$ (non diminution des effectifs) et $(n_{k,t-1} - n_{kt})$, dans le cas contraire;

*** Contraintes**

– Équations de conservation des stocks:

• Stock de plis *Grand Format* mécanisables $(i = 1)$:

$$S_{1\,du,\,t+1} = S_{1\,dut} + x_{1\,dut} - \sum_{k=1}^{3} p_{1\,dukt} \tag{1}$$

• Stock de plis *Grand Format* non mécanisables $(i = 2)$:

$$S_{2\,du,\,t+1} = S_{2\,dut} + x_{2\,dut} - p_{2\,du\,3\,t} \tag{2}$$

• Stock de plis *Grand Format* pour *Tri par côtés* $(i = 3)$:

$$S_{3\,du,\,t+1} = S_{3\,dut} + \alpha_{2\,d} \sum_{k=1}^{2} [p_{6\,duk,\,t-1} + p_{1\,duk,\,t-1}] + \alpha_3 (p_{1\,du3,\,t-1} + p_{2\,du3,t\,-1}) - p_{3\,du\,4\,t} \tag{3}$$

• Stock de plis *Petit Format* mécanisables non indexés ($i = 4$):

$$S_{4\,du,\,t+1} = S_{4\,dut} + x_{4\,dut} - \sum_{k=9}^{10} p_{4\,dukt} - p_{4\,du\,5\,t} \tag{4}$$

• Stock de plis *Petit Format* mécanisables indexés ($i = 5$):

$$S_{5\,du,\,t+1} = S_{5\,dut} + x_{5\,dut} + (1-\beta_7)\,p_{7\,du7,\,t-1} - \sum_{k=9}^{10} p_{5\,dukt} - p_{5\,du\,5\,t} \tag{5}$$

• Stock de plis *Petit Format* non mécanisables ($i = 6$):

$$\begin{aligned} S_{6\,du,\,t+1} = {} & S_{6\,dut} + x_{6\,dut} + \beta_{7\,du7,\,t-1} \\ & + \beta_6 (p_{5\,du9,\,t-1} + p_{5\,du10,\,t-1}) - p_{6\,du\,5\,t} \end{aligned} \tag{6}$$

• Stock de plis *Petit Format* à indexer ($i = 7$):

$$S_{7\,du,\,t+1} = S_{7\,dut} + \beta_{1\,d}\,(p_{4\,du9,\,t-1} + p_{4\,du10,\,t-1}) - p_{7\,du\,7\,t} \tag{7}$$

• Stock de plis *Petit Format* pour tri par côtés ($i = 8$):

$$S_{8\,du,\,t+1} = S_{8\,dut} + \beta_2 \sum_{i=4}^{6} p_{idu5,\,t-1} + x_{8\,dut} - p_{8\,du\,6\,t} \tag{8}$$

• Stock de plis *Petit Format* mécanisables pour reprise tri réflexe en TG1 ($i = 9$ et $d = 1$):

$$\begin{aligned} S_{9,1\,u,\,t+1} = {} & S_{9,1\,ut} + \beta_{3,1}\,[(1-\beta_{1,1})\,p_{4,1\,u10,\,t-1} \\ & + (1-\beta_{6,1})\,p_{5,1\,u10,\,t-1}] - p_{9,1\,u\,10\,t} \end{aligned} \tag{9}$$

• Stock de plis *Petit Format* mécanisables pour reprise CEDEX dans le cas de tri autonome TG2 ($i = 9$ et $d = 2$):

$$S_{9,2\,u,\,t+1} = S_{9,2\,ut} + \beta_{3,2}\,[(1-\beta_{1,2})p_{4,2\,u9,\,t-1} + (1-\beta_{6,2})p_{5,2\,u9,\,t-1}] \tag{10}$$

• Stock de plis *Petit Format* mécanisables pour reprise TG3 dans le cas de tri autonome TG2 ($i = 10$ et $d = 2$):

$$S_{10,2\,u,\,t+1} = S_{10,2\,ut} + \beta_{4,2}\,[(1-\beta_{1,2})p_{4,2\,u9,\,t-1} + (1-\beta_{6,2})p_{5,2\,u9,\,t-1}] \tag{11}$$

• Stock de plis *Petit Format* à trier « au pouce » $(i = 11)$:

$$S_{11\,du,\,t+1} = S_{11\,dut} + x_{11\,dut} - p_{11\,du\,8\,t} \tag{12}$$

– contrainte d'utilisation des antennes manuelles de la TOP (les antennes manuelles de la TOP ne sont utilisables que si les 3 antennes automatiques le sont déjà):

$$n_{1\,t} \leq 3\,n_{2\,t} \tag{13}$$

– contrainte d'utilisation du tri réflexe (si le centre ne dispose que de 2 ELIT, il est impossible d'utiliser individuellement les ELIT si un tri réflexe est utilisé):

$$n_{9\,t} + 2\,n_{10,\,t} \leq 2 \tag{14}$$

– contrainte d'utilisation des effectifs (l'effectif disponible est égal ou supérieur à l'effectif occupé):

$$\sum_{h} \mu_{ht} \cdot v_h \geq \sum_{k=1}^{8} n_{kt} + 3\,(n_{9\,t} + 2\,n_{10\,t}) \tag{15}$$

– contrainte de niveau de service:

• pour le trafic urgent $(u = 1)$, le stock résiduel doit être nul au début de la période $t = \zeta$ suivant immédiatement la coupure de la classe de destination d:

$$S_{id\,1\,\zeta} = 0 \tag{16}$$

• pour le trafic non urgent $(u = 2)$, l'utilisation du modèle dans une logique de régime de croisière nécessite d'avoir un stock de début de journée $(t = 0)$ égal au stock final de fin de journée, à la période $t = T$:

$$S_{id2,0} = S_{id2,\,T} \tag{17}$$

– Contraintes de capacité de production

• du chantier TOP $(k = 1$ à $2)$:

$$\sum_{k=1}^{2} \sum_{d=1}^{2} \sum_{u=1}^{2} [p_{6\,dukt} + p_{1\,dukt}] \leq \sum_{k=1}^{2} \pi_k\, n_{kt} \tag{18}$$

• du chantier *Tri général-Grand Format* $(k = 3)$:

$$\sum_{d=1}^{2} \sum_{u=1}^{2} (p_{1\,du\,3\,t} + p_{2\,du\,3\,t}) \leq \pi_3\, n_{3\,t} \tag{19}$$

• du chantier *Tri par côtés-Grand Format* ($k = 4$):

$$\sum_{d=1}^{2} \sum_{u=1}^{2} p_{3\,du\,4\,t} \leq \pi_4\, n_{4t} \tag{20}$$

• du chantier *Tri général-Petit Format* ($k = 5$):

$$\sum_{i=4}^{6} \sum_{d=1}^{2} \sum_{u=1}^{2} p_{idu\,5\,t} \leq \pi_5\, n_{5t} \tag{21}$$

• du chantier *Tri par côtés Petit Format* ($k = 6$):

$$\sum_{d=1}^{2} \sum_{u=1}^{2} p_{8\,du\,6\,t} \leq \pi_6\, n_{6t} \tag{22}$$

• du chantier PIM ($k = 7$):

$$\sum_{d=1}^{2} \sum_{u=1}^{2} p_{7\,du\,7\,t} \leq \pi_7\, n_{7t} \tag{23}$$

• du chantier tri au pouce ($k = 8$):

$$\sum_{d=1}^{2} \sum_{u=1}^{2} p_{11,\,du\,8\,t} \leq \pi_8\, n_{8t} \tag{24}$$

• des chantiers ELIT travaillant indépendamment ($k = 9$) ou en tri réflexe ($k = 10$)

. pour le TG2, où les ELIT fonctionnent de manière autonome:

$$\sum_{u=1}^{2} p_{4,2,u,9,t} + p_{5,2,u,9,t} + \sum_{i=9}^{10} p_{i,2,u,9,t} \leq \pi_9\, n_{9t} \tag{25}$$

. pour le TG1, où le tri réflexe est obligatoire:

$$\sum_{u=1}^{2} p_{4,1,u,10,t} + p_{5,1,u,10,t} + p_{9,1,u,10,t} \leq 2\,\pi_{10}\, n_{10t} \tag{26}$$

– contraintes découlant de l'unicité du plan de tri dans un chantier, au cours d'une période:

$$\sum_{d} \delta_{dkt} \leq 1 \tag{27}$$

avec, la valeur de δ_{dkt} déterminée par la contrainte :

$$\delta_{dkt} \geq \frac{\sum_i \sum_u p_{idukt}}{H} \tag{28}$$

où H est une valeur quelconque supérieure à la plus grande valeur possible que peut prendre la production de la période $(\sum_i \sum_u p_{idukt})$.

– contraintes de non-utilisation de machine découlant du vidage des cases lors du changement de plan de tri (vidage sur une période) :

$$\delta_{1\,kt} + \delta_{2\,k,\,t+1} \leq 1 \tag{29}$$

$$\delta_{2\,kt} + \delta_{1\,k,\,t+1} \leq 1 \tag{30}$$

– contraintes liées au lissage des effectifs (lissage pris en compte dans la fonction-objectif) :

$$n_{kt} = n_{k,\,t-1} + \eta^+_{kt} - \eta^-_{kt} \tag{31}$$

*** Fonction-objectif**

Il s'agit, *a priori*, de minimiser le coût de fonctionnement du système :

$$\text{Min}\,.\,\{\sum_h \Gamma_h \cdot v_h\} \tag{32}$$

On introduira les variables η^+_{kt} dans la fonction objectif, avec une pénalité ψ_1, faible par rapport au coût d'un opérateur supplémentaire, ce qui forcera cette variable à prendre une valeur la plus faible possible, afin de *privilégier les solutions lissant les effectifs*. Dans ces conditions, la relation 32 devient :

$$\text{Min}\,.\left\{\sum_h \Gamma_h \cdot v_h + \psi_1 \cdot \sum_t \left[\sum_{k=1}^{8} \eta^+_{kt} + 3\,(\eta^+_{9\,t} + 2\,\eta^+_{10\,t})\right]\right\} \tag{33}$$

RÉFÉRENCES

A. Brooke, D. Kendrick et A. Meeraus, *GAMS: A User's Guide*, Scientific Press, Redwood City, CA, 1988.

M. E. Cebry, A. H. De Silva et F. J. Di Lisio, Management Science in Automating Postal Operations: Facility and Equipment Planning in the United States Postal Service, *Interfaces*, Vol. *22*, January-February 1992, p. 110-130.

M. F. Clark, WITNESS: unlocking the power of visual interactive simulation, *European Journal of Operational Research*, Vol. *54*, n° 3, October 1991, p. 293-298.

F. H. MURPHY, E. A. STOHR et A. ASTHANA, Representation schemes for linear programming models, *Management Science*, Vol. *38*, n° 7, July 1992, p. 964-991.

J. OH, Mail production operations planning analysis: a modeling approach, *Actes des Premières Journées Européennes sur les Technologies Postales*, Vol. *2*, Nantes, 14-16 juin 1993, p. 685-692.

S. D. WERT, J. F. BARD, A. H. DE SILVA et T. A. FEO, A simulation Analysis of Advanced Concepts for Semi-Automated Mail Processing, *Journal of Operations Research Society*, Vol. *42*, n° 12, 1991, p. 1071-1086.