HEC MONTRÉAL
École affiliée à l'Université de Montréal

Gestion de l'incertitude dans les problèmes de localisation et application à la gestion des services préhospitaliers d'urgence

par
Valérie Bélanger

Thèse présentée en vue de l'obtention du grade de Ph.D. en administration
(option Méthodes quantitatives)

Novembre 2015

HEC MONTRÉAL
École affiliée à l'Université de Montréal

Cette thèse intitulée :

Gestion de l'incertitude dans les problèmes de localisation et application à la gestion des services préhospitaliers d'urgence

Présentée par :

Valérie Bélanger

a été évaluée par un jury composé des personnes suivantes :

Raf Jans
HEC Montréal
Président-rapporteur

Patrick Soriano
HEC Montréal
Directeur de recherche

Walter Rei
Université du Québec à Montréal
Membre du jury

Frédéric Semet
École Centrale de Lille
Examinateur externe

Olivier Gerbé
HEC Montréal
Représentant du directeur de HEC Montréal

RÉSUMÉ

Plusieurs travaux parus jusqu'à maintenant se sont intéressés à l'élaboration de stratégies et de méthodes afin de prendre en compte l'incertitude lors de la phase de planification statique, c'est-à-dire *a priori*. Néanmoins, dans certains contextes, même en essayant de considérer au mieux les réalisations éventuelles de l'incertitude, il est possible que la situation évolue ou qu'un événement imprévisible se produise de sorte que des solutions établies *a priori* ne permettent plus d'obtenir de performances adéquates. Pourtant, beaucoup moins de travaux ont considéré de manière plus explicite et formelle l'évolution du système dans le temps et les différentes façons de gérer ces systèmes en temps réel. L'étude et le développement de modèles pour réagir et s'adapter aux événements incertains qui surviennent en cours d'opération et à l'évolution du système dans le temps présentent un défi important, d'où l'interêt de cette recherche.

La présente thèse s'intéresse donc au développement et à l'analyse de stratégies de gestion qui permettront de réagir et de s'adapter, dans un cadre clair et formel, aux réalisations possibles de l'incertitude en cours d'opération et à l'évolution du système dans le temps. Dans le cadre de cette thèse, la gestion de l'incertitude est étudiée dans le contexte d'un problème de localisation général, puis dans le contexte particulier de la localisation des véhicules ambulanciers. Plus concrètement, cette thèse vise le développement de modèles mathématiques représentatifs prenant en compte les différentes sources d'incertitude et le dynamisme présents dans le contexte de la localisation, puis celui de méthodes de résolution capables de fournir des solutions de bonne qualité à ces modèles. Enfin, elle vise à analyser et à discuter des stratégies développées et de leurs implications du point de vue de la pratique, tant au niveau des conséquences économiques qu'au niveau de la gestion des ressources humaines.

Les résultats obtenus à la suite des analyses menées dans le cadre de cette thèse ont permis de montrer que le fait de maintenir les décisions initialement établies lorsque l'état du système évolue ou qu'un événement imprévisible se produit n'est pas sans conséquence, et peut devenir rapidement coûteux pour une organisation. Ensuite, dès que l'on modifie un peu les décisions établies, la qualité de la solution s'améliore dans une plus grande proportion, puis cette amélioration diminue au fur et à mesure que les modifications apportées deviennent plus importantes. L'utilisation de stratégies permettant de réagir à l'incertitude en cours d'opération présente donc des bénéfices clairs, mais qui ne sont pas sans effort. Le compromis entre le gain en performance et les efforts qui y sont associés dépendra fortement du contexte d'application et de la nature des décisions considérées. Les outils développés dans le cadre de cette thèse se sont toutefois montrés efficaces pour mener différentes analyses qui permettront de mieux

comprendre le compromis entre le gain en performance et les efforts associés. Ils permettront aussi de soutenir une prise de décision mieux informée en pratique.

Mots clés : Incertitude, réoptimisation, localisation, affectation, déploiement, redéploiement, services préhospitaliers d'urgence, programmation mathématique, simulation, génération de colonnes, matheuristique.

Méthodes de recherche : Modélisation mathématique, méthode quantitative.

ABSTRACT

Up to now, many studies have focused on the development of methods and strategies to deal with decision problems under uncertainty during the static planning phase, i.e. a priori. However, in several cases, the situation might evolve in such an unpredictable way that solutions taken a priori might not be able to achieve good performances anymore. Some adjustments might be needed to regain or maintain adequate performances. Unfortunately, less effort has been devoted to the development of strategies and formalized procedures to react to uncertainty dynamically. The study and the development of models to adapt the system during operations when uncertain events occur or when the system evolves through time still present many challenges.

This thesis is interested in the development and analysis of strategies to formally adjust decisions taken a priori in order to account for the realization of uncertain events during operations. In this thesis, uncertainty management is studied in the context of a general location problem, and in the specific context of ambulance location. In particular, this thesis aims to develop mathematical models capable to react to the different sources of uncertainty that will affect a system during operations, and the development of solution approaches able to provide good quality solutions to these models in real-life situations. Finally, it aims to analyze and discuss the proposed strategies and their implications in practice, both in terms of economic consequences and human resource management impacts.

The results obtained through the various analyses presented in this thesis have shown that keeping decisions as initially planned when the system state evolves or an unpredictable event occurs is not without consequence. It can indeed becomes very expensive for an organization. Then, when one agrees to change the established decisions slightly, the quality of the solution is improved in a significant proportion. This improvement then continues as decisions are modified further, but in a decreasing manner as changes grow in importance. Results also have shown that the use of strategies to react to uncertainty during operations lead to clear benefits, but that cannot be achieved without effort. The right tradeoff between system performances and the effort required to modify decisions will strongly depend on the context under study and the nature of decisions taken. Nevertheless, approaches and methodologies developed in this thesis have proven to be effective to conduct such a tradeoff analysis as well as to support decision making in practical contexts.

Keywords: Uncertainty, reoptimization, location, assignment, deployment, redeployment, emergency medical services, mathematical programming, simulation, column generation, matheuristic.

Research methods: Mathematical modeling, quantitative research.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

À mon grand-papa, avec qui je jouais aux mathématiques.

REMERCIEMENTS

Cette thèse n'aurait pas été possible sans la contribution des personnes que j'ai côtoyées au cours des dernières années. Je désire donc prendre quelques lignes afin de leur exprimer mes plus sincères remerciements. Tout d'abord, je tiens à remercier M. Patrick Soriano et M. Angel Ruiz, mes directeur et codirecteur de thèse, pour leur soutien, leur disponibilité et leur entière collaboration tout au long du projet. Votre aide a été précieuse et je vous en suis très reconnaissante. Je tiens aussi à remercier M. Roberto Wolfler Calvo, M. Yannick Kergosien et à M. Ettore Lanzarone avec qui j'ai collaboré à différents moments de ma thèse. Enfin, je désire remercier M. Gilbert Laporte et M. Walter Rei qui ont accepté de faire partie de mon comité de thèse. Vous avez été de bon conseil à des moments importants de ce parcours. Sur une note moins formelle, je voudrais remercier mes parents et amis qui m'ont soutenue dans ce long processus. Vous avez tous contribué, à votre manière, à la réalisation de cette thèse. Un merci tout spécial à Dominique et Émilie qui m'ont aidé avec la révision de la thèse, et à Hanna qui a dû partager sa mère (avec un ordinateur) dans les derniers mois de la rédaction.

INTRODUCTION

La gestion d'une entreprise ou d'une organisation comporte un certain nombre d'enjeux et de défis. Parmi ces défis : la gestion de l'incertitude. En effet, la prise de décision, et ce, à tous les niveaux, tant stratégique, tactique, qu'opérationnel, peut être grandement influencée par les différentes sources d'incertitude auxquelles sont confrontées les organisations. Dans plusieurs contextes, on considérera pourtant que toutes les données sont connues avec certitude au moment de la prise de décision, ce qui n'est pas nécessairement le cas en réalité. Les décisions sont alors prises lors de la phase de planification statique, c'est-à-dire *a priori*, sans remise en question durant la phase d'exécution. Ce faisant, les performances du système peuvent être affectées de manière considérable lorsque l'incertitude se réalise. Ainsi, pour ces organisations, il deviendra important, voire nécessaire, de se doter de méthodes et de stratégies pour considérer plus explicitement les différentes sources d'incertitude de manière proactive ou pour y réagir en cours d'opération. Naturellement, intégrer l'incertitude dans les processus de gestion requiert des méthodes et des stratégies plus complexes, mais devrait, à terme, mener à des bénéfices importants.

L'un des problèmes abondamment traités jusqu'à maintenant et qui, comme tant d'autres, peut être affecté de manière importante par un certain nombre de phénomènes incertains, est la localisation des installations ou des points de service d'une organisation. En effet, une organisation devra, à un moment ou à un autre, déterminer comment déployer ses ressources et où les localiser afin de satisfaire adéquatement la demande de ses clients ou d'offrir un service approprié à la population. Dans plusieurs cas, les décisions de localisation devront être déterminées sans une connaissance parfaite des données du problème et de leur évolution dans le temps. D'une part, certaines décisions sont établies pour une longue période : une connaissance parfaite des données futures est rarement disponible. D'autre part, certaines données peuvent être amenées à évoluer dans le temps, parfois de manière imprévisible, ce qui amène des difficultés supplémentaires. Bien intégrer l'incertitude, tant dans la planification *a priori* que dans l'établissement de stratégies réactives, devient alors critique pour ces organisations. Ceci est d'autant plus vrai pour des systèmes oeuvrant dans un environnement en constante évolution et pour lesquels la capacité à servir la population de manière efficiente peut avoir un impact important sur son bien-être. C'est le cas, par exemple, des systèmes de santé et de sécurité publique.

Les organisations responsables des services préhospitaliers d'urgence (SPU) sont des organisations importantes au sein des systèmes de santé et de sécurité publique. Leur mission principale consiste à répondre adéquatement aux appels de détresse en prodiguant les premiers soins aux

personnes concernées et leur transport, au besoin, vers le service des urgences du centre hospitalier approprié. Afin d'offrir ce service à la population, les SPU doivent mobiliser un ensemble de ressources (personnel soignant, véhicules ambulanciers, centre d'appels) puis les gérer de façon adéquate. L'utilisation efficiente des véhicules ambulanciers disponibles, de même que leur localisation sur le territoire à desservir, représentent donc des aspects critiques de la gestion des SPU. Les SPU opèrent dans un environnement en constante évolution (c'est-à-dire fortement dynamique) étant donné la nature incertaine de la demande, des temps d'intervention et de la disponibilité des ressources, ce qui complexifie le prise de décision. La capacité d'une telle organisation à répondre adéquatement aux appels de détresse est donc en lien étroit avec sa capacité à anticiper puis à réagir aux différentes sources d'incertitude. Le développement de stratégies de gestion dynamique permettant aux SPU de considérer convenablement ces différentes sources d'incertitude devient alors tout à fait justifié.

Jusqu'à présent, il a été possible d'observer qu'un grand nombre de travaux se sont intéressés à la prise en compte de l'incertitude *a priori*, et ce, tant dans le contexte général de la localisation, que dans le cadre particulier de la gestion des services préhospitaliers d'urgence. Les organisations essaient alors de se préparer de manière adéquate aux réalisations possibles des événements incertains en cours d'opération et à l'évolution du système dans le temps. Néanmoins, dans certains contextes, même en essayant d'anticiper au mieux les réalisations éventuelles de l'incertitude, il est possible que la situation change de manière importante ou qu'un événement imprévisible se présente de sorte que les performances du système se dégradent considérablement. Il pourra alors devenir nécessaire de réagir en cours d'opération afin de retrouver un niveau de performance adéquat. Pourtant, beaucoup moins d'efforts ont été consacrés à l'analyse et au développement de stratégies pour réagir à l'incertitude et adapter les décisions en cours d'opération, c'est-à-dire quoi faire lorsque le système n'est plus en mesure de fonctionner adéquatement à la suite d'un événement imprévisible ou d'un changement de données importants. De plus, de manière générale, ces stratégies sont beaucoup moins claires et formellement définies. L'étude et le développement de modèles pour réagir et s'adapter aux différents événements incertains en cours d'opération présentent un défi important, d'où l'interêt de cette recherche.

La présente thèse s'intéresse donc au développement et à l'analyse de stratégies de gestion qui permettront de réagir et s'adapter, dans un cadre clair et formel, aux réalisations possibles de l'incertitude en cours d'opération et à l'évolution du système dans le temps. Dans le cadre de cette thèse, la gestion de l'incertitude est étudiée dans le contexte d'un problème de localisation général, puis dans le contexte particulier de la localisation des véhicules ambulanciers. Plus

concrètement, cette thèse vise le développement de modèles mathématiques représentatifs prenant en compte les différentes sources d'incertitude et le dynamisme présents dans le contexte de la localisation, puis celui de méthodes de résolution capables de fournir des solutions de bonne qualité à ces modèles, mais qui pourront également être utilisées en pratique, c'est-à-dire pour résoudre des problèmes de taille réelle. Enfin, elle vise à analyser et à discuter des stratégies développées et de leurs implications du point de vue de la pratique, tant au niveau des conséquences économiques qu'au niveau de la gestion des ressources humaines.

Les contributions de la thèse s'orientent selon trois perspectives : tout d'abord, d'un point de vue théorique, en proposant de nouveaux modèles qui permettront de formaliser et de définir clairement des stratégies de gestion afin de réagir et s'adapter aux différentes sources d'incertitude en cours d'opération et à l'évolution du système dans le temps, et ce, dans le contexte de la localisation ; d'un point de vue méthodologique, par le développement de méthodes de résolution efficaces et performantes pouvant être utilisées pour déterminer des solutions réalisables et pertinentes pour des problèmes de taille réelle ; et enfin, d'un point de vue pratique, par l'analyse des stratégies proposées et l'étude du compromis entre les bénéfices de l'utilisation de telles stratégies et les inconvénients qui y sont associés, entre autres au niveau de la gestion du personnel.

Cette thèse s'organise comme suit. Tout d'abord, le *Chapitre 1* présente une revue et une classification des différentes stratégies proposées afin de considérer l'incertitude dans les problèmes de gestion. Plus précisément, les méthodes de résolution, tant *proactives* que *réactives*, sont répertoriées, puis discutées en portant une attention particulière à leurs forces, limites et distinctions. Dans le cadre de cette thèse, nous nous intéresserons plus en profondeur à l'élaboration et à l'analyse de stratégies réactives pour la gestion dynamique de différents systèmes, considérée en cours d'opération. Cette revue des différentes stratégies pour la prise en compte de l'incertitude permettra donc de situer nos travaux par rapport aux différentes stratégies proposées jusqu'à maintenant.

Le *Chapitre 2* s'intéresse ensuite à la modélisation, à la résolution et à l'analyse d'un problème de localisation perturbé considéré en temps réel. Lorsque différentes perturbations du système se produisent, il peut devenir nécessaire de revoir une solution déterminée initialement de manière à assurer sa faisabilité ou à améliorer l'atteinte des objectifs, tout en limitant l'impact des perturbations sur les opérations prévues. Dans ce contexte, ce chapitre vise donc à formaliser le concept de la réoptimisation contrôlée, puis à l'appliquer à la formulation d'un problème de localisation perturbé. Plus concrètement, la réoptimisation contrôlée a pour but de limiter ou de contrôler la différence entre deux solutions successives. Ce chapitre présente et introduit ensuite

une méthode générique et flexible pour la résolution du problème étudié, qui pourra être intégrée aussi au sein d'un outil d'aide à la décision afin de soutenir la prise de décision en temps réel. Enfin, il analyse plus en détail le compromis entre le contrôle de la solution et les coûts impliqués, et ce, pour différents types de perturbations. Ce chapitre présente ainsi une première étude d'un problème de localisation perturbé, dans un contexte très générique. Il permettra d'introduire le concept de réoptimisation contrôlée, puis d'amorcer une analyse du compromis entre l'amélioration des performances d'un système et les inconvénients ou les coûts, tant financiers que sociaux, engendrés pour y parvenir. Cette analyse sera étoffée tout au long de la thèse, dans le cadre plus particulier de la gestion des SPU.

Le *Chapitre 3* propose une revue des différents travaux en lien avec les problèmes de déploiement et le redéploiement des véhicules ambulanciers. Ces problèmes s'intéressent principalement à la localisation des véhicules sur le territoire à desservir de manière à assurer un service adéquat à la population. Ainsi, dans un premier temps, ces problèmes sont décrits puis situés par rapport aux problèmes généraux de localisation. Les approches de modélisation proposées, de même que les méthodes de résolution développées dans ce contexte, sont ensuite abordées. Les règles de répartition, c'est-à-dire la sélection du véhicule à affecter à un appel, sont également discutées. Les problèmes de redéploiement et de répartition étudiés dans ce chapitre font partie intégrante de la gestion dynamique des véhicules ambulanciers. Différentes études ont été proposées jusqu'à maintenant en ce sens, mais il reste encore beaucoup à faire dans ce registre. Les trois derniers chapitres de la thèse sont donc dédiés à l'analyse et au développement de stratégies pour la gestion dynamique d'un SPU. La revue des travaux importants en lien avec la gestion de SPU permettra de bien situer nos travaux, de même que de faire ressortir les besoins en la matière, notamment au niveau du redéploiement dynamique.

Le *Chapitre 4* propose une évaluation et une analyse comparative de différentes stratégies de gestion des véhicules ambulanciers, et ce, par une étude de simulation. Ce chapitre présente la définition de quatre stratégies de gestion opérationnelle prenant en compte l'aspect dynamique du problème, puis les modélise en considérant un cadre commun. Il vise ensuite à quantifier les bénéfices des stratégies de gestion plus dynamiques et flexibles par rapport aux stratégies statiques, mais aussi à comparer ces stratégies dans différents contextes pouvant se distinguer par leurs niveaux de congestion et par leurs profils de demande. Une attention particulière est alors portée au compromis entre les gains en performance et les inconvénients liés à l'utilisation de stratégies dynamiques, notamment ceux associés aux aspects économiques et à la gestion du personnel. À la suite de cette analyse, il sera possible de valider la pertinence d'une gestion dynamique afin d'adapter les décisions à l'évolution du système au cours de journée. Les constats

effectués permettront également de justifier et de guider le développement de modèles mathématiques et de méthodes de résolution pour le déploiement et le redéploiement des véhicules ambulanciers, sujets des deux derniers chapitres de la thèse.

Enfin, le *Chapitre 5* présente un modèle de décision pour la gestion des SPU considérant le redéploiement des véhicules, mais également une préaffectation anticipative des demandes éventuelles aux véhicules disponibles. Ce modèle vise à minimiser le temps de réponse espéré, une mesure définie dans ce chapitre, tout en limitant les efforts liés au redéploiement. La capacité réelle des véhicules y sera également prise en compte. Une analyse détaillée du modèle est ensuite présentée, suivie d'une discussion sur le compromis entre les performances du système en utilisant une telle stratégie et les inconvénients qui y sont associés. Malgré les bénéfices potentiels, l'implantation d'une stratégie basée sur le modèle proposé en pratique requiert une méthode de résolution rapide et performante qui permettra de trouver une solution au problème étudié dans des délais raisonnable. Le développement d'une approche de type matheuristique pour résoudre des instances de taille réelle est proposée au *Chapitre 6*.

CHAPITRE 1

CONSIDÉRER L'INCERTITUDE DANS LES PROBLÈMES DE GESTION

L'incertitude est inhérente à la vie. Elle est présente partout, de la date précise de la prochaine catastrophe naturelle aux fluctuations du coût de l'essence. Et personne n'y échappe. En effet, bien peu de choses sont fixées, connues à l'avance. Il faut donc composer inévitablement avec l'incertitude lorsque des décisions doivent être prises, aussi banales soient elles : quel itinéraire suivre pour se rendre au travail, à quel moment doit-on acheter les billets d'avion pour les prochaines vacances en famille, doit-on se faire vacciner contre la grippe ou non ? Évidemment, les conséquences de considérer ou non l'incertitude lors de la prise de décision ou de se doter de moyens pour y réagir ne sont pas toujours critiques. Une compagnie de transport aérien sera beaucoup plus touchée par une tempête de neige que le pauvre citoyen qui devra rester à la maison cette journée-là en raison de la météo. Pour certaines organisations, il devient toutefois important, voire crucial, d'estimer au meilleur de leur connaissance les différentes sources d'incertitude et de les prendre en compte au sein de la prise de décision afin d'en limiter les effets négatifs. Néanmoins, intégrer l'incertitude dans les problèmes de gestion n'est pas une mince affaire. Cela entraîne généralement un processus plus complexe, requiert des analyses plus poussées, mais au bout du compte, on y gagne ou du moins, on devrait y gagner.

Malgré les enjeux importants liés à l'intégration de l'incertitude dans les différents problèmes de gestion, diverses avancées technologiques telles que l'amélioration des technologies de l'information et de communication ainsi que la puissance accrue des ordinateurs la rendent désormais envisageable. En effet, il est maintenant possible d'accéder à une quantité importante de données en temps réel, de les traiter dans des délais de temps raisonnables puis de retransmettre, au besoin, l'information sur les décisions qui ont été prises presqu'instantanément. Il n'y a pas si longtemps, cela n'était envisageable que théoriquement. Ainsi, conscientes des possibilités et des bénéfices potentiels, les organisations se questionnent davantage sur les différentes manières de mieux intégrer l'incertitude dans leur processus décisionnel. Différentes communautés de chercheurs et de praticiens (en recherche opérationnelle, en informatique, en mathématiques appliquées, en ingénierie, en sciences de la gestion) s'intéressent à développer des approches permettant de soutenir ce processus et de mieux considérer l'incertitude dans les problèmes de gestion, de leur modélisation à leur résolution, puis lors de l'implantation des solutions retenues.

Mais au fait, qu'entend-on concrètement par incertitude ? L'incertitude, c'est l'état de ce qui est incertain, qui n'est pas fixé, déterminé à l'avance, qui n'est pas connu avec certitude[1]. Lors de la prise de décision, pour des applications réelles telles que l'établissement des tournées pour une entreprise de transport adapté, la confection des horaires du personnel infirmier dans une clinique de soins médicaux ou encore le choix de la localisation des installations pour une compagnie spécialisée en communication, l'incertitude se traduit généralement selon deux perspectives : la qualité et l'évolution de l'information disponible (Pillac *et al*., 2013).

La *qualité* de l'information reflète l'incertitude sur les données disponibles au moment de la prise de décision. Par exemple, lors de la gestion d'un système préhospitalier d'urgence (SPU), la localisation initiale des véhicules ambulanciers sur le territoire à desservir est généralement établie en utilisant une estimation de la demande réelle et des temps de déplacement. L'*évolution* de l'information traduit plutôt le fait que, dans certains problèmes, une partie ou la totalité de l'information change ou se révèle durant la phase d'exécution. Par exemple, toujours dans le contexte de la gestion d'un SPU, les demandes d'aide sont placées en temps réel. L'information concernant le nombre exact de demandes, leur localisation et leur sévérité n'est connue que pendant les opérations. Ainsi, plusieurs problèmes de gestion, dans des domaines aussi variés que le transport de marchandises et de personnes, la logistique humanitaire, la production industrielle et le domaine hospitalier, pour n'en nommer que quelques-uns, peuvent être vus selon l'une ou l'autre de ces perspectives, voire les deux. Ces distinctions mènent à l'identification de différentes classes de problèmes en lien avec le type d'incertitude considéré et le moment où les décisions sont prises ou revues.

Tout d'abord, certains problèmes *déterministes* et *statiques* considèrent l'incertitude implicitement selon la perspective de la qualité de l'information. Dans le cas d'un problème *déterministe* et *statique*, toutes les données sont connues avec certitude au moment où les décisions sont prises. De plus, toutes les décisions sont prises *a priori*, aucune décision n'étant remise en question durant la phase d'exécution. Ainsi, les problèmes *déterministes* et *statiques* ne permettent pas de considérer explicitement l'incertitude, mais visent plutôt à approximer les données incertaines sous la forme de valeurs moyennes ou de valeurs probables basées sur la connaissance et l'expérience des gestionnaires. Le fait d'approximer ainsi l'incertitude permet de considérer l'hypothèse d'une connaissance parfaite et complète de l'information. Dans certains cas, cette approximation relativement « grossière » de l'incertitude est la seule avenue possible. Lorsqu'un problème *déterministe* et *statique* est résolu, on présume qu'il existe, dans l'ensemble des solutions réalisables, une solution qui possède la meilleure valeur d'un critère connu et mesurable.

[1] Grand Robert de la langue française

La solution d'un problème *déterministe* et *statique* est donc choisie de manière à optimiser un objectif unique, bien défini et mesurable. La même remarque s'applique dans le cas multicritère où un ensemble de solutions non-dominées sont choisies en fonction de critères connus et mesurables. Néanmoins, les solutions ainsi déterminées peuvent devenir peu robustes lorsque les données réelles sont connues, d'où l'intérêt de développer des méthodes et des stratégies pour considérer plus explicitement l'incertitude. Les solutions stochastiques présenteront alors des caractéristiques différentes, que les solutions déterministes ne possèdent pas, et qui permettront de mieux faire face à l'incertitude.

Les problèmes *statiques* et *stochastiques* considèrent explicitement l'incertitude selon la perspective de la *qualité* de l'information. Ils se caractérisent par le fait que les différentes données du problème se présentent sous la forme de variables aléatoires pour lesquelles la réalisation n'est connue avec certitude que pendant la phase d'exécution. Les principales décisions ne sont prises qu'une seule fois, avant la phase d'exécution. Ces décisions, prises avant la réalisation de l'incertitude, sont aussi connues sous le nom de décisions de première étape. Des actions de recours sont ensuite entreprises pendant la phase d'exécution, basées sur des stratégies établies à l'avance, de manière à adapter les décisions une fois l'incertitude levée. Différentes méthodes permettront de déterminer des solutions ou des stratégies aux caractéristiques différentes. Parmi les méthodes les plus communément utilisées, la programmation stochastique avec contraintes probabilistes permet de déterminer des décisions de première étape telles que les performances ou la faisabilité de la solution trouvée soient garanties avec une probabilité donnée. La programmation stochastique avec recours cherche des décisions de première étape telles que la valeur espérée pour les actions de recours soit optimale. L'optimisation robuste vise plutôt à déterminer une solution qui demeure réalisable pour toutes les réalisations possibles de l'incertitude, même dans le pire des cas. Dans la plupart de circonstances, il est toutefois peu probable de présumer que tout ira mal en même temps. On imposera alors un « budget d'incertitude » ou, autrement dit, une borne sur ce qui pourra mal tourner. La solution obtenue est donc la meilleure solution possible à l'intérieur d'un budget d'incertitude donné. Dans le cas de l'optimisation robuste, l'incertitude est modélisée sous la forme d'un ensemble de réalisations ou de scénarios possibles plutôt que représentée par des distributions de probabilités. Tous les concepts discutés ici sont applicables pour les problèmes *statiques* et *stochastiques* multi-critères, mais des adaptations sont nécessaires afin de prendre en compte adéquatement les différents objectifs visés.

Selon le contexte d'application étudié, plusieurs données peuvent être affectées par l'incertitude lors de la définition d'un problème *statique* et *stochastique*. Par exemple, pour le problème de

tournées de véhicules, plusieurs cas ont été étudiés : les problèmes avec clients stochastiques où un client doit être servi avec une probabilité donnée ; les problèmes avec temps stochastiques où les temps de service ou de transport sont modélisés comme des variables aléatoires ; et les problèmes avec demandes stochastiques (Pillac *et al*., 2013; Gendreau *et al*., 2015). Le problème de localisation des ambulances est un second exemple où diverses données du problème ne sont pas connues avec certitude au moment de la prise de décision. En effet, les organisations responsables des services préhospitaliers d'urgence oeuvrent dans un environnement hautement incertain où la fréquence et la localisation des demandes de service peuvent varier selon l'heure, la journée, la semaine. Différents cas de figure ont alors été étudiés afin de résoudre le problème de localisation des ambulances considérant la disponibilité des ambulances (Daskin, 1983; ReVelle et Hogan, 1988, 1989), l'arrivée des demandes de service (Ball et Lin, 1993; Beraldi *et al*., 2004; Beraldi et Bruni, 2009) et le temps de déplacement d'une localisation donnée vers une demande de service (Daskin, 1987; Goldberg *et al*., 1990b; Ingolfsson *et al*., 2008) comme étant stochastiques.

Les problèmes *dynamiques* considèrent quant à eux l'incertitude selon la perspective de l'*évolution* de l'information. Ainsi, une partie ou la totalité des données est inconnue au départ puis se révèle en temps réel durant la phase d'exécution. Comme pour les problèmes statiques, il est possible de distinguer deux types de problèmes dynamiques : les problèmes *dynamiques* et *déterministes* et les problèmes *dynamiques* et *stochastiques*. Dans le cas des problèmes *dynamiques* et *déterministes*, toutes les données nécessaires sont connues avec certitude au moment où les décisions sont prises. Aucune information concernant les réalisations futures n'est considérée. Un problème déterministe est donc résolu à chaque étape de décision. Tel que discuté dans Berbeglia *et al*. (2010), la solution d'un problème dynamique ne peut pas être statique. Elle constitue plutôt une stratégie ou une politique utilisant l'information qui se révèle dans le temps et spécifiant clairement les actions à entreprendre au fur et à mesure que le temps avance. On cherche alors la meilleure politique possible, plutôt que de déterminer une bonne solution pour une réalisation possible des événements incertains. Dans le cas des problèmes *dynamiques* et *stochastiques*, l'information concernant les distributions de probabilités des différentes données et événements futurs est considérée lors de la prise de décision. Un problème stochastique est donc résolu à chaque étape de décision. Le même principe que dans le cas *statique* s'applique ici : on cherchera la meilleure stratégie possible plutôt qu'une bonne solution pour une réalisation particulière de l'incertitude.

Différents contextes d'application impliquent la résolution de problèmes dynamiques. Par exemple, toujours dans le domaine du transport, plusieurs compagnies de messagerie express sont confron-

tées à chaque jour à la résolution de problèmes dynamiques. En effet, les requêtes sont reçues tout au long de la journée et généralement, un faible pourcentage des demandes est connu à l'avance. L'affectation des requêtes aux véhicules, la détermination des tournées de même que l'ordonnancement des différentes requêtes doivent être effectués et revus en temps réel (Mitrović-Minić *et al*., 2004). Dans le contexte de la gestion des services préhospitaliers d'urgence, la localisation des ambulances sur le territoire à desservir peut être revue de manière dymanique. Le problème de redéploiement dynamique vise alors à relocaliser les véhicules en temps réel, lorsque l'état du système, c'est-à-dire le nombre de véhicules disponibles, change ou le requiert (selon des critères prédéfinis), notamment à la suite de l'affectation d'un ou de plusieurs véhicules à des appels d'urgence (Gendreau *et al*., 2001; Andersson et Värbrand, 2007; Gendreau *et al*., 2006; Mason, 2013).

Plusieurs approches ont été proposées à ce jour afin de considérer les problèmes de gestion sous incertitude. La sélection d'une approche ou d'une méthode de résolution appropriée pour aborder un problème dépendra de plusieurs facteurs : le type d'incertitude, la disponibilité de l'information, le niveau de décision, le temps disponible pour la résolution du problème, les technologies disponibles, etc. ll importe de bien connaître les différentes options possibles afin de faire un choix judicieux dans le contexte considéré. Ce chapitre vise donc à identifier et à définir les différentes méthodes proposées afin de considérer l'incertitude dans les problèmes de gestion en faisant ressortir leurs forces et leurs limites respectives.

De manière générale, les différentes méthodes proposées peuvent être classées en deux grandes familles : les méthodes de résolution *proactives* et les méthode de résolution *réactives*. Les méthodes *proactives* visent à prendre en compte l'incertitude, avant même la phase d'exécution, de manière à en limiter les impacts négatifs. Elles sont généralement appliquées une seule fois, *a priori*, avant la phase d'exécution. Ces méthodes sont tout indiquées pour résoudre les problèmes *statiques*. Les méthodes *réactives* visent, quant à elles, à réagir à l'incertitude plutôt qu'à l'anticiper. En effet, dans certains contextes, les données peuvent changer ou se révéler durant la phase d'exécution. Il peut alors devenir intéressant, voire nécessaire, de revoir la solution déterminée précédemment ou de la mettre à jour afin d'améliorer l'atteinte des objectifs visés. Les méthodes *réactives* sont appliquées en cours d'opération, *a posteriori*, de manière périodique ou lorsque le système le requiert. Les méthodes *réactives* représentent des choix naturels pour résoudre les problèmes dynamiques.

Dans ce chapitre, les méthodes *proactives* sont d'abord abordées, suivies des méthodes *réactives*. Bien que ces deux types de méthodes soient traditionnellement utilisées indépendamment, il existe une dualité intéressante entre elles qui nous porte à croire que, dans certaines circons-

tances, il serait intéressant de les considérer conjointement. Par exemple, lorsqu'une méthode réactive est appliquée, un modèle de programmation stochastique pourrait être résolu. Ainsi, les efforts de réoptimisation entre deux étapes de décision successives pourraient être limités. De plus, dans certains cas, même une solution déterminée *a priori* en considérant l'incertitude pourrait, à un certain point, devenir impraticable, nécessitant ainsi le recours à une méthode réactive. L'application des deux types de méthodes pourrait certes amener des bénéfices importants. Une discussion sur les avantages potentiels et les enjeux liés à l'utilisation seule ou conjointe des méthodes présentées viendra clore ce chapitre.

L'objectif général de ce chapitre consiste donc à présenter une revue et une classification des différentes méthodes de résolution permettant de considérer l'incertitude dans les problèmes de gestion. Plus précisément, les différentes méthodes proposées pour chaque type de problème, statique ou dynamique, seront répertoriées puis discutées en mettant l'accent sur leurs forces, leurs limites et leurs distinctions. Chaque méthode sera également illustrée grâce à un problème classique tiré du domaine de la distribution : le problème de tournées de véhicules. Ainsi, pour chaque méthode considérée, une variante du problème de tournées de véhicules sera présentée afin d'en faire ressortir les principales caractéristiques. D'autres applications dans divers domaines, notamment dans le domaine du transport et de la localisation, seront également répertoriées pour chaque méthode. Naturellement, nous ne prétendons pas couvrir toutes les méthodes et les applications qui ont été considérées jusqu'à maintenant, mais plutôt dresser un portrait réaliste des méthodologies qui ont été utilisées plus fréquemment ou qui semblent émerger dans la littérature, puis d'en illustrer le potentiel grâce à des exemples d'application plus concrets.

De toute évidence, peu importe la méthode choisie pour prendre en compte l'incertitude, le modèle résultant devra être résolu de manière à déterminer une bonne solution, idéalement optimale, au problème traité. Considérer explicitement l'incertitude dans un problème de gestion peut toutefois amener des difficultés au niveau de la résolution. En effet, dans la plupart des cas, le fait de considérer l'incertitude engendre des modèles plus complexes, et donc plus difficiles à résoudre. De plus, dans d'autres cas, notamment dans le cas de méthodes réactives, le temps disponible pour la résolution est limité. Les problèmes traités devront donc être résolus très rapidement. La variété des problèmes considérés de même que les défis reliés à leur résolution ont mené au développement d'algorithmes de résolution variés, parfois exacts, mais surtout heuristiques, afin de déterminer de bonnes solutions dans des délais de temps raisonnables. L'objectif de ce chapitre consistant plutôt à présenter les différentes approches proposées pour chaque type de problème, statique ou dynamique, du point de vue de la modélisation, le recensement exhaustif des algorithmes développés pour la résolution proprement dit ne sera pas présenté ici.

Quelques références seront toutefois proposées, pour chaque type de méthode, afin d'orienter le lecteur vers les algorithmes de résolution appropriés.

1.1 Problèmes statiques : méthodes de résolution proactives

Traditionnellement, plusieurs problèmes d'optimisation sous incertitude ont été résolus en substituant les données incertaines par leurs valeurs moyennes ou par des valeurs déterminées grâce à des modèles de prévision ou des données historiques. L'incertitude est alors approximée avant la modélisation. Une telle substitution mène à la résolution de problèmes *déterministes* et *statiques*. Elle mène alors à des problèmes plus faciles à modéliser, de plus petite dimension et, conséquemment, plus faciles à résoudre ; la solution choisie étant celle qui possède la meilleure valeur d'un critère connu et mesurable. En contrepartie, les solutions générées sont souvent moins flexibles et robustes face aux variations des données, pouvant ainsi mener à des solutions de piètre qualité, voire irréalisables, lorsque confrontées aux données réelles. Intégrer l'incertitude plus explicitement lors du développement des modèles représentant le contexte traité constitue alors une alternative intéressante pour améliorer la flexibilité et la robustesse des décisions implantées. Cela entraînera des solutions stochastiques aux caractéristiques différentes des solutions déterministes qui permettront de mieux faire face à l'incertitude.

Les différentes méthodes proactives présentées dans cette section visent donc, chacune à leur manière, à intégrer l'incertitude liée à la qualité des données dans la phase de modélisation. Chaque méthode est alors appliquée une seule fois de manière à estimer les réalisations futures de l'incertitude avant de mettre le système en opération. L'incertitude est prise en compte *a priori*. Il en occasionnera une solution ou une politique unique, généralement plus robuste que celle déterminée grâce au modèle déterministe équivalent et possédant des caractéristiques différentes. Cette solution sera appliquée telle quelle en suivant les règles établies à l'avance ou moyennant quelques modifications durant la phase d'exécution. On parlera aussi de planification à l'avance (Yu et Qi, 2004). Les méthodes proactives sont considérées plus traditionnellement dans le cadre de décisions à caractère stratégique ou tactique (Van Hentenryck et Bent, 2006).

Considérer une méthode proactive présente divers avantages. Tout d'abord, comme l'application de méthodes proactives ne nécessite pas la prise de décision en temps réel, le temps de calcul ne constitue généralement pas un enjeu critique. De plus, puisque le problème est résolu une seule fois avant la phase d'exécution, le plan d'opération est connu à l'avance. Sa mise en oeuvre en est donc simplifiée. D'une part, le plan établi sera plus facilement accepté du point de vue des ressources humaines et les activités connexes pourront aussi être planifiées à l'avance. D'autre

part, aucun support technologique pour la communication en temps réel n'est nécessaire.

Néanmoins, afin de considérer explicitement l'incertitude dans le cadre d'une méthode proactive, il est nécessaire d'en distinguer les différentes sources, de les décrire, puis de quantifier leurs réalisations possibles de même que leurs conséquences. Meilleure est la connaissance de l'incertitude, meilleures sont les décisions. En pratique, bien qu'il soit souhaitable d'identifier et d'estimer à l'avance toutes les sources d'incertitudes, il ne sera pas toujours possible de le faire. En effet, le nombre de sources d'incertitude et/ou leurs réalisations possibles peut devenir trop important. Dans ce cas, il sera nécessaire de poser certaines hypothèses simplificatrices afin de réduire la taille du problème de manière à pouvoir le résoudre. Ainsi, la solution déterminée permettra de se prémunir contre les sources d'incertitude qui auront été considérées. De plus, même en essayant d'anticiper au mieux les différentes sources d'incertitude, le système peut évoluer de manière imprévisible, forçant ainsi le recours à des méthodes réactives. Enfin, les méthodes proactives mènent généralement à des modèles plus complexes quand on les compare aux modèles déterministes équivalents. La résolution de ces modèles requiert donc le développement d'algorithmes performants afin de pouvoir résoudre le problème dans des délais de temps raisonnables. Malgré tout, dans plusieurs circonstances, les efforts de modélisation et de résolution rapportent en flexibilité et en robustesse.

Dans cette section, trois méthodes proactives sont présentées plus en détail : la programmation stochastique, l'optimisation robuste et la programmation dynamique. Ces trois méthodes permettent de modéliser explicitement l'incertitude et ont été appliquées à plusieurs reprises dans les domaines liées à la recherche opérationnelle. La théorie de la décision et le contrôle stochastique optimal permettent aussi de considérer l'incertitude dans les problèmes de gestion lorsque l'incertitude est représentée sous la forme de distributions de probabilités. Leurs applications sont toutefois plus limitées dans les domaines qui nous intéressent. Pour cette raison, ces méthodologies ne sont pas abordées ici.

1.1.1 Programmation stochastique

La programmation stochastique est une méthodologie qui étudie la prise de décision optimale sous incertitude. Elle allie de manière étroite les aspects théoriques issus de la programmation mathématique et les différentes procédures algorithmiques d'optimisation. Les modèles issus de la programmation stochastique sont en fait des généralisations des modèles déterministes où certaines données ne sont pas connues avec certitude. L'incertitude y est représentée sous forme d'événements aléatoires ; les paramètres et données incertaines étant représentés par des variables aléatoires pour lesquelles une description des distributions de probabilités est suppo-

sée connue. Les paramètres incertains peuvent être représentés par des distributions discrètes ou continues. Les valeurs réelles des différentes données ne sont révélées qu'à la suite de la réalisation des événements aléatoires. Les décisions se divisent donc en étapes : les décisions de première étape, prises avant la réalisation des événements aléatoires, et les actions de recours, déterminés à la suite de la réalisation des événements aléatoires (Birge et Louveaux, 2011).

Deux principaux types de modèles stochastiques sont considérés en pratique : les modèles avec fonction de recours et les modèles avec contraintes probabilistes. Les premiers travaux en lien avec la programmation stochastique, que ce soit la programmation stochastique avec fonction de recours ou avec contraintes probabilistes, datent des années 1950. En effet, tel que discuté dans Birge (1997), les premiers modèles impliquant une observation suivie d'une réaction (ou action de recours) sont apparus dans Beale (1955) et Dantzig (1955). Les modèles avec contraintes probabilistes ont été développés quelques années plus tard, principalement par Charnes et Cooper (1959). Depuis l'apparition de ces travaux fondateurs, de nombreux modèles et applications basés sur le principe de la programmation stochastique ont vu le jour et continuent d'apparaître régulièrement.

Formellement, la programmation stochastique avec fonction de recours vise à modéliser explicitement les décisions de première et de deuxième étape. Les décisions de première étape sont considérées avant la réalisation des événements aléatoires, puis des actions de recours ; les décisions de deuxième étape sont quant à elles entreprises à la suite de leur réalisation. En effet, en considérant une décision donnée x, la réalisation d'une variable aléatoire w peut entraîner la violation des contraintes, nécessitant ainsi le recours à différentes actions afin de retrouver la faisabilité. La mise en place de ces actions de recours engendre alors un coût ou une pénalité $Q(x,w)$. En considérant $f(x)$ et $g(x)$, les fonctions représentant respectivement les coûts et les contraintes associées aux décisions de premières étape et $F(w,x,y)$ et $G(w,x,y)$, les fonctions représentant respectivement les coûts et les contraintes associées aux décisions de deuxième étape conditionnellement aux décisions de première étape et à une réalisation de la variable aléatoire w, le modèle avec fonction de recours se formule de la manière suivante (Kall, 1982; Kall et Wallace, 1995; Birge et Louveaux, 2011) :

$$\min f(x) + E_{w\in\Omega}[Q(x,w)] \tag{1.1}$$

sous les contraintes :

$$g(x) \leq 0, x \in X, \tag{1.2}$$

avec

$$Q(x,w) = \min F(w,x,y) \tag{1.3}$$

sous les contraintes :

$$G(w,x,y) \geq 0, y \in Y, \tag{1.4}$$

où $X \in \mathscr{R}^{n_1}$, $Y \in \mathscr{R}^{n_2}$, w est une variable aléatoire appartenant à l'espace des probabilités $(\Omega, \mathscr{F}, \mathscr{P})$ où $\Omega \in \mathscr{R}^k$. Il est possible de constater en observant ce modèle que les décisions de première étape x sont fortement influencées par les décisions de deuxième étape y et vice versa. Les décisions sont également affectées par la réalisation des variables aléatoires. De plus, il est possible de constater qu'une description adéquate de la fonction de recours $Q(x,w)$ doit être disponible. Dans certaines circonstances, formuler la fonction de recours représente un défi en soi. La formulation présentée pour un modèle à deux étapes de décisions peut être étendue directement aux modèles stochastiques multi-étapes.

Le deuxième type de modèles stochastiques étudiés, les modèles avec contraintes probabilistes, considèrent, tout comme les modèles avec fonction de recours, les décisions de première et de deuxième étape. Néanmoins, lors de la formulation d'un modèle probabiliste, la description des décisions ou des actions de recours n'est pas nécessaire. Dans ce cas, plutôt que de décrire formellement la fonction de recours, certaines contraintes sont exprimées de manière à garantir les performances du système ou la faisabilité du problème en fonction des décisions de première étape avec un niveau de fiabilité donné α. Le modèle probabiliste se formule donc de la manière suivante (Kall, 1982) :

$$\min f(x) \tag{1.5}$$

sous les contraintes :

$$P(\{g_i(x) \leq 0, i = 1,...,m\}) \geq \alpha, \tag{1.6}$$

ou

$$P(\{g_i(x) \leq 0\}) \geq \alpha_i, \;\; i = 1,...,m, \tag{1.7}$$

et

$$x \in X, \tag{1.8}$$

où $g_i(x)$ représente une contrainte à satisfaire pour le problème en question. En général, il est possible d'écrire, à partir de la formulation du modèle probabiliste, un modèle déterministe équivalent, beaucoup plus facile à traiter qu'un modèle avec recours. L'utilisation d'une telle

formulation est particulièrement intéressante lorsque les coûts ou les bénéfices liés aux décisions de deuxième étape sont difficiles à déterminer.

Bien qu'elle présente des avantages certains, notamment au niveau de la robustesse des solutions obtenues, l'utilisation de la programmation stochastique amène aussi un certain nombre de défis. Tout d'abord, une description des variables aléatoires doit être disponible sous la forme d'une distribution de probabilité ou une représentation approximative de celles-ci sous la forme d'un ensemble de scénarios représentatif du contexte étudié. Obtenir de telles distributions ou définir un ensemble de scénarios afin de représenter une situation réelle n'est pas toujours une tâche facile. Dans plusieurs cas, il sera nécessaire de faire certaines hypothèses simplificatrices afin de représenter adéquatement les sources d'incertitude par leurs distributions de probabilité ou par un ensemble de scénarios. De plus, dans le cas de la programmation stochastique avec fonction de recours, la description des actions de recours, puis la formulation de la fonction de recours associée, peuvent devenir complexes. La formulation de la fonction de recours va influencer fortement le niveau de difficulté d'un problème. Ainsi, les problèmes peuvent rapidement devenir difficiles à résoudre. Conséquemment, le temps de calcul pour la résolution de programmes stochastiques est généralement long. Plusieurs approches ont été proposées afin de résoudre de manière exacte ou heuristique les différents modèles de programmation stochastique (Birge, 1997). Parmi les plus utilisées notons la méthode *L-shape* (Laporte et Louveaux, 1993; Birge et Louveaux, 2011), le *progressive hedging* (Rockafellar et Wets, 1991; Lokketangen et Woodruff, 1996) et le *sample average approximation* (Kleywegt *et al*., 2001).

Malgré les bénéfices potentiels, le fait d'intégrer l'incertitude nécessite donc des efforts tant au niveau de la modélisation que de la résolution, et c'est d'autant plus vrai pour les modèles avec fonction de recours. Une question se pose alors : est-il vraiment intéressant de considérer de tels modèles ? Deux mesures ont été définies afin de répondre à cette question et ainsi valider l'utilisation d'un modèle stochastique avec fonction de recours pour aborder un problème donné : la valeur espérée de l'information parfaite, ou *expected value of perfect information* (EVPI) en anglais, et la valeur de l'information stochastique, ou *value of stochastic information* (VSS) en anglais. Dans un premier temps, la EVPI vise à déterminer le montant maximal qu'un preneur de décision est prêt à payer pour avoir une connaissance parfaite du futur. Plus précisément, la EVPI calcule la différence entre une solution déterminée grâce à un modèle avec fonction de recours et l'espérance des solutions prises une fois que l'incertitude est révélée. Naturellement, pour chaque réalisation possible de l'incertitude, une solution optimale devra être déterminée afin de calculer l'espérance, ce qui constitue un travail non-négligeable. Plus la différence entre les deux solutions est importante, plus on sera prêt à payer pour connaître le futur. La VSS

mesure, quant à elle, la qualité d'une solution déterminée grâce à un modèle déterministe où les paramètres incertains sont approximés par leurs valeurs moyennes. La VSS calcule donc la différence entre une solution trouvée grâce à un modèle déterministe et une solution trouvée grâce à un modèle avec fonction de recours. Plus la différence entre les solutions est grande, moins la qualité de la solution trouvée grâce au modèle déterministe est grande (Birge et Louveaux, 2011).

La programmation stochastique a été employée pour traiter différents problèmes, et ce, dans divers domaines tel que la distribution de produits. Depuis son introduction par Dantzig et Ramser (1959), le problème de tournées de véhicules (PTV) et ses variantes ont été considérés dans de nombreux contextes d'application où des biens et produits doivent être déplacés vers ou à partir de différents points. Le PTV se définit généralement sur un graphe $G = (V, E, C)$ où $V = \{v_0, ..., v_n\}$ représente un ensemble de sommets, $E = \{(v_i, v_j) : i \neq j, v_i, v_j \in V\}$, un ensemble d'arcs, et où $C = (c_{ij})_{(v_i,v_j)\in E}$ représente la matrice des distances, des coûts ou des temps de déplacement pour l'ensemble des arcs appartenant à E. Traditionnellement, le sommet v_0 représente un dépôt et les sommets restants représentent les différents clients (ou requêtes) à desservir. Le PTV consiste à déterminer un ensemble de tournées pour un nombre donné de véhicules identiques de capacité limitée et basés au dépôt de manière à servir adéquatement chaque sommet en le visitant exactement une fois tout en minimisant les coûts totaux. Plusieurs variantes de ce problème ont été étudiées à ce jour (Toth et Vigo, 2015; Golden *et al.*, 2008). Dans ce chapitre, nous présenterons plusieurs versions du PTV afin d'illustrer les différentes méthodologies abordées.

Dans plusieurs contextes pratiques, certaines composantes ou paramètres du problème de tournées de véhicules ne peuvent être connus avec certitude. Par exemple, le volume de lait à transporter d'une entreprise laitière vers une usine de transformation pourrait varier considérablement d'une journée à l'autre. Le problème de tournées de véhicules stochastique (PTVS) a donc été formulé de manière à considérer la variabilité liée à l'une ou l'autre de ses composantes. Selon la nature de l'incertitude, différentes versions du PTVS ont été proposées. Une des versions les plus étudiées jusqu'à maintenant est le problème de tournées de véhicules avec demandes stochastiques. Dans ce cas, seule une approximation de la demande est connue au moment où les tournées sont planifiées. La demande réelle d'un client n'est connue qu'au moment où le véhicule arrive chez le client en question. Les tournées sont alors planifiées à l'avance en considérant les distributions de probabilités des demandes des clients et demeurent généralement inchangées pour toute la période de planification. Puisque la demande réelle n'est pas connue lors de la planification des tournées et que chaque véhicule dispose d'une capacité limi-

tée, dans certaines circonstances, il peut devenir impossible de satisfaire la demande d'un client. On parlera alors de *route failures* en anglais. Dans ce cas, différentes stratégies sont possibles. La stratégie la plus simple consiste à retourner au dépôt pour s'approvisionner de manière à pouvoir finir de servir adéquatement le client. D'autres recours sont également envisageables. Par exemple, le véhicule pourra faire un retour préventif au dépôt si la probabilité de ne pas pouvoir servir adéquatement le prochain client devient importante. Selon le type de recours considéré et la distribution de probabilités des demandes, différents modèles de programmation stochastique ont été proposés. De manière générale, les modèles avec contraintes probabilistes chercheront à déterminer un ensemble de tournées pour lesquelles la probabilité d'excéder la capacité du véhicule demeure en-deçà d'un seuil donné. Les modèles avec fonction de recours considèreront plutôt explicitement les deux étapes de décision, la séquence de clients à visiter sera déterminée et les actions de recours correspondantes seront évaluées. Pour plus de détails sur les différents formulations du PTVS, nous invitons le lecteur à consulter Gendreau *et al*. (2015).

Outre la confection des tournées de véhicules, la programmation stochastique a été utilisée pour considérer différents problèmes liés à la confection de réseaux et de chaînes logistiques (Klibi *et al*., 2010; Crainic *et al*., 2011) ; la localisation d'entrepôt et de véhicules dont la localisation des ambulances (Laporte *et al*., 1989; Snyder, 2006; Bélanger *et al*., 2012; Boloori Arabani et Farahani, 2012) ; la planification dans le domaine hospitalier (Beraldi et Bruni, 2009) ; et la planification de la logistique humanitaire (Balcik et Beamon, 2008; Rawls et Turnquist, 2011; Mu *et al*., 2011; Noyan, 2012; Klibi *et al*., 2013), pour ne nommer que ceux-là. Ceci ne constitue pas une liste exhaustive de toutes les applications possibles de la programmation stochastique, son nombre étant trop vaste, mais bien un aperçu de certains contextes où elle a été utilisée. Nous invitons le lecteur à se référer aussi à Sahinidis (2004) et Birge (1997) pour d'autres exemples d'applications de la programmation stochastique.

1.1.2 Optimisation robuste

L'optimisation robuste est une méthodologie de modélisation, combinée à des outils de calcul, permettant de traiter des problèmes d'optimisation en présence d'incertitude. Dans le cas de l'optimisation robuste, les données incertaines appartiennent à un ensemble d'incertitude dont les caractéristiques et les propriétés sont connues. L'incertitude y est donc modélisée sous la forme d'un ensemble de réalisations ou de scénarios possibles plutôt que représentée par des distributions de probabilités. L'idée principale de l'optimisation robuste consiste à déterminer

une solution optimale qui respectera l'ensemble des contraintes du problème, et ce, pour toutes les réalisations éventuelles de l'incertitude. Aucune violation des contraintes n'est permise. L'optimisation robuste permet donc de trouver *a priori* une solution réalisable pour toutes les variations des paramètres à l'intérieur de l'ensemble prédéfini. De cette manière, la solution établie initialement pourra être maintenue peu importe les perturbations des données (Ben-Tal et Nemirovski, 2002; Beyer et Sendhoff, 2007; Ben-Tal *et al.*, 2009; Bertsimas *et al.*, 2011). Tout comme dans le cas de la programmation stochastique, les paramètres incertains peuvent être modélisés comme étant discrets ou continus. Les paramètres discrets mèneront à des approches basées sur un ensemble de scénarios. Dans le cas de paramètres continus, on considérera plutôt qu'ils appartiennent à un intervalle prédéfini, défini comme l'intervalle ou l'ensemble d'incertitude (Snyder, 2006).

Tel que discuté dans Bertsimas et Sim (2004), l'optimisation robuste proprement dite a vu le jour au début des années 1970. Les travaux de Soyester (1973) présentent alors un modèle d'optimisation linéaire qui permet de déterminer une solution réalisable pour toutes les données possibles appartenant à un ensemble convexe. Lorsque considéré en pratique, le modèle proposé par Soyester (1973) peut générer des solutions trop conservatrices dans le sens où elles perdent trop en optimalité, lorsqu'on les compare aux solutions optimales déterminées en utilisant la valeur nominale des paramètres, afin d'en assurer la robustesse. Ainsi, afin de pallier à ce problème et d'assurer un bon compromis entre les performances et la robustesse d'une solution, plusieurs approches ont vu le jour jusqu'à présent. Parmi ces approches notons celles de Ben-Tal et Nemirovski (1998, 1999, 2000) et de Bertsimas et Sim (2004). Ces approches présument qu'il est peu probable que tout puisse mal tourner en même temps. On cherchera alors à trouver la meilleure solution possible à l'intérieur d'un « budget » d'incertitude donné.

En considérant $f_0(x)$, la fonction objectif d'un problème d'optimisation donné et $f_i(x,u_i) \leq 0$, l'ensemble des contraintes associées à ce problème, un modèle d'optimisation robuste se formule de la manière suivante (Bertsimas *et al.*, 2011) :

$$\min f_0(x) \tag{1.9}$$

sous les contraintes :

$$f_i(x,u_i) \leq 0, \quad \forall u_i \in U_i, \;\; i = 1,...,m, \tag{1.10}$$

où $x \in \mathbb{R}^n$ est le vecteur représentant l'ensemble des variables de décision et $u_i \in \mathbb{R}^n$, le vecteur représentant les paramètres incertains prenant des valeurs arbitraires à l'intérieur des différents ensembles d'incertitude fermés $U_i \subseteq \mathbb{R}^k$. Ce modèle vise donc à minimiser la fonction

objectif $f_0(x)$ tout en satisfaisant l'ensemble des contraintes $f_i(x, u_i)$, et ce, pour toutes les valeurs possibles de $u_i \in U_i$. Notons que pour un tel modèle, lorsque l'incertitude n'affecte que les contraintes impliquées, c'est-à-dire la faisabilité du problème, l'optimisation robuste vise à déterminer la meilleure solution pour toutes les réalisations possibles de l'incertitude. Dans certains cas, l'incertitude affectera plutôt les paramètres de la fonction objectif, c'est-à-dire l'optimalité du problème. L'optimisation robuste visera alors à obtenir une solution de bonne qualité bien peu importe la réalisation de paramètres incertains de la fonction objectif. De toute évidence, la fonction objectif, les contraintes, de même que l'ensemble d'incertitude, peuvent prendre différentes formes. Cela entraînera des modèles de différents niveaux de complexité mathématique qui pourront devenir, dans certains cas, difficiles à traiter et à résoudre. La forme de l'ensemble d'incertitude permettra aussi de prendre en compte différents compromis entre les performances et la robustesse d'une solution. Pour plus de détails sur les différents types de modèles d'optimisation robuste, leurs caractéristiques respectives et les méthodes pour les résoudre, nous invitons le lecteur à se référer aux livres de Ben-Tal *et al*. (2009) et Kouvelis et Yu (1997) de même qu'à la récente revue de Gabrel *et al*. (2014).

Puisqu'elle vise à se prémunir contre toutes les réalisations possibles de l'incertitude, la solution d'un problème d'optimisation robuste est généralement conservatrice. En effet, la solution d'un problème d'optimisation robuste doit demeurer réalisable pour toutes les réalisations possibles à l'intérieur de l'ensemble d'incertitude. Pour pallier à ce problème et gagner en flexibilité, Ben-Tal *et al*. (2004) ont proposé une extension de l'optimisation robuste classique : l'optimisation robuste ajustable ou *adjustable robust optimization* en anglais. L'optimisation robuste classique considère que toutes les décisions doivent être fixées *a priori*, aucune action de recours n'étant permise afin de maintenir la faisabilité. Bien que cela permette de représenter certains contextes d'applications, Ben-Tal *et al*. (2004) observent que, dans plusieurs circonstances, seule une partie des décisions sont réellement fixées avant la réalisation des événements aléatoires, les décisions restantes étant considérées ensuite. Ainsi, de manière similaire à la programmation stochastique avec fonction de recours, ils proposent de diviser l'ensemble des variables de décision en deux groupes : les variables *non-ajustables*, fixées avant la réalisation de l'incertitude et les variables *ajustables* ou actions de recours, considérées à la suite de la réalisation des variables aléatoires. Les décisions liées aux variables ajustables permettront de maintenir la faisabilité en tout temps, en fonction des décisions liées aux variables non-ajustables et à la réalisation des événements aléatoires. En considérant de cette manière deux étapes de décision, il sera possible de générer des solutions plus flexibles et souvent de meilleure qualité, qui permettront de satisfaire en tout temps les différentes contraintes du problème. L'utilisation de

l'optimisation robuste ajustable est fortement justifiée lorsque l'optimisation robuste classique fournit une solution trop conservatrice et/ou peu représentative de la réalité.

Un des avantages certains de l'optimisation robuste classique ou ajustable repose sur la robustesse des solutions proposées. En effet, la solution fournie par un modèle d'optimisation déterministe sans considération explicite de l'incertitude est généralement peu robuste, bien qu'optimale pour l'ensemble des données considérées. De telles solutions deviennent rapidement non-réalisables lorsque les données du problème changent légèrement. L'optimisation robuste permet de pallier à ce problème en assurant la faisabilité d'une solution dans toutes les circonstances définies par l'ensemble d'incertitude. Elle permet également de considérer les problèmes où l'incertitude ne peut être représentée sous forme d'événements aléatoires dont les distributions de probabilités sont supposées connues ou lorsque l'information caractérisant les variables aléatoires n'est pas disponible. L'ensemble d'incertitude doit toutefois être clairement spécifié. Un ensemble d'incertitude raisonnable, décrivant adéquatement les différents scénarios probables, mais qui demeure traitable d'un point de vue mathématique et gérable en terme de temps de résolution, n'est pas toujours facile à définir. De plus, le choix de l'ensemble d'incertitude influencera fortement la solution finale. En effet, la solution obtenue pourrait devenir extrêmement conservatrice afin de se prémunir contre un scénario particulier ayant une très faible probabilité d'occurrence. Des performances moyennes devront donc être sacrifiées afin de gagner en robustesse face aux différentes perturbations des données. Le compromis entre robustesse et performance est alors inévitable.

Récemment, l'optimisation robuste a été considérée afin de traiter différentes variantes du problème de tournées de véhicules (PTV) où les demandes d'un ensemble fixe de clients et les temps de déplacement peuvent être incertains, mais caractérisés par un ensemble d'incertitude connu. Ainsi, Sungur *et al*. (2008) ont formulé le problème de tournée de véhicules robuste (PTVR). Le PTVR consiste à déterminer un ensemble de tournées optimales permettant de minimiser les coûts de transport tout en assurant la satisfaction des demandes des clients, et ce, peu importe les demandes qui se réaliseront. Les demandes exactes ne sont connues qu'à l'arrivée du véhicule chez le client en question. Le PTVR identifie une solution *a priori* qui minimise la distance totale parcourue pour le pire cas et qui demeurera réalisable en tout temps, aucun recours n'étant considéré dans ce cas. Puisque qu'aucune action de recours n'est permise afin de maintenir la faisabilité de la solution, il est possible qu'aucune solution réalisable robuste ne puisse être identifiée. Inévitablement, dans ces circonstances, différentes actions de recours devront être envisagées. Par exemple, un véhicule pourrait retourner au dépôt afin de se réapprovisionner pendant la phase d'exécution. À cet effet, Morales (2006), Erera *et al*. (2009) et

Agra *et al*. (2013) ont présenté différentes variantes robustes du PTV sous incertitude où différentes actions de recours ont été considérées. Le fait de considérer de telles actions lors de la formulation d'un modèle d'optimisation robuste amène inévitablement une complexité mathématique plus importante. Bien qu'ils puissent être plus difficiles à traiter, ces modèles avec recours représentent généralement plus fidèlement la réalité.

Bertsimas *et al*. (2011) ont répertorié différentes applications de l'optimisation robuste dans des domaines aussi variés que la gestion de porte-feuille (Gregory *et al*., 2011), la gestion de la chaîne d'approvisionnement et l'ingénierie. Des modèles d'optimisation robuste ont également été développés pour traiter différents problèmes liés à la localisation (Baron *et al*., 2011), la planification de la production (Alem et Morabito, 2012) et l'ordonnancement (Bohle *et al*., 2010; Li et Ierapetritou, 2008). De toute évidence, d'autres contextes pourraient justifier l'utilisation de l'optimisation robuste telle que proposée par Ben-Tal *et al*. (2009) ; les applications présentées dans cette section ne représentant qu'un bref aperçu des différents problèmes considérés jusqu'à présent. Nous invitons le lecteur à se référer à Kouvelis et Yu (1997), Ben-Tal et Nemirovski (2002), Ben-Tal *et al*. (2009), Bertsimas *et al*. (2011) et Gabrel *et al*. (2014) pour différentes revues des applications de l'optimisation robuste.

D'autres méthodes liées à l'obtention de solutions robustes ont également été proposées. Elle se distinguent néanmoins, par certains aspects, du cadre théorique proposé pour l'optimisation robuste tel que décrit dans Bertsimas *et al*. (2011). En effet, tel que discuté dans Ben-Tal et Nemirovski (1998), la programmation mathématique robuste (PMR) proposée par Mulvey *et al*. (1995) permet aussi de considérer l'incertitude dans les problèmes de gestion en fournissant des solutions robustes. La PMR allie la programmation par objectifs et les approches basées sur la définition de scénarios. Dans ce cas, plusieurs scénarios sont donc considérés. La solution candidate pour le problème étudié pourra toutefois violer certaines contraintes lorsque chaque scénario indépendant est considéré. Ces violations de contraintes seront incluses dans la fonction objectif sous la forme d'un terme de pénalité qui assurera la stabilité de la solution déterminée. Lorsque la réalisation des scénarios devient obligatoire et en considérant certaines hypothèses, la PMR devient un cas particulier de l'optimisation robuste telle que présentée précédemment.

1.1.3 Programmation dynamique et processus de décision markoviens

La programmation dynamique est une méthodologie développée pour la prise de décision séquentielle. Elle regroupe un ensemble d'outils mathématiques utilisés pour analyser et modéliser les décisions séquentielles. Plus spécifiquement, un processus de décisions séquentielles

implique une séquence d'actions et de décisions dans le but d'atteindre un objectif commun. Dans ce contexte, le compromis qui existe entre les coûts associés aux décisions immédiates et ceux reliés aux décisions futures (qui dépendent généralement des décisions immédiates) peut être pris en compte. La programmation dynamique peut être *déterministe*, c'est-à-dire qui s'intéresse aux situations où toutes les données et l'évolution du système sont connus avec certitude, ou *stochastique*, c'est-à-dire qui considère la prise de décision sous incertitude. Dans ce cas, les données incertaines sont représentées sous la forme de distributions de probabilités comme dans le cas de la programmation stochastique. La programmation dynamique peut appliquer un contrôle *discret* ou *continu*. Elle pourra également considérer un horizon de planification *fini* ou *infini* (Bellman, 1957; Denardo, 1982).

Le terme *programmation dynamique* a été introduit par Bellman au début des années 1950 alors qu'il étudiait les processus de décision multi-étapes à la RAND Corporation à Santa Monica en Californie (Dreyfus, 2002). Les contributions de Bellman à la programmation sont multiples, de la formulation de l'équation de Bellman, un résultat central de la programmation dynamique qui formule un problème d'optimisation en une forme récursive (Bellman, 1957), à son application dans différents champs d'application. Depuis, la programmation dynamique est devenue une discipline importante en mathématiques appliquées, en recherche opérationnelle et en informatique, de même qu'une méthode de résolution de problèmes dans différents domaines de l'ingénierie, de l'économie, du commerce et de la gestion, pour ne nommer que quelques exemples (Sniedovich, 2010).

Plus formellement, tout processus de décisions séquentielles considéré par programmation dynamique implique une chaîne d'événements, de décisions et d'observations. À chaque instant où une décision doit être prise, généralement à chaque étape de décision n, l'état s du système est observé. L'état du système consiste en un résumé de l'historique du processus permettant une évaluation adéquate des différentes décisions admissibles. Tout état s fait partie de l'ensemble des états possibles S. À chaque état s est associé un ensemble de décisions admissibles $D(s)$. Ainsi, à chaque étape de décision n, l'état $s \in S$ du système est observé, puis une décision $d \in D(s)$ est sélectionnée. Cette décision engendrera un gain $r(s,d)$ puis une transition vers l'état $t(s,d)$. L'état $t(s,d)$ est alors observé, puis la chaîne d'événements se répète jusqu'à la fin du processus de décision. L'objectif final du problème consiste à maximiser les gains totaux pour l'ensemble du processus décisionnel. Afin de respecter la notation proposée dans Denardo (1982), on dira que l'état s_n est un élément de S_n et qu'une décision d_n est liée à l'étape n. Ainsi, en définissant $f(s_n)$, la fonction représentant le gain total maximal obtenu à partir de l'étape n jusqu'à l'étape $N+1$ si le système se trouve dans l'état s_n, il est possible d'écrire l'équation

fonctionnelle suivante :

$$f(s_n) = \max_{d_n \in D(s_n)} \{r(s_n, d_n) + f[t(s_n, d_n)]\}. \tag{1.11}$$

En programmation dynamique, l'équation fonctionnelle permet de caractériser l'ensemble des solutions à un problème donné. La solution à une équation fonctionnelle est généralement déterminée par récursion. La récursion consiste à déterminer la solution optimale pour chaque état, en suivant une séquence prédéterminée. Un problème de programmation dynamique n'est donc pas résolu directement, mais plutôt représenté par un ensemble de problèmes, un problème par état. Le modèle présenté en (1.11) n'est valide que pour un problème déterministe avec contrôle discret et horizon fini. Il permet toutefois de bien illustrer les principes généraux de la programmation dynamique. Naturellement, la programmation dynamique sous toutes ses formes a suscité beaucoup d'intérêt jusqu'à présent (Bellman, 1957; Howard, 1960; Denardo, 1982; Sniedovich, 2010; Bertsekas, 1987, 2012). Nous invitons le lecteur à consulter ces ouvrages pour une description plus détaillée des différentes catégories des modèles de programmation dynamique et des algorithmes pour les résoudre.

Le modèle présenté ci-haut pose l'hypothèse selon laquelle toutes les décisions sont prises avec une connaissance parfaite du système et de son évolution. Malheureusement, dans bien des contextes, ce n'est pas nécessairement le cas. En effet, l'état dans lequel se trouve le système à la suite d'une décision donnée n'est pas toujours connu avec certitude. Il est toutefois possible de faire un choix éclairé en se basant sur la programmation dynamique *stochastique* lorsque les distributions de probabilités des événements aléatoires sont connues. La réalisation des événements aléatoires pourra avoir un impact sur le gain immédiat et sur l'état futur du système. Ainsi, à chaque étape de décision n, l'état du système s_n est observé puis une décision d_n est sélectionnée. Une variable aléatoire w suivant une loi de probabilité $P(s_n, d_n)$ est ensuite générée. À la suite de l'observation de w, un gain $r(s_n, d_n, w)$ est engendré et l'état du système à la prochaine étape de décision $t(s_n, d_n, w)$ est connu. La programmation dynamique stochastique consiste alors à déterminer l'ensemble des décisions admissibles qui permettront de maximiser l'espérance mathématique des gains totaux ou l'utilité espérée. Dans ce cas, l'équation fonctionnelle s'écrit de la manière suivante :

$$g(s_n) = \max_{d_n \in D(s_n)} E_w\{r(s_n, d_n, w) + g[t(s_n, d_n, w)]\} \tag{1.12}$$

où $g(s_n)$ représente le gain maximal espéré de la période n à la période N lorsque le système est dans l'état s_n. Dans le contexte de la programmation dynamique stochastique, lorsque l'espace

des états possibles est fini, on montre qu'il s'agit d'un processus de décision markovien (Howard, 1960; Puterman, 1994). Dans ce cas, l'évolution entre les états est alors gouvernée par des probabilités de transition, comme dans le cas des chaînes de Markov. Ainsi, même si la prise de décision est considérée sans une connaissance parfaite du futur, les probabilités de transition sont toujours connues. Les gains immédiats et les transitions entre les états ne dépendront que de l'état actuel du système et de la décision choisie. L'équation fonctionnelle pour un processus de décision markovien s'écrit alors de la manière suivante :

$$g(s_n) = \max_{d_n \in D(s_n)} \{r(s_n, d_n) + \sum_{s_{n+1} \in S_{n+1}} p_{s_n, s_{n+1}}(s_n, d_n) g(s_{n+1})\} \tag{1.13}$$

où s_n représente l'état du système à l'étape n, s_{n+1}, l'état du système à l'étape $n+1$, $p_{s_n,s_{n+1}}$, la probabilité de transition de l'état s_n vers l'état s_{n+1}, S_{n+1}, l'ensemble des états possibles à l'étape $n+1$, et $g(s_n)$, le gain maximal espéré de la période n à la période N si le système est dans l'état s_n.

La programmation dynamique sous toutes ses formes a été appliquée afin d'aborder divers problèmes de décisions séquentielles, et ce, dans des domaines aussi variés que la gestion des inventaires et la finance. Lors de la formulation d'un modèle de programmation dynamique, tant déterministe que stochastique, une des étapes cruciales consiste à déterminer l'ensemble des variables d'état permettant de caractériser le système étudié. Selon le cas, le nombre de variables d'état et, conséquemment, le nombre d'états possibles peut devenir très grand. La taille des problèmes à traiter augmente alors rapidement, ce qui constitue malheureusement l'un des inconvénients de la programmation dynamique. La résolution de tels problèmes, de manière exacte, devient difficile, voire impossible, dans des délais de temps raisonnables. Différentes méthodes heuristiques ont été proposées afin de permettre la résolution de modèles de programmation dynamique pour lesquels le nombre d'états possibles devient important. Ces méthodes se basent généralement sur différentes techniques de réduction de l'espace d'états possibles et stratégies d'approximation de la fonction valeur de manière à fournir une représentation adéquate du problème considéré (Bertsekas et Tsitsiklis, 1996; Powell, 20011). Enfin, en présence d'incertitude, lorsque le système est représenté comme processus de décision markovien par exemple, une connaissance des événements ou des états futurs de même que leurs distributions de probabilité est nécessaire. Caractériser adéquatement chaque événement possible en fonction des décisions admissibles, puis déterminer les distributions de probabilités associées peut représenter, dans certains contextes, un défi de taille.

La programmation dynamique permet de considérer différents types de problèmes pouvant être représentés comme un processus de décision séquentiel. Comme beaucoup de problèmes, le

problème de tournées de véhicules peut être représenté par un modèle de programmation dynamique, chaque étape de décision correspondra alors à la sélection du prochain client à visiter. Jusqu'à présent, la plupart des études qui se sont intéressées au PTV avec demandes stochastiques par programmation dynamique ont considéré l'élaboration d'une tournée pour un seul véhicule. En effet, les applications au cas de plusieurs véhicules sont plutôt limitées puisque, même dans le cas d'un véhicule, le nombre d'états possibles devient rapidement très grand, ce qui engendre des difficultés au niveau de la résolution.

Dror *et al*. (1989) ont présenté une première formulation du problème de tournées pour un seul véhicule avec demandes stochastiques sous la forme d'un processus de décision markovien. Dans ce cas, la demande exacte n'est connue qu'à l'arrivée du véhicule chez le client. À l'arrivée du véhicule, deux décisions sont possibles : servir le client puis se déplacer vers une autre localisation ou ne pas servir le client immédiatement et se déplacer vers une autre localisation. Par exemple, le véhicule pourra retourner au dépôt afin de se réapprovisionner. Ainsi, lorsque le véhicule arrive à une localisation donnée, l'état du système est observé, puis une décision est prise. L'objectif consiste alors à déterminer une politique optimale permettant de minimiser l'ensemble des coûts espérés. La politique optimale décrira, pour chaque état possible, la prochaine localisation à visiter. La séquence de clients à visiter fournie grâce à un processus de décision markovien est amenée à évoluer dynamiquement en fonction de l'état du système, mais en suivant une politique optimale déterminée *a priori* contrairement à la séquence de clients fournie par programmation stochastique ou par optimisation robuste qui demeure fixe. Dans le cas du modèle de Dror *et al*. (1989), l'état du système est décrit par un vecteur comportant $n+2$ variables d'état qui inclut la position courant du véhicule, le stock restant à l'intérieur du véhicule et l'état de la demande pour chacun des n clients. Ce problème croît donc rapidement avec le nombre de clients à visiter, complexifiant ainsi sa résolution. Différents auteurs se sont intéressés à la formulation et aux développements d'algorithmes afin de résoudre ce problème de manière heuristique (Secomandi, 2000, 2003; Novoa et Storer, 2009).

White (1985, 1988, 1993) a répertorié différentes applications de la programmation dynamique stochastique en finance et en gestion de porte-feuille, puis en gestion de la production et des inventaires de même que pour la résolution de problèmes variées en agriculture et en localisation, pour n'en nommer que quelques-unes. Ces revues ne présentent qu'un bref aperçu des problèmes qui ont été considérés grâce à cette méthodologie. Nous invitons le lecteur à se référer à Begen (2011) pour un ouvrage plus récent sur la programmation dynamique stochastique et ses applications.

1.2 Problèmes dynamiques : méthodes de résolution réactives

Dans plusieurs contextes d'application, toutes les données nécessaires à la prise de la décisions ne sont pas connues avec certitude au moment où les décisions doivent être considérées. Les méthodes proactives consistent alors à estimer au mieux les réalisations possibles de l'incertitude, lors de la planification statique, de façon à déterminer une solution de bonne qualité ou qui demeure réalisable pour un ensemble de scénarios probables. Pourtant, dans certaines situations, les données sont sujettes à évoluer également dans le temps, souvent de manière imprévisible, rendant la planification à l'avance difficile, voire impossible. En effet, pour plusieurs applications, les données ne se révèlent que pendant la phase d'exécution. Pensons, par exemple, aux appels de détresse pour un service préhospitalier d'urgence ou aux requêtes de transport pour un service de taxi. Ainsi, bien qu'une stratégie de gestion ou un plan d'opération initial puisse être établi avec l'information disponible *a priori*, celle-ci étant, en général, assez limitée, il sera toutefois nécessaire d'intégrer l'information qui se révèlera dans le temps de manière à adapter, à mettre à jour dynamiquement, la solution ou la stratégie déjà établie. Dans certaines autres circonstances, des événements imprévisibles pourront rendre la solution établie à l'avance non-réalisable ou de très mauvaise qualité de sorte qu'une révision des décisions soit nécessaire afin d'assurer le bon fonctionnement et les performances du système. Par exemple, une tempête de neige pourrait forcer une compagnie aérienne à réorganiser ses opérations de manière à détourner certains vols. Un événement imprévisible pourrait également forcer une organisation à planifier rapidement ses opérations sans une connaissance parfaite de la situation réelle puis à les revoir au fur et à mesure que l'information se précise et que la situation progresse. Pensons à la planification de l'aide humanitaire lors d'une catastrophe naturelle. Dans l'une ou l'autre de ces circonstances, les objectifs et les enjeux liés à la prise de décision ne sont pas les mêmes que dans le cadre de la planification statique *a priori*. Il est donc intéressant d'étudier les particularités de ces contextes d'application et des méthodologies proposées afin de les considérer.

Les méthodes réactives présentées dans cette section visent à intégrer l'incertitude, lors de la prise de décision, selon la perspective de l'évolution des données, de la perturbation des données pendant la phase d'exécution à l'arrivée d'un événement imprévisible. Elles sont appliquées en cours d'opération, une fois que le système a commencé à opérer. La prise en compte de l'incertitude se fait donc *a posteriori*. À chaque étape de décision, un problème est résolu, puis le système est modifié, mis à jour, en fonction de la solution déterminée. On parlera aussi de planification en temps réel (Yu et Qi, 2004). Les méthodes réactives pourront être appliquées périodiquement, lorsque les performances du système le justifient, c'est-à-dire lorsque les per-

formances se détériorent au-delà d'un seuil pré-établi, lorsque la solution courante devient non réalisable, ou encore à chaque fois qu'un événement donné survient. Lorsque la solution est remise en question, un problème déterministe ou stochastique pourra être résolu. Dans ce dernier cas, on parlera alors plutôt d'une méthode réactive-proactive puisqu'en plus de réagir, on essaiera d'anticiper le futur de manière à se prémunir contre l'incertitude.

L'utilisation de méthodes réactives pour réagir aux différentes sources d'incertitude présente plusieurs avantages. Tout d'abord, les méthodes réactives sont toujours applicables, ce qui n'est pas nécessairement le cas pour les méthodes proactives. En effet, afin d'appliquer une méthode proactive, il est nécessaire d'avoir une connaissance minimale des données du système et de leurs réalisations futures. Les méthodes réactives peuvent toujours être appliquées. Une solution peut toujours être construite puis mise à jour lorsque l'information se révèle. De plus, certains événements sont si imprévisibles qu'ils sont impossibles à anticiper, mais il sera toujours possible d'y réagir. L'application de méthodes réactives ne se fait toutefois pas sans effort. En effet, modifier un plan en cours d'opérations implique certaines difficultés. Du point de vue des ressources humaines, la modification d'une stratégie déjà établie peut amener des inconvénients certains aux personnes touchées. De plus, des technologies de l'information et de communication doivent être disponibles afin de transmettre, en temps réel, l'information concernant les nouvelles décisions. Le temps de réaction, depuis l'occurrence d'un événement jusqu'à l'implémentation des nouvelles décisions, peut parfois être très court. Dans ce cas, le temps disponible pour la résolution d'un problème étant limité, les choix au niveau de la modélisation et des algorithmes de résolution deviennent des enjeux majeurs.

Afin de répondre aux besoins de la prise de décision en temps réel, deux courants principaux ont été identifiés. Ces deux courants seront discutés plus en détail dans la présente section. La première méthode que nous aborderons est l'optimisation dynamique. L'optimisation dynamique est justifiée lorsque très peu de données sont disponibles au moment où le système est mis en opération. Les données se révèlent une à une pendant la phase d'exécution. La deuxième méthode que nous décrirons est la gestion des perturbations ou *disruption management* en anglais. La gestion des perturbations est pertinente lorsque, à la suite d'un événement donné, le système n'est plus en mesure de fonctionner adéquatement. Différentes décisions devront alors être prises de manière à maintenir un niveau de performance acceptable, ou du moins, à retrouver la faisabilité. La gestion des perturbations intègre différents types de perturbations et de modifications, de la variation des paramètres du système à la suite de l'occurence d'événements rares et imprévisibles. Ces deux courants sont deux manières différentes d'aborder les problèmes dynamiques, de la modélisation au développement d'algorithmes de résolution.

1.2.1 Optimisation dynamique

L'optimisation dynamique s'inscrit dans un processus décisionnel qui permet de réagir et de s'adapter aux événements extérieurs lorsqu'une partie de l'incertitude est révélée. En effet, lorsqu'un problème d'optimisation dynamique est considéré, certaines données, voire la totalité des données, ne se révèlent que pendant la phase d'exécution. Les données deviennent donc disponibles dynamiquement de sorte qu'une séquence de décisions doivent être prises en temps réel de manière à intégrer l'information disponible au moment de la prise de décision, mais sans connaissance ou avec une connaissance imparfaite du futur (Albers, 2003; Van Hentenryck et Bent, 2006; Mahdian *et al.*, 2012). Une solution initiale est alors construite grâce aux données disponibles *a priori*, si cela est possible et pertinent. Ensuite, à chaque fois qu'un événement survient ou que les données se précisent, une ou plusieurs décisions sont prises, puis la solution initiale est modifiée en conséquence. Il y a donc une alternance constante entre l'observation et la prise de décision. La solution déterminée n'est pas statique, elle évolue dans le temps. L'optimisation dynamique permet de considérer des problèmes avec un horizon de planification borné ou non-borné. Elle est habituellement utilisée afin d'aborder des décisions de nature opérationnelle. Afin de mesurer le niveau de dynamisme d'un problème d'optimisation dynamique, une mesure appelée « degré de dynamisme » a été considérée. Le degré de dynamisme se définit comme le ratio entre le nombre de décisions prises dynamiquement, c'est-à-dire qui se réalisent en temps réel, et le nombre total de décisions (Lund *et al.*, 1996). Le degré de dynamisme permet donc de mesurer les efforts nécessaires à la prise de décision en temps réel. Plus le degré de dynamisme est élevé, plus les efforts sont grands.

Tel que discuté précédemment, l'optimisation dynamique est une approche pour aborder des problèmes dynamiques, utilisant ou développant différents outils de résolution propres au problème considéré. L'optimisation dynamique est utilisée depuis plusieurs décennies afin de résoudre différents problèmes pratiques où les données se révèlent en cours d'opération et où une réoptimisation complète ou totale de la solution est nécessaire en temps réel. Le spectre de l'optimisation dynamique est donc très large. Conséquemment, il est difficile d'identifier les premières applications de l'optimisation dynamique, le nombre d'applications étant vaste et propre à chaque domaine d'étude. Ainsi, contrairement aux autres méthodes décrites jusqu'à maintenant, nous ne présenterons pas de discussion à propos de premiers travaux en lien avec l'optimisation dynamique. En contrepartie, nous rapporterons plutôt, à la fin de la présente section, différentes applications possibles de l'optimisation dynamique dans les domaines qui nous intéressent particulièrement.

Traditionnellement, l'optimisation dynamique ne considère pas l'incertitude selon la perspective de la qualité de l'information disponible. Les décisions sont établies en considérant seulement l'information disponible au moment de la prise de décision, de même que l'ensemble des décisions passées. La variabilité des données ou la probabilité des événements futurs n'est pas intégrée. La résolution d'un problème d'optimisation dynamique ne nécessite donc aucune connaissance des événements futurs, ce qui représente un avantage certain. Les différents algorithmes développés pour la résolution de tels problèmes d'optimisation dynamique ont donc été conçus de manière à intégrer progressivement l'information concernant les nouvelles données. Un problème sera résolu à chaque nouvel événement, en temps réel. Les algorithmes proposés devront donc être conçus de manière à limiter le temps de calcul, comme c'est d'ailleurs le cas pour toutes les approches réactives.

Différents algorithmes dynamiques ou *online algorithms* ont été conçus afin de résoudre des problèmes d'optimisation dynamique. Le ratio de compétitivité est une mesure proposée afin d'en évaluer la qualité. L'idée de la compétitivité consiste à comparer une solution générée grâce à un algorithme dynamique à celle trouvée grâce à un algorithme statique équivalent, en posant l'hypothèse que toutes les données sont connues à l'avance. Mieux l'algorithme dynamique approxime la solution optimale, plus l'algorithme est compétitif (Albers, 2003). Ainsi, en considérant $\mathscr{I}$, l'ensemble des instances possibles pour un problème d'optimisation P, $z^\star(I_s)$, la valeur optimale pour une instance statique $I_s \in \mathscr{I}$ pour laquelle tous les événements futurs sont connues *a priori* et $z_A(I) = z(x_A(I))$, la valeur de la solution finale trouvée grâce à un algorithme dynamique A pour une instance $I \in \mathscr{I}$ donnée, le ratio de compétitivité se formule de la manière suivante (Van Hentenryck et Bent, 2006) :

$$\max_{I \in \mathscr{I}} \frac{z^\star(I_s)}{z_A(I)}. \tag{1.14}$$

Il est important de noter que l'équation (14) demeure valide pour tout problème de minimisation. Le ratio de compétitivité correspond donc au pire ratio entre $z^\star(I_s)$ et $z_A(I)$ parmi toutes les instances appartenant à l'ensemble $\mathscr{I}$. Un algorithme dynamique est dit c-compétitif s'il existe une constante c telle que $z_A(I) \leq cz^\star(I_s), \forall I \in \mathscr{I}$. La notion de compétitivité des algorithmes a été abordée dans plusieurs études jusqu'à maintenant. Nous invitons le lecteur à se référer à Albers (2003) et à Fiat et Woeginger (1998) pour plus d'informations à ce sujet.

L'étude de la compétitivité telle que présentée ci-haut requiert, en général, une analyse approfondie du problème traité. Cela peut devenir rapidement complexe dans le cadre d'applications réelles. La valeur de l'information a donc été proposée afin de mesurer la qualité des algorithmes dynamiques. Elle constitue une mesure plus flexible et plus pratique pour l'évaluation des al-

gorithmes dynamiques, dans des contextes d'applications plus réalistes (Pillac *et al.*, 2013). En considérant maintenant $z_A(I_s)$, la valeur de la fonction objectif déterminée grâce à un algorithme dynamique A pour une instance statique I_s, la valeur $V_A(I)$ de l'information se calcule de la manière suivante (Mitrović-Minić *et al.*, 2004) :

$$V_A(I) = \frac{z_A(I) - z_A(I_s)}{z_A(I_s)}. \tag{1.15}$$

Cette mesure représente donc l'écart entre une solution déterminée grâce à un algorithme A pour une instance donnée I et la solution déterminée avec le même algorithme, mais lorsque toutes les données sont connues à l'avance.

Malgré les enjeux relatifs à la modélisation et à la résolution des problèmes d'optimisation dynamique, la prise de décision en temps réel est justifiée, voire nécessaire, dans plusieurs contextes applications réelles. Pensons, par exemple, à une entreprise de messagerie express où les requêtes arrivent en temps réel et ce, tout au long de la journée. Dans ces cas, l'ensemble des requêtes dévoilées en temps réel doivent être servies adéquatement. Ces différents problèmes sont généralement représentés par des variantes du problème de tournées de véhicules dynamiques (PTVD) où l'arrivée des requêtes de clients se fait pendant la phase d'exécution. Le PTVD vise donc à planifier et à mettre à jour, en temps réel, les tournées établies de manière à considérer les requêtes dévoilées tout en minimisant les coûts ou encore en maximisant le service offert. En général, un ensemble de tournées partielles est planifié *a priori* en utilisant l'information disponible, puis les tournées sont mises à jour dynamiquement au fur et à mesure que les requêtes sont reçues. Dans certains contextes, différentes actions autres que la mise à jour des tournées pourront être envisagées. Par exemple, lorsque la garantie de service n'est pas exigée, une requête pourrait être refusée. De plus, un véhicule pourrait être détourné de sa route actuelle si cette stratégie permet de réduire les coûts. Le fait de considérer ces actions/décisions supplémentaires amène un niveau de difficulté accru tant au niveau de la résolution que de l'implémentation. Plusieurs auteurs se sont intéressés à différentes variantes du PTVD. Nous invitons le lecteur à se référer à la revue de Pillac *et al.* (2013) pour une description plus étoffée des différents modèles pour le PTVD, de leurs applications et des méthodes de résolution.

L'optimisation dynamique se présente dans différents contextes d'application dans le domaine du transport de biens et de personnes, notamment dans le cas du problème de collecte et de livraison dynamique (Berbeglia *et al.*, 2010) et de la planification du transport adapté (Cordeau et Laporte, 2007). Des applications ont également été repertoriées dans le domaine hospitalier, pour le transport planifié de patients (Kergosien *et al.*, 2011; Beaudry *et al.*, 2009; Hanne *et al.*, 2009) ou pour la localisation et l'affectation des ambulances (Gendreau *et al.*, 2001; Andersson

et Värbrand, 2007; Bélanger *et al.*, 2012; Schmid, 2012) par exemple. Enfin, différents problèmes d'ordonnancement en temps réel ont aussi été considérés dans la littérature (Durbin et Hoffman, 2008). Naturellement, ces applications de l'optimisation dynamique ne représentent qu'un bref aperçu des contextes où le développement de modèles et d'algorithmes pour la gestion en temps réel est pertinent.

Optimisation combinatoire dynamique et stochastique

Tel que mentionné précédemment, l'optimisation dynamique ne considère traditionnellement pas la variabilité des données ou l'anticipation d'événements futurs. Pourtant, comme dans le cas de la planification *a priori*, il peut devenir pertinent, dans plusieurs contextes, de considérer plus explicitement les différentes sources d'incertitude lors de la prise de décision. Les solutions générées pourraient devenir plus robustes, limitant ainsi les efforts de réoptimisation dans le futur. L'optimisation dynamique sous incertitude a suscité beaucoup d'intérêt récemment. Paru en 2006, l'ouvrage de Van Hentenryck et Bent est d'ailleurs dédié à l'optimisation combinatoire dynamique et stochastique ou *online stochastic combinatorial optimization* en anglais. L'optimisation combinatoire dynamique et stochastique est une classe de problèmes d'optimisation dynamique avec prise en compte exogène de l'incertitude. Plus précisément, l'optimisation combinatoire stochastique et dynamique est une méthodologie qui permet de réagir et de s'adapter aux événements extérieurs efficacement sous contraintes de temps, en anticipant le futur et en apprenant du passé afin de produire des solutions plus robustes et efficientes. Elle allie de manière étroite la programmation stochastique, la théorie sur les algorithmes dynamiques et l'optimisation combinatoire pour la prise de décision séquentielle sous incertitude.

L'optimisation stochastique et dynamique requiert, comme dans le cas de la programmation stochastique, une connaissance des distributions de probabilités des événements futurs. La réalisation des différents événements aléatoires ne devra toutefois pas dépendre du processus de prise de décision. De plus, comme dans le cas de l'optimisation dynamique, seule une partie des données est connue au départ. Une solution initiale est alors construite, si cela est possible et pertinent de le faire, à partir des données disponibles *a priori*, puis la solution est mise à jour au fur et à mesure que la situation se précise. Il y a donc toujours une alternance entre l'observation et la prise de décision. Toutefois, dans le cadre de l'optimisation stochastique et dynamique, les décisions prises de cette manière sont irrévocables. Il n'y a pas de recours possible pour assurer la faisabilité de la solution comme en programmation stochastique par exemple.

Les algorithmes dynamiques anticipatifs ou *online anticipatory algorithms* ont été développés pour la résolution des problèmes d'optimisation dynamique et stochastique. Van Hentenryck et

Bent (2006) ont présenté une classe d'algorithmes dynamiques anticipatifs visant à prendre des décisions en temps réel en représentant le futur par un échantillonage des réalisations possibles des événements futurs. Les décisions sont alors déterminées en résolvant différents problèmes d'optimisation déterministes représentant plusieurs réalisations éventuelles du futur. Van Hentenryck et Bent (2006) proposent d'analyser la performance de ces algorithmes en mesurant la perte espérée ou *expected loss* en anglais. Ainsi, en considérant $z_A(w)$, une solution trouvée grâce à un algorithme dynamique A à un problème d'optimisation P où w est une réalisation des données aléatoires, et $z_O(w)$, la solution optimale à ce même problème d'optimisation, la perte espérée se calcule de la manière suivante :

$$\mathbb{E}_{w\in\Omega}(z_A(w) - z_O(w)) \tag{1.16}$$

où Ω représente la distribution de probabilité des données aléatoires. Notons que cette équation est valide pour tout problème de minimisation. Naturellement, le temps de résolution pour les algorithmes développés dans le contexte des problèmes d'optimisation dynamique et stochastique demeure un aspect critique, comme c'est le cas pour les problèmes plus généraux d'optimisation dynamique.

Dans plusieurs contextes, considérer l'optimisation dynamique et stochastique peut mener à des bénéfices potentiels. Tel que noté par Van Hentenryck et Bent (2006), c'est notamment le cas, par exemple, lors de l'affectation et de la relocalisation des véhicules ambulanciers, pour la gestion des défaillances dans les réseaux électriques, pour la gestion d'une pandémie ou pour la gestion des feux de brousse. Plus formellement, Van Hentenryck et Bent (2006) ont présenté trois champs d'application où l'optimisation dynamique et stochastique est justifiée : les problèmes d'ordonnancement dynamique et stochastique, les problèmes de réservation dynamique et stochastique et les problèmes de routage dynamique et stochastique. Les problèmes de routage étudiés sont similaires aux problèmes présentés pour l'optimisation dynamique. Toutefois, dans ce cas, les distributions de probabilités des demandes devront être considérées lors de la modélisation et de la résolution du problème traité. L'incertitude y est généralement représentée sous la forme d'un ensemble de scénarios probables générés à partir des distributions de probabilités données tel que c'est le cas en programmation stochastique. Pour ces trois types de problèmes, différentes variantes sont présentées de même que différents algorithmes pour les résoudre. Nous invitons le lecteur à se référer à cet ouvrage pour une description plus détaillée des différents algorithmes et une analyse de leurs performances respectives.

1.2.2 Gestion des perturbations

La gestion des perturbations ou *disruption management* en anglais est un processus visant à revoir un plan d'opération à la suite d'une ou plusieurs perturbations, de manière à améliorer les performances du système ou encore à rétablir la faisabilité. Les perturbations considérées peuvent être de différentes natures : changements dans l'environnement du système, changements dans les paramètres du système, changements dans les ressources disponibles, arrivées d'événements imprévisibles, apparitions de nouvelles restrictions, apparitions de nouvelles considérations ou incertitude dans les performances du système. Lorsqu'un problème de gestion des perturbations est considéré, une solution initiale doit d'abord être déterminée à partir des données disponibles au moment de la prise de décision et de l'information accessible concernant les réalisations éventuelles des événements futurs. Cette solution constituera le plan d'opération initial. Les décisions qui constituent le plan d'opération initial sont appliquées, puis le système subit un certain nombre de perturbations causées par différents facteurs internes et/ou externes incontrôlables. Ces perturbations vont venir affecter, de manière plus ou moins significative, les performances du système. Le plan d'opération initial pourra alors devenir sous-optimale ou simplement non-réalisable. Naturellement, toutes les perturbations considérées par la gestion des perturbations n'auront pas le même impact. Par exemple, le déraillement d'un train aura généralement un impact beaucoup plus considérable sur le réseau ferroviaire avoisinant que le retard d'un employé.

La gestion des perturbations telle que considérée ici est une méthode plus récente que celles présentées jusqu'à maintenant. Elle a pris naissance dans les années 1990, principalement dans le cadre de la résolution de problèmes en lien avec l'industrie aérienne (Bard *et al.*, 2001; Clausen *et al.*, 2001). Elle a notamment connu un succès important lors du développement d'un outil de planification et de replanification des équipes de travail chez Continental Airlines (Yu *et al.*, 2003). Les principes de la gestion des perturbations continuent d'être appliquées dans l'industrie aérienne, mais aussi dans d'autres domaines tels que l'ordonnancement, la gestion de la production et la gestion des chaînes d'approvisionnement.

Formellement, la gestion des perturbations est une méthodologie qui vise à revoir dynamiquement un plan d'opération déterminé initialement de manière à obtenir un nouveau plan d'opérations reflétant les contraintes et les objectifs du nouvel environnement dans lequel le système évolue, tout en minimisant l'impact des différentes perturbations (Yu et Qi, 2004). Le problème perturbé considèrera alors la déviation ou la différence entre la solution trouvée initialement et la solution déterminée à la suite des perturbations. Afin d'intégrer adéquatement cette déviation, les problèmes de gestion des perturbations peuvent comporter plusieurs objectifs. D'une

part, ils considèreront les objectifs du problème initial, comme par exemple la maximisation des profits. D'autre part, ils pourront considérer aussi la minimisation des coûts associés à la déviation de la solution, aussi appelés coûts de rétablissement des opérations ou *recovery cost* en anglais. Ces coûts de déviation peuvent prendre différentes formes : délai supplémentaire, coût supplémentaire, changement de la capacité de production, déplacement supplémentaire, etc. La solution d'un problème perturbé devra généralement être établie de manière à minimiser le temps de rétablissement ou *recovery time* en anglais, défini comme le temps écoulé entre l'occurence d'une perturbation et le moment où le plan d'opération initial est rétabli. Naturellement, revenir rapidement aux opérations initialement planifiées peut devenir coûteux. Le compromis entre le coût et le temps de rétablissement des opérations doit alors être considéré lors de la prise de décision. La gestion des perturbations se distingue donc de l'optimisation dynamique par deux aspects principaux. Tout d'abord, elle vise à revenir le plus rapidement possible aux opérations normales, c'est-à-dire aux opérations planifiées initialement. De plus, elle considère explicitement les coûts associés à la modification de la solution déjà établie, ce qui n'est généralement pas le cas en optimisation dynamique.

Deux modèles généraux ont été proposés afin de représenter différents problèmes de gestion des perturbations (Yu et Qi, 2004). Ces deux modèles sont basés sur la programmation par objectifs. Afin de respecter la notation proposée par Yu et Qi (2004), nous définirons x^o comme l'ensemble des décisions initiales, établies avant les perturbations considérées et x, comme l'ensemble des variables de décisions associées au problème courant. De plus, on considérera que les décisions x appartiennent à l'ensemble des décisions admissibles $\hat{X}$ défini de manière à satisfaire les différentes contraintes du problème traité sous perturbations. Enfin, nous définirons $g(a^+, a^-)$ une fonction permettant de calculer les coûts de déviation pour des valeurs données de a^+ et a^-, variables correspondant respectivement aux déviations positives et négatives des nouvelles valeurs des variables de décision par rapport à leurs valeurs initiales. Ainsi, en considérant les variables et les fonctions définies ci-haut, le premier modèle proposé par Yu et Qi (2004) se formule de la manière suivante :

$$\min g(a^+, a^-) \tag{1.17}$$

sous les contraintes :

$$x \in \hat{X}, \tag{1.18}$$

$$x + a^+ - a^- = x^0, \tag{1.19}$$

$$a^+, a^- \geq 0. \tag{1.20}$$

Ce modèle cherche à déterminer une solution réalisable qui soit le plus proche possible de la solution initialement établie. De cette manière, les coûts de déviation sont minimisés.

Le deuxième modèle proposé par Yu et Qi (2004) vise, quant à lui, à déterminer une solution réalisable permettant de minimiser la valeur de $\hat{f}(x)$, la valeur de la fonction objectif dans l'environnement perturbé. Ce modèle se formule de la façon suivante :

$$\min_{x \in \hat{X}} \hat{f}(x). \tag{1.21}$$

Ce modèle se distingue d'un modèle d'optimisation classique par le fait qu'il considère une fonction objectif et un ensemble de contraintes modifiées à la suite des différentes perturbations. De plus, par rapport au premier modèle présenté, il ne considère pas explicitement les coûts de déviation, c'est-à-dire qu'il ne cherche pas à minimiser la déviation entre la solution x^o et la solution x. Il est toutefois possible de les intégrer lors de la description de l'ensemble des solutions admissibles en éliminant, par exemple, les décisions qui engendrent une déviation au-delà d'une limite acceptable. Yu et Qi (2004) ont présenté différentes approches afin de considérer simultanément les deux objectifs décrits ci-haut. La programmation par objectifs lexicographique et la mise en place d'une fonction objectif unique formée par la somme pondérée des différentes objectifs font partie des méthodologies envisagées. Naturellement, d'autres approches peuvent être utilisées afin de considérer différents problèmes de gestion des perturbations. Néanmoins, dans tous les cas, il sera nécessaire d'élaborer différents mécanismes afin de limiter l'impact des perturbations et de permettre le retour aux activités normales.

La gestion des perturbations constitue donc un cadre théorique permettant de prendre en compte un grand nombre de perturbations de différentes natures. Lorsqu'une telle approche est considérée, un plan d'opération initial doit d'abord être établi. Une bonne connaissance du système étudié est alors nécessaire afin de déterminer un plan d'opération initial adéquat, quelle que soit la méthode utilisée pour le faire. Lorsque le système fait face à des perturbations, un nouveau modèle considérant les différents coûts de déviation impliqués est conçu puis résolu. Considérer ainsi les coûts de déviation lors de l'établissement de la nouvelle solution permet d'en faciliter l'implémentation tant au niveau opérationnel que du point de vue des ressources humaines et organisationnelles. Néanmoins, identifier correctement les coûts de déviation, les définir, puis les quantifier peut représenter un exercice difficile. Enfin, comme pour toutes les méthodes réactives, le temps de résolution devient rapidement un enjeu critique. En effet, un nouveau plan d'opération doit généralement être livré et implémenté dans des délais très courts. Le temps disponible pour la résolution d'un problème de gestion des perturbations donné est fortement influencé par la nature du problème et le type de perturbations considérées. Dans certaines cir-

constances, le problème perturbé pourra être résolu de manière efficiente en utilisant le même algorithme que celui considéré pour la résolution du problème initial. Dans d'autres circonstances, un nouvel algorithme devra être développé afin de résoudre adéquatement le problème perturbé dans les délais de temps disponibles. À cet effet, les différentes approches de résolution considérées traditionnellement en recherche opérationnelle, tant exactes qu'heuristiques, pourront être considérées pour la gestion des perturbations. La sélection des approches de résolution pertinente dépendra du problème traité, mais aussi du temps disponible pour sa résolution.

Récemment, Mu *et al*. (2011) ont présenté un modèle de gestion des perturbations afin de considérer une nouvelle classe de problèmes de tournées de véhicules : les problèmes de tournées de véhicules perturbés (PTVP) ou *disrupted vehicle routing problem* (DVRP) en anglais. Dans le cas du PTVP, différents types de perturbations peuvent survenir lors de la phase d'exécution : panne de véhicules, accidents et congestion, départ retardé du dépôt ou d'un point de service, apparition de nouvelles demandes, annulation de requêtes. Ainsi, lorsque différentes perturbations surviennent, la qualité des tournées planifiées peut se dégrader de manière importante, voire même devenir non-réalisables, et ce, peu importe la qualité des tournées planifiées à l'avance. Il sera donc nécessaire de revoir rapidement la solution établie initialement afin de réagir aux différentes perturbations dans le but d'en limiter les impacts négatifs. Le PTVP se distingue du PTV notamment parce qu'il doit considérer, parmi ses objectifs, la limitation des inconvénients liés à la modification des tournées. Mu *et al*. (2011) ont considéré plus particulièrement le problème où un bris de véhicule peut survenir pendant les livraisons, c'est-à-dire entre le moment où un véhicule quitte le dépôt et le moment où il visite le dernier client de sa tournée. Dans ce cas, l'objectif consiste à minimiser le nombre de véhicules utilisés et la distance pour assurer l'ensemble des livraisons à la suite d'un bris en considérant qu'un véhicule de remplacement est toujours disponible au dépôt. Naturellement, une solution utilisant le véhicule de remplacement pour terminer les livraisons est de moins bonne qualité qu'une solution qui assure le service grâce à la flotte de véhicules déjà en place. Tel que noté par les auteurs, le PTVP est, en fait, un cas particulier des problèmes de tournées de véhicules dynamiques (Pillac *et al*., 2013).

Outre le domaine du transport routier, la gestion des perturbations a suscité beaucoup d'intérêts dans le domaine du transport aérien (Filar *et al*., 2001; Kohl *et al*., 2007; Ball *et al*., 2007; Clausen *et al*., 2010) de même que pour la planification du transport ferroviaire (Jespersen-Groth *et al*., 2009; Acuna-Agost *et al*., 2011; Narayanaswami et Rangaraj, 2011; Nielsen *et al*., 2012; Narayanaswami et Rangaraj, 2013). Yu et Qi (2004) ont également répertorié différentes applications de la gestion des perturbations en ordonnancement (Qi *et al*., 2006), en planification de

la production (Clausen *et al*., 2001; Xia *et al*., 2004; Yang *et al*., 2005) et en coordination de chaîne d'approvisionnement (Qi *et al*., 2004; Xiao *et al*., 2005, 2007; Xiao et Qi, 2008; Chen et Xiao, 2009) pour n'en nommer que quelques-unes. Nous invitons de lecteur à se référer à Yu et Qi (2004) pour une description plus détaillée de ces différentes applications.

1.3 Outils d'évaluation de solutions pour des problèmes stochastiques ou dynamiques

Les différentes méthodes présentées jusqu'à maintenant permettent de déterminer une solution à un problème donné, en fonction du modèle conçu afin de le représenter et de l'approche utilisée pour le résoudre. Elles permettent la construction d'une solution complète, c'est-à-dire de fournir les meilleures décisions possibles pour l'ensemble des décisions considérées par le problème, à partir de zéro. On parlera alors de méthodes prescriptives. Toutefois, dans divers contextes, l'application de méthodes descriptives plutôt que prescriptives peut être justifiée. Une méthode descriptive consiste à évaluer une solution ou une politique déjà établie dans un contexte donné. L'application de telles méthodes peut s'avérer pertinente lorsque le système étudié devient trop complexe ou lorsque que les hypothèses nécessaires à l'élaboration d'un modèle mathématique prescriptif sont peu réalistes. Elles sont également justifiées lorsqu'un problème présente un ensemble restreint d'alternatives à évaluer. Chaque alternative considérée sera alors évaluée, puis la meilleure solution ou la meilleure stratégie sera sélectionnée en fonction des mesures de performance choisies.

Les méthodes descriptives incluent un ensemble de méthodes analytiques, telles que la théorie des files d'attente, et numériques, telles que la simulation. Les méthodes analytiques permettent de déterminer, de manière exacte, un ensemble de mesures à partir de modèles théoriques. Elles peuvent être considérées lorsque le système étudié se représente adéquatement par un des modèles théoriques connus. Dans certaines circonstances, le système étudié peut devenir trop complexe ou encore le modèle analytique le décrivant trop difficile à résoudre, de sorte qu'une analyse par simulation est de préférence. La simulation ne fournira alors qu'une estimation des performances réelles du système. L'utilisation d'une méthode analytique est donc préférable lorsque le contexte d'application et les objectifs du problème le permettent. La simulation présente toutefois un intérêt certain, notamment pour sa grande flexibilité. Elle permettra aussi d'évaluer différentes stratégies ou politiques qui pourront être appliquées en temps réel. Dans cette section, nous aborderons d'abord la théorie des files d'attente, puis la simulation. Bien qu'il existe d'autres modèles analytiques permettant de prévoir les performances d'un système en présence d'incertitude, nous nous limiterons ici à la présentation de la théorie des files d'attente, puisqu'elle a été utilisée dans différents domaines qui nous intéressent.

1.3.1 Théorie des files d'attente

La théorie des files d'attente est une méthode analytique qui fournir un ensemble d'équations permettant d'analyser le comportement d'un système pouvant être représenté comme un système de files d'attente. La théorie des files d'attente prend ses origines dans les travaux de Erlang, un mathématicien danois, qui s'est intéressé à la modélisation du système téléphonique de la ville de Copenhagen, au Danemark (Sundarapandian, 2009). Tel que discuté dans Kingman (2009), dans son premier article paru en 1909 au sujet de la théorie des files d'attente, Erlang argumente que le nombre d'appels arrivant à un centre de communication téléphonique pour une période donnée suit une distribution de Poisson. Il y montre aussi comment cette hypothèse mène à l'expression de la distribution des temps d'attente lorsque les appels ont une durée égale et fixe. Depuis, la théorie des files d'attente a permis d'étudier des problèmes dans différents contextes, de la gestion des hôpitaux aux contrôles routiers.

Plus concrètement, un système de files d'attente est composé de clients et de serveurs. Les clients arrivent dans un système à un taux d'arrivée donné. Si un serveur est libre à l'arrivée du client, ce dernier est servi sans délai. Dans le cas contraire, le client est placé en file et attend jusqu'à ce qu'un serveur se libère et que son tour vienne. Une fois servi, le client quitte le système. Différents éléments principaux permettent de caractériser un système de file d'attente : le processus d'arrivée des clients dans le système, le mécanisme en place pour assurer le service, c'est-à-dire le nombre de serveurs et le temps de service et la discipline de la file. Des mesures telles que le nombre moyen de clients dans le système, le temps moyen d'un client dans le système ou le temps moyen d'un client en file, pour n'en nommer que quelques-uns, pourront ensuite être calculées à partir des caractéristiques du système (Gross, 2008; Gautam, 2012). Des hypothèses relatives au taux d'arrivée, à la distribution du temps de service et aux règles régissant la file d'attente sont requises afin d'utiliser adéquatement les diverses équations issues de la théorie des files d'attente. Bien que les modèles de base proposés par la théorie des files d'attente puissent représenter de nombreuses applications, pensons par exemple à l'affection du personnel dans un centre d'appels ou à la planification d'une salle d'urgence dans un hôpital (Gautam, 2012), plusieurs contextes sont beaucoup plus complexes et vont au-delà du modèle décrit ici. Différentes adaptations ont alors été proposées afin de considérer des problèmes plus complexes, comme c'est le cas pour la planification d'un service préhospitalier d'urgence. Nous invitons le lecteur à se référer au *Chapitre 3* de la présente thèse pour une description plus détaillée des modèles développés dans ce contexte.

1.3.2 Simulation

La simulation est une méthodologie visant à imiter le comportement et les opérations d'un système. Un système se définit, dans le contexte de la simulation, comme un ensemble d'entités qui agissent et interagissent ensemble de manière à atteindre un objectif commun. Comme dans le cas de la programmation dynamique, l'état du système, caractérisé par un ensemble de variables d'états, permet de décrire le système étudié à tout instant. La simulation permet donc de représenter un grand nombre de systèmes. Elle permet de représenter des systèmes discrets ou continus. Dans le cas d'un système discret, les variables d'état changent à différents points séparés dans le temps. Dans un système continu, les variables d'état sont plutôt amenées à changer de manière continue dans le temps. La simulation peut également être employée afin de considérer différents problèmes statiques ou dynamiques. Un modèle de simulation statique permet de traiter un problème qui survient à un instant donné ou encore un problème pour lequel le temps est négligeable. La simulation de Monte Carlo représente un exemple de simulation statique (Rubinstein, 1981; Manno, 1999). La simulation dynamique considère, quant à elle, des systèmes pour lesquels le temps doit être pris en compte. Enfin, la simulation peut être utilisée aussi bien dans un contexte déterministe que stochastique. Dans le cas déterministe, aucune composante probabiliste n'est considérée. Le modèle de simulation est lancé une seule fois de manière à évaluer la réponse du système. La simulation devient particulièrement intéressante dans les contextes stochastiques où une ou plusieurs composantes sont aléatoires. Les variables aléatoires vont alors influencer les événements futurs et l'état du système, et conséquemment, les performances du système (Law, 2006; Ross, 2013).

La simulation à événements discrets a été considérée jusqu'à présent dans plusieurs contextes d'application qui nous intéressent. La simulation à événements discrets telle qu'on la connaît aujourd'hui semble avoir été utilisée pour la première fois en Grande-Bretagne à la fin des années 1950, pour traiter des problèmes en lien avec l'industrie de l'acier. La première publication scientifique à aborder ce concept a été écrite par Tocher (1962). Dans ses travaux, Tocher considère la simulation comme un moyen temporaire pour résoudre des problèmes qui pourront éventuellement être résolus grâce aux statistiques mathématiques (Pidd, 2014). Depuis, la simulation à événements discrets est devenue une technique employée dans un grand nombre de champ d'applications tels que l'ingénierie, l'économie et la gestion, pour ne nommer que quelques exemples.

Plus précisément, la simulation à événements discrets vise la modélisation de systèmes qui évolue dans le temps et où les variables d'état changent de manière instantanée à différents points discrets. Un événement survient alors à un point donné dans le temps venant affecter l'état du

système et menant éventuellement à la planification d'événements futurs. Tout au long de la simulation, des observations aléatoires sont générées de manière à représenter adéquatement les différentes sources d'incertitude. La réalisation des variables aléatoires aura aussi un impact sur la planification des événements et l'état du système. La simulation à événements discrets s'intéresse donc à l'évolution d'un système à différents points dans le temps. Les points considérés de même que les sources d'incertitude pertinentes dépendront fortement de l'objectif de la simulation.

Utilisée seule, la simulation à événements discrets permet d'évaluer les performances d'un système confronté à différentes sources d'incertitude. Elle permet de fournir une estimation des performance de systèmes complexes dans un contexte généralement plus réaliste que celui considéré lors de l'utilisation de méthodes prescriptives telles que la programmation mathématique. Comme tout processus de modélisation, la conception d'un modèle de simulation requiert un certain nombre d'hypothèses. Une bonne connaissance du système étudié, des éléments qui le composent et des relations pertinentes entre ses éléments est nécessaire afin de construire un modèle valide qui pourra répondre adéquatement aux questions posées. Une fois le développement du modèle achevé, ce dernier doit être implémenté. Le codage d'un modèle de simulation représente généralement une étape coûteuse en terme de ressources et de temps. Bien que plusieurs logiciels de simulation, comme par exemple Arena (Kelton *et al*., 2006; Altiok et Melamed, 2007), ProModel (Harrell *et al*., 2011), Simul8 (Concannon *et al*., 2007) ou FlexSim (Lavery *et al*., 2011), aient été développés afin de simplifier le développement de simulateurs, ces outils sont parfois limités en termes de flexibilité. Le codage complet du simulateur, à partir de zéro, est souvent nécessaire afin d'obtenir une représentation adéquate du système, ce qui peut représenter des efforts considérables. Enfin, une fois le modèle de simulation conçu et implémenté, différentes alternatives pourront être testées en modifiant les paramètres d'entrée, tout en gardant un bon contrôle sur l'ensemble des conditions expérimentales.

La théorie des files d'attente de même que la simulation sont traditionnellement utilisées afin d'évaluer une solution déjà établie. Elles peuvent aussi être utilisées au sein d'un processus d'optimisation. Par exemple, une méthode analytique ou un modèle de simulation pourraient être utilisés afin d'évaluer une solution dans un processus de recherche locale. Tout un pan de la littérature s'est d'ailleurs intéressé à l'utilisation conjointe de la simulation et des différentes techniques d'optimisation (Andradóttir, 1998; Tekin et Sabuncuoglu, 2004; Fu *et al*., 2005). L'utilisation de la simulation pour évaluer une solution dans un processus de recherche augmente généralement le temps de calcul de manière notable. La pertinence d'utiliser de telles techniques dépendra donc fortement du type d'application traité. Néanmoins, l'utilisation

conjointe de la simulation et de l'optimisation dans des contextes où l'incertitude est présente peut permettre d'évaluer une solution dans un contexte plus réaliste et ainsi améliorer la qualité et la flexibilité des solutions choisies lorsqu'appliquées en réalité.

1.4 Conclusion

Dans plusieurs contextes d'applications, considérer l'incertitude lors de la prise de décision est un enjeu crucial afin de garantir l'atteinte des différents objectifs. On pourra alors essayer de se prémunir au mieux contre les différentes sources d'incertitude ou encore développer des stratégies afin d'y réagir adéquatement. Deux types de méthodes principales ont été identifiées afin de mieux considérer l'incertitude dans les problèmes de gestion : les méthodes proactives et les méthodes réactives. Les méthodes proactives permettent d'anticiper l'incertitude, avant même la phase d'exécution, de manière à en limiter les effets négatifs. Elles sont tout indiquées pour la résolution de problèmes *statiques*. Les méthodes réactives visent plutôt à réagir à l'incertitude de manière à modifier dynamiquement la solution initialement déterminée. Elles permettront la résolution de problèmes *dynamiques*.

Trois méthodes proactives ont été présentées ici : la programmation stochastique, l'optimisation robuste et la programmation dynamique. La programmation stochastique est une méthodologie qui considère explicitement les décisions prises avant et après la réalisation des événements aléatoires, appelées aussi actions de recours. Elle vise à déterminer les meilleures décisions possibles en considérant les réalisations éventuelles de l'incertitude, celle-ci étant représentée sous la forme de variables aléatoires dont les distributions de probabilités sont connues, de même que les différentes actions de recours envisageables. L'optimisation robuste vise, quant à elle, à déterminer une solution satisfaisant les contraintes du problème traité et ce, pour toutes les réalisations possibles de l'incertitude, aucune violation des contraintes n'étant permise. Contrairement à la programmation stochastique, l'incertitude y est représentée sous la forme d'un ensemble des réalisations possibles. Les solutions ainsi générées sont généralement robustes puisqu'elles permettent de se prémunir contre l'incertitude, même dans le pire cas. Néanmoins, l'optimisation robuste peut mener à des solutions très conservatrices, parfois même trop. Pour cette raison, différentes approches ont été proposées afin de pallier le trop grand niveau de conservatisme de l'optimisation robuste classique. Enfin, la programmation dynamique est une technique qui permet aussi de considérer l'incertitude dans un problème de gestion. La programmation dynamique est surtout utilisée pour traiter des problèmes séquentiels ou des problèmes qui peuvent être représentés comme tel, aussi bien déterministes que stochastiques. Lorsque le problème est stochastique et que l'espace des états possibles est

fini, on parlera de processus de décision markovien. La programmation dynamique diffère des deux premières méthodes présentées essentiellement par sa manière d'aborder et de formuler un problème grâce à des équations fonctionnelles. Les trois méthodes présentées ici possèdent un ensemble de caractéristiques communes. Tout d'abord, afin de considérer l'incertitude lors du processus de modélisation, une représentation adéquate de l'incertitude et des données qui y sont associées devra être disponible. De plus, le fait de considérer l'incertitude explicitement lors du processus de modélisation mène généralement à la conception de modèles plus complexes, et souvent, plus difficiles à résoudre. Le temps de résolution peut alors devenir un enjeu critique.

Deux approches réactives ont ensuite été présentées : l'optimisation dynamique et la gestion des perturbations. L'optimisation dynamique est une approche qui considère différents problèmes où seule une partie des données est connue au moment de la prise de décision. Une solution initiale est alors déterminée à partir des données connues *a priori*, puis la solution est mise à jour dynamiquement pendant la phase d'exécution, au fur et à mesure que les données se révèlent. L'optimisation combinatoire dynamique et stochastique a également été proposée afin de considérer ce type de problème. Dans ce cas, les décisions sont prises de manière à considérer aussi les événements futurs lors de la mise à jour de la solution. Une connaissance des distributions de probabilités associées aux événements futurs est alors nécessaire. La gestion des perturbations est, quant à elle, une approche visant à revoir un plan d'opérations initial à la suite de différentes perturbations. Dans le contexte de la gestion de perturbations, la solution du problème perturbé devra refléter les contraintes de même que les objectifs du problème initial. La différence entre la solution déterminée initialement et la solution revue à la suite des perturbations devra donc être considérée lors de la prise de décision. Dans la plupart des cas, il sera également souhaitable de revenir à la solution initialement planifiée le plus rapidement possible. Enfin, les approches réactives, que ce soit l'optimisation dynamique ou la gestion des perturbations, requièrent généralement un temps de réaction très court. Dans plusieurs cas, il sera nécessaire de concevoir des modèles qui pourront être résolus, souvent de manière heuristique, dans de courts délais. La complexité mathématique acceptable de même que les méthodes employées pour la résolution de ces modèles dépendra aussi du type d'applications considérées.

Enfin, deux outils d'évaluation de solutions pour des problèmes stochastiques ou dynamiques ont aussi été abordés. En effet, la simulation et la théorie des files d'attente peuvent également être employées afin d'évaluer une solution déterminée *a priori* ou d'évaluer une politique de gestion dynamique. Contrairement aux autres méthodes présentées, la simulation et la théorie des files d'attente ne permettent pas la construction d'une solution, mais bien l'évaluation d'une

solution déjà établie. La théorie des files d'attente fournit un ensemble d'équations afin de déterminer les performances d'un système pouvant être représenté par une file d'attente. La simulation vise, quant à elle, à imiter les opérations et les processus d'un système donné. Bien qu'elle ne fournisse qu'une estimation des performances du système étudié, la simulation est généralement plus flexible que les méthodes analytiques telles que la théorie des files d'attente, d'où son intérêt certain.

Dans ce chapitre, les méthodes proactives et réactives ont été présentées indépendamment. Pourtant, dans bien des contextes, leur utilisation conjointe pourrait amener des bénéfices importants. Tout d'abord, une méthode proactive pourrait être utilisée afin de planifier les opérations *a priori*, puis, au besoin, une méthode réactive pourrait être employée. En considérant ainsi une méthode proactive afin de déterminer une solution initiale, les efforts de réoptimisation pendant la phase d'exécution pourraient être limités. De plus, dans plusieurs contextes, toutes les sources d'incertitude ne pourront être considérées *a priori*, d'où l'intérêt de considérer aussi une stratégie afin d'y réagir. Enfin, lorsqu'une méthode réactive sera appliquée, un modèle de type proactif pourra être résolu. Un modèle d'optimisation robuste pourrait être considéré de manière réactive par exemple. Ainsi, les efforts de réoptimisation entre deux étapes de décision successives pourraient aussi être limités.

Bien qu'à notre avis la synergie entre les deux types de méthodes puisse mener à de meilleures solutions, plus robustes et mieux adaptées à la gestion en contexte incertain, cette utilisation conjointe ne peut se réaliser sans effort. D'une part, le temps de calcul peut devenir important lorsqu'une approche proactive est considérée, ce qui ne constitue pas un enjeu majeur lorsque les problèmes sont résolus *a priori*, mais qui peut devenir très problématique lorsqu'ils sont considérés en temps réel. D'autre part, l'utilisation de méthodes réactives, proactives ou non, peut présenter différentes difficultés, notamment la nécessité de se doter de technologies d'information et de communication permettant leur mise en oeuvre adéquate, de même que le gérer l'impact de modifications plus fréquentes de la solution, du point de vue des ressources humaines. Dans certains contextes, l'utilisation de méthodes proactives et/ou réactives ne sera donc tout simplement pas envisageable. Néanmoins, dans plusieurs cas, il sera nécessaire, malgré les enjeux et les difficultés, de réfléchir aux différentes manières d'intégrer l'une ou l'autre de ces approches, voire les deux, dans le processus décisionnel.

Les avancées technologiques, la puissance accrue des ordinateurs et la prise de conscience des gestionnaires quant à la nécessité de considérer l'incertitude lors de la prise de décision sont des signes encourageants qui permettent d'espérer un développement plus marqué de telles méthodes dans le but d'améliorer la qualité des solutions pour des problèmes de gestion en

présence d'incertitude. De plus en plus, des méthodes réactives proactives pourront être envisagées et appliquées pour des problèmes avec différents niveaux de dynamisme. Cela constitue, à notre avis, une avenue de recherche des plus intéressantes. Bien que nous sommes convaincus de l'intérêt de l'utilisation conjointe des deux types de méthodologies présentées dans ce chapitre, nous nous limiterons dans le cadre de cette thèse à l'analyse et aux développement de méthodes réactives pour traiter différents problèmes perturbés et dynamiques, d'abord dans le contexte plus général d'un problème de localisation, puis dans le contexte particulier de la gestion des services préhospitaliers d'urgence. Nous envisageons toutefois de considérer l'utilisation conjointe des deux types de méthodes présentées ici lors de recherches futures, en particulier dans le domaine de la gestion des soins de santé.

CHAPITRE 2

LE PROBLÈME DE LOCALISATION PERTURBÉ : MODÈLES, ANALYSE ET RÉSOLUTION PAR UNE APPROCHE DE RÉOPTIMISATION CONTRÔLÉE

La localisation des différentes installations d'une organisation, ports, usines, hôpitaux ou stations de métro pour ne nommer que quelques exemples, peut avoir un impact important sur ses performances. La localisation adéquate de ces installations peut donc représenter un enjeu majeur. Ainsi, les sites ou installations devront être localisés de manière à répondre adéquatement à une demande ou à fournir un service approprié à la population. Cette famille de problèmes est mieux connue sous le nom de problèmes de localisation. Beaucoup d'intérêt a été porté jusqu'à présent à la modélisation et à la résolution de différentes variantes de problèmes de localisation déterministes, c'est-à-dire pour lesquels toutes les données du problème sont connues au moment où les décisions de localisation sont prises (ReVelle et Eiselt, 2005; ReVelle *et al*., 2008; Daskin, 2008; Drezner et Hamacher, 2001; Laporte *et al*., 2015).

Dans le contexte de la localisation, comme dans bien d'autres contextes d'ailleurs, toutes les données d'un problème ne sont généralement pas connues de manière exacte au moment de la prise de décision. D'une part, les décisions de localisation sont généralement établies pour une longue période et une connaissance parfaite de la situation future est rarement disponible. D'autre part, les données du problème peuvent être amenées à évoluer dans le temps, parfois de manière imprévisible, ce qui engendre des difficultés supplémentaires. L'incertitude dans les problèmes de localisation peut donc être vue selon les deux perspectives présentées au *Chapitre* 1 soit la qualité et l'évolution de l'information (Pillac *et al*., 2013). Rappelons que la qualité de l'information considère l'incertitude sur les données disponibles au moment de la prise de décision. L'évolution de l'information traduit plutôt le fait que, dans certains problèmes, une partie ou la totalité des données change ou se révèle durant la phase d'exécution. Négliger ces deux aspects peut avoir des conséquences néfastes majeures. Les solutions déterminées en posant l'hypothèse que toutes les données et les événements futurs sont connus avec certitude au moment de la prise de décision peuvent être bonnes en moyenne, mais peuvent se dégrader rapidement, voire devenir non réalisables, lorsque les données se précisent ou qu'un événement incertain survient. Il est donc crucial de se doter de méthodes afin de déterminer des solutions robustes, mais également d'élaborer des stratégies qui permettront de réagir face à l'incertitude.

Différentes approches ont été proposées jusqu'à maintenant afin de considérer plusieurs variantes du problème de localisation en présence de données incertaines. Ces problèmes consi-

déreront généralement l'incertitude liée aux demandes, aux distances et aux temps de déplacement de même qu'aux coûts d'opération et d'ouverture des sites (Current *et al*., 2001). On parlera alors de problèmes de localisation stochastiques ou robustes (Snyder, 2006). Dans le cas d'un problème de localisation *stochastique*, les paramètres incertains sont représentés par des variables aléatoires pour lesquelles les distributions de probabilités sont connues. Les problèmes de localisation *stochastiques* sont généralement formulés comme des modèles de programmation stochastique avec contraintes probabilistes ou avec fonction de recours (Birge et Louveaux, 2011). Dans un modèle probabiliste, certaines contraintes sont exprimées de manière à garantir les performances du système avec un niveau de fiabilité donné. Dans le cas d'un modèle avec fonction de recours, les décisions de première étape déterminent d'abord la localisation des installations puis, les décisions de deuxième étape ou actions de recours sont considérées ensuite. Ces actions de recours viseront, par exemple, à affecter les clients aux sites sélectionnés. L'optimisation robuste permet également de formuler différents problèmes de localisation sous incertitude. Cette approche de modélisation est habituellement considérée lorsqu'aucune information concernant les distributions de probabilités des paramètres incertains n'est disponible. Un ensemble de scénarios est alors défini de manière à représenter adéquatement la situation étudiée. Snyder (2006) a répertorié différentes applications de la localisation stochastique et robuste dans des domaines tels que la gestion des services d'urgence, la gestion d'un réseau public de bibliothèques ou la gestion de déchets et de matières dangereuses.

L'incertitude dans les problèmes de localisation peut également être considérée selon la perspective de l'évolution des données. En effet, dans ce contexte, les données peuvent être amenées à évoluer, parfois de manière imprévisible, tout au long de l'horizon de planification. On parlera alors de problèmes de localisation dynamiques (Current *et al*., 2001; Owen et Daskin, 1998; Boloori Arabani et Farahani, 2012). Les problèmes de localisation dynamiques considèrent généralement un horizon de planification divisé en un nombre donné de périodes de manière à représenter adéquatement les variations possibles des paramètres du problème. La demande, les coûts, les capacités, les temps de déplacement ou le nombre de sites à localiser pourront ainsi varier d'une période à l'autre. Les problèmes de localisation dynamiques demeurent toutefois déterministes puisque les paramètres sont connus à l'avance bien que dépendant du temps ou de la période. Current *et al*. (2001) ont identifié deux catégories de problème de localisation dynamiques : les problèmes dynamiques *implicites* et les problèmes dynamiques *explicites*. Dans le premier cas, toutes les décisions de localisation sont déterminées au début de l'horizon de planification en considérant l'évolution des données d'une période à l'autre. Ces décisions sont toutefois irrévocables, aucune action n'est permise entre les périodes de manière à amé-

liorer la solution. Dans le deuxième cas, différentes actions sont admises entre les périodes : ouverture et fermeture de sites, expansion de la capacité, relocalisation de certains sites. Ces actions dépendront fortement du contexte d'application. Les problèmes de localisation dynamiques constituent des extensions naturelles des problèmes statiques équivalents où plusieurs périodes sont considérées et où différentes actions et coûts permettent de lier les périodes entre elles. Boloori Arabani et Farahani (2012) ont répertorié différentes applications de la localisation dynamique pour la gestion d'un réseau d'écoles, la gestion d'un système d'hôpital ou encore pour la localisation dans le domaine des télécommunications et de l'électronique.

Les deux grandes familles de problèmes présentés ici, soit les problèmes de localisation stochastiques et les problèmes de localisation dynamiques, permettent de considérer les différentes sources d'incertitude *a priori*. Une solution est donc déterminée à l'avance de manière à prendre en compte l'incertitude avant même la phase d'exécution. Cette solution est ensuite appliquée telle quelle pendant la phase d'exécution. Au *Chapitre 1*, ces approches ont été présentées sous le nom de méthodes de résolution *proactives*. Pourtant, dans certains contextes, même en essayant d'anticiper au mieux les différentes réalisations de l'incertitude, il est possible que la situation change ou qu'un événement imprévisible se présente de sorte que les performances du système se dégradent considérablement. Il pourra alors devenir nécessaire de réagir *a posteriori* afin de retrouver un niveau de performance adéquat. Cette situation pourrait se présenter dans le contexte de la localisation de centres d'aide humanitaire. Par exemple, lorsqu'une campagne de vaccination est lancée (ce fut le cas en 2009 au Québec avec la grippe H1N1 (The Toronto Star, 2009; Audet, 2009; Cameron, 2009)), il est primordial de déterminer rapidement la localisation des centres de vaccination, de même que l'affectation du personnel et des zones de population aux centres, afin de freiner au mieux l'épidémie. De manière générale, ces décisions sont prises sans connaître avec exactitude la réaction de la population à la campagne de vaccination et l'évolution de la maladie traitée. Ainsi, après un certain temps, certains centres de vaccination peuvent devenir peu utilisés tandis que d'autres peuvent demeurer très achalandés. De plus, certaines localisations peuvent devenir inutilisables, ce pourrait être le cas d'une école à la suite de la rentrée scolaire par exemple. Il pourrait donc être intéressant de revoir le plan d'action lorsque l'information se précise de manière à maintenir un bon service à la population. On parlera alors de problèmes de localisation dynamiques en temps réel.

Les problèmes de localisation dynamiques s'apparentent aux problèmes statiques équivalents, mais s'en distinguent aussi par divers aspects. Tout d'abord, les problèmes de localisation statiques sont généralement résolus au niveau tactique/stratégique. Les problèmes de localisation dynamiques sont plutôt considérés au niveau opérationnel et résolus en temps réel. De plus,

les problèmes de localisation dynamiques en temps réel intègrent généralement un ensemble de considérations pratiques de manière à assurer la stabilité de la solution. Cela permettra notamment de faciliter la mise en oeuvre de la nouvelle solution et d'en limiter les coûts. Enfin, les problèmes en temps réel se distingueront aussi des problèmes résolus *a priori* puisqu'ils privilégieront généralement l'obtention rapide d'une solution en opposition à la recherche de l'optimalité. Les problèmes de localisation dynamiques en temps réel ont été moins étudiés dans la littérature que les deux types de problèmes présentés précédemment. Il présente toutefois un intérêt certain dans les domaines tels que la gestion d'urgences ou de crises en contexte urbain (par exemple pour la gestion des véhicules d'urgence) ou en contexte humanitaire (par exemple à la suite d'un désastre).

Ce chapitre s'intéresse à la modélisation, à la résolution et à l'analyse d'un problème de localisation dynamique en temps réel. Ainsi, lorsque différentes perturbations surviennent, une solution déterminée initialement devra être revue de manière à assurer sa faisabilité ou à améliorer l'atteinte des objectifs, tout en limitant l'impact des perturbations sur les opérations prévues. La réoptimisation se fait de manière contrôlée. Plus précisément, ce chapitre s'intéresse à étudier et à proposer une approche formelle afin de réagir à l'incertitude en cours d'opération en adaptant de manière contrôlée le plan d'opération initialement élaboré. Les contributions de ce chapitre sont multiples. Tout d'abord, d'un point de vue théorique, il formalise le concept de la réoptimisation contrôlée, basé sur les principes de la gestion des perturbations (Yu et Qi, 2004), puis l'applique à la formulation d'un problème de localisation perturbé. Au meilleur de notre connaissance, l'idée du contrôle d'une solution n'a pas été considérée directement dans le cadre de problèmes de localisation tels que ceux considérés dans le présent chapitre. Dans un deuxième temps, d'un point de vue méthodologique, ce chapitre présente et introduit une méthode générique et flexible basée sur la génération de colonnes pour la résolution du problème étudié, mais qui pourrai aussi être intégrée au sein d'un outil d'aide à la décision afin de soutenir la prise de décision dans différents contextes. Enfin, d'un point de vue pratique, il propose une analyse détaillée du compromis entre le contrôle de la solution et les coûts impliqués, et ce, pour différents types de perturbations.

Le présent chapitre se structure comme suit. Tout d'abord, une brève revue de la gestion des perturbations dans le contexte de la localisation est présentée. Le problème de localisation perturbé est ensuite décrit, puis une méthode heuristique générique est proposée afin de le résoudre. Enfin, différentes expérimentations visant à analyser le compromis entre le contrôle de la solution et les coûts impliqués et à valider la méthode proposée sont présentées. Une conclusion, de même que différentes perspectives de recherche, viennent clore ce chapitre.

2.1 Gestion de perturbations

La gestion des perturbations, ou *disruption management* en anglais, est un processus visant à revoir un plan d'opération lorsqu'une ou plusieurs perturbations se présentent de manière à améliorer les performances du système ou encore à rétablir la faisabilité. Les perturbations considérées peuvent être de différentes natures : changements dans l'environnement du système, changements dans les paramètres du système, changements dans les ressources disponibles, arrivées d'événements imprévisibles, apparitions de nouvelles restrictions, apparitions de nouvelles considérations ou incertitude dans les performances du système. La gestion des perturbations cherche donc à modifier dynamiquement une solution déterminée initialement de manière à obtenir une nouvelle solution reflétant les contraintes et les objectifs du nouvel environnement dans lequel le système évolue, tout en minimisant l'impact des différentes perturbations (Yu et Qi, 2004). Le problème perturbé considérera, lors de la recherche de la nouvelle solution, la déviation ou la différence entre la solution trouvée initialement et la solution déterminée de manière à réagir aux perturbations. Ces « coûts » de déviation peuvent prendre différentes formes : délai supplémentaire, coût supplémentaire, changement de la capacité de production, déplacement supplémentaire, etc. Différents mécanismes pourront alors être mis en place afin de limiter les coûts de déviation. Le temps de rétablissement des opérations, ou *recovery time* en anglais, défini comme le temps écoulé entre l'occurrence d'une perturbation et le moment où le système revient au plan d'opération initial, est un autre aspect important de la gestion des perturbations. Ainsi, la solution d'un problème perturbé devra souvent être établie de manière à revenir le plus rapidement possible aux opérations initialement planifiées, c'est-à-dire de façon à minimiser le temps de rétablissement. Nous invitons de lecteur à se référer à Yu et Qi (2004) pour une description plus détaillée des applications possibles de la gestion de perturbations.

2.1.1 Gestion proactive des perturbations dans le contexte de la localisation

Trois types de perturbations principales ont été considérées jusqu'à présent dans les différents problèmes de localisation : la défaillance d'un site ou d'une installation ou *facility disruption* en anglais, l'inopérabilité d'un lien entre un site et un client ou *link failure* en anglais et la variation de la demande. Dans les deux premières situations, en cas de défaillance d'un site ou d'un lien, le système n'est plus en mesure d'acheminer adéquatement les biens et services vers les points de demande. Cette situation se présente par exemple lorsqu'une route devient impraticable à la suite d'un désastre (Altay et Green, 2006; Anaya-Arenas *et al.*, 2012; Galindo et Batta, 2013).

Dans ces cas, on essaiera de considérer les perturbations potentielles lors de la planification de manière à assurer la fiabilité du réseau de distribution. Dans la troisième situation, soit lorsque la demande est amenée à varier, tant au niveau de l'ensemble des clients à desservir que de la quantité demandée, les perturbations viendront plutôt affecter la capacité d'une organisation à fournir la bonne quantité, au bon moment, au bon endroit. En effet, lorsque la demande varie, le système peut ne plus être en mesure de répondre adéquatement à la demande, même si la structure même du réseau demeure utilisable. Bien qu'une variation de la demande n'a généralement pas le même impact que la fermeture inattendue d'un site ou d'un dépôt, il peut devenir crucial de considérer la variabilité associée à la demande lors de la modélisation de différentes variantes de problèmes de localisation afin de déterminer des solutions robustes et efficientes. Cet aspect a été considéré plus traditionnellement dans la littérature en lien avec les problèmes de localisation stochastiques et robustes. Afin de limiter l'impact des perturbations éventuelles, différentes approches *proactives* ont été proposées.

Drezner (1987) fut le premier à considérer la possibilité qu'un site ne soit plus en mesure de fournir un service ou de répondre à une demande. Il propose alors deux modèles qui permettront d'assurer la fiabilité du système en cas de défaillance d'un ou d'un certain nombre de sites. Les deux modèles proposés sont des extensions des problèmes p-médiane et p-centre. Snyder et Daskin (2005) ont, quant à eux, introduit deux modèles de localisation basés sur une formulation bi-objectif. Le premier objectif de ces modèles considère alors les coûts associés au problème original sans perturbation et le deuxième objectif, les coûts relatifs aux actions encourues pour faire face aux perturbations du système. Les deux modèles proposés présument que chaque site a la même probabilité de devenir inutilisable. Afin de relaxer cette hypothèse et ainsi considérer les cas où chaque site possède une probabilité donnée de défaillance, différentes approches ont été proposées : génération de scénarios (Shen *et al*., 2011; Snyder *et al*., 2006), ajout de termes non-linéaires (Berman *et al*., 2007; Cui *et al*., 2010; Shen *et al*., 2011; Aboolian *et al*., 2013), considération d'une double affectation (Lim *et al*., 2010, 2013) ou méthode de *continuum approximation* (Daganzo et Newell, 1986; Cui *et al*., 2010; Li et Ouyang, 2010). Nous invitons le lecteur à se référer à Snyder *et al*. (2014) pour une description plus détaillée de chacune des approches et des modèles qui en découlent.

Les problèmes de localisation où un lien entre un site et un point de demande peut devenir inopérable ont aussi été considérés. Néanmoins, ils sont plutôt limités à des cas particuliers. Nel et Colbourn (1990) ont étudié la localisation d'une unique station de diffusion dans un réseau de communication peu fiable. Dans ce cas, le problème est représenté par un graphe où chaque arête possède une probabilité donnée et indépendante d'être utilisable. Le problème

vise alors à identifier le sommet pour lequel le nombre espéré de voisins connectés est le plus grand possible en présence d'arêtes potentiellement défectueuses. Eiselt *et al*. (1992) ont plutôt considéré le problème de localisation d'un nombre donné de sites dans un réseau où une seule arête peut devenir inopérable à la fois. Le modèle proposé vise alors à minimiser le nombre total de demandes déconnectées des sites sélectionnés en considérant qu'une arête puisse devenir inutilisable. Enfin, Melachrinoudis et Helander (1996) ont étudié la localisation d'un site sur un arbre où chaque arête présente une probabilité donnée et indépendante de devenir inutilisable. L'objectif de ce problème consiste alors à maximiser le nombre espéré de points de demande atteignables par un chemin utilisable.

Enfin, différentes extensions stochastiques de problèmes de localisation classiques tels que le problème p-médiane, le problème p-centre et le problème de localisation avec coût fixe ont été proposées dans la littérature. Dans ces cas, la demande est considérée incertaine. Tout d'abord, Weaver et Church (1983) ont présenté une formulation stochastique pour le problème p-médiane où la réalisation des demandes et les coûts de transport sont considérés incertains, mais possèdent une probabilité d'occurrence donnée. Louveaux (1986) a, quant à lui, étudié les versions stochastiques du problème de p-médiane avec capacité et du problème de localisation avec capacité et coût fixe. Les deux modèles proposés par Louveaux (1986) visent à déterminer la localisation des sites de même que l'affectation des clients aux sites de manière à maximiser le profit espéré lorsque les demandes et les coûts sont représentés par des variables aléatoires. Une pénalité est alors considérée lorsque la demande des clients ne pourra être satisfaite. Le problème de localisation stochastique sans capacité a été étudié par Ravi et Sinha (2004). Dans ce contexte, on considérera que les installations pourront être ouvertes à la première ou à la deuxième étape, moyennant des coûts fixes. Enfin, Listes et Dekker (2005) ont étudié la localisation dans le contexte de la logistique inverse en prenant en compte l'incertitude liée à la demande. Différents modèles basés sur l'optimisation robuste ont également été proposés afin de considérer les problèmes de localisation en présence de demandes incertaines, généralement représentées sous la forme d'intervalles. Les problèmes considérés par optimisation robuste visent principalement deux objectifs soit la minimisation du coût maximum ou la minimisation du regret maximum. Puisque ces objectifs de type *minimax* augmentent le niveau de difficulté des problèmes traités, leur analyse et leur application sont plutôt limitées à des cas où un seul site est considéré ou encore à l'étude de la localisation dans des réseaux présentant une structure particulière (Labbé *et al*., 1991; Chen et Lin, 1998; Averbakh et Berman, 2000; Carrizosa et Nickel, 2003). La liste des travaux présentée ici ne prétend pas être exhaustive. Elle vise plutôt à donner un aperçu des travaux illustrant bien les méthodologies employées pour considérer la

variabilité liée à la demande. Nous invitons le lecteur à se référer à Snyder (2006), pour une revue plus détaillée des différents modèles proposés pour la localisation en présence de demandes incertaines.

2.1.2 Gestion réactive des perturbations dans le contexte de la localisation

Les approches présentées à la section 2.1.1 ont toutes été appliquées *a priori*. Elles cherchent donc à déterminer une solution qui permettra de se prémunir contre les différentes sources d'incertitude et ainsi assurer le maintien d'un certain niveau de fiabilité. Bien que ces approches permettent de limiter l'impact des perturbations éventuelles, elles n'entrent pas dans le cadre de la gestion des perturbations tel que présentée par Yu et Qi (2004) qui vise plutôt à réagir aux différentes sources de perturbations. En effet, à la suite des variations de la demande, à l'occurrence d'événements imprévisibles ou encore à l'apparition de nouvelles possibilités, il peut devenir intéressant voire nécessaire de revoir la solution courante, et ce, même si l'incertitude a déjà été prise en compte lors de la construction d'une solution initiale. Tous les types de perturbations n'auront naturellement pas le même impact sur le système. Cela mènera donc à différentes stratégies de réoptimisation, propres aux problèmes étudiés. Par exemple, la relocalisation fréquente d'un site de production est sans doute difficilement envisageable alors que la relocalisation d'un site de distribution temporaire d'aide humanitaire pourra être envisagée plus facilement. Bien que la gestion des perturbations telle qu'elle a été présentée par Yu et Qi (2004) ne semble pas avoir été employée explicitement pour traiter les problèmes de localisation, l'idée de résoudre dynamiquement, en temps réel, un problème de manière à considérer l'évolution des données, puis d'appliquer différents mécanismes afin de limiter l'impact des réoptimisations successives a été considérée dans des cas d'applications particuliers. Ces différents cas d'applications présentent certaines similitudes avec la gestion des perturbations.

Par exemple, dans le domaine de la localisation des véhicules ambulanciers, différents modèles ont été proposés afin de localiser et de relocaliser les véhicules sur le territoire à desservir de manière à considérer l'évolution de la demande et de la disponibilité des véhicules. Certains de ces modèles ont intégré la notion de contrôle de la solution lors de la résolution des problèmes de redéploiement multi-période et dynamique des ambulances. Carpentier (2006) a proposé d'ajouter un terme dans la fonction objectif afin de limiter la distance totale parcourue pour le redéploiement des véhicules entre les périodes. Schmid et Doerner (2010) ont plutôt intégré une pénalité dans la fonction objectif de manière à contrôler le nombre de véhicules déplacés entre les périodes. Ces deux problèmes sont résolus, une seule fois, *a priori*, en considérant les coûts de déviation sous la forme d'un terme de pénalité dans la fonction objectif.

Gendreau *et al*. (2001) ont, quant à eux, étudié le problème de redéploiement dynamique des véhicules ambulanciers, qu'ils ont résolu en temps réel. Afin de tenir compte des différents coûts associés au redéploiement, ils proposent d'intégrer un terme de pénalité dans la fonction objectif de manière à tenir compte de l'historique des mouvements. Ce terme de pénalité vise à décourager les déplacements fréquents, les distances de redéploiement trop longues et les allers-retours entre un même poste d'attente. Enfin, Andersson et Värbrand (2007) ont considéré les coûts de déviation au sein du problème de redéploiement dynamique en minimisant le temps maximal de redéploiement dans la fonction objectif et en ajoutant des contraintes visant à limiter le temps de déplacement de même que le nombre de véhicules relocalisés. Tous ces modèles ont donc intégré d'une manière ou d'une autre l'idée de contrôle des coûts de déviation. Néanmoins, la manière de définir ces coûts de déviation dépend fortement du problème traité. Une définition différente des coûts de déviation pourra mener à des solutions différentes. Identifier adéquatement les coûts de déviation, les définir puis les quantifier représente donc un enjeu important. De plus, il est possible d'observer que ces modèles n'intègrent pas de mécanismes permettant de revenir à des opérations dites normales. En général, dans un système en constante évolution, tel que c'est le cas lors de la gestion des services préhospitaliers d'urgence, la notion d'opérations dites normales peut devenir moins pertinente.

Dans ce chapitre, nous proposerons une approche formelle pour réagir à différentes sources d'incertitude en cours d'opération. Cette approche sera ensuite appliquée à la formulation d'un problème de localisation perturbé. Plus précisément, nous considérerons le problème de localisation avec capacité et affectation unique en présence de différentes sources de perturbations : variation de la demande et de l'ensemble de clients à desservir, nouvelles possibilités quant à la sélection des sites et fermeture inattendue de sites. Le modèle proposé afin de considérer ce problème s'inspirera de la gestion de perturbations en introduisant des mécanismes qui permettront de contrôler la différence entre la nouvelle solution et la solution déterminée initialement, que nous appellerons solution de référence. Il s'en distinguera néanmoins puisqu'il ne cherchera pas à revenir à des opérations dites normales. Le modèle proposé pourra être résolu une fois en réaction à un événement particulier ou encore s'inscrire dans une séquence de décisions. Une réoptimisation dite contrôlée sera donc lancée à chaque fois que la situation change, se précise et le requiert. La réoptimisation contrôlée visera à déterminer une solution de bonne qualité pour le problème perturbé en tenant compte de la solution du problème de référence, c'est-à-dire en contrôlant la différence entre les solutions du problème modifié et du problème de référence. La réoptimisation contrôlée se définira alors comme suit :

Définition 1. *Soit P, un problème d'optimisation tel que* $z_P = min_{x \in \overline{A}(i)} \overline{f}(x,i)$ *où* $\overline{A}(i)$ *représente l'ensemble des solutions réalisables pour une instance donnée* $i \in \overline{I}$, $\overline{I}$ *l'ensemble des instances possibles pour P et* $\overline{f}(x,i)$, *la valeur de la fonction objectif pour une solution x et une instance donnée i et soit* $x^\star$ *une solution de bonne qualité au problème P et considérée comme solution de référence, le problème de réoptimisation contrôlée RP se définit comme un problème d'optimisation tel que* $z_{RP} = min_{x \in A(i)} f(x,x^\star,i)$ *où* $f(x,x^\star,i) = \overline{f}(x,i) + \xi h(x,x^\star,i)$. *Dans ce cas,* ξ *représente la pénalité dans la fonction objectif associée au changement de la solution de référence et* $A(i)$ *l'ensemble des solutions réalisables pour un instance donnée* $i \in I$, *I l'ensemble des instances possibles pour le problème RP, et satisfaisant la contrainte de changement admissible* $h(x,x^\star,i) \leq \Delta$ *où* $h(x,x^\star,i)$ *est une fonction visant à mesurer le changement entre une solution x et la solution de référence* $x^\star$ *et* Δ, *le changement maximal admis.*

En observant cette définition, il est possible de constater que le problème de réoptimisation contrôlée conserve les contraintes et la fonction objectif du problème de référence, mais intègre maintenant deux mécanismes afin de limiter la différence entre deux solutions successives : l'ajout de contraintes, nommées *contraintes de changement admissible*, et l'introduction d'une *pénalité* dans la fonction objectif. L'ajout de contraintes visera à exclure de l'ensemble des solutions réalisables, les solutions qui engendrent une différence trop importante. La pénalité dans la fonction objectif visera plutôt à pénaliser progressivement la différence entre deux solutions : plus la distance entre la nouvelle solution et la solution de référence est grande, plus la pénalité sera importante. Ces deux mécanismes pourront être utilisés simultanément ou de façon indépendante. La définition de $h(x,x^\star,i)$ et la valeur choisie pour Δ dépendront fortement du contexte d'application et seront fixées par l'utilisateur. Néanmoins, il est important d'observer que si les contraintes de changement admissible sont trop sévères, il est possible que *RP* n'admette pas de solutions réalisables. L'ajout d'une pénalité dans la fonction objectif pourra alors devenir une alternative intéressante. De plus, il est possible de constater que si l'on fixe $\xi = 0$ et que l'on pose $\Delta = \infty$, $P = RP$.

Comme pour la plupart des méthodes réactives, le temps disponible pour la résolution de tels problèmes peut rapidement devenir un enjeu critique. En effet, dans certains contextes, une nouvelle solution devra être identifiée et implémentée dans des délais très courts. C'est notamment le cas lors du redéploiement des véhicules ambulanciers. Le temps disponible pour la résolution dépendra donc fortement de la nature du problème et du type de perturbation considéré. Ainsi, dans certains cas, le problème de réoptimisation pourra être résolu de manière efficiente en utilisant le même algorithme ou un algorithme similaire à celui considéré pour la résolution du problème initial. Dans d'autres circonstances, un nouvel algorithme devra être développé afin

de tenir compte de la structure particulière de RP ou afin de le résoudre adéquatement dans les délais de temps disponibles. La sélection de la méthode de résolution adéquate dépendra donc du problème traité, mais aussi du temps disponible à sa résolution.

2.2 Le problème de localisation avec capacité et affectation unique

Les problèmes de localisation ont été grandement étudiés jusqu'à présent (Laporte *et al*., 2015). Ils permettent de considérer différentes décisions liées à la localisation dans plusieurs domaines, tels que la conception de réseau de télécommunication, de production ou de distribution pour ne nommer que ceux-ci. Plus précisément, le problème de localisation avec capacité et affection unique (PLCAU), ou *single-source capacitated facility location problem* (SSCLP) en anglais, consiste à sélectionner un ensemble de dépôts puis à affecter chaque client à un seul et unique dépôt de façon à minimiser l'ensemble des coûts de localisation et d'affectation tout en respectant la capacité des dépôts ouverts. Le PLCAU est un problème d'optimisation combinatoire appartenant à la classe des problèmes N-P difficiles (Garey et Johnson, 1979).
De manière générale, deux formulations ont été proposées afin de représenter adéquatement ce problème : une formulation classique qui associe une variable à chaque décision de localisation et d'affectation (Barceló et Casanovas, 1984; Klincewick et Luss, 1986; Cortinhal et Captivo, 2003) et une formulation basée sur le problème de partitionnement qui associe plutôt une variable à la sélection d'une colonne, généralement définie comme un ensemble de clients et un dépôt (Neebe et Rao, 1983; Diaz et Fernandez, 2002; Ceselli *et al*., 2009). Différentes méthodes de résolution, heuristiques et exactes, ont aussi été développées afin de résoudre ce problème. Le PLCAU constitue donc un problème connu, structuré et applicable dans différents contextes. Il offre un cadre intéressant et bien défini pour l'étude du problème de localisation pertubé et de la réoptimisation contrôlée. Ainsi, dans cette section, nous présenterons brièvement les deux formulations principales proposées afin de représenter le PLCAU ainsi que les méthodes développées afin de les résoudre.

2.2.1 Formulation classique

Soit J, l'ensemble des dépôts potentiels, I, l'ensemble des clients, C_{ij}, le coût d'affectation du client $i \in I$ au dépôt $j \in J$, F_j, le coût fixe d'utilisation du dépôt $j \in J$, d_i, la demande du client $i \in I$ et b_j, la capacité du dépôt $j \in J$ et soit x_{ij}, une variable binaire qui vaut 1 si le client i est affecté au dépôt j, et 0 autrement, et y_j, une variable binaire qui vaut 1 si le dépôt j est utilisé, et 0 autrement, le problème de localisation avec capacité et affectation unique se formule de la façon suivante :

(P_c)

$$z_{P_c} = \min \sum_{j \in J} \sum_{i \in I} C_{ij} x_{ij} + \sum_{j \in J} F_j y_j \tag{2.1}$$

sous les contraintes :

$$\sum_{j \in J} x_{ij} = 1, \ \forall i \in I, \tag{2.2}$$

$$\sum_{i \in I} d_i x_{ij} \leq b_j y_j, \ \forall j \in J, \tag{2.3}$$

$$x_{ij} \in \{0,1\}, \ \forall i \in I, \ \forall j \in J, \tag{2.4}$$

$$y_j \in \{0,1\}, \ \forall j \in J. \tag{2.5}$$

La fonction objectif (2.1) du problème vise alors à minimiser l'ensemble des coûts associés à l'utilisation des dépôts et à l'affectation des clients. Les contraintes (2.2) et (2.4) garantissent que chaque client soit servi par exactement un dépôt. La contrainte (2.3) assure que la capacité des dépôts ouverts soit respectée. L'intégralité des variables de localisation est assurée par la contrainte (2.5).

2.2.2 Formulation basée sur le problème de partitionnement

Soit R, l'ensemble des colonnes réalisables où chaque colonne l inclut un dépôt et un ensemble de clients tel que la contrainte de capacité du dépôt est respectée, R_i, l'ensemble des colonnes couvrant le client i, R_j, l'ensemble des colonnes associées au dépôt j et C_l, le coût associé à la colonne l et soit x_l, une variable binaire qui vaut 1 si la colonne l est sélectionnée, et 0 autrement, le problème de localisation avec capacité et affectation unique se formule de la façon suivante :

(P_p)

$$z_{P_p} = \min \sum_{l \in R} C_l x_l \tag{2.6}$$

sous les contraintes :

$$\sum_{l \in R_i} x_l = 1, \forall i \in I, \tag{2.7}$$

$$\sum_{l \in R_j} x_l \leq 1, \forall j \in J, \tag{2.8}$$

$$x_l \in \{0,1\}. \tag{2.9}$$

où I représente toujours l'ensemble des clients et J, l'ensemble des dépôts. La fonction objectif (2.6) vise alors à minimiser les coûts de sélection des colonnes. La contrainte (2.7) vise à assurer

la couverture de chaque client par exactement une colonne tandis que la contrainte (2.8) assure que chaque dépôt soit sélectionné au plus une fois. Enfin, la contrainte (2.9) garantit l'intégralité des variables de sélection. Dans le cas du PLCAU, le coût C_l associé à une colonne l est donné par $F_{j_l} + \sum_{i \in N_l} C_{i j_l}$ où j_l est le dépôt associé à la colonne l et N_l, l'ensemble des clients inclus dans la colonne l.

2.2.3 Méthodes de résolution

Plusieurs méthodes ont été proposées afin de résoudre le PLCAU de manière exacte ou heuristique. Dans un premier temps, la formulation classique du problème (P_c) a généralement été résolue de manière heuristique au moyen d'approches basées sur la relaxation lagrangienne Ces approches se distinguent principalement par le choix de la ou des contraintes relaxées puis par la technique employée afin de déterminer une solution réalisable. Barceló et Casanovas (1984), Pirkul (1987) et Sridharan (1993) ont choisi de considérer la relaxation de la contrainte d'affectation des clients (2.2). Le problème relaxé consistera alors en un nombre donné de problèmes de sac à dos indépendants. Klincewick et Luss (1986) ont plutôt considérer la relaxation lagrangienne des contraintes de capacité (2.3). Cette relaxation mène à un problème de localisation sans capacité qu'ils ont résolu grâce à une méthode d'ascension duale. Une solution réalisable est ensuite déterminée par une heuristique gloutonne, puis l'affectation des clients est améliorée. Enfin, Beasley (1993) a développé un cadre général pour le développement d'heuristiques lagrangiennes où les contraintes de capacité (2.3) et d'affectation (2.2) sont dualisées. Ce cadre méthodologique a ensuite été amélioré par Agar et Salhi (1998). Les approches de Beasley (1993) et d'Agar et Salhi (1998) se distinguent principalement par la manière de générer des solutions réalisables.

D'autres méthodes heuristiques ont été considérées pour la résolution du PLCAU. Ainsi, Hindi et Pienkosz (1999) ont combiné la relaxation lagrangienne et la recherche locale à voisinage restreint. Dans cette méthode, l'optimisation du sous-gradient est utilisée afin de déterminer une borne inférieure, puis la recherche locale est employée afin d'identifier une bonne borne supérieure. Rönnqvist *et al*. (1999) ont plutôt proposé l'utilisation d'une heuristique de type *repeated matching* afin de déterminer une solution réalisable au PLCAU. Delmaire *et al*. (1999) ont développé quatre heuristiques pour la résolution du PLCAU : une heuristique réactive de type GRASP, une heuristique de recherche avec tabous, puis deux approches hybrides combinant des éléments méthodologiques des deux heuristiques précédentes. Cortinhal et Captivo (2003) ont, quant à eux, utilisé conjointement la relaxation lagrangienne et la recherche avec tabous afin de déterminer respectivement une borne inférieure et supérieure au PLCAU. Cortinhal et

Captivo (2003) ont ensuite eu recours à différents algorithmes génétiques pour la résolution du même problème. Les différents algorithmes génétiques testés se distinguent principalement par la manière de gérer les différentes contraintes, et, par conséquent, les solutions non-réalisables. Malheureusement, les recherches de Cortinhal et Captivo (2003) ont permis de montrer que les algorithmes génétiques semblaient peu intéressants pour la résolution du problème étudié. Ahuja *et al*. (2004) ont, quant à eux, présenté un algorithme de recherche à très grand voisinage afin de résoudre le PLCAU. La particularité de cette méthode repose sur le fait qu'elle permet d'effectuer à chaque itération plusieurs échanges d'un ensemble de clients. Chen et Ting (2008) ont plutôt proposé l'intégration d'une heuristique lagrangienne et d'un système de colonies de fourmis. Enfin, Contreras et Diaz (2008) ont développé un algorithme basé sur le *scatter search* pour la résolution du PLCAU.

Différentes méthodes exactes ont aussi été proposées pour résoudre PLCAU et ce, pour les deux types de formulation présentées. Tout d'abord, Neebe et Rao (1983) ont formulé le PLCAU comme un problème de partitionnement (P_p) qu'ils ont résolu grâce à un algorithme alliant la génération de colonnes et un schéma de séparation et d'évaluation progressive. Dans ce cas, la relaxation linéaire du problème est résolu afin de déterminer une borne inférieure puis, si la solution déterminée est non-entière, une procédure de séparation et d'évaluation progressive est lancée afin de déterminer une solution optimale entière. Holmberg *et al*. (1999) ont, quant à eux, résolu la formulation classique du problème (P_c) en relaxant de manière lagrangienne les contraintes d'affectation des clients afin de trouver une borne inférieure au problème. Ils ont ensuite appliqué une heuristique de type *repeated matching* afin de déterminer une solution réalisable. Ces deux bornes sont ensuite incorporées à un schéma de séparation et d'évaluation progressive afin de déterminer une solution optimale au PLCAU. Diaz et Fernandez (2002) ont plutôt développé un algorithme exact pour résoudre le PLCAU où une procédure de génération de colonnes est intégrée à un schéma de type *branch-and-price*. Une borne inférieure est alors déterminée par la relaxation lagrangienne des contraintes d'affectation des clients, à partir de la formulation de type partitionnement, puis une heuristique primale est utilisée afin de trouver une solution réalisable au problème, à partir de la solution de la relaxation lagrangienne. Cette solution est ensuite améliorée par recherche locale. Plus récemment, Ceselli *et al*. (2009) ont développé un cadre de résolution général de type *branch-and-price* pour un ensemble de problèmes de localisation avec capacité, dont le problème avec affectation unique. Dans ce schéma algorithmique, Ceselli *et al*. (2009) proposent de déterminer une borne inférieure par relaxation linéaire ou relaxation lagrangienne. Ils proposent ensuite deux heuristiques, respectivement basées sur la solution de la relaxation lagrangienne et sur la solution de la relaxation linéaire du

problème de partitionnement, afin de déterminer une solution réalisable au problème. Enfin, Yang *et al*. (2012) ont développé un algorithme exact de type *cut-and-solve* pour la résolution de P_c. À chaque niveau de l'arbre de branchement, deux problèmes sont résolus. Le *Sparse Problem*, dont l'espace de solution est réduit par la fixation de certaines variables, est d'abord résolu afin de déterminer une borne supérieure au problème, puis la relaxation linéaire du *Dense Problem* est résolue afin d'identifier une borne inférieure. Différentes inégalités sont ajoutées afin de renforcer la relaxation linéaire du problème.

En conclusion, il est possible de constater que différentes approches, tant heuristiques qu'exactes, ont été proposées pour résoudre le PLCAU. La plupart des approches intégraient, du manière ou d'une autre, une relaxation lagrangienne d'un ou plusieurs groupes de contraintes, puis différents mécanismes ont été mis en place afin de déterminer une solution au problème. La méthode de résolution proposée par Ceselli *et al*. (2009) qui utilise la formulation de type partitionnement pour développer un cadre de résolution général qui permet de traiter un ensemble de problème de localisation nous paraît particulièrement intéressante. En effet, la flexibilité de la méthode et son caractère générique est un aspect important dans le contexte du problème de localisation perturbé. Nous en discuterons davantage à la section 2.4.

2.3 Le problème de localisation perturbé

Dans le contexte de la localisation, différents types de modifications ou de perturbations des données peuvent survenir. La demande de certains clients peut varier, la capacité voire la disponibilité de certains dépôts peut changer ou encore certains clients peuvent apparaître ou disparaître. Le problème de localisation avec capacité et affectation unique perturbé (PLCAUP) consiste donc à localiser ou à relocaliser, en cas de perturbations, les dépôts de capacité donnée, puis à réaffecter les zones de demandes, clients ou ressources aux différents dépôts de manière à ce que chacun d'entre eux se rapporte à une seule et unique localisation, en considérant la solution de référence. Ce problème se distingue du problème de localisation sous-jacent, le PLCAU, puisqu'il intègre les mécanismes de contrôle issus de la réoptimisation contrôlée de manière à limiter la différence entre la solution élaborée initialement et la solution du problème perturbé. Dans cette section, deux formulations sont présentées pour le PLCAUP, visant toutes deux à définir puis à intégrer aux formulations présentées à la sections précédentes les mécanismes issus de la réoptimisation contrôlée.

2.3.1 Formulation classique

Selon les applications, différentes définitions de la contrainte de changement admissible peuvent être proposées. Par exemple, dans le contexte du redéploiement des véhicules ambulanciers, la contrainte de changement admissible pourrait viser à limiter le nombre de véhicules à déplacer ou encore à contrôler la distance maximale de relocalisation. Dans le cas qui nous intéresse, on considérera la différence entre deux solutions comme la différence pondérée entre les décisions de localisation et d'affectation déjà établies et les nouvelles décisions de localisation et d'affectation. Ainsi, afin de tenir compte de la hiérarchie entre les deux types de décisions, des poids β et α seront respectivement associées à la modification du statut d'un dépôt et à la modification de l'affectation d'un client. Différents cas de figure pourront alors être considérés.

Dans tous les cas, les poids β et α sont choisis par l'utilisateur de manière à représenter le contexte dans lequel il évolue. Par exemple, en fixant $\beta = 100$ et $\alpha = 1$, on représentera un contexte où le changement de statut d'un dépôt implique un coût beaucoup plus important que le changement d'affectation d'un client. La valeur choisie pour Δ, le changement maximal permis ou autrement dit le « budget » de la contrainte de changement admissible, est aussi fixée par l'utilisateur. D'un point de vue pratique, la valeur de Δ correspond au prix maximum que l'utilisateur est prêt à payer pour changer la solution initialement élaborée. Les valeurs sélectionnées pour β et α auront nécessairement un impact sur la valeur choisie pour Δ. Il reviendra donc à l'utilisateur de fixer ces valeurs de manière à représenter la situation qui l'intéresse. En changeant les valeurs des paramètres, des solutions aux caractéristiques différentes, en termes de qualité de la solution et d'efforts liés au changement de la solution, pourront être obtenues.

Ainsi, en définissant $\bar{I}$ l'ensemble des clients dans le problème de référence, $\bar{J}$ l'ensemble des dépôts dans le problème de référence, $J^\star$ l'ensemble de dépôts utilisés dans la solution de référence $S^\star$, I l'ensemble des clients dans le problème perturbé, J, l'ensemble des dépôts dans le problème perturbé,A l'ensemble des affectations (i,j) réalisables pour le problème perturbé, $\overline{A}$ l'ensemble des (i,j) réalisables pour le problème de référence, $A^\star$ l'ensemble des (i,j) tels que $x^\star_{ij} = 1$ dans la solution de référence $S^\star$, A^+ l'ensemble des (i,j) tels que $i \in I \cap \bar{I}$ et $j \in J \setminus \bar{J}$ et en considérant Δ le changement maximal admissible, la contrainte de changement admissible proposée se formule de la manière suivante :

$$\sum_{(i,j)\in(A\cap\overline{A}\setminus A^\star)\cup A^+} \alpha x_{ij} + \sum_{j\in J^\star\cap J} \beta(1-y_j) + \sum_{j\in J\setminus J^\star} \beta y_j \leq \Delta. \tag{2.10}$$

Il est important de mentionner ici qu'au niveau des variables d'affectation, seules les variables qui sont toujours considérées et qui prennent une valeur de 0 dans la solution de référence sont

comptabilisées ici afin d'éviter le double comptage. En effet, comme un client doit nécessairement être affecté à un dépôt, si la variable d'affectation d'un client et d'un dépôt particulier passe de 1 à 0, nécessairement une autre variable d'affectation passera de 0 à 1. Seul le deuxième changement est comptabilisé ici. De plus, la contrainte de changement admissible ne comptabilise pas l'affectation de nouveaux clients, c'est-à-dire les clients tels que $i \in I \setminus \bar{I}$.

Tel que discuté précédemment, dans certains cas, la contrainte de changement admissible peut devenir trop rigide ou encore, pour diverses raisons, il peut être nécessaire de pénaliser la différence entre la nouvelle solution et la solution de référence. La pénalité attribuée à la différence entre les deux solutions pourra naturellement varier selon le type d'application. Dans le contexte qui nous intéresse, nous souhaitons proposer une pénalité générale qui pourra s'appliquer au problème de localisation perturbé, mais aussi à d'autres types de problèmes où la réoptimisation contrôlée pourra être justifiée. La différence pondérée entre les décisions de localisation et d'affectation déjà établies et les nouvelles décisions de localisation et d'affectation sera pénalisée ici, dans le même esprit que la contrainte de changement admissible proposée précédemment. Ainsi, en définissant ξ, la pénalité associée à la modification de la solution de référence, toujours choisie par l'utilisateur, la nouvelle fonction objectif peut s'écrire de la façon suivante :

$$\sum_{(i,j)\in A} C_{ij}x_{ij} + \sum_{j\in J} F_j y_j + \xi[\sum_{(i,j)\in(A\cap\bar{A}\setminus A^\star)\cup A^+} \alpha x_{ij} + \sum_{j\in J^\star\cap J} \beta(1-y_j) + \sum_{j\in J\setminus J^\star} \beta y_j] = \tag{2.11}$$

$$\sum_{(i,j)\in A} C_{ij}x_{ij} + \sum_{(i,j)\in(A\cap\bar{A}\setminus A^\star)\cup A^+} \xi\alpha x_{ij} + \sum_{j\in J} F_j y_j + \xi\beta|J^\star\cap J| - \sum_{j\in J^\star\cap J} \xi\beta y_j + \sum_{j\in J\setminus J^\star} \xi\beta y_j =$$

$$\xi\beta|J^\star\cap J| + \sum_{(i,j)\in(A^\star\cap A)\cup(A\setminus A^+\cup\bar{A})} C_{ij}x_{ij} + \sum_{(i,j)\in(A\cap\bar{A}\setminus A^\star)\cup A^+} (C_{ij}+\xi\alpha)x_{ij} + \sum_{j\in J^\star\cap J} (F_j-\xi\beta)y_j + \sum_{j\in J\setminus J^\star} (F_j+\xi\beta)y_j.$$

En posant, c_{ij} égal à C_{ij}, $\forall (i,j) \in (A^\star\cap A)\cup(A\setminus A^+\cup\bar{A})$ et à $C_{ij}+\alpha\xi$, $\forall (i,j) \in (A\cap\bar{A}\setminus A^\star)\cup A^+$ et en posant f_j égal à $F_j-\xi\beta$, $\forall j \in J^\star\cap J$ et à $F_j+\xi\beta$, $\forall j \in J\setminus J^\star$, on peut réécrire la fonction objectif de la façon suivante :

$$\xi\beta|J^\star\cap J| + \sum_{j\in J}\sum_{i\in I} c_{ij}x_{ij} + \sum_{j\in J} f_j y_j \tag{2.12}$$

Ainsi, en réécrivant la fonction objectif pour prendre en compte la pénalité proposée, la nouvelle fonction objectif présente la même forme que (2.1) à une constante près, et en considérant des

coûts fixes et variables modifiés. Ceci demeure vrai lorsque le terme de pénalité est linéaire.
En intégrant la contrainte de changement admissible et le terme de pénalité proposés, le problème de localisation avec capacité et affectation unique perturbé se formule de la façon suivante :

(RP_c)

$$z_{RP_c} = \min \sum_{j\in J}\sum_{i\in I} c_{ij}x_{ij} + \sum_{j\in J} f_j y_j \tag{2.13}$$

sous les contraintes :

$$\sum_{j\in J} x_{ij} = 1,\ \forall i \in I, \tag{2.14}$$

$$\sum_{i\in I} d_i x_{ij} \leq b_j y_j,\ \forall j \in J, \tag{2.15}$$

$$\sum_{(i,j)\in(A\cap\overline{A}\setminus A^\star)\cup A^+} \alpha x_{ij} + \sum_{j\in J^\star\cap J} \beta(1-y_j) + \sum_{j\in J\setminus J^\star} \beta y_j \leq \Delta, \tag{2.16}$$

$$x_{ij} \in \{0,1\},\ \forall i \in I,\ \forall j \in J, \tag{2.17}$$

$$y_j \in \{0,1\},\ \forall j \in J. \tag{2.18}$$

2.3.2 Formulation de type partitionnement

Tout comme la formulation du problème sous-jacent, il est possible de proposer une formulation alternative, basée sur le problème de partitionnement, pour les problèmes de localisation perturbés. Ainsi, soit h_l, la contribution de la colonne l à la contrainte de changement admissible, c'est-à-dire la différence entre la colonne l et l'ensemble des colonnes appartenant à la solution de référence $S^\star$ et en considérant toujours ξ, la pénalité associée à la différence entre la solution du problème perturbé et la solution de référence, le problème de localisation avec capacité et affectation unique perturbé se formule maintenant de la manière suivante :

(RP_p)

$$z_{RP_p} = \min \sum_{l\in R} C_l x_l + \xi \sum_{l\in R} h_l x_l \tag{2.19}$$

sous les contraintes :

$$\sum_{l\in R_k} x_l = 1, \forall k \in K, \tag{2.20}$$

$$\sum_{l\in R} h_l x_l \leq \Delta, \tag{2.21}$$

$$x_l \in \{0,1\},\ \forall l \in R. \tag{2.22}$$

K représente ici l'union de l'ensemble des clients et des dépôts, $K = I \cup J$. Il est possible de constater que la formulation de RP_p est légèrement différente de la formulation équivalente proposée pour le problème de localisation P_p. Naturellement, elle intègre maintenant les mécanismes propres à la réoptimisation contrôlée, mais elle considère aussi une contrainte d'égalité pour la sélection des dépôts plutôt qu'une inégalité comme c'était le cas dans la formulation présentée en 2.2.2. De cette manière, il est possible de considérer de manière efficace l'impact de la fermeture des dépôts. Ainsi, en générant des colonnes correspondant aux dépôts inutilisés, c'est-à-dire des colonnes correspondant à un dépôt et à un ensemble vide de clients, en posant les valeurs de h_l et C_l de façon appropriée et en forçant le système à sélectionner exactement une fois chaque dépôt, l'impact de la non-utilisation d'un dépôt peut être considéré adéquatement. De plus, en posant $c_l = C_l + \xi h_l$, où C_l est calculé par $F_{j_l} + \sum_{i \in I} C_{i j_l} a_l^i$, on peut réécrire $z_{RP_p} = \min \sum_{l \in R} c_l x_l$. La fonction objectif présente alors la même forme que la fonction objectif originale (2.6), mais avec des coûts associés aux colonnes modifiées.

2.3.2.1 Calcul de h_l

Le calcul du paramètre h_l, paramètre visant à mesurer la différence entre deux solutions demeure une étape importante lors de la formulation d'un problème basé sur la réoptimisation contrôlée. Bien que la façon de calculer h_l puisse dépendre du contexte d'application, elle ne modifie en rien la formulation du problème de réoptimisation en soi. Dans le cas du problème de localisation perturbé étudié ici, nous avons vu précédemment que la différence entre deux solutions correspond à la différence pondérée entre les décisions de localisation et d'affectation déjà établies et les nouvelles décisions de localisation et d'affectation. Ainsi, afin de tenir compte de la hiérarchie entre les deux types de décision, des poids β et α sont toujours associés à la modification du statut d'un dépôt et à la modification de l'affectation d'un client. De la même manière que ce qui a été décrit pour la formulation classique, les valeurs de β, α et Δ sont sélectionnées par l'utilisateur. Deux cas sont alors possible pour le calcul de h_l. Ces deux cas sont détaillés ici.

Cas 1 : Le dépôt j_l, dépôt associé à la colonne l, appartient à $J^\star$, l'ensemble de dépôts utilisés dans la solution de référence $S^\star$.

$$h_l = \alpha \sum_{i \in (I \cap \bar{I}) \setminus P} a_l^i, \tag{2.23}$$

où a_l^i est un paramètre qui vaut 1 si le client i est inclus dans la colonne l, et 0 autrement, et P,

l'ensemble des clients tels que $a_p^i = 1$ où a_p^i vaut 1 si le client i est inclus dans la colonne p, et 0 autrement. Dans le contexte ici, la colonne p représente la colonne associée au dépôt j_l dans la solution de référence $S^\star$. P représente donc l'ensemble de clients servis par le dépôt j_l dans la solution de référence $S^\star$.

Cas 2 : Le dépôt j_l, dépôt associé à la colonne l, n'appartient pas à $J^\star$, l'ensemble de dépôts utilisés dans la solution référence $S^\star$.

$$h_l = \beta + \alpha \sum_{i \in I \cap \bar{I}} a_l^i. \tag{2.24}$$

2.4 Approche de résolution

La formulation du problème de localisation perturbé basée sur le partitionnement telle qu'elle a été exposée à la section 2.3.2 présente, à notre avis, certains avantages par rapport la formulation classique présentée à la section 2.3.1. Tout d'abord, la façon de calculer h_l peut être adaptée au contexte d'application sans pour autant modifier la formulation du problème de réoptimisation en soi. L'expression de la contrainte de changement admissible (2.21) est donc très flexible. De plus, la formulation de type partitionnement est générale. En effet, elle permet de considérer des problèmes différents du problème étudié ici en modifiant la définition des colonnes. Elle s'adapte donc facilement à différents problèmes où la réoptimisation contrôlée pourra être justifiée. Pour ces raisons, nous avons choisi de développer une approche basée sur la formulation de type partitionnement afin de résoudre le problème considéré. Le problème de localisation perturbé, tel qu'il a été formulé en 2.3.2, comporte toutefois un très grand nombre de variables. La génération de toutes les colonnes réalisables constitue en soi une tâche extrêmement difficile, même pour des problèmes de taille relativement petite. Nous proposons donc une méthode qui se base sur la génération de colonnes afin de le résoudre.

À notre avis, la génération de colonnes se prête bien à la réoptimisation contrôlée. Tout d'abord, elle permet de considérer l'information recueillie lors de la résolution du problème de référence ou tirée de la solution de référence elle-même. En effet, les colonnes appartenant à la solution de référence et qui sont toujours réalisables pour le problème perturbé pourront être intégrées à l'ensemble de colonnes initial. Il en est de même pour un certain nombre de colonnes qui auraient pu être générées lors de la résolution du problème de référence. De plus, l'ajout de la contrainte de changement admissible nous porte à croire qu'un plus petit nombre de colonnes seront générées, puisqu'elle restreint l'espace des solutions réalisables. Enfin, différentes méthodes basées sur la génération de colonnes ont été proposées pour la résolution du problème de localisation sous-jacent (Neebe et Rao, 1983; Diaz et Fernandez, 2002; Ceselli *et al.*, 2009).

Cette section présente donc certains aspects théoriques reliés au développement de l'approche de résolution proposée. Elle s'intéresse d'abord à la résolution du problème maître puis à la présentation du sous-problème. L'approche développée pour la résolution du problème de localisation perturbé sera ensuite présentée plus formellement de même que les différents mécanismes mis en place afin d'améliorer la qualité de la solution finale.

2.4.1 Résolution du problème maître

Une approche de résolution basée sur la génération de colonnes comporte deux étapes principales : la résolution du problème maître et la résolution du sous-problème, c'est-à-dire la génération de nouvelles colonnes. La résolution du problème maître permet de déterminer les variables duales associées au problème. Ces variables duales servent à formuler le sous-problème qui permettra de déterminer une ou plusieurs nouvelles colonnes de coût réduit négatif. Lorsqu'il n'y a plus de colonnes de coût réduit négatif, le problème dans son ensemble est résolu à l'optimalité (Gilmore et Gomory, 1961).

Le problème de localisation perturbé considéré ici a été formulé comme un problème de partitionnement. Le problème maître consiste donc en la résolution d'un problème de partitionnement. Récemment, Boschetti *et al*. (2008) ont proposé une méthode d'ascension duale ou *dual ascent* en anglais afin de résoudre le problème de partitionnement classique en alliant relaxation paramétrique et relaxation lagrangienne pour la détermination des variables duales et d'une borne inférieure. Grâce aux différentes expérimentations effectuées dans Boschetti *et al*. (2008), les auteurs ont constaté que cette méthode était rapide et stable. La méthode d'ascension duale proposée par Boschetti *et al*. (2008) nous paraît donc une approche intéressante pour l'approximation des variables duales utilisées lors de la génération des colonnes.

Le problème de localisation perturbé se distingue toutefois du problème de partionnement classique puisqu'il considère la contrainte de changement admissible, apport majeur de notre approche. Le fait de considérer cette contrainte a nécessairement un impact sur la méthode d'ascension duale. En effet, cela entraîne l'apparition d'une nouvelle variable duale qu'il faudra approximer, puis montrer que cette approximation permet bien de déterminer une solution duale réalisable pour le problème considéré. Il est donc important de revenir sur certains aspects théoriques reliés à la méthode d'ascension duale et à son adaptation afin de considérer la contrainte de changement admissible.

Tout d'abord, Boschetti *et al*. (2008) proposent le changement de variable suivant afin de formuler une relaxation paramétrique du problème considéré : $x_l = \sum_{k \in N_l} \frac{q_k}{q(N_l)} y_l^k$ où N_l est l'ensemble des éléments (dépôt et clients) appartenant à la colonne l, q_k est un paramètre donné et $q(N_l) = \sum_{k \in R_l} q_k$, le poids associé à la colonne $l \in R$. Ainsi, en appliquant ce changement de variables au problème *RP*, nous obtenons la formulation suivante :

($RP(q)$)

$$z_{RP(q)} = \min \sum_{l \in R} c_l \sum_{i \in N_l} \frac{q_i}{q(N_l)} y_l^i \tag{2.25}$$

sous les contraintes :

$$\sum_{l \in R_k} \sum_{i \in N_l} \frac{q_i}{q(N_l)} y_l^i = 1, \forall k \in K, \tag{2.26}$$

$$\sum_{l \in R} h_l \sum_{i \in N_l} \frac{q_i}{q(N_l)} y_l^i \leq \Delta, \tag{2.27}$$

$$\sum_{l \in R_k} y_l^k = 1, \forall k \in K, \tag{2.28}$$

$$y_l^k \in \{0,1\},\ \forall k \in I,\ \forall l \in R. \tag{2.29}$$

Boschetti *et al*. (2008) proposent ensuite de relaxer de manière lagrangienne les contraintes (2.26). Dans le cas qui nous intéresse, les contraintes (2.26) et (2.27) sont relaxées simultanément donnant lieu au problème suivant :

($LRP(q,\mu,\lambda)$)

$$z_{LRP(q,\mu,\lambda)} = \min \sum_{l \in R} c_l \sum_{i \in N_l} \frac{q_i}{q(N_l)} y_l^i + \mu(\Delta - \sum_{l \in R} h_l \sum_{i \in N_l} \frac{q_i}{q(N_l)} y_l^i) + \sum_{k \in K} \lambda_k (1 - \sum_{l \in R_k} \sum_{i \in N_l} \frac{q_i}{q(N_l)} y_l^i) \tag{2.30}$$

$$z_{LRP(q,\mu,\lambda)} = \min \sum_{k \in K} [\sum_{l \in R_k} (c_l - \mu h_l - \lambda(N_l)) \frac{q_k}{q(N_l)} y_l^k] + \mu\Delta, \tag{2.31}$$

$$\sum_{l \in R_k} y_l^k = 1, \forall k \in K, \tag{2.32}$$

$$y_l^k \in \{0,1\},\ \forall k \in K,\ \forall l \in R. \tag{2.33}$$

où λ_k et μ représentent les multiplicateurs de lagrange associés respectivement aux contraintes (2.26) et (2.27), $\lambda_k \in \mathbb{R}$ et $\mu \leq 0$ et où $\lambda(N_l) = \sum_{k \in N_l} \lambda_k$.

Il est possible de constater que la structure du problème $LRP(q,\mu,\lambda)$ est similaire à la structure du problème $LRP(q,\lambda)$ étudié par Boschetti *et al*. (2008). Le problème $LRP(q,\mu,\lambda)$ est toujours décomposable en $|K|$ problèmes, un pour chaque client et un pour chaque dépôt. Chaque problème peut alors être résolu par inspection. Ainsi, en considérant l_k, l'index de la colonne

couvrant le client ou le dépôt k telle que :

$$\frac{q_k(c_{l_k} - \mu h_{l_k} - \lambda(N_{l_k}))}{q(N_{l_k})} = \min_{l \in R_k} \frac{q_k(c_l - \mu h_l - \lambda(N_l))}{q(N_l)}, \tag{2.34}$$

la solution optimale du problème associé au client ou au dépôt k est $y^k_{l_k} = 1$ et $y^k_l = 0,\ \forall l \in R_k \setminus \{l_k\}$. Le coût de la solution optimale est alors donné par :

$$z'_{LRP(\lambda,\mu,q)} = \sum_{k \in K} [\frac{q_k(c_{l_k} - \mu h_{l_k} - \lambda(N_{l_k}))}{q(N_{l_k})} + \lambda_k], \tag{2.35}$$

$$z_{LRP(\lambda,\mu,q)} = \mu\Delta + z'_{LRP_{rel}(\lambda,\mu,q)}. \tag{2.36}$$

Le théorème 1 de Boschetti *et al*. (2008) montre que toute solution optimale du problème $LRP(\lambda,q)$ fournit une solution réalisable pour le problème dual original de coût $z_D = z^\star_{LRP(\lambda,q)}$, en posant $u_k = \frac{q_k(c_{l_k} - \mu h_{l_k} - \lambda(N_{l_k}))}{q(N_{l_k})} + \lambda_k$ où u_k est la variable duale associée à la contrainte (2.20). Nous allons donc postuler le même théorème, mais en considérant maintenant le problème $LRP(\lambda,\mu,q)$ et la nouvelle variable w.

Théorème 1. *Toute solution $LRP(\lambda,\mu,q)$, pour $\lambda \in R^{(|I|\times|J|)}$, μ et $q > 0$ fournit une solution réalisable pour le problème dual de coût $z_D = z^\star_{LRP(\lambda,\mu,q)}$ en posant les variables duales selon les expressions suivantes :*

$$u_k = \frac{q_k(c_{l_k} - \mu h_{l_k} - \lambda(N_{l_k}))}{q(N_{l_k})} + \lambda_k,\ \forall k \in K, \tag{2.37}$$

$$w = \mu. \tag{2.38}$$

Preuve 1. *On va d'abord montrer qu'en posant $u_k = \frac{q_k(c_{l_k} - \mu h_{l_k} - \lambda(N_{l_k}))}{q(N_{l_k})} + \lambda_k$ et $w = \mu$, $z_D = z^\star_{LRP(\lambda,\mu,q)}$.*

$$z_D = \sum_{k \in K} u_k + \Delta w = \sum_{k \in K} [\frac{q_k(c_{l_k} - \mu h_{l_k} - \lambda(N_{l_k}))}{q(N_{l_k})} + \lambda_k] + \Delta\mu = z^\star_{LRP(\lambda,\mu,q)} \tag{2.39}$$

Il faut maintenant montrer qu'en posant ainsi u_k et w, on obtient une solution réalisable pour le problème dual correspondant. Cette preuve suit la preuve présentée par Boschetti et al. (2008). En considérant la contrainte du problème dual associée à chaque colonne l, comme $l \in R_k\ \forall k \in N_l$, l'inégalité suivante tient toujours :

$$q_k(c_{l_k} - \mu h_{l_k} - \lambda(N_{l_k}))/q(N_{l_k}) \leq q_k(c_l - \mu h_l - \lambda(N_l))/q(N_l),\ \forall k \in N_l. \tag{2.40}$$

Alors :

$$u_k \leq [q_k(c_l - \mu h_l - \lambda(N_l))/q(N_l)] + \lambda_k,\ \forall k \in N_l, \tag{2.41}$$

$$\sum_{k \in N_l} u_k \leq \sum_{k \in N_l} [q_k(c_l - \mu h_l - \lambda(N_l))/q(N_l)] + \sum_{k \in N_l} \lambda_k,\ \forall l \in R, \tag{2.42}$$

$$\sum_{k \in N_l} u_k \leq [(c_l - \mu h_l - \lambda(N_l))/q(N_l)] \sum_{k \in N_l} q_k + \sum_{k \in N_l} \lambda_k,\ \forall l \in R, \tag{2.43}$$

$$\sum_{k \in N_l} u_k \leq [(c_l - \mu h_l - \lambda(N_l))/q(N_l)]q(N_l) + \lambda(N_l),\ \forall l \in R, \tag{2.44}$$

$$\sum_{k \in N_l} u_k \leq c_l - \mu h_l,\ \forall l \in R. \tag{2.45}$$

Il faut montrer que $\sum_{k \in N_l} u_k + h_l w \leq c_l,\ \forall l \in R$. *On a déjà montré que* $\sum_{k \in N_l} u_k \leq c_l - \mu h_l,\ \forall l \in R$, *alors :*

$$\sum_{k \in N_l} u_k + h_l w \leq c_l - \mu h_l + h_l w,\ \forall l \in R. \tag{2.46}$$

En posant $w = \mu$ *:*

$$\sum_{k \in N_l} u_k + h_l \mu \leq c_l - \mu h_l + h_l \mu,\ \forall l \in R. \tag{2.47}$$

Alors :

$$\sum_{k \in N_l} u_k + h_l \mu = \sum_{k \in N_l} u_k + h_l w \leq c_l,\ \forall l \in R. \tag{2.48}$$

La méthode d'ascension duale proposée par Boschetti *et al*. (2008) et adaptée pour considérer la contrainte de changement admissible pourra donc être utilisée afin de résoudre le problème maître relaxé. De cette manière, une borne inférieure de même qu'une approximation des variables duales pour le PLCAUP pourront être déterminées. Ces variables duales pourront ensuite être utilisées lors de la formulation du sous-problème au sein du processus de génération de colonnes.

2.4.2 Sous-problème : Résolution de $|J|$ problèmes de sac à dos

Une colonne est définie comme un ensemble de clients affectés à un dépôt. Le coût d'une colonne se calcule donc en additionnant le coût d'utilisation du dépôt associé à la colonne et les coûts d'affectation des clients couverts par cette même colonne :

$$c_l = f_{j_l} + \sum_{i \in I} c_{i j_l} a_l^i, \tag{2.49}$$

où j_l représente le dépôt associé à la colonne l et a_l^i, un paramètre qui vaut 1 si le client i est couvert par la colonne l, et 0 autrement.

En considérant u_k et w, les variables duales associées respectivement aux contraintes (2.20) et (2.21), le coût réduit d'une colonne se calcule de la façon suivante :

$$\hat{c}_l = c_l - \sum_{i \in I} u_i a_l^i - u_{j_l} - w h_l = f_{j_l} - u_{j_l} + \sum_{i \in I} (c_{i j_l} - u_i) a_l^i - w h_l. \tag{2.50}$$

Le sous-problème consiste donc à générer des colonnes de coût réduit négatif qui respectent les contraintes de capacité du dépôt auquel elles sont associées, c'est-à-dire à déterminer les valeurs de a_l^i en fonction des variables duales courantes de manière à respecter la capacité du dépôt, et ce, pour chaque dépôt. Toutefois, comme h_l varie en fonction de a_l^i, il est nécessaire d'amener quelques précisions quant au calcul de h_l et de $\hat{c}_l$ pour les deux cas présentés précédemment. De plus, selon le cas, une constante $\xi\beta$ devra être ajoutée lors du calcul de $\hat{c}_l$.

Cas 1 : Le dépôt j_l appartient à l'ensemble de dépôts de la solution de référence $J^\star$. Dans ce cas, comme $h_l = \alpha \sum_{i \in I \setminus P} a_l^i$, le coût réduit peut s'écrire de la manière suivante :

$$\hat{c}_l = f_{j_l} + \xi\beta - u_{j_l} + \sum_{i \in I} (c_{i j_l} - u_i) a_l^i - w(\alpha \sum_{i \in I \setminus P} a_l^i), \tag{2.51}$$

$$\hat{c}_l = f_{j_l} + \xi\beta - u_{j_l} + \sum_{i \in (I \setminus \bar{I}) \cup (P \cap I)} (c_{i j_l} - u_i) a_l^i + \sum_{i \in (I \cap \bar{I}) \setminus P} (c_{i j_l} - u_i - \alpha w) a_l^i. \tag{2.52}$$

Le terme $\xi\beta$ apparaît ici car la colonne appartient à $S^\star$. En posant, $\phi_i = c_{i j_l} - u_i$, $\forall i \in (I \setminus \bar{I}) \cup (P \cap I)$ et $\phi_i = c_{i j_l} - u_i - \alpha w$, $\forall i \in (I \cap \bar{I}) \setminus P$, l'expression du coût réduit se réécrit de la manière suivante :

$$\hat{c}_l = f_{j_l} + \xi\beta - u_{j_l} + \sum_{i \in I} \phi_i a_l^i. \tag{2.53}$$

Il est alors possible de constater qu'afin de générer une nouvelle colonne et de déterminer son coût réduit, il suffit de minimiser les coûts associés à la sélection des clients en considérant la capacité du dépôt, ce qui équivaut à la résolution d'un problème de sac à dos pour un dépôt donné.

Cas 2 : Le dépôt j_l n'appartient pas à l'ensemble de dépôts de la solution de référence $J^\star$. Dans ce cas, comme $h_l = \beta + \alpha \sum_{i \in I} a_l^i$, le coût réduit s'écrit de la manière suivante :

$$\hat{c}_l = f_{j_l} - u_{j_l} + \sum_{i \in I} (c_{i j_l} - u_i) a_l^i - w(\alpha \sum_{i \in I \cap \bar{I}} a_l^i + \beta), \tag{2.54}$$

$$\hat{c}_l = f_{j_l} - u_{j_l} - w\beta + \sum_{i \in I} \phi_i a_l^i, \tag{2.55}$$

où $\phi_i = c_{i j_l} - u_i$, $\forall i \in I \setminus \bar{I}$ et $\phi_i = c_{i j_l} - u_i - w\alpha$, $\forall i \in I \cap \bar{I}$. Le coût réduit d'une colonne l

présente, à quelques constantes près, la même forme que dans le *Cas 1*.
Le sous-problème consiste donc à résoudre, à chaque itération, $|J|$ problèmes de sac à dos en considérant les variables duales obtenues lors de la résolution du problème maître. Le problème de sac à dos suivant est alors résolu pour chaque dépôt :

(SP_j)

$$z_{sp_j} = \min \sum_{i \in I} \phi_i a_l^i \tag{2.56}$$

sous les contraintes :

$$\sum_{i \in I} d_i a_l^i \leq b_j, \tag{2.57}$$

$$a_l^i \in \{0,1\}. \tag{2.58}$$

Les valeurs de ϕ_i sont calculées selon le cas. Ainsi, en considérant $z_{sp_{j_l}}$, la solution du problème de sac à dos associé au dépôt j_l, le coût réduit d'une colonne l se calcule de la manière suivante :

$$\hat{c}_l = \begin{cases} f_{j_l} + \xi\beta - u_{j_l} + z_{sp_{j_l}}, & \text{si } j_l \in J^\star \\ f_{j_l} - u_{j_l} - w\beta + z_{sp_{j_l}}, & \text{si } j_l \notin J^\star \end{cases}. \tag{2.59}$$

À chaque itération, $|J|$ colonnes sont générées suite à la résolution du problème SP_j grâce à un algorithme de programmation dynamique développé par Pisinger (1997). Le coût réduit de chaque colonne est ensuite calculé (2.59) et seules les colonnes de coût réduit négatif sont conservées et ajoutées à l'ensemble de colonnes existant. Afin d'accélérer le processus de génération de colonnes, il pourrait être possible d'ajouter plus de $|J|$ colonnes par itération en sélectionnant des colonnes de coût réduit supérieures aux colonnes générées. Le *multiple pricing* n'est toutefois pas utilisé ici.

2.4.3 Algorithme proposé

Tous les éléments théoriques et conceptuels nécessaires à la présentation de l'approche heuristique proposée pour la résolution du problème de localisation pertubé ont été exposés précédemment. Il est alors possible de présenter l'algorithme de résolution en soi. Plus formellement, l'algorithme de résolution comporte cinq phases principales :

- **Phase 1 :** Génération d'un ensemble de colonnes initiales
- **Phase 2 :** Génération d'une borne supérieure initiale
- **Phase 3 :** Génération de colonnes par ascension duale (*dual ascent*)

- **Phase 4 :** Génération de colonnes avec CPLEX
- **Phase 5 :** Génération d'une borne supérieure finale

La phase 1 permet de générer un ensemble de colonnes initiales afin de démarrer la recherche d'une solution réalisable et l'algorithme de génération de colonnes en soi. La phase 2 est ensuite lancée. Cette phase vise à déterminer une première borne supérieure pour le problème étudié. Cette borne est nécessaire au bon fonctionnement de la Phase 3. La phase 3 constitue le coeur de la méthode. Elle permet de déterminer un ensemble de colonnes intéressantes par la résolution de différents sous-problèmes formulés à partir valeurs duales déterminées grâce à la méthode d'ascension duale. Puisque les valeurs duales sont approximatives, la phase 4 est lancée de manière à compléter l'ensemble de colonnes par la résolution de différents sous-problèmes formulés à partir de vraies valeurs duales trouvées grâce à CPLEX. La phase 4 assure ainsi la génération de bonnes colonnes qui n'auraient pu être générées à la phase 3. Enfin, si la solution du problème maître relaxé n'est pas entière, une solution entière réalisable est déterminée en résolvant le problème maître entier grâce à CPLEX. L'ensemble de colonnes générées aux phases précédentes sera utilisé lors de la résolution du problème entier à la phase 5. La solution de ce problème devient alors la borne supérieure finale. Ces cinq phases sont présentées plus en détail dans les sous-sections qui suivent.

2.4.3.1 PHASE 1 : Génération d'un ensemble de colonnes initiales

La phase de génération de colonnes initiales vise, comme son nom l'indique, à générer un ensemble de colonnes initiales afin de démarrer la recherche d'une solution réalisable et l'algorithme de génération de colonnes en soi. Ces colonnes initiales sont générées de quatre façons. Tout d'abord, $|J|$ colonnes sont générées grâce à la résolution du problème de sac à dos suivant : (PSD_{j_l})

$$z_{PSD_{j_l}} = \max \sum_{i \in I} (M - c_{ij}) a_l^i \tag{2.60}$$

sous les contraintes :

$$\sum_{i \in I} d_i a_l^i \leq b_{j_l}, \tag{2.61}$$

$$a_l^i \in \{0,1\}, \tag{2.62}$$

où $M = \max_i c_{ij_l}$. Un tel problème de sac à dos est donc résolu pour chaque dépôt potentiel, toujours en utilisant l'algorithme de Pisinger (1997). Chaque colonne générée grâce à la résolution du PSD_{j_l} comporte le dépôt j_l considéré et l'ensemble des clients pour lesquels $a_l^i = 1$. En observant les $|J|$ colonnes ainsi générées, il est possible de constater que certains clients

peuvent demeurer non-couverts par l'ensemble de colonnes initiales, ce qui n'est pas souhaitable. Afin de remédier à cette situation, $|I| \times |J|$ colonnes triviales sont également générées, une pour chaque couple (client, dépôt). Ces colonnes visent à garantir que chaque client soit couvert par au moins une colonne. Dans le cas où une solution de référence est prise en compte, les colonnes de la solution de référence, et qui sont encore réalisables, c'est-à-dire qui respectent les contraintes de capacité, sont aussi ajoutées à l'ensemble de colonnes initiales. Enfin, les colonnes correspondant au fait de ne pas utiliser un dépôt sont également générées. On construira alors $|J|$ colonnes incluant le dépôt considéré et un ensemble vide de clients. Une fois les colonnes initiales générées, leur coût et leur déviation sont calculés.

PHASE 1 : Génération d'un ensemble de colonnes initiales

Poser $\hat{R} = \emptyset$.

- **Générer les colonnes par résolution du** PSD_{j_l}

  Pour chaque dépôt j**, générer une colonne** l :

  Résoudre PSD_{j_l} ;
  Générer la colonne formée de j_l et des clients tels que $a_l^i = 1$.

  Considérer R_1^0, l'ensemble des colonnes ainsi générées.

- **Générer les colonnes triviales**

  Pour chaque client i **faire** :

  Pour chaque dépôt j **faire** :
  Générer la colonne formée du dépôt j et du client i.

  Considérer R_2^0, l'ensemble des colonnes ainsi générées.

- **Générer les colonnes à partir de la solution de référence**

  Pour chaque colonne l appartenant à la solution de référence **faire** :

  Si la contrainte de capacité est respectée,
  Conserver la colonne l.

  Considérer R_3^0, l'ensemble des colonnes de la solution de référence conservées.

- **Générer les colonnes correspondant aux dépôts inutilisés**

  Pour chaque dépôt j **faire** :

Générer la colonne formée du dépôt j_l et pour laquelle $a_l^i = 0,\ \forall i \in I$.

Considérer R_4^0, l'ensemble des colonnes correspondant aux dépôts inutilisés.

Poser $\hat{R} = R_1^0 \cup R_2^0 \cup R_3^0 \cup R_4^0$.
Calculer le coût et la déviation des colonnes $l \in \hat{R}$.

Passer à la PHASE 2.

Différentes méthodes pourraient être imaginées afin de générer les colonnes initiales. De plus, si le problème de référence a été résolu par génération de colonnes, il pourrait être intéressant d'ajouter certaines colonnes visitées lors de la résolution de ce problème et qui semble intéressante pour le problème de localisation perturbé. Néanmoins, pour le moment, nous nous en tiendrons aux quatre types de colonnes initiales présentés précédemment.

2.4.3.2 PHASE 2 : Génération d'une borne supérieure initiale

Une solution réalisable pour le problème *RP* doit également être déterminée, si une telle solution existe. Cette borne supérieure sera utilisée lors de la mise à jour des multiplicateurs lagrangiens lors de la phase de la génération de colonnes par ascencion duale. Deux heuristiques ont été proposées ici afin de déterminer une borne supérieure.
La première méthode proposée est une heuristique gloutonne visant à minimiser l'ensemble des coûts d'utilisation des dépôts et d'affectation des clients. Dans ce cas, la contrainte de changement admissible est relaxée. En considérant I', l'ensemble des clients non affectés et J', l'ensemble des dépôts non utilisés, cette heuristique fonctionne de la manière suivante :

Heuristique de minimisation des coûts

$I' = I$, $J' = J$, $\hat{I} = \emptyset$ et $z_{BS_{\text{coût}}} = 0$.
Tant que $I' \neq \emptyset$, **faire** :

Sélectionner le dépôt $j \in J'$ tel que $j = \arg\min \frac{f_j}{b_j}$
$capacite_{residuelle} = b_j$, $J' = J'/\{j\}$, $\hat{I} = I'$
$z_{BS_{\text{coût}}} = z_{BS_{\text{coût}}} + f_j$.

Tant que $capacite_{residuelle} > 0$ et que $\hat{I} \neq \emptyset$ **faire** :

Sélectionner le client $i \in \hat{I}$ tel que $i = \arg\min c_{ij}$;

Si $d_i \leq capacite_{residuelle}$:

Le client i est couvert par le dépôt j, $I' = I'/\{i\}$.

$$capacite_{residuelle} = capacite_{residuelle} - d_i$$
$$z_{BS_{\text{coût}}} = z_{BS_{\text{coût}}} + c_{ij}.$$
$$\hat{I} = \hat{I} / \{i\}.$$

Comme la contrainte de changement admissible est relaxée, il est possible que la solution ainsi générée ne soit pas réalisable pour le problème *RP*.

La deuxième heuristique proposée se base plutôt sur la minimisation du changement entre la solution de référence et la nouvelle solution. Ainsi, plutôt que de résoudre directement le problème *RP*, le problème suivant sera résolu :

(BS_Δ)

$$\min \sum_{l \in R} h_l x_l \tag{2.63}$$

sous les contraintes :

$$\sum_{l \in R_k} x_l = 1, \forall k \in K, \tag{2.64}$$

$$x_l \in \{0,1\}. \tag{2.65}$$

Ce problème vise donc à minimiser le changement admissible tout en respectant les contraintes du problème orginal. Une fois le changement minimal admissible déterminé, le coût z_{BS_Δ} associé à cette solution est calculé. En plus de fournir une solution réalisable au problème, si cette solution existe, la résolution du problème BS_Δ permet de déterminer une borne inférieure sur la valeur de Δ. Néanmoins, la solution déterminée grâce à cette méthode peut être de piètre qualité au niveau du coût.

Le problème BS_Δ est résolu ici par génération de colonnes. Ainsi, la relaxation linéaire du problème maître est résolue grâce à CPLEX afin de déterminer les variables duales. Le sous-problème consistera toujours en un problème de sac à dos. Toutefois, dans ce cas, les coefficients associés à chaque client sont les suivants : si $j_l \in J^\star$, alors $\phi_i = -u_i$, $\forall i \in (I \setminus \bar{I}) \cup (P \cap I)$ et $\phi_i = \alpha - u_i$, $\forall i \in (I \cap \bar{I}) \setminus P$, et si $j_l \notin J^\star$, $\phi_i = -u_i$, $\forall i \in I \setminus \bar{I}$ et $\phi_i = \alpha - u_i$, $\forall i \in I \cap \bar{I}$. Ainsi, à partir de la solution du problème SP_j correspondant, le coût réduit d'une colonne se calcule alors comme suit :

$$\hat{h}_l = \begin{cases} z_{sp_{j_l}} - u_{j_l}, & \text{si } j_l \in J^\star \\ z_{sp_{j_l}} - u_{j_l} + \beta, & \text{si } j_l \notin J^\star \end{cases}. \tag{2.66}$$

Une fois le processus de génération de colonnes terminé, une solution entière est déterminée en résolvant le problème maître entier avec CPLEX, en considérant l'ensemble de colonnes gé-

nérées précédemment. La solution déterminée n'est pas nécessairement optimale, mais permet de fournir une borne supérieure sur la valeur de Δ nécessaire afin de retrouver une solution réalisable. De plus, la solution de la relaxation linéaire du problème BS_Δ permet d'en déterminer une borne inférieure. Dans le cas où la valeur Δ fixée est plus petit que la solution de la relaxation linéaire du problème BS_Δ l'utilisateur permet de trouver, le problème *RP* ne possède pas de solution réalisable. L'algorithme s'arrête alors et l'utilisateur est invité à modifier la valeur de Δ si c'est possible de le faire. L'algorithme redémarre ensuite à nouveau en prenant en compte la nouvelle valeur de Δ.
La meilleure solution réalisable fournie par ces deux heuristiques est alors utilisée comme borne supérieure pour le problème *RP*.

PHASE 2 : Génération d'une solution réalisable

Poser $Vmin_{\text{coût}} = \infty$ et $z_{BS} = \infty$.

- **Déterminer une solution en utilisant l'heuristique de minimisation des coûts**

  Si la solution est réalisable **faire** :

  Soit $z_{BS_{\text{coût}}}$ le coût de la solution ainsi déterminée, $Vmin_{\text{coût}} = z_{BS_{\text{coût}}}$;
  Soit R_5, les colonnes correspondant à cette solution, $\hat{R} = \hat{R} \cup R_5$.

- **Déterminer une solution en utilisant l'heuristique de minimisation du changement admissible**

  Jusqu'à ce que $R' = \emptyset$ **faire** :

  Résoudre la relaxation linéaire du problème BS_Δ grâce à CPLEX

  Soit z_{BS_Δ}, la solution obtenue et u_k, les variables duales correspondantes :

  Pour chaque dépôt j **faire** :

  Résoudre le problème de sac à dos correspondant SP_j ;

  Calculer le coût réduit de chaque colonne générée $\hat{h}_l$.

  Déterminer R', l'ensemble des nouvelles colonnes de coût réduit négatif ;
  Poser $\hat{R} = \hat{R} \cup R'$.

  Résoudre le problème BS_Δ avec CPLEX.

$z_{BS} = \min(Vmin_{\text{coût}}, z_{BS_\Delta})$.

Passer à la PHASE 3.

2.4.3.3 PHASE 3 : Génération de colonnes par ascension duale (*dual ascent*)

La génération de colonnes par ascension duale vise à générer un ensemble de colonnes intéressantes pour le problème *RP*. Elle pourra, ou non, inclure différentes phases de diversification de manière à améliorer la qualité de la solution trouvée. Dans ce qui suit, la phase de génération de colonnes par ascension duale de même que les deux phases de diversification proposées sont décrites plus en détail. Les paramètres et variables nécessaires à l'implémentation de ces différentes phases sont résumés au Tableau 2.1.

Phase principale de génération de colonnes par ascension duale (PHASE 3)	
ε^0	Valeur initiale du pas pour la mise à jour des multiplicateurs de lagrange fixée par l'utilisateur
ε	Valeur du pas pour la mise à jour des multiplicateurs de lagrange
$MaxItMacro$	Nombre maximal d'itérations de génération de colonnes par ascension duale fixé par l'utilisateur
$MaxItMicro$	Nombre maximal d'itérations lors de la recherche d'une solution duale fixé par l'utilisateur
$NbItSa$	Nombre d'itérations sans amélioration de la solution duale
$NbItSa_{max}$	Nombre maximal d'itérations sans amélioration de la solution duale fixé par l'utilisateur
$NbChgt$	Nombre de changements du pas (ε) effectué dans le processus de recherche
$NbChgt_{max}$	Nombre maximal de changements du pas (ε) permis avant sa réinitialisation fixé par l'utilisateur
κ	Facteur de réduction du pas (ε) fixé par l'utilisateur
Phase de diversification par modification de la valeur de Δ (PHASE 3.0)	
$\Delta_{recherche}$	Valeur de Δ utilisé lors du processus de recherche d'une solution duale
$Multi_{\varepsilon_{D0}}$	Valeur du multiplicateur permettant de calculer la valeur du pas pour la phase 3.0
$MultiMacro_{D0}$	Valeur du multiplicateur permettant de déterminer le nombre maximal d'itérations de génération de colonnes par ascension duale pour la phase 3.0 fixée par l'utilisateur
$MultiMicro_{D0}$	Valeur du multiplicateur permettant de déterminer le nombre maximal d'itérations lors de la recherche d'une solution duale pour la phase 3.0 fixée par l'utilisateur
$MaxItMacro_{D0}$	Nombre maximal d'itérations de génération de colonnes par ascension duale pour la phase 3.0
$MaxItMicro_{D0}$	Nombre maximal d'itérations lors de la recherche d'une solution duale pour la phase 3.0
$NbChgt_{\Delta}$	Nombre de fois où la valeur de $\Delta_{recherche}$ a été modifiée pendant le processus de recherche
$NbChgtTot_{\Delta}$	Nombre total de fois où la valeur de $\Delta_{recherche}$ devra être modifiée pendant processus de recherche fixé par l'utilisateur
$NbIt_{\Delta}$	Nombre d'itérations pour lesquelles la valeur de $\Delta_{recherche}$ demeure la même
Phase de diversification par modification de la valeur de ε (PHASE 3.1)	
$Multi_{\varepsilon_{D1}}$	Valeur du multiplicateur permettant de calculer la valeur du pas pour la phase 3.0
$MultiMacro_{D1}$	Valeur du multiplicateur permettant de déterminer le nombre maximal d'itérations de génération de colonnes par ascension duale pour la phase 3.1 fixée par l'utilisateur
$MultiMicro_{D1}$	Valeur du multiplicateur permettant de déterminer le nombre maximal d'itérations lors de la recherche d'une solution duale pour la phase 3.1 fixée par l'utilisateur
$MaxItMacro_{D1}$	Nombre maximal d'itérations de génération de colonnes par ascension duale pour la phase 3.1
$MaxItMicro_{D1}$	Nombre maximal d'itérations lors de la recherche d'une solution duale pour la phase 3.1

Tableau 2.1 – Paramètres nécessaires pour la génération de colonnes par ascension duale

Tout d'abord, la génération de colonnes par ascension duale se divise elle-même en trois étapes : l'initialisation, la recherche d'une bonne solution duale et la génération de colonnes. La phase d'initialisation consiste, comme son nom l'indique, à fixer les valeurs initiales de q_k, λ_k et μ. Tout d'abord, il a été montré que la valeur de q_k n'influence pas la qualité de la solution finale déterminée (Boschetti *et al*., 2008). Pour cette raison, les valeurs de $q_k, \forall k \in I \cup J$, sont initialisées respectivement à la distance moyenne d'un client k aux dépôts et à la distance moyenne d'un dépôt k aux clients. Les multiplicateurs de lagrange, μ et $\lambda_k, \forall k \in K$, de même que la borne inférieure LB sont, quant à eux, initialisés à 0.

Ensuite, pour un nombre donné d'itérations *MaxItMacro*, l'algorithme itère entre la recherche d'une bonne solution duale et la génération de colonnes. Afin de déterminer une bonne solution duale, deux étapes sont effectuées successivement : tout d'abord, une solution duale est calculée à partir des multiplicateurs de lagrange initiaux ou déterminés à l'itération précédente, puis les multiplicateurs de lagrange sont mis à jour. Ces deux étapes sont effectuées pour un nombre donné d'itérations *MaxItMicro* ou jusqu'à ce que $\sum_k \theta_k^2 = 0$ et que $\psi^2 \geq 0$ où θ_k et ψ mesurent les violations des contraintes dualisées dans l'objectif et calculées par : $\theta_k = 1 - \sum_{l \in R_k} \sum_{k \in R_l} \frac{q_k}{q(R_l)} y_l^k$ et $\psi = \Delta - \sum_{l \in R} b_l \sum_{i \in R_l} \frac{q_k}{q(R_l)} y_l^k$. Enfin, les meilleures valeurs duales trouvées lors de la recherche d'une bonne solution duale sont utilisées afin de générer de nouvelles colonnes. Si aucune colonne de coût réduit négatif ne peut être générée, une borne inférieure valide est obtenue. L'algorithme itère ainsi entre la recherche d'une bonne solution duale et la génération de colonnes jusqu'à ce que *MaxItMacro* soit atteint ou qu'une borne valide puisse être déterminée.

PHASE 3 : Génération de colonnes par ascension duale (dual ascent)

Initialisation

Poser $\lambda_k = 0, \forall k \in K$, $q_k = \frac{\sum_{j \in J} c_{kj}}{|J|}, \forall k \in I$, $q_k = \frac{\sum_{i \in I} c_{ik}}{|I|}, \forall k \in J$, $\mu = 0$, $\varepsilon = \varepsilon^0$, $NbItSa = 0$, $NbChgt = 0$, $z^\star_{L\hat{R}P(q,\mu,\lambda)} = 0$ et $z_{BI} = 0$.

Pour *MaxItMacro* **faire** :

Recherche de la meilleure solution duale :

En considérant q_k, λ_k, μ et $\hat{R}$,

Pour *MaxItMicro* ou **jusqu'à ce que** $\sum_k \theta_k^2 = 0$ et $\psi^2 \geq 0$ **faire** :

- Déterminer la solution du problème $L\hat{R}P(q, \mu, \lambda)$, $z_{L\hat{R}P(q,\mu,\lambda)}$, par inspection (2.34,

2.36)

et en déduire les variables duales u_k et w (2.37).

- Vérifier la solution obtenue $z_{L\hat{R}P(q,\mu,\lambda)}$.

  Si $z_{L\hat{R}P(q,\mu,\lambda)} > z^{\star}_{L\hat{R}P(q,\mu,\lambda)}$ **alors** :

  $z^{\star}_{L\hat{R}P(q,\mu,\lambda)} = z_{L\hat{R}P(q,\mu,\lambda)}$ et $NbItSa = 0$.

  Sinon :

  $NbItSa = NbItSa + 1$.

  Si $NbItSa \geq NbItSa_{max}$ **alors** :

  $\varepsilon = \varepsilon/\kappa$, $NbChgt = NbChgt + 1$ et $NbItSa = 0$.

  Si $NbChgt > NbChgt_{max}$ **alors** :

  $\varepsilon = \varepsilon^0$ et $NbChgt = 0$.

- Mettre à jour les multiplicateurs de lagrange :

  $\lambda_k = \lambda_k + \varepsilon \frac{zz}{\sum_i \theta_i^2 + \psi^2} \theta_k$, $\mu = \mu + \varepsilon \frac{zz}{\sum_k \theta_k^2 + \psi^2} \psi$ où $zz = z_{BS} - z_{L\hat{R}P(q,\mu,\lambda)}$.

$z^{\star}_D = z^{\star}_{L\hat{R}P(q,\mu,\lambda)}$.

Génération de colonnes

Soit $z^{\star}_D$, la meilleure solution duale obtenue après $MaxItMicro$ ou lorsque le critère d'arrêt est atteint et $u^{\star}_k$ et $w^{\star}$, les variables duales correspondantes :

- Générer les nouvelles colonnes :

  Pour chaque dépôt j **faire** :

  Résoudre le problème de sac à dos correspondant SP_j ;

  Calculer le coût réduit de chaque colonne générée $\hat{c}_l$.

- Vérifier si la solution duale est réalisable (borne inférieure valide) :

  Soit $R^{'}$, l'ensemble des nouvelles colonnes de coût réduit négatif,

  Si $R^{'} = \emptyset$, la solution duale est réalisable :

  Si $z^{\star}_D > z_{BI}$ **poser** $z_{BI} = z^{\star}_D$.

  Si $R^{'} \neq \emptyset$, la solution duale n'est pas réalisable :

  Poser $\hat{R} = \hat{R} \cup R^{'}$.
  $\varepsilon = \varepsilon^0, NbItSa = 0, NbChgt = 0$ et $z^{\star}_{L\hat{R}P(q,\mu,\lambda)} = 0$.

Si $z_{BI} = 0$, aucune borne valide n'a été trouvée :

Poser $MaxItMacro$=$MaxItMacro$ + 25.

Passer à la PHASE 4.

La génération de colonnes par ascension duale permet de déterminer une bonne borne inférieure au problème étudié. Elle générera donc un ensemble de colonnes afin d'atteindre cet objectif. Néanmoins, l'obtention d'un ensemble de colonnes plus riche et diversifié peut permettre d'améliorer la qualité de la borne supérieure finale (voir PHASE 5). Deux mécanismes ont alors été proposés afin d'améliorer la diversité de l'ensemble de colonnes $\hat{R}$. Le premier mécanisme vise à modifier la valeur de Δ fixée par l'utilisateur lors de la recherche d'une bonne solution duale. En perturbant ainsi la valeur du membre de droite de la contrainte de changement admissible, on souhaite visiter un ensemble de solutions à la limite de la frontière réalisable, dans l'espace des solutions réalisables et non-réalisables. Le second mécanisme consiste plutôt à travailler avec une valeur de ε différente de celle fixée initialement afin de tirer profit de l'approximation des valeurs duales et ainsi, générer des colonnes qui ne pourraient être générées autrement et qui pourraient amener de la diversité à l'ensemble de colonnes $\hat{R}$. Des détails supplémentaires au niveau de la justification de ces deux phases de diversification sont présentés à la section 2.5.2. Ces deux phases, si elles sont implantées, seront lancées avant la génération de colonnes en soi (PHASE 3). De plus, elles ne contribueront pas à la mise à jour de la borne inférieure LB.

PHASE 3.0 : Diversification par modification de la valeur de Δ

Initialisation

Poser $\lambda_k = 0$, $q_k = \frac{\sum_{j \in J} c_{kj}}{|J|}$ $\forall k \in I$, $q_k = \frac{\sum_{i \in I} c_{ik}}{|I|}$ $\forall k \in J$, $\mu = 0$, $z^{\star}_{L\hat{R}P(q,\mu,\lambda)} = 0$,

$\varepsilon = Multi_{\varepsilon_{D0}} * \Delta$, $\Delta_{recherche} = UNIF[0,8\Delta; 1,8\Delta]$,

$MaxItMacro_{D0} = MultiMacro_{D0} * MaxItMacro$,

$MaxItMicro_{D0} = MultiMicro_{D0} * MaxItMicro$,

$NbIt_{\Delta} = \frac{MaxItMacro_{D0}}{NbChgtTot_{\Delta}+1}$ et $NbChgt_{\Delta} = 1$.

Pour $MaxItMacro_{D0}$ **faire** :

Recherche de la meilleure solution duale :

En considérant q_k, λ_k, μ et $\hat{R}$,

Pour $MaxItMicro_{D0}$ ou **jusqu'à ce que** $\sum_k \theta_k^2 = 0$ et $\psi^2 \geq 0$ **faire** :

• Déterminer la solution du problème $L\hat{R}P(q,\mu,\lambda)$, $z_{L\hat{R}P(q,\mu,\lambda)}$, par inspection (2.34, 2.36)

et en déduire les variables duales u_k et w (2.37).

• Vérifier la solution obtenue $z_{L\hat{R}P(q,\mu,\lambda)}$.

Si $z_{L\hat{R}P(q,\mu,\lambda)} > z^{\star}_{L\hat{R}P(q,\mu,\lambda)}$ **alors** :

$z^{\star}_{L\hat{R}P(q,\mu,\lambda)} = z_{L\hat{R}P(q,\mu,\lambda)}$.

• Mettre à jour les multiplicateurs de lagrange :

$\lambda_k = \lambda_k + \varepsilon \frac{zz}{\sum_i \theta_i^2 + \psi^2} \theta_k$, $\mu = \mu + \varepsilon \frac{zz}{\sum_k \theta_k^2 + \psi^2} \psi$

où $\psi = \Delta_{recherche} - \sum_{l \in R} b_l \sum_{i \in N_l} \frac{q_k}{q(N_l)} y_l^k$ et $zz = z_{BS} - z_{L\hat{R}P(q,\mu,\lambda)}$.

$z^{\star}_D = z^{\star}_{L\hat{R}P(q,\mu,\lambda)}$.

Génération de colonnes

Soit $z^{\star}_D$, la meilleure solution duale obtenue après $MaxItMicro_{D0}$ ou lorsque le critère d'arrêt est atteint et $u_k^{\star}$ et $w^{\star}$, les variables duales correspondantes :

• Générer les nouvelles colonnes :

Pour chaque dépôt j **faire** :

Résoudre le problème de sac à dos correspondant SP_j ;

Calculer le coût réduit de chaque colonne générée $\hat{c}_l$.

Soit R', l'ensemble des nouvelles colonnes de coût réduit négatif,

Si $R' \neq \emptyset$:

Poser $\hat{R} = \hat{R} \cup R'$.

$z^{\star}_{L\hat{R}P(q,\mu,\lambda)} = 0$.

Modification de $\Delta_{recherche}$

Si le numéro de l'itération en cours est égal à $NbIt_{\Delta} \times NbChgt_{\Delta}$ **faire :**

$\Delta_{recherche} = UNIF[0,8\Delta; 1,6\Delta]$, $NbChgt_{\Delta} = NbChgt_{\Delta} + 1$ et $z^{\star}_{L\hat{R}P(q,\mu,\lambda)} = 0$

Passer à la PHASE 3.

PHASE 3.1 : Diversification par modification de la valeur de ε

Initialisation

Poser $\lambda_k = 0$, $q_k = \frac{\sum_{j\in J} c_{kj}}{|J|}\ \forall k \in I$, $q_k = \frac{\sum_{i\in I} c_{ik}}{|I|}\ \forall k \in J$, $\mu = 0$, $z^{\star}_{L\hat{R}P(q,\mu,\lambda)} = 0$,

$\varepsilon = Multi_{\varepsilon_{D1}} * \Delta$, $MaxItMacro_{D1} = MultiMacro_{D1} * MaxItMacro$

et $MaxItMicro_{D1} = MultiMicro_{D1} * MaxItMicro$.

Pour $MaxItMacro_{D1}$ **faire** :

Recherche de la meilleure solution duale :

En considérant q_k, λ_k, μ et $\hat{R}$,

Pour $MaxItMicro_{D1}$ ou **jusqu'à ce que** $\sum_k \theta_k^2 = 0$ et $\psi^2 \geq 0$ **faire** :

- Déterminer la solution du problème $L\hat{R}P(q,\mu,\lambda)$, $z_{L\hat{R}P(q,\mu,\lambda)}$, par inspection (2.34, 2.36) et en déduire les variables duales u_k et w (2.37).
- Vérifier la solution obtenue $z_{L\hat{R}P(q,\mu,\lambda)}$.

  Si $z_{L\hat{R}P(q,\mu,\lambda)} > z^{\star}_{L\hat{R}P(q,\mu,\lambda)}$ **alors** :

  $z^{\star}_{L\hat{R}P(q,\mu,\lambda)} = z_{L\hat{R}P(q,\mu,\lambda)}$.
- Mettre à jour les multiplicateurs de lagrange :

  $\lambda_k = \lambda_k + \varepsilon \frac{zz}{\sum_i \theta_i^2 + \psi^2} \theta_k$, $\mu = \mu + \varepsilon \frac{zz}{\sum_k \theta_k^2 + \psi^2} \psi$ où $zz = z_{BS} - z_{L\hat{R}P(q,\mu,\lambda)}$.

$z^{\star}_D = z^{\star}_{L\hat{R}P(q,\mu,\lambda)}$.

Génération de colonnes

Soit $z^{\star}_D$, la meilleure solution duale obtenue après $MaxItMicro_{D1}$ ou lorsque le critère d'arrêt est atteint et $u^{\star}_k$ et $w^{\star}$, les variables duales correspondantes :

- Générer les nouvelles colonnes :

  Pour chaque dépôt j **faire** :

  Résoudre le problème de sac à dos correspondant SP_j ;

  Calculer le coût réduit de chaque colonne générée $\hat{c}_l$.

  Soit R', l'ensemble des nouvelles colonnes de coût réduit négatif,

  Si $R' \neq \emptyset$:

  Poser $\hat{R} = \hat{R} \cup R'$.

$$z^{\star}_{L\hat{R}P(q,\mu,\lambda)} = 0$$

Passer à la PHASE 3.

2.4.3.4 PHASE 4 : Génération de colonnes grâce à CPLEX

La méthode d'ascension duale fournit une approximation des variables duales associées au problème *RP*. Le fait d'utiliser ainsi une approximation peut permettre de générer un certain nombre de colonnes intéressantes, du point du vue du problème RP, qui ne pourraient être générées par un processus de génération de colonnes exact. Néanmoins, il est aussi possible que certaines colonnes intéressantes ne soient pas comprises dans l'ensemble de colonnes déterminé au cours de la PHASE 3. Afin de remédier à cette situation, une seconde phase de génération de colonnes peut être initialisée et résolue grâce à CPLEX. La phase de génération de colonnes par CPLEX visera alors à résoudre la relaxation linéaire du problème *RP* à partir de l'ensemble de colonnes $\hat{R}$ disponible une fois la PHASE 3 terminée. Ainsi, il sera possible d'obtenir les variables duales exactes et de vérifier si de nouvelles colonnes de coût réduit négatif peuvent toujours être générées. Si tel est le cas, un nouveau processus de génération de colonnes est entamé. Sinon, la phase de génération de colonnes prend fin. Il est important de mentionner que cette phase de génération de colonnes avec CPLEX n'est pas nécessaire à la détermination d'une borne inférieure ou supérieure, mais pourrait contribuer à leur amélioration.

PHASE 4 : Génération de colonnes grâce à CPLEX

Considérer $\hat{R}$, l'ensemble de colonnes obtenu une fois la PHASE 3 terminée.

Jusqu'à ce que $R^{'} = \emptyset$ **faire** :

Résoudre la relaxation linéaire du problème $\hat{RP}$ **grâce à CPLEX**

Soit $z_{\hat{RP}}$, la solution obtenue et u_k et w, les variables duales correspondantes :

Pour chaque dépôt j **faire** :

Résoudre le problème de sac à dos correspondant SP_j ;

Calculer le coût réduit de chaque colonne générée $\hat{c}_l$.

Déterminer $R^{'}$, l'ensemble des nouvelles colonnes de coût réduit négatif ;

Poser $\hat{R} = \hat{R} \cup R^{'}$.

Passer à la PHASE 5.

2.4.3.5 PHASE 5 : Génération d'une borne supérieure finale

Les quatre étapes précédentes ont permis de déterminer l'ensemble de colonnes nécessaires afin d'obtenir la solution optimale du problème linéaire. L'objectif premier de l'approche proposée consiste toutefois à déterminer une solution réalisable au problème concerné, dans le cas présent, le problème de localisation avec capacité et affectation unique perturbé. Si la solution de la relaxation considérée est entière, aucun problème ne se pose puisqu'on a alors la garantie d'avoir trouvé la solution optimale. Si ce n'est pas le cas, il est nécessaire de déterminer une méthode afin d'obtenir une solution entière et réalisable. Différentes avenues sont alors possibles. Tout d'abord, une solution entière pourrait être obtenue à partir de la solution de la relaxation considérée puis améliorée par recherche locale ou grâce à une métaheuristique par exemple. Dans ce cas, la méthode heuristique développée gagnera à être adaptée au problème étudiée.

Il est également possible d'obtenir une solution entière réalisable en résolvant le problème entier, avec CPLEX par exemple, en utilisant l'ensemble de colonnes obtenu après la PHASE 4. Un problème se pose alors : même si une solution entière optimale pour le problème $\hat{RP}$ est déterminée, il est possible que cette solution ne soit pas optimale pour le problème original RP, c'est-à-dire qu'il manque un certain nombre de colonnes afin de trouver une solution entière optimale. CPLEX fournira alors la meilleure solution possible en considérant l'ensemble de colonnes disponible. Ainsi, en résolvant le problème à la racine, la solution obtenue n'est pas nécessairement optimale. Afin d'obtenir une garantie d'optimalité, il faudrait lancer un schéma de résolution de type *branch-and-price*, tel que celui utilisé pour le développement de méthodes de résolution exactes pour le problème de localisation sous-jacent (Neebe et Rao, 1983; Diaz et Fernandez, 2002; Ceselli *et al.*, 2009), ou encore générer toutes les colonnes manquantes, c'est-à-dire celles dont $\hat{c}_l \leq z_{BS} - z_{BI}$, puis résoudre à nouveau le problème à l'optimalité.

Dans le cas ici, nous nous limiterons à la résolution du problème $\hat{RP}$ à la racine. D'une part, cela nous permettra de conserver la flexibilité et la généralité de l'approche proposée tout en limitant le temps de calcul. Différents mécanismes ont toutefois été mis en place afin d'augmenter la richesse et la diversité de l'ensemble de colonnes déterminées une fois la PHASE 4 terminée et ainsi tenter d'améliorer la qualité de la borne supérieure finale. Néanmoins, nous croyons qu'il pourrait être intéressant de développer différentes méthodes heuristiques propres au contexte étudié afin de déterminer une bonne borne supérieure ou d'améliorer la borne actuelle. De plus, aucun schéma de résolution de type *branch-and-price* ne sera présenté ici. En effet, à notre avis, dans plusieurs contextes où l'incertitude et le niveau de dynamisme est important, il est moins pertinent de résoudre le problème étudié de manière exacte. Résoudre de manière exacte un tel problème peut être coûteux en termes de temps de calcul. Ainsi, si la situation est amenée

à changer fréquemment, la recherche d'une solution optimale peut s'avérer moins intéressante par rapport à la recherche rapide d'une bonne solution. Le développement de méthodes exactes, si justifié, pourra toutefois faire l'objet de recherches futures.

PHASE 5 : Génération d'une borne supérieure finale

Considérer $\hat{R}$, l'ensemble de colonnes obtenu une fois la PHASE 4 terminée.

Poser $TmpsLimite_{CPLEX} = Tmps_{max}$ puis :

Résoudre le problème $\hat{RP}$ grâce à CPLEX

Soit $z_{\hat{RP}}$, la solution obtenue, **poser** $z_{BS} = z_{\hat{RP}}$.

Retourner z_{BS} et z_{LB}.

2.5 Expérimentation

L'expérimentation présentée dans ce chapitre poursuit deux objectifs principaux. D'une part, elle vise à analyser le compromis entre le contrôle de la solution et la qualité de la solution finale obtenue, et ce, pour différents types de perturbations et différentes valeurs de α et β. D'autre part, elle souhaite montrer la flexibilité et la généralité de la méthode, c'est-à-dire montrer que l'approche proposée offre de bonnes performances pour un ensemble de variantes du problème de localisation perturbé.

Afin de mener l'expérimentation, différents groupes d'instances ont été générés à partir des instances du PLCAU décrites dans Delmaire *et al*. (1999) et disponibles en ligne au *http ://www-eio.upc.es/∼elena*. Les instances utilisées par Delmaire *et al*. (1999) ont été divisées en sept groupes selon la taille (n,m) des problème considérés où n représente le nombre de clients et m le nombre de dépôts potentiels. Ainsi, G_1 comporte six problèmes $(20,10)$, G_2 comporte onze problèmes $(30,15)$, G_3 comporte huit problèmes $(40,20)$, G_4 comporte huit problèmes $(50,20)$, G_5 comporte huit problèmes $(60,30)$, G_6 comporte huit problèmes $(75,30)$ et G_7 comporte huit problèmes $(90,30)$. Ces problèmes possèdent les caractéristiques suivantes (voir Delmaire *et al*. (1999) pour plus de détails) :

- **Demande de client (d_i)** : Les demandes des clients sont entières et générées à partir d'une distribution uniforme centrée en $D_{prom} = 20$ et de limite inférieure $D_{min} = 10$.
- **Capacité des dépôts (b_j)** : Les capacités sont entières et générées à partir d'une distribu-

tion uniforme centrée en $B_{prom} = \frac{\sum_{i=1}^{n} d_i}{D_{coef} k}$ et de limite inférieure $B_{min} = D_{prom}$. Dans ce cas, k est généré aléatoirement dans un intervalle $[\lfloor 0,1n \rfloor ; n]$ et D_{coef} est un paramètre donné, fixé à environ 1,01, selon les instances.

- **Coût de transport (C_{ij})** : Les coûts de transport sont entiers et générés à partir d'une distribution uniforme centrée en $C_{prom} = 50$ et de limite inférieure $C_{min} = 0$.

- **Coût fixe d'utilisation des dépôts (F_j)** : Les coûts fixes sont entiers et générés à partir d'une distribution uniforme centrée en $F_{prom} = RB_{prom}$ et de limite inférieure $F_{min} = RB_{min}$ où R est un paramètre donné, compris entre 10 et 100, selon les instances.

Afin de représenter adéquatement le problème de localisation perturbé, les sept groupes d'instances présentés ci-haut ont été modifiés afin de prendre en compte différents types de perturbations. Sept types de perturbations ont été analysés soit l'ajout de dépôts (M0), le retrait de dépôts (M1), l'apparition de nouveaux clients (M2), l'élimination de clients existants (M3), l'augmentation de la quantité demandée (M4), la réduction de la quantité demandée (M5), la variation de la quantité demandée comportant des augmentations et des réductions (M6). En considérant N_p, le niveau de perturbations du problème, exprimé en pourcentage (%), les instances ont été modifiées de la manière suivante :

- **Ajout de dépôts (M0)** : Le nombre de dépôts ajoutés $\lceil N_p m \rceil$ est d'abord calculé puis les coûts fixes, les capacités et les coûts de transport qui leur sont associés sont déterminés tels qu'ils ont été décrits précédemment.

- **Retrait de dépôts (M1)** : Un nombre $\lceil N_p |J^\star| \rceil$ de dépôts sont retirés aléatoirement parmi l'ensemble des dépôts ouverts dans la solution de référence $J^\star$, chaque dépôt ouvert ayant la même probabilité d'être sélectionné.

- **Apparition de nouveaux clients (M2)** : Le nombre de nouveaux clients $\lceil N_p n \rceil$ est d'abord calculé puis les demandes et les coûts de transport qui leur sont associés sont déterminés tels qu'ils ont été décrits précédemment.

- **Élimination de clients existants (M3)** : Un nombre $\lceil N_p n \rceil$ de clients à éliminer sont sélectionnés aléatoirement parmi l'ensemble de clients I, chaque client ayant la même probabilité d'être sélectionné.

- **Augmentation de la quantité demandée (M4)** : $\lceil N_p n \rceil$ de clients dont la quantité demandée va augmenter sont sélectionnés aléatoirement parmi l'ensemble de clients I, chaque client ayant la même probabilité d'être sélectionné. La demande de chaque client

choisi est ensuite fixée à $d_i + d_{aug}$ où d_{aug} est généré aléatoirement dans un intervalle $[0, Prop_{dem} d_i]$, $Prop_{dem}$ étant un paramètre donné, fixé ici à 0,5.

- **Réduction de la quantité demandée (M5)** : $\lceil N_p n \rceil$ de clients dont la quantité demandée va diminuer sont sélectionnés aléatoirement parmi l'ensemble de clients I, chaque client ayant la même probabilité d'être sélectionné. La demande de chaque client choisi est ensuite fixée à $d_i - d_{aug}$ où d_{aug} est généré aléatoirement dans un intervalle $[0, Prop_{dem} d_i]$, $Prop_{dem}$ étant un paramètre donné, fixé ici à 0,5.

- **Variation de la quantité demandée (M6)** : $\lceil N_p n \rceil$ de clients dont la quantité demandée va varier sont sélectionnés aléatoirement parmi l'ensemble de clients I, chaque client ayant la même probabilité d'être sélectionné. On déterminera ensuite si le client sélectionné fait face à une augmentation ou à une réduction, chaque cas ayant une probabilité d'occurence égale à 0.5. Dans le cas d'une augmentation, la demande est fixée à $d_i + d_{aug}$ tandis que dans le cas d'une diminution, la demande est fixée à $d_i - d_{aug}$.

Il est important de mentionner que les différents types de perturbations ont été analysés indépendamment afin d'éliminer les effets croisés et de mieux comprendre ce qui se passe dans les diverses circonstances. De plus, il est important de constater que selon le type et le niveau de perturbations considérés, certaines instances peuvent devenir non-réalisables. Ces informations seront mentionnées tout au long de l'expérimentation.

2.5.1 Analyse du compromis entre le contrôle de la solution et la qualité de la solution finale.

Afin d'analyser le compromis entre le contrôle de la solution et la qualité de la solution finale obtenue, différentes courbes ont été tracées et comparées. Les différentes courbes tracées présentent le pourcentage de déviation d'une solution par rapport à la solution optimale, $\frac{z_\Delta}{z_{opt}} - 1$, exprimé en pourcentage (%), où z_Δ et z_{opt} représentent respectivement, pour le problème considéré, la valeur optimale pour une valeur de Δ donnée et la valeur de la solution optimale sans contrainte de changement admissible. Dans un premier temps, afin de déterminer la valeur de z_{opt}, la formulation classique du problème P est résolu à l'optimalité avec CPLEX. La valeur optimale de Δ correspondant à cette solution, Δ_{opt}, est ensuite déterminée en résolvant un problème similaire où l'objectif consiste à minimiser la valeur de Δ en imposant que les coûts totaux demeurent égaux à z_{opt}. Dans un deuxième temps, le problème consistant à déterminer la valeur minimale du changement, Δ_{min}, afin d'obtenir une solution réalisable pour le problème étudié est résolu à l'optimalité avec CPLEX si la solution de référence n'est pas réalisable.

Le problème RP_c correspondant est ensuite résolu en fixant $\Delta = \Delta_{min}$, afin de déterminer z_{min}. Enfin, un nombre donné de problèmes, un pour chaque valeur de Δ choisie dans l'intervalle $[\Delta_{min}, \Delta_{opt}]$, sont résolus à l'optimalité, toujours avec CPLEX. Les valeurs associées aux points (Δ_{min}, z_{min}) et (Δ_{opt}, z_{opt}) de même qu'à l'ensemble des points intermédiaires sont calculées puis utilisées afin de tracer chacune des courbes présentées ici. Dans ce contexte, la solution de référence utilisée correspond à la solution optimale du problème original, sans perturbation, tel que présenté dans Delmaire *et al*. (1999), à partir duquel le problème perturbé a été défini.

Dans cette section, les courbes présentant la valeur de la solution finale en fonction du contrôle ou des efforts admis de la solution sont d'abord tracées pour les 8 instances appartenant à G_4, et ce, pour $\alpha = 1, \beta = 5, \xi = 0$ et pour tous les types de perturbations. Les courbes obtenues pour les 8 instances appartenant à G_4 pour quatre couples (α, β) sont ensuite présentées. Dans ce cas, le type de perturbations $M6$ et une valeur de $\xi = 0$ sont considérés. Il est important de mentionner que la valeur de ξ, la pénalité associée dans la fonction objectif au changement de la solution, a été fixée à 0 afin de mesurer indépendamment l'impact de la contrainte de changement admissible. Enfin, les courbes de compromis pour différentes valeurs de ξ, en considérant toujours $\alpha = 1$, $\beta = 5$ et les instances appartenant à G_4, sont présentées.

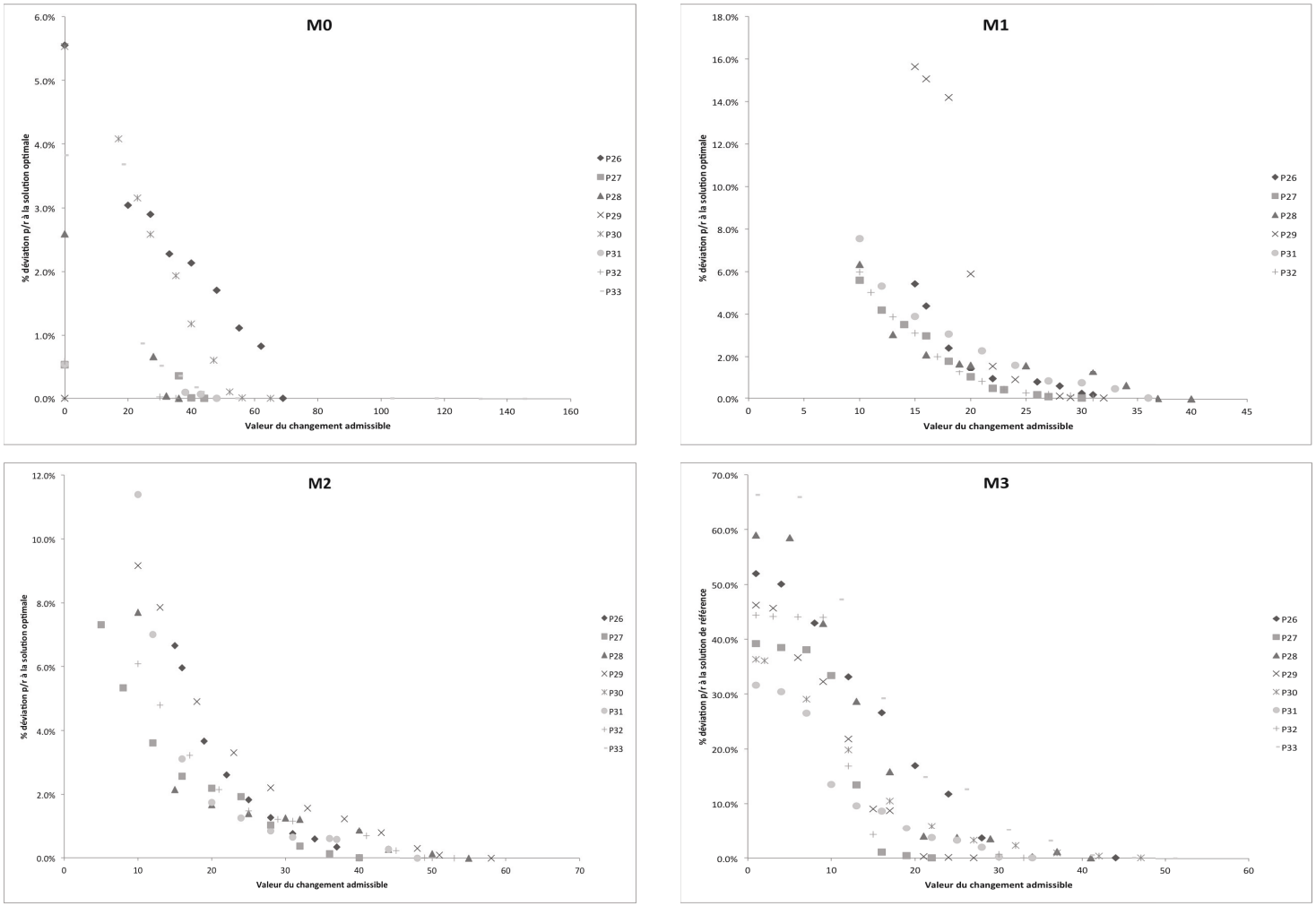
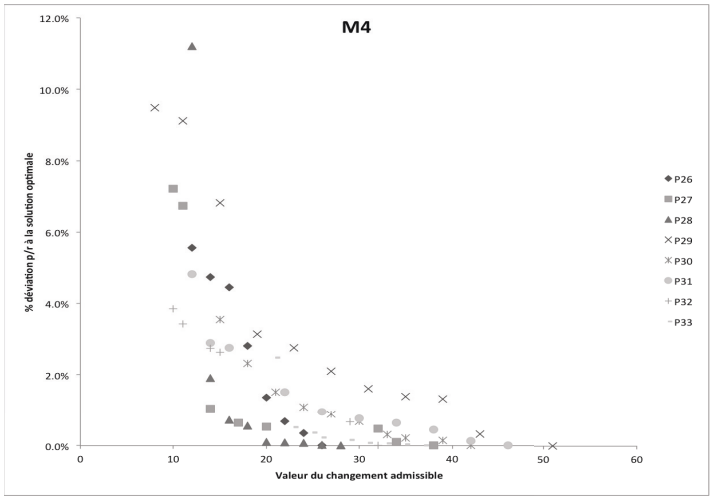
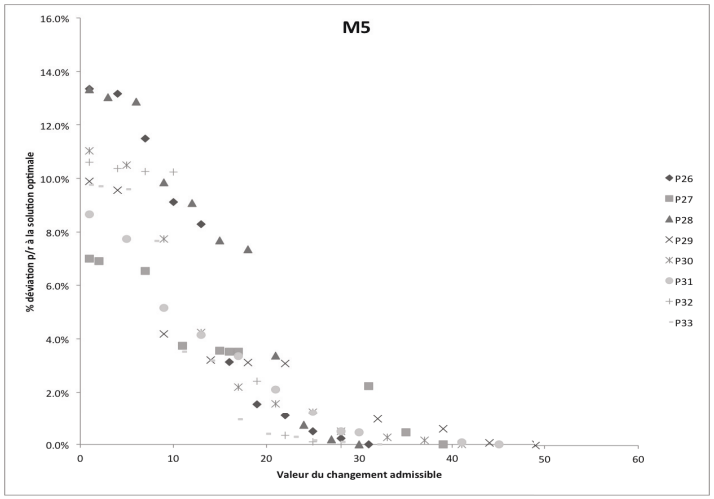


Figure 2.1 – Compromis pour différents types de perturbations avec $\alpha = 1$ et $\beta = 5$ (I)





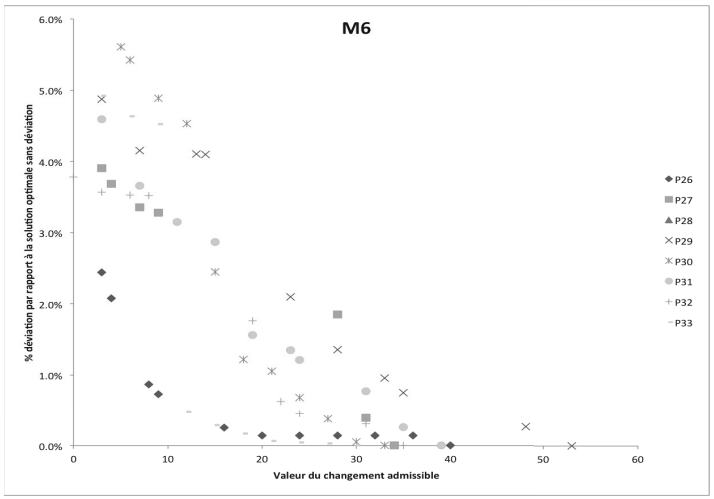


Figure 2.2 – Compromis pour différents types de perturbations avec $\alpha = 1$ et $\beta = 5$ (II)

2.5.1.1 Types de perturbations

En observant les résultas présentés aux figures 2.1 et 2.2, il est possible d'observer que, dans le cas des types de perturbations $M1$, $M2$, $M4$ et $M6$, les solutions de référence ne sont généralement pas réalisables. En effet, dans les deux premiers cas, soit le retrait de dépôts et l'ajout de nouveaux clients, les perturbations peuvent avoir un impact important sur la faisabilité du problème considéré. En effet, pour deux instances ($P30$ et $P33$), les perturbations sont telles que le problème perturbé ne peut admettre de solutions réalisables. Pour les autres instances, il faudra un Δ_{min} variant entre 10 et 15 unités de changement afin de retrouver la faisabilité, ce qui correspond, dans le contexte étudié au changement de statut de 2 ou 3 dépôts, au changement d'affectation de 10 à 15 clients, ou une combinaison des deux. Les solutions correspondantes présentent une déviation par rapport à la solution optimale sans contrainte de changement admissible variant entre 5 et 16 % pour $M1$ et variant entre 6 et 11 % pour $M2$. Pour $M4$ et $M6$, soit l'augmentation et la variation de la quantité demandée, les solutions de référence sont aussi non-réalisables, à l'exception de l'instance $P32$ et du type de perturbation $M6$. Néanmoins, toutes les instances demeurent réalisables lorsque confrontées aux perturbations. Pour $M4$, les valeurs de Δ_{min} se situent entre 8 et 20. Les efforts de changement de la solution de référence afin d'obtenir une solution réalisable sont donc équivalents aux deux premiers cas discutés ici. Les solutions correspondantes varient, pour leur part, entre 2 et 10 % par rapport à la solution optimale sans contrainte de changement admissible. Enfin, dans le cas de la variation de la demande, l'impact des perturbations est moins important. On parle alors de valeurs de Δ_{min} variant entre 0 et 5. Dans ce cas, puisqu'une diminution et une augmentation de la quantité demandée sont considérées simultanément, l'effet de l'augmentation de la demande de certains clients est contrebalancé par la diminution de la demande d'autres clients. Pour $M6$, la déviation des solutions $z_{\Delta_{min}}$ par rapport à z_{opt} variera de 2 à 6 %.

Pour $M1$, $M2$, $M4$ et $M6$, il est possible de constater que la première solution réalisable, c'est-à-dire celle qui correspond à la plus petite valeur de Δ possible, est relativement bonne et que l'amélioration potentielle de cette solution est limitée. En effet, en retirant des dépôts ou en augmentation la demande, l'espace des solutions réalisables pour le problème perturbé est plus restreint que dans le cas du problème de référence. Conséquemment, les choix sont plus limités en termes de sélection de dépôts et d'affectation des clients. En contrepartie, dans le cas des types de perturbations $M0$, $M3$ et $M5$, soit respectivement l'ajout de dépôts, le retrait de clients et la diminution de la quantité demandée, les solutions de référence sont toujours réalisables. Maintenir le *statu quo* peut toutefois mener à une augmentation importante des coûts lorsqu'on les compare aux coûts correspondant à une réoptimisation complète. En effet, il est

possible d'observer que le fait de toujours considérer la solution de référence comme solution au problème perturbé amène une déviation par rapport à la solution optimale sans contrainte de changement admissible variant entre 0 et 5,5 % pour $M0$, entre 30 et 66 % pour $M3$ et entre 7 et 13 % pour $M5$. Ne pas considérer de changement de la solution est donc relativement coûteux, principalement dans le cas $M3$. En effet, dans ce cas, le fait de retirer un certain nombre de clients amène une réduction de la quantité totale demandée et engendre ainsi une réduction du nombre de dépôts nécessaires afin de satisfaire la demande, ce qui peut avoir un impact important sur les coûts totaux. Rappelons ici que les coûts d'utilisation des dépôts sont considérablement plus élevés que les coûts d'affectation des clients aux dépôts ouverts. Cet effet est également présent dans le cas $M5$, mais beaucoup moins marqué dû à une moins grande réduction de la demande : seuls certains clients voient leur demande diminuer. Dans le cas du type de perturbation $M0$, la différence entre la solution de référence et la solution optimale sans considération aux changements admissibles est plus petite. En effet, la capacité et les coûts fixes associés aux nouveaux dépôts potentiels ont été déterminés de manière à représenter la structure actuelle du problème. Les coûts fixes et la capacité de ces nouveaux dépôts sont donc du même ordre que ceux présents dans le problème de référence. De plus, comme il n'y a pas de changement au niveau de la demande, un nombre de dépôts similaire devrait être sélectionné lors d'une réoptimisation complète. Comme les coûts fixes représentent la plus grande part de coûts totaux, les gains possibles sont moins grands. Cela se reflète bien dans l'analyse des figures 2.1 et 2.2. Néanmoins, dans la plupart des cas énoncé ci-haut, il serait intéressant de modifier la solution de référence, bien qu'elle reste réalisable pour le problème perturbé.

Enfin, dans tous les cas, il est possible de constater en observant les courbes présentées aux figures 2.1 et 2.2 que le fait de modifier la première solution réalisable améliore d'abord la solution dans une plus grand proportion, puis cette amélioration diminue au fur et à mesure que la valeur de Δ augmente. Autrement dit, dès que l'on accepte de modifier un peu le plan d'opération, la qualité de la solution s'améliore dans une plus grande proportion. Ce comportement semble plus marqué dans les cas où la solution de référence est toujours réalisable puisque l'espace de solutions réalisables est généralement plus grand.

2.5.1.2 Valeurs de α et β

Différentes valeurs de α et de β ont été utilisées afin d'analyser leur impact sur le compromis entre le contrôle de la solution ou les efforts admis et la qualité de la solution finale. Les valeurs de α et de β choisies sont reportées entre parenthèses sur les figures. Ces valeurs ont été choisies de manière à faire varier l'importance relative accordée aux décisions de sélection des

dépôts et aux décisions d'affectation des clients. Ainsi, en posant $\alpha = 0$ et $\beta = 1$, il est possible de représenter la situation où aucune pénalité n'est attribuée aux changements des décisions d'affectation. À l'opposé, le cas où $\alpha = 1$ et $\beta = 1$ représente un contexte où une importance égale est accordée aux décisions de localisation et d'affectation.

En observant les courbes présentées à la Figure 2.3, il est possible de constater qu'en faisant varier les valeurs de α et de β, les différents constats discutés précédemment sont toujours vrais. En effet, les différences entre les valeurs $z_{\Delta_{min}}$ et z_{opt} sont équivalentes et le comportement des courbes de compromis entre les efforts de changement de la solution et la qualité de la solution finale sont similaires. Les valeurs de α et de β n'ont pas d'impact significatif sur l'allure générale des courbes. Naturellement, les valeurs de Δ_{min} et Δ_{opt} varieront en fonction des valeurs de α et de β fixées.

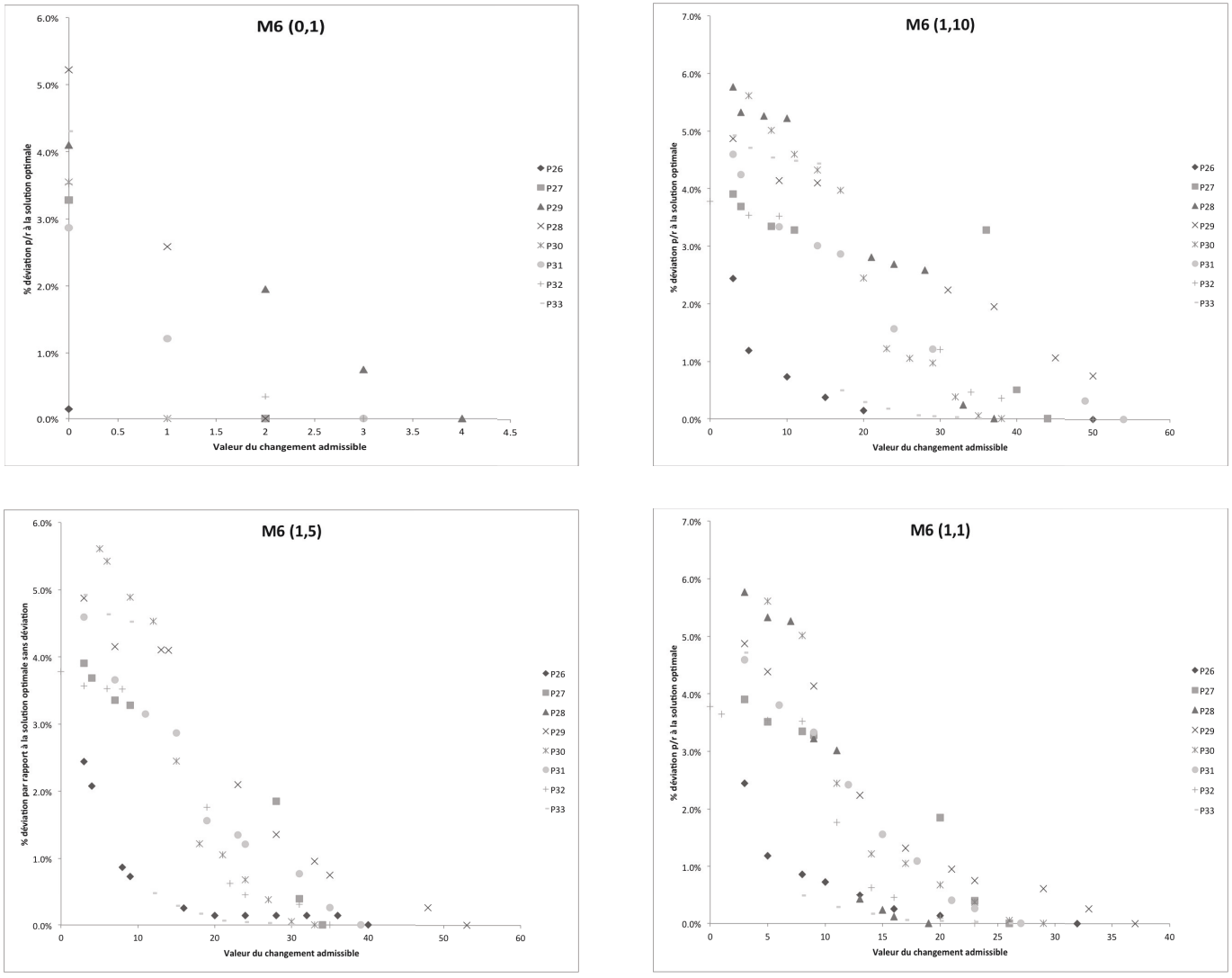


Figure 2.3 – Compromis pour différentes valeurs de (α, β)

2.5.1.3 Valeurs de ξ

Différentes valeurs de ξ ont également été utilisées afin d'analyser leur impact sur le compromis entre le contrôle ou les efforts de changement de la solution et la qualité de la solution finale. Néanmoins, avant de discuter les courbes de compromis en soi, il importe de discuter brièvement la sélection des différentes valeurs de ξ utilisées. En effet, lorsqu'une pénalité est imposée au changement de la solution dans la fonction objectif, c'est-à-dire $\xi > 0$, deux termes non-nuls doivent être additionnés : le terme associé aux coûts réels de la solution, exprimé en unités de coût et le terme associé au changement de la solution, exprimé en unités de changement. Afin de pouvoir additionner ces deux termes sur une base commune, il est judicieux de choisir ξ de manière à représenter un coût par unité de changement. Ici, la valeur de ξ est fixée en considérant les valeurs de Δ_{opt}, Δ_{min}, z_{opt} et z_{min} telles que déterminées précédemment, pour chaque problème traité. Ainsi, en considérant γ, un paramètre représentant l'importance relative accordée aux coûts de changement de la solution par rapport aux coûts originaux, la valeur de ξ est donnée par $\gamma\frac{\Delta_{min}-\Delta_{opt}}{z_{opt}-z_{min}}$. Les valeurs de γ sont reportées entre parenthèses sur les figures. Naturellement, dans un cas réel, la valeur de ξ devra être choisie par l'utilisateur en fonction des données du problème et du contexte étudié. Dans le cas où $\xi > 0$, les deux mécanismes issus de la réoptimisation contrôlée, soit la pénalité dans la fonction objectif et la contrainte de changement admissible, sont pris en compte simultanément.

L'analyse des courbes tracées pour différentes valeurs de γ, et conséquemment de ξ, ont permis d'observer que plus la valeur de ξ est grande, plus la valeur de Δ_{opt}, la valeur de Δ associée à la solution optimale sans contrainte de changement admissible, est petite. En effet, en imposant une pénalité plus importante dans la fonction objectif au changement de la solution par rapport à la solution de référence, les solutions dont les valeurs de Δ correspondantes sont plus petites sont favorisées. Ce phénomène s'observe bien à la Figure 2.4. Conséquemment, puisque la valeur de Δ_{opt} est plus petite, l'écart entre z_{opt} et z_{min} diminue d'où la réduction du pourcentage de déviation des solutions obtenues par rapport à z_{opt}. De plus, comme l'écart entre Δ_{min} et Δ_{opt} est plus petit pour de grandes valeurs de ξ. Il y a nécessairement moins de solutions intermédiaires, d'où le nombre de points restreint lorsque $\gamma = 1$ et $\gamma = 2$. Enfin, en ce qui concerne l'allure générale des courbes, la valeur de ξ ne semble pas avoir d'impact significatif, du moins dans les cas où $\gamma = 0$, $\gamma = 0,5$, et $\gamma = 1$, cas qui disposent de suffisamment de points pour pouvoir apprécier l'allure des courbes.

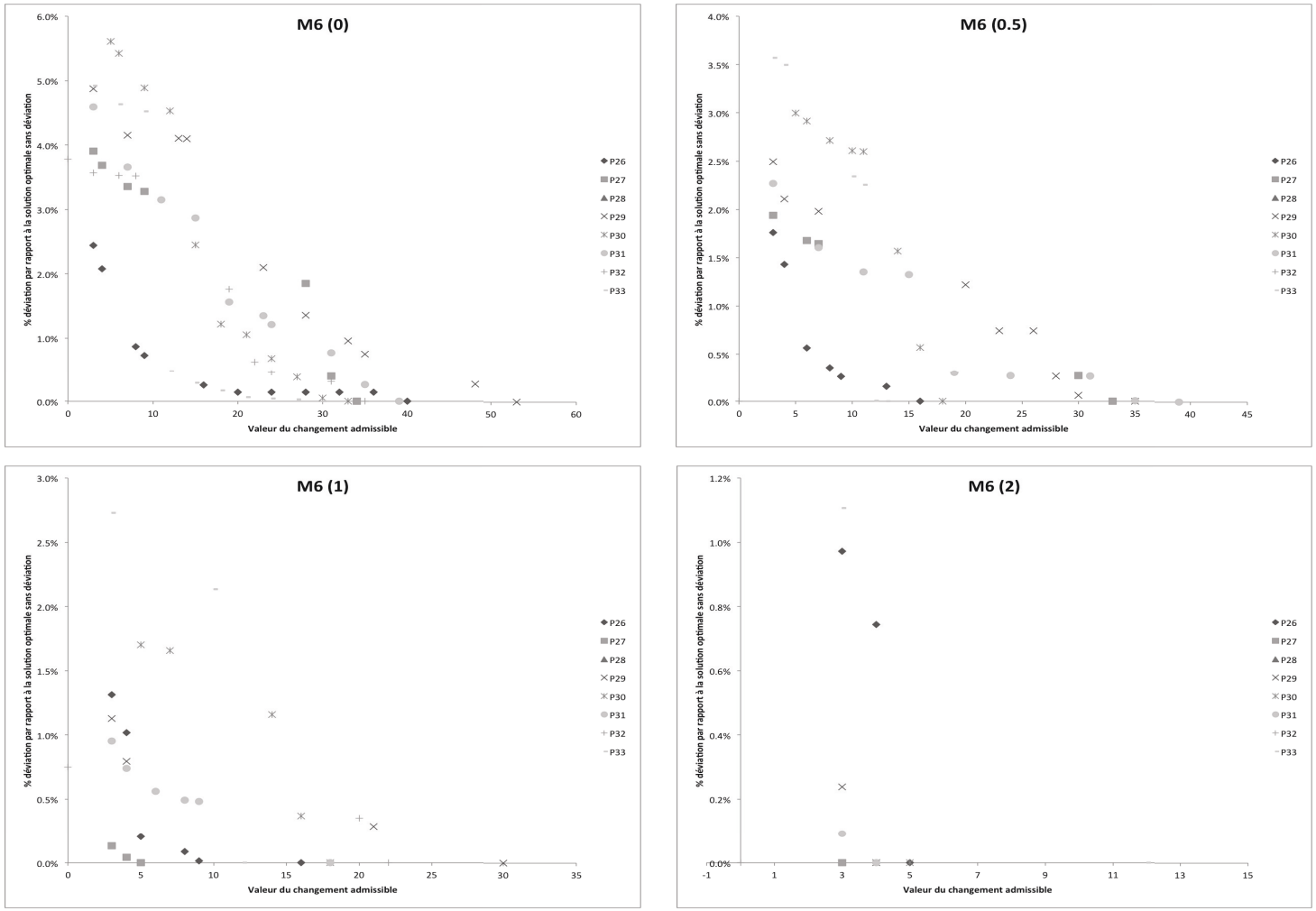


Figure 2.4 – Compromis pour différentes valeurs de ξ (variation du paramètre γ)

2.5.2 Analyse des performances de la méthode proposée

Dans un premier temps, différents tests ont été menés afin de calibrer la méthode proposée et d'en analyser la sensibilité par rapport aux différentes valeurs des paramètres. Cette première étape permettra également de déterminer un ensemble de paramètres standards qui pourront être utilisés lors de la validation finale de la méthode. Ensuite, différents tests ont été effectués afin de montrer que l'approche proposée offre de bonnes performances bien pour un ensemble de variantes du problème de localisation perturbé. Ainsi, différents cas incluant différents types de perturbations, différentes valeurs de α, β, ξ et Δ et différentes tailles de problèmes sont présentés et analysés dans cette section.

Afin de comparer les résultats obtenus, la déviation de la borne inférieure par rapport à la solution optimale du problème concerné, $\frac{z_{BS}}{z^{\star}_{opt}} - 1$, exprimée en %, de même que la déviation de la borne supérieure finale par rapport à la solution optimale du problème concerné, $\frac{z_{BS}}{z^{\star}_{opt}} - 1$, toujours exprimée en %, sont présentées. Ainsi, $z^{\star}_{opt}$ représente la valeur de la solution optimale pour une valeur Δ fixée par l'utilisateur. Les valeurs de $z^{\star}_{opt}$ ont été déterminées par la résolution du problème RP_c équivalent, grâce à CPLEX, telles que déterminées en 2.5.1. Les déviations moyennes ($\%_{moy}$), maximales ($\%_{max}$) et minimales ($\%_{min}$) présentées ont été calculées sur l'ensemble des instances appartenant aux différents groupes analysés lors des tests. L'information concernant les différents groupes d'instances et les paramètres utilisés pour effectuer les tests sera fournie tout au long de la présentation des résultats. Enfin, les temps de calcul, lorsque présentés, sont exprimés en secondes.

En pratique, le type de perturbation est une conséquence du contexte étudié et les valeurs de α, β, ξ et Δ sont des choix de l'utilisateur. Les valeurs de α et de β représentent respectivement les coûts ou les poids attribués au changement du plan d'opération initialement établi. Ces coûts constituent une conséquence directe du contexte étudié ou encore un choix de gestion. La même remarque s'applique pour les paramètres ξ et Δ. Les valeurs choisies dépendront donc fortement du contexte et de l'organisation concernée. Néanmoins, l'analyse effectuée précédemment, de même que celle présentée ici, pourront guider un utilisateur potentiel dans le choix de paramètres à sélectionner lorsque l'approche proposée est utilisée comme outil d'aide à la décision.

2.5.2.1 Calibration des paramètres et analyse de sensibilité

Différents tests ont été effectués afin de valider la méthodologie et d'en faire le calibration. À cet effet, il est possible de diviser l'ensemble des tests menés selon deux objectifs soit l'amélioration de la borne inférieure (A, B, C, D) et l'amélioration de la borne supérieure (E, F, G).

Au cours de cette série d'expérimentations, nous tenterons de déterminer des valeurs adéquates pour les différents paramètres et d'identifier un schéma algorithmique qui permet de déterminer de bons résultats dans des délais de temps raisonnables.

Tous les tests en lien avec la calibration ont été effectués en considérant les instances du groupe G_4, pour le type de perturbation $M6$ et en considérant $\alpha = 1$, $\beta = 5$, $\xi = 0$ et $\Delta = 1,2\Delta_{opt}$. Rappelons que Δ_{opt} est la valeur de Δ correspondant à la solution optimale, telle que déterminée en 2.5.1. Il est toutefois important de rappeler que Δ est une valeur arbitraire qui devra être fixée par l'utilisateur et qui dépendra fortement des valeurs de α et β. En posant de celle manière la valeur de Δ, il est possible de représenter le cas où la contrainte de changement admissible a été relaxée. L'impact de la valeur de Δ sera analysé plus en détail en 2.5.2.2. Enfin, à moins d'indication contraire, les paramètres suivants ont été utilisés :

- **En A, B et C** : $\varepsilon = 1, \kappa = 1, MaxItMac = 100, MaxItMic = 250, NbItSa_{max} = 1000$ et $NbChgt_{max} = 0$. Dans ce cas, le mécanisme de réduction du pas automatique a été désactivé.
- **En D** : $\varepsilon = 2, \kappa = 2, MaxItMac = 100, MaxItMic = 250$ et $NbChgt_{max} = 100$.
- **En E et F** : $\varepsilon = 2, \kappa = 2, MaxItMac = 100, MaxItMic = 250, NbItSa_{max} = 1000$ et $NbChgt_{max} = 100$.

Analyse des paramètres pouvant influencer la borne inférieure

A) Valeur du pas (ε). En observant les résultats rapportés Tableau 2.2, il est possible de constater que plus la valeur du pas ε est petite, meilleure est la borne inférieure. Pour de grandes valeurs de ε, la borne inférieure moyenne s'améliore dans une plus grande mesure lorsque la valeur de ε réduit, puis, à partir de $\varepsilon = 0,5$, la borne inférieure moyenne ne s'améliore que de $0,06\%$. De plus, la valeur de ε ne semble pas avoir d'impact considérable sur la valeur moyenne de la borne supérieure. On constate néanmoins que les valeurs moyennes semblent légèrement meilleures pour des valeurs de ε un peu plus grandes, soit pour $\varepsilon = 2$ et $\varepsilon = 1$.

	z_{BS}			z_{BI}		
ε	$\%_{moy}$	$\%_{max}$	$\%_{min}$	$\%_{moy}$	$\%_{max}$	$\%_{min}$
5	2,40%	4,90%	0,40%	−10,79%	−1,66%	−45,35%
2	2,09%	4,90%	0,00%	−3,92%	−1,49%	−11,36%
1	2,29%	4,90%	0,21%	−3,16%	−1,20%	−11,33%
0,5	2,66%	4,90%	0,00%	−3,01%	−0,78%	−11,31%
0,25	2,82%	4,90%	0,00%	−2,95%	−0,58%	−11,30%
0,1	2,74%	4,90%	0,16%	−2,89%	−0,44%	−11,30%

Tableau 2.2 – Variation du pas (ε) lors de la mise à jour des multiplicateurs de lagrange

B) Nombre maximal d'itérations lors de la recherche d'une solution duale (*MaxItMicro*). Les valeurs présentées au Tableau 2.3 indiquent que la borne inférieure moyenne s'améliore de manière plus ou moins notable lorsque la valeur de *MaxItMicro* augmente. En effet, il est possible d'observer une amélioration moyenne de 0.17% lorsque *MaxItMicro* passe de 50 itérations à 1000. Le temps de calcul augmente pour sa part beaucoup plus considérablement, on parlera d'un temps de calcul 20 à 30 fois plus important pour 1000 itérations par rapport à 50. En ce qui concerne la borne supérieure moyenne, le comportement est beaucoup plus difficile à prédire. La richesse et la diversité de l'ensemble de colonnes obtenues à la suite de la génération de colonnes par ascension duale (PHASE 3) est tout à fait imprévisible. Enfin, il a été possible d'observer que le temps de calcul augmente environ d'un facteur 1 avec *MaxItMicro*. Ainsi, si on double *MaxItMicro*, on doublera le temps de calcul.

	z_{BS}			z_{BI}		
MaxItMicro	$\%_{moy}$	$\%_{max}$	$\%_{min}$	$\%_{moy}$	$\%_{max}$	$\%_{min}$
50	2,19%	4,90%	0,09%	−1,98%	−0,73%	−3,73%
100	2,56%	4,90%	0,00%	−1,92%	−0,72%	−3,63%
250	2,29%	4,90%	0,21%	−1,83%	−0,73%	−3,63%
500	2,45%	4,90%	0,09%	−1,84%	−0,72%	−3,57%
1000	2,66%	4,90%	0,24%	−1,81%	−0,72%	−3,56%

Tableau 2.3 – Variation du nombre maximal d'itérations *micro* (*MaxItMicro*)

C) Nombre maximal d'itérations de génération de colonnes par ascension duale (*MaxItMacro*). Il est possible de constater au Tableau 2.4 qu'en augmentant la valeur de *MaxItMacro*, la borne inférieure moyenne s'améliore. Cette amélioration est toutefois très petite, 0,08% en moyenne, par rapport au temps de calcul supplémentaire nécessaire. De plus, lorsque la valeur de *MaxItMacro* augmente, la borne supérieure moyenne s'améliore légèrement ou demeure inchangée. En effet, en augmentant *MaxItMacro*, certaines colonnes qui n'améliorent pas la valeur de la relaxation linéaire de $\hat{RP}$, mais qui peuvent mener à l'amélioration de la qualité de la solution finale entière, peuvent être générées. Ce comportement est toutefois plutôt difficile à prédire. Enfin, il est possible d'observer que le temps de calcul augmente toujours d'un facteur 1 avec *MaxItMacro*, c'est-à-dire que si on double *MaxItMacro*, le temps de calcul doublera aussi.

	z_{BS}			z_{BI}		
MaxItMacro	$\%_{moy}$	$\%_{max}$	$\%_{min}$	$\%_{moy}$	$\%_{max}$	$\%_{min}$
50	2,70%	4,90%	0,00%	−1,80%	−0,43%	−3,68%
100	2,56%	4,90%	0,00%	−1,76%	−0,42%	−3,58%
250	2,53%	4,90%	0,00%	−1,74%	−0,42%	−3,55%
500	2,53%	4,90%	0,00%	−1,72%	−0,42%	−3,54%
1000	2,53%	4,90%	0,00%	−1,72%	−0,42%	−3,53%

Tableau 2.4 – Variation du nombre maximal d'itérations *macro* (*MaxItMacro*)

D) Analyse du mécanisme de réduction du pas automatique. Nous avons observé en A) que la réduction du pas permettait d'améliorer la borne inférieure. Néanmoins, un grand pas peut également amener de la diversité à l'ensemble de colonnes générées ce qui est souhaitable du point de vue de la borne supérieure finale. Ainsi, tel que présenté précédemment, un mécanisme de réduction automatique du pas a été implanté afin de permettre l'utilisation d'un pas initial relativement grand puis de le réduire progressivement, d'un facteur κ, à chaque fois que la solution duale ne s'améliore pas pour un nombre donné d'itérations ($NbItSa_{max}$). Au Tableau 2.5, différentes valeurs de $NbItSa_{max}$ ont été testées en fixant $\varepsilon^0 = 2$ et $\kappa = 2$. En comparant les résultats obtenus ici aux résultats obtenus en A, B et C, il est possible de constater que la borne inférieure moyenne déterminée en utilisant le mécanisme de réduction du pas automatique est de meilleure qualité que celles obtenues en A, B et C. Le mécanisme implanté permet donc d'améliorer la valeur de la borne inférieure moyenne. Néanmoins, lorsque l'on compare les solutions obtenues en utilisant le mécanisme, mais pour différentes valeurs $NbItSa_{max}$, la qualité de la borne inférieure semble insensible à la valeur de $NbItSa_{max}$. Le mécanisme implanté ne semble pas non avoir d'impact important sur la borne supérieure moyenne.

	z_{BS}			z_{BI}		
$NbItSa_{max}$	$\%_{moy}$	$\%_{max}$	$\%_{min}$	$\%_{moy}$	$\%_{max}$	$\%_{min}$
10	$2,65\%$	$4,90\%$	$0,29\%$	$-1,54\%$	$-0,38\%$	$-3,41\%$
25	$2,43\%$	$4,90\%$	$0,34\%$	$-1,54\%$	$-0,37\%$	$-3,40\%$
50	$3,12\%$	$5,87\%$	$0,35\%$	$-1,54\%$	$-0,37\%$	$-3,40\%$
100	$2,50\%$	$4,90\%$	$0,00\%$	$-1,59\%$	$-0,37\%$	$-3,45\%$

Tableau 2.5 – Variation du nombre d'itérations sans amélioration ($NbItSa_{max}$)

Analyse des mécanismes pour l'amélioration de la borne supérieure

E) Analyse de la phase de diversification basée sur la modification de la valeur de Δ (Phase 3.0). Des expérimentations préliminaires ont permis d'observer qu'en modifiant la valeur de Δ pendant la recherche d'une bonne solution duale, la borne supérieure peut changer et s'améliorer. En effet, *a priori*, afin de représenter le problème de localisation perturbé sans contrainte de changement admissible, la valeur de Δ a été fixée à $1,2\Delta_{opt}$. Néanmoins, il été possible d'observer que, dans certains cas, en modifiant la valeur de Δ, la borne supérieure s'améliore. Ce comportement est toutefois très difficile à prévoir. Afin de tirer profit de cette caractéristique, une phase de diversification a été proposée. Cette phase visera donc à utiliser pendant la phase de recherche une valeur de Δ différente de celle qui a été fixée par l'utilisateur.

Tout d'abord, différentes valeurs de $\Delta \geq \Delta_{opt}$ ont été testées, et ce, de manière formelle. Malheureusement, aucune tendance n'a pu être observée en ce qui concerne l'amélioration ou la

détérioration de la borne supérieure. Puisque le comportement est difficile à prédire et peut varier d'un problème à un autre, la valeur de Δ utilisée lors de la recherche d'une bonne solution duale, $\Delta_{recherche}$, est fixée aléatoirement. Tel que décrit lors de la présentation de l'algorithme, la valeur de $\Delta_{recherche}$ est tirée une distribution uniforme de limite inférieure $0,8\Delta$ et de limite supérieure $1,8\Delta$. Une valeur initiale pour $\Delta_{recherche}$ est fixée au début de la PHASE 3.0, puis cette valeur est modifiée un certain nombre de fois tout au long du processus de génération de colonnes. En modifiant ainsi la valeur du membre de droite de la contrainte de changement admissible, on cherche à visiter un ensemble de solutions à la limite de la frontière réalisable, dans l'espace des solutions réalisables et non-réalisables. Cela pourra permettre la génération de colonnes plus diversifiées. Cette phase de diversification s'inspire donc de certains mécanismes issus du développement de métaheuristiques. Les différents paramètres utilisés lors de la PHASE 3.0 sont fixés à partir des paramètres de la phase principale de génération de colonnes par ascension duale (PHASE 3). Ainsi, à moins d'indication contraire, les paramètres associés à la phase de diversification 3.0 sont fixés de la manière suivante : $Multi_{\varepsilon_{D0}} = 0,1$, $NbCgt_{D0} = 4$, $MultiMacro_{D0} = 0,75$ et $MultiMicro_{D0} = 0,25$. Notons qu'ici, seuls les résultats concernant la borne supérieure sont présentés puisque la borne inférieure n'est pas influencée par ce mécanisme. De plus, les valeurs moyennes rapportées aux tableaux 2.6 et 2.7 ont été déterminées à partir des meilleures valeurs trouvées pour 5 réplications.

E.1) Multiplicateur associé à $MaxItMac_{D0}$ ($MultiMacro_{D0}$). En observant les résultats présentés au Tableau 2.6, il est possible de constater que le mécanisme de diversification implanté n'améliore pas systématiquement la borne supérieure. Toutefois, pour certaines instances, un ensemble de colonnes qui n'aurait pas pu être trouvées autrement a été identifié menant ainsi à de meilleures solutions finales. De plus, en augmentant la valeur de $MultiMacro_{D0}$, et donc $MaxItMacro_{D0}$, la borne supérieure moyenne peut aussi s'améliorer, comme c'était le cas pour la PHASE 3. Néanmoins, la valeur de la borne supérieure moyenne rapportée pour $MultiMacro_{D0} = 1$ est de moins bonne qualité que pour $MultiMacro_{D0} = 0,75$ et $MultiMacro_{D0} = 0,5$. Ceci peut s'expliquer par le fait que les valeurs de $\Delta_{recherche}$ sont déterminées aléatoirement et que les valeurs moyennes présentées ont été calculées à partir des meilleures valeurs trouvées pour 5 réplications, ce qui est peu. Enfin, plus $MaxItMacro_{D0}$ est grand, plus le temps de calcul augmente, sans gain important pour les instances testées. Il est donc important de considérer le compromis entre entre le temps de calcul et le gain potentiel sur la borne supérieure lors de la sélection des paramètres finaux.

	z_{BS}		
$MultiMacro_{D0}$	$\%_{moy}$	$\%_{max}$	$\%_{min}$
0,25	2,79%	4,90%	0,27%
0,5	2,43%	4,90%	0,28%
0,75	2,41%	4,90%	0,26%
1	2,67%	4,90%	0,19%

Tableau 2.6 – Variation du multiplicateur de $MaxItMac$ ($MultiMacro_{D0}$) - PHASE 3.0

E.2) Nombre de changement de $\Delta_{recherche}$ ($NbChgt_{P0}$). Les résultats présentés au *Tableau* 2.7 permettent de tirer les mêmes conclusions que dans le cas précédent : le mécanisme implanté n'améliore pas systématiquement la borne supérieure moyenne, mais, pour certaines instances, il permet de trouver des colonnes qui ne peuvent être générées autrement, et par conséquent, de déterminer de meilleures solutions. Il est donc intéressant de faire varier la valeur de $\Delta_{recherche}$ un certain nombre de fois.

	z_{BS}		
$NbChgt_{P0}$	$\%_{moy}$	$\%_{max}$	$\%_{min}$
2	2,51%	4,90%	0,26%
4	2,41%	4,90%	0,26%
10	2,45%	4,90%	0,24%
20	2,57%	4,83%	0,22%

Tableau 2.7 – Nombre de changements de Δ pendant la PHASE 3.0

F) Analyse de la phase de diversification basée sur la modification de ε (Phase 3.1). Un des constats que nous avons pu tiré des analyses menées en A, B, C et D est que le paramètre qui semble influencer le plus la qualité de la borne supérieure est le pas ε. En effet, une valeur de ε plus grande, et souvent une valeur de $MaxItMax$ plus petite, mène généralement à des approximations moins précises des valeurs duales, permettant ainsi la génération de colonnes qui ne pourraient être trouvées autrement. L'idée ici est donc de tirer profit de l'inexactitude des valeurs duales et du bruit issu de la relaxation lagrangienne afin d'introduire de la diversité dans l'ensemble de colonnes générées. Les différents paramètres de la PHASE 3.1 sont fixés à partir des paramètres de la phase principale de génération de colonnes par ascension duale. Ainsi, à moins d'indication contraire, les paramètres associés à la phase de diversification 3.1 ont été fixés de la manière suivante : $Multi_{\varepsilon_{D1}} = 0,05$, $MultiMacro_{D1} = 0,5$ et $MultiMicro_{D1} = 0,5$. Dans ce cas aussi, seuls les résultats concernant la borne supérieure sont présentés puisque la borne inférieure n'est pas affectée par cette phase de diversification.

F.1) Multiplicateur du pas ($Mutli_{\varepsilon_{D1}}$). Il est possible d'observer au *Tableau* 2.8 que la phase de diversification basée sur la modification du ε peut contribuer à l'amélioration de la borne supérieure. En effet, en comparant les résultats obtenus ici aux résultats obtenus en A, B, C et D,

il est possible de constater que la borne supérieure moyenne est généralement de meilleure qualité. De plus, plus la valeur de $Mutli_{\varepsilon_{D1}}$ augmente, meilleure est la borne supérieure moyenne. La meilleure valeur moyenne est obtenue lorsque $Multi_{\varepsilon_{D1}} = 0,1$.

	z_{BS}		
$Multi_{\varepsilon_{D1}}$	$\%_{moy}$	$\%_{max}$	$\%_{min}$
0,025	2,65%	4,90%	0,16%
0,05	2,53%	4,90%	0,00%
0,075	2,40%	4,90%	0,00%
0,1	2,14%	4,90%	0,00%

Tableau 2.8 – Variation du multiplicateur du pas ($Mutli_{\varepsilon_{D1}}$)

F.2) Multiplicateur associé au $MaxItMac_{D1}$ ($MultiMacro_{D1}$). L'expérimentation menée pour différentes valeurs de $MultiMacro_{D1}$ a permis d'observer que le mécanisme proposé peut contribuer à l'amélioration de la borne supérieure. Toutefois, dans ce cas, on constate que les résultats obtenus sont en moyenne de moins bonne qualité que ceux rencontrés précédemment. Dans les faits, si on observe plus en détail les résultats pour chaque instance, on peut observer qu'une des instances obtient de mauvais résultats par rapport aux autres, ce qui a pour effet d'augmenter de manière importante la moyenne. Aucune des analyses effectuées n'a cependant permis d'expliquer les raisons spécifiques du mauvais comportement pour cette instance. Pour les autres instances, les résultats demeurent similaires. Tel que mentionné précédemment, augmenter $MultiMacro_{D1}$ et donc, $MaxItMac_{D1}$, peut mener à une amélioration de la borne supérieure, mais le temps de calcul devient nécessairement plus important. Pour les différents tests menés ici, le gain ne justifie pas nécessairement le temps de calcul.

	z_{BS}		
$MultiMacro_{D1}$	$\%_{moy}$	$\%_{max}$	$\%_{min}$
0,25	3,11%	5,61%	0,00%
0,5	3,05%	5,10%	0,00%
0,75	3,04%	5,10%	0,00%
1	3,04%	5,10%	0,00%

Tableau 2.9 – Variation du multiplicateur de $MaxItMac$ ($MultiMacro_{D1}$) - PHASE 3.1

F.3) Multiplicateur associé au $MaxItMic_{D1}$ ($MultiMicro_{D1}$). Les résultats présentés au *Tableau* 2.10 permettent de constater que le mécanisme de diversification implanté ici peut contribuer à l'amélioration de la borne supérieure. En effet, la valeur de la borne supérieure moyenne peut varier en fonction de $MultiMicro_{D1}$, mais le comportement est difficile à prédire. De manière générale, des valeurs de $MultiMicro_{D1}$ plus petite permettent de générer plus de diversité, mais ce n'est pas toujours le cas. De plus, plus la valeur de $MultiMicro_{D1}$ est grande, plus le temps de calcul est important, sans gain notable sur les instances testées.

	z_{BS}		
$MultiMicro_{D1}$	$\%_{moy}$	$\%_{max}$	$\%_{min}$
0,25	2,30%	4,90%	0,00%
0,5	2,53%	4,90%	0,00%
0,75	2,42%	4,90%	0,24%
1	2,37%	4,90%	0,00%

Tableau 2.10 – Variation du multiplicateur de $MaxItMic$ ($MultiMicro_{D1}$) - PHASE 3.1

G) Comparaison des différents schémas algorithmiques. Différents schémas algorithmiques ont été testés ici afin de mesurer l'impact de chacune des phases décrites précédemment. Les résultats présentés sont basés sur un ensemble de paramètres standards soit :

- **PHASE 3.0** : $Multi_{\varepsilon_{D0}} = 0,1$, $NbCgt_{D0} = 4$, $MultiMacro_{D0} = 0,5$ et $MultiMicro_{D0} = 0,25$.
- **PHASE 3.1** : $Multi_{\varepsilon_{D1}} = 0,1$, $MultiMacro_{D1} = 0,5$ et $MultiMicro_{D1} = 0,25$.
- **PHASE 3** : $\varepsilon = 2, \kappa = 2, MaxItMac = 100, MaxItMic = 250, NbItSa_{max} = 25$ et $NbChgt_{max} = 100$.
- **PHASE 5** : $Tmps_{max} = 3600$.

Ces paramètres standards seront également utilisés pour le paramétrage des tests finaux présentés en 2.5.2.2. Naturellement, pour certaine instance, il est possible de déterminer une meilleure solution en ajustant les paramètres à l'instance concernée, mais ces paramètres n'ont pas été utilisées ici.

	z_{BS}		
Génération de colonnes	$\%_{moy}$	$\%_{max}$	$\%_{min}$
Phases 4 et 5	2,73%	4,90%	0,33%
Phases 3 et 5	3,12%	5,87%	0,35%
Phases 3, 4 et 5	3,12%	5,87%	0,35%
Phases 3.0, 3 et 5	2,21%	4,53%	0,00%
Phases 3.1, 3 et 5	2,14%	4,90%	0,23%
Phases 3.1, 3, 4 et 5	2,14%	4,90%	0,23%
Phases 3.0, 3.1, 3 et 5	2,09%	4,90%	0,16%
Phases 3.0, 3.1, 3, 4 et 5	1,76%	4,90%	0,00%

Tableau 2.11 – Comparaison des schémas algorithmiques

En observant les valeurs moyennes présentées au Tableau 2.11, il est possible de constater que chacune des phases contribue, dans une certaine mesure, à l'amélioration de la solution. De plus, lorsque l'on observe les résultats obtenus pour chaque instance appartenant à $G4$, la génération de colonnes par ascension duale obtient généralement de meilleurs résultats que la génération de colonnes avec CPLEX utilisée seule. La phase de génération de colonnes avec CPLEX (PHASE 4) utilisée conjointement avec la phase 3 n'améliore pas systématiquement la

valeur de la solution finale, ce qui porte à croire que l'ensemble de colonnes générées à la suite de la PHASE 3 est suffisamment bon et que la PHASE 4 n'est en mesure d'ajouter que très peu de colonnes, le cas échéant. Enfin, les deux phases de diversification étudiés en E et F utilisés seules ou conjointement permettent également d'améliorer la solution finale. Néanmoins, chacune des phases de diversification engendre du temps supplémentaire qu'il faudra considérer plus en détail lors de la sélection du schéma algorithme final.

Les résultats présentés sont basés sur le type de perturbations $M6$. Une analyse plus poussée des schémas algorithmiques, et ce, pour tous les types de perturbations, a également été effectuée. Cette analyse de même que les temps de calcul pour chaque schéma algorithmique sont présentés et discutés en détail à la section suivante.

2.5.2.2 Résultats finaux

Différents tests ont été menés afin de valider que l'approche proposée permet de déterminer de bons résultats pour un éventail de problèmes. Pour ce faire, trois séries de tests ont été menées. La première série d'expérimentations vise à montrer que l'algorithme développé peut considérer adéquatement différents types de perturbations. Elle permettra également de faire ressortir les difficultés relatives du traitement des différents types de perturbations. La deuxième série de tests vise à évaluer la méthode pour différentes valeurs de α, β, ξ et Δ. Tout d'abord, en faisant varier α et β, il est possible d'analyser l'impact de l'importance relative accordée aux décisions de localisation et d'affectation des clients sur l'efficacité de la méthode proposée. De plus, en imposant différentes valeurs de $\xi > 0$, il est possible de valider que la méthode fonctionne toujours bien en considérant une pénalité sur le changement entre la solution du problème considéré et la solution de référence dans la fonction objectif. Il est important de rappeler que, pour toutes les expérimentations menées jusqu'à présent, seul le cas où $\xi = 0$ a été considéré. Enfin, en testant différentes valeurs de Δ, il est possible d'observer l'impact d'un contrôle plus ou moins grand de la solution ou, autrement dit, d'une contrainte de changement admissible plus ou moins serrée. La troisième série d'expérimentations vise, quant à elle, à valider la méthode pour différentes tailles de problèmes, allant de 20 clients et 10 dépôts potentiels jusqu'à 90 clients et 30 dépôts potentiels.

Les déviations moyennes des bornes inférieures et supérieures par rapport à la solution optimale, exprimées en %, sont présentées pour tous les tests effectués. La déviation moyenne de la borne inférieure lorsque la PHASE 4 est utilisée correspond à la valeur de la relaxation linéaire du problème RP_p. Tel que mentionné dans Boschetti *et al.* (2008), la borne inférieure obtenue grâce à la méthode d'ascension duale proposée dans leur étude est presque égale à la valeur de

la relaxation linéaire du problème et domine strictement la borne inférieure obtenue par la relaxation lagrangienne classique du problème de partitionnement. Lorsque la PHASE 4 n'est pas lancée, la borne inférieure z_{BI} est déterminée grâce à la méthode d'ascension duale (PHASE 3). L'information à propos du temps de calcul est également rapportée dans cette section. Il est toutefois important de mentionner que l'algorithme décrit n'a pas été optimisé du point de vue du temps de calcul. Les temps rapportés permettront toutefois d'avoir une idée de l'effort de calcul relatif entre les différents schémas algorithmiques. De plus, à moins d'indication contraire, les paramètres standards présentés en 6.3.1 sont employés ici.

A) Types de perturbations. Les tests menés afin de valider la méthode pour différents types de perturbations ont été effectués en considérant le groupe d'instances G_4 et en prenant en compte $\alpha = 1$, $\beta = 5$, $\xi = 0$ et $\Delta = 1.2\Delta_{opt}$. De plus, différentes versions de l'algorithme ont été testées afin d'analyser l'impact des différentes phases de génération de colonnes sur la qualité de solution finale obtenue et le temps de calcul.

	Phases	M0 Nouveaux dépôts	M1 Fermeture dépôts	M2 Ajout clients	M3 Retrait clients	M4 Augmentation demande	M5 Réduction demande	M6 Variation demande
z_{BS} $(\%_{moy})$	**4 et 5**	0.73%	5,51%	4,89%	10,67%	2,54%	5,91%	2,73%
	3 et 5	1,07%	5,83%	4,03%	12,65%	1,64%	5,69%	3,12%
	3, 4 et 5	1,07%	5,83%	4,03%	10,12%	1,52%	5,60%	3,12%
	3.0, 3 et 5	0,57%	3,99%	3,68%	7,89%	1,51%	5,37%	2,21%
	3.1, 3 et 5	1,00%	5,06%	3,46%	9,02%	1,69%	4,63%	2,14%
	3.0, 3.1, 3 et 5	0,56%	4,07%	2,94%	3,72%	1,50%	3,60%	2,09%
	3.0, 3.1, 3, 4 et 5	0,57%	3,09%	2,56%	3,45%	1,42%	3,03%	1,76%
z_{BI} $(\%_{moy})$	**4 et 5**	−1,26%	−2,34%	−1,77%	−2,33%	−1,06%	−1,37%	−1,53%
	3 et 5	−1,26%	−2,35%	−1,81%	−2,44%	−1,06%	−1,40%	−1,54%
	3, 4 et 5	−1,26%	−2,34%	−1,77%	−2,33%	−1,06%	−1,37%	−1,53%
	3.0, 3 et 5	−1,27%	−2,34%	−1,82%	−2,35%	−1,06%	−1,38%	−1,54%
	3.1, 3 et 5	−1,27%	−2,34%	−1,84%	−2,37%	−1,06%	−1,38%	−1,54%
	3.0, 3.1, 3 et 5	−1,27%	−2,34%	−1,78%	−2,34%	−1,06%	−1,38%	−1,54%
	3.0, 3.1, 3, 4 et 5	−1,26%	−2,34%	−1,77%	−2,33%	−1,06%	−1,37%	−1,53%
Tmps (s)	**4 et 5**	2	45	423	1	7	6	2
	3 et 5	5	58	291	3	19	6	17
	3 et 4	5	38	163	3	19	6	16
	3.0, 3 et 5	33	304	1625	21	107	38	81
	3.1, 3 et 5	6	104	172	5	27	10	17
	3.0, 3.1 et 3	41	600	1511	30	136	54	90
	3.0, 3.1, 3, 4 et 5	39	307	1495	32	123	54	99

Tableau 2.12 – Résultats pour différents types de perturbations

En observant les résultats présentés au Tableau 2.12, il est possible de constater que la phase de génération de colonnes par ascension duale (PHASE 3), employée seule, ne permet pas d'améliorer systématiquement la borne supérieure moyenne trouvée grâce à la génération de colonnes par CPLEX (PHASE 4). De plus, l'implantation d'une phase de génération de colonnes avec CPLEX à la suite de la PHASE 3 n'améliore pas systématiquement la qualité de la borne supérieure trouvée. Par contre, les phases de diversification 3.0 et 3.1, utilisée seule ou

conjointement, semblent améliorer la qualité de la borne supérieure par rapport à l'utilisation seule des phases 3 ou 4. On parlera alors d'une amélioration variant entre 1 à 4 %. Les mêmes conclusions peuvent être tirées pour tous les types de perturbations ce qui confirme bien les résultats obtenus lors de l'analyse précédente. Enfin, l'implantation de la phase de génération de colonnes avec CPLEX à la suite du schéma complet (phases 3.0, 3.1 et 3) améliore aussi la borne supérieure. Dans ce cas, comme la borne supérieure finale est de meilleure qualité, on présume que CPLEX a un effet intéressant en termes de diversité.

Le schéma algorithmique, et plus particulièrement les phases de diversification 3.0 et 3.1, ne semble pas avoir d'impact sur la borne inférieure. Le schéma algorithmique pourra toutefois avoir un impact notable sur le temps de calcul. De manière générale, l'ajout des phases, et plus précisément de phases de diversification, augmente nécessairement le temps de calcul. Ceci est particulièrement vrai pour la phase 3.0. En effet, comme la phase 3.0 comporte une composante aléatoire, 5 réplications sont lancés et le meilleur résultat parmi les 5 réplications est enregistré. Le fait de considérer 5 réplications a pour effet d'augmenter considérablement le temps de calcul. Les résultats obtenus peuvent toutefois justifier l'augmentation du temps de calcul à condition que le temps disponible à la résolution du problème soit suffisamment grand. Cela dépendra fortement du contexte d'application. Dans certains cas, on observe l'effet inverse, c'est-à-dire que considérer plusieurs phases mène à une réduction du temps de calcul. Par exemple, pour le type de modification $M2$, le temps de calcul correspondant aux phases 3 et 4 est plus petit que celui correspondant à la phase 3. Il est important de mentionner que, dans tous les cas, une grande part du temps de calcul est dédié à la résolution du problème maître entier (PHASE 5). Dans le cas énoncé ci-haut, le fait de générer un ensemble de colonnes différent en utilisant à la fois les phases 3 et 4 a pour effet de réduire le temps nécessaire pour la phase 5.

En observant maintenant les résultats obtenus pour chaque type de perturbation, il est possible de constater que les meilleurs résultats sont obtenus pour les types de perturbations $M0$, $M4$ et $M6$, soit respectivement l'ajout de dépôts, l'augmentation et la variation de la demande. Dans ces cas, le schéma complet a permis de déterminer une borne supérieure ayant une déviation moyenne de 0,57 %, 1,42% et 1,76% par rapport à la solution optimale. Les moins bons résultats ont, pour leur part, été obtenus pour les types de perturbations $M1$, $M3$ et $M5$, soit la fermeture de dépôts, le retrait de clients et la réduction de la demande, où des déviations moyennes de 3,09% 3,45% et 3,03 % ont été observées. Enfin, les résultats obtenus pour le type de perturbations $M2$, soit l'ajout de clients, sont de l'ordre de 2,56%, ce qui les situent entre ceux obtenus pour les deux groupes décrits précédemment en termes de qualité de la borne supérieure. Néan-

moins, en analysant plus en détail les résultats trouvés pour chaque instance, il est possible de constater que des valeurs maximales de 7,78 %, 9,12 % et de 10,05% ont obtenues pour $M1$, $M3$ et $M5$ contre des valeurs maximales de 2,59 %, 3,97 %, 4,04 % et 4,90 % pour $M0$, $M2$, $M4$ et $M6$. De plus, pour $M3$, deux instances ont obtenus une borne supérieure à plus de 5% de l'optimum, trois pour $M5$. Pour les autres types de perturbations, aucune instance n'a obtenue un résultat aussi élevé, à l'exception d'une instance pour $M1$. Les instances ayant obtenu de moins bons résultats peuvent donc avoir un impact important sur l'écart de la borne supérieure moyenne.

Il a également été possible de constater que, pour toutes les instances pour lesquelles la borne supérieure déterminée est de moins bonne qualité, le ratio entre la demande totale et la capacité totale des dépôts potentiels, $\frac{D_{tot}}{B_{tot}}$ est petit. On parle, en général, d'une valeur en-deçà de 40 %. Lorsque le ratio $\frac{D_{tot}}{B_{tot}}$ est petit, le nombre de solutions réalisables a tendance à augmenter. Nécessairement, en retirant des clients ou en réduisant leur demande, la valeur de D_{tot} diminue, et conséquemment le ratio $\frac{D_{tot}}{B_{tot}}$, ce qui peut expliquer en partie les moins bonnes performances obtenues pour les types de perturbations $M3$ et $M5$. Pour ces instances, une plus longue phase de diversification pourrait être bénéfique. Il est important de mentionner que, bien que les instances ayant obtenues de moins bons résultats ont toutes de petites valeurs de $\frac{D_{tot}}{B_{tot}}$, l'approche proposée a aussi été en mesure d'obtenir de très bons résultats pour de telles instances. D'autres mesures ont été calculées afin d'analyser les structures des instances, notamment le ratio entre le coût fixe moyen de sélection des dépôts et le coût d'affectation moyen des clients aux dépôts et le ratio entre le coût fixe moyen de sélection et la capacité moyenne des dépôts. Il semble toutefois que seul le ratio $\frac{D_{tot}}{B_{tot}}$ puisse expliquer la difficulté à résoudre certaines instances.

En ce qui concerne la borne inférieure moyenne, les types de modifications $M1$ et $M3$ semblent donner une borne inférieure légèrement moins serrée, particulièrement pour certaines instances, ce qui a pour effet de dégrader la valeur moyenne. En effet, pour M1, l'instance P28 donne une borne inférieure de -4,18% et pour M3, l'instance P27 donne une borne inférieure d'environ -6,5%. Pour ces deux instances, on note également des petites valeurs de $\frac{D_{tot}}{B_{tot}}$ soit respectivement 46% et 11%. Enfin, le temps de calcul est beaucoup plus important pour les types de modifications M1, M2 et M4, c'est-à-dire les cas où la solution de référence est non-réalisable et où les perturbations ont un impact plus important sur le problème. Dans ces cas, l'augmentation du temps de calcul est principalement due à la phase de détermination d'une borne supérieure finale avec CPLEX (PHASE 5). Il semble que la valeur du ratio $\frac{D_{tot}}{B_{tot}}$ ait aussi un impact sur l'augmentation du temps de calcul lors de la PHASE 5. Les instances ayant de petites valeurs de $\frac{D_{tot}}{B_{tot}}$ semblent être plus longues à résoudre, et ce, pour tous les types de perturbations, vu

l'espace de solutions grandissant.

B) Valeurs de α, β, ξ et Δ. Les tests menés pour différentes valeurs de α, β, ξ et Δ ont été effectués en considérant le groupe d'instances G_4 et le type de perturbations $M6$, soit la variation de la demande des clients. Trois versions de l'algorithme sont présentées ici soit la version comportant la phase de génération de colonnes avec CPLEX (PHASE 4) seule, celle qui inclut la génération de colonnes par ascension duale et celle par CPLEX (phases 3 et 4) et le schéma complet. Puisque les mêmes conclusions peuvent être tirées ici, le schéma algorithmique ne sera pas discuté de manière approfondie dans cette section. Les résultats sont plutôt présentés à titre indicatif. Des commentaires sur les schémas algorithmiques sont toutefois ajoutées à la discussion lorsqu'il sera pertinent de le faire.

B1) Valeurs de α et β. Dans ce cas, la valeur de ξ a été fixée à 0 et la valeur de Δ à $1,2\Delta_{opt}$.

	Phases	$(0,1)$	$(1,10)$	$(1,5)$	$(1,1)$
$z_{BS}(\%_{moy})$	**4 et 5**	$5,77\%$	$2,15\%$	$2,73\%$	$2,89\%$
	3, 4 et 5	$4,20\%$	$2,59\%$	$3,12\%$	$2,52\%$
	3.0, 3.1, 3, 4 et 5	$4,11\%$	$2,14\%$	$1,76\%$	$1,61\%$
$z_{BI}(\%_{moy})$	**4 et 5**	$-1,53\%$	$-1,53\%$	$-1,53\%$	$-1,53\%$
	3, 4 et 5	$-1,54\%$	$-1,54\%$	$-1,54\%$	$-1,54\%$
	3.0, 3.1, 3, 4 et 5	$-1,54\%$	$-1,54\%$	$-1,54\%$	$-1,54\%$
$Tmps(s)$	**4 et 5**	8	3	2	5
	3, 4 et 5	11	13	16	10
	3.0, 3.1, 3, 4 et 5	57	89	99	75

Tableau 2.13 – Résultats pour différentes valeurs de (α,β)

En observant les résultats présentés au *Tableau* 2.13, il est possible de constater que la borne supérieure moyenne semble s'améliorer au fur et à mesure que le ratio entre α et β augmente. Autrement dit, la borne supérieure moyenne s'améliore lorsqu'une importance relative plus grande est accordée au changement de l'affectation des clients par rapport au changement de statut des dépôts. En effet, lorsque $\alpha = 0$ et $\beta = 1$, le changement des décisions d'affectation des clients aux dépôts ne contribue pas à la contrainte de changement admissible. Plusieurs solutions peuvent alors être équivalentes du point de vue de la contrainte de changement admissible. Ceci peut expliquer, de moins partiellement, la difficulté à déterminer une meilleure solution dans ce cas. Ensuite, au fur et à mesure que le ratio $\frac{\alpha}{\beta}$ augmente, la borne supérieure moyenne déterminée s'améliore. Les valeurs de α et de β ne semblent toutefois pas avoir d'impact significatif sur la valeur de la borne inférieure moyenne. Les temps de calcul, quant à eux, demeurent sensiblement les mêmes pour toutes les valeurs (α,β).

B2) Valeurs de ξ. Dans ce cas, les valeurs de α, β et Δ ont été posées respectivement à 1, 5 et $1,2\Delta_{opt}$. En fixant ainsi la valeur de Δ, il est possible d'analyser indépendamment l'impact de différentes valeurs de la pénalité associée au changement admissible dans la fonction objectif sur les performances de la méthode proposée. De plus, la valeur de ξ a été déterminée comme à la *Section* 2.5.1 en faisant varier la valeur de γ.

	Phases	$\gamma = 0$	$\gamma = 0,5$	$\gamma = 1$	$\gamma = 2$
$z_{BS}(\%_{moy})$	**4 et 5**	2.73%	1.00%	0.55%	0.35%
	3, 4 et 5	3,12%	1,34%	0,90%	0,65%
	3.0, 3.1, 3, 4 et 5	1,76%	0,75%	0,55%	0,28%
$z_{BI}(\%_{moy})$	**4 et 5**	−1,53%	−1,87%	−1,74%	−1,51%
	3, 4 et 5	−1,54%	−1,87%	−1,74%	−1,52%
	3.0, 3.1, 3, 4 et 5	−1,54%	−1,87%	−1,75%	−1,56%
$Tmps(s)$	**4 et 5**	2	2	1	1
	3, 4 et 5	16	11	10	10
	3.0, 3.1, 3, 4 et 5	99	69	64	62

Tableau 2.14 – Résultats pour différentes valeurs de ξ

En observant les résultats obtenus pour différentes valeurs de γ, il est possible de remarquer que la borne supérieure moyenne trouvée s'améliore avec la valeur de γ et donc, de ξ. En augmentant γ, une pénalité plus grande est accordée au changement de la solution dans la fonction objectif, favorisant ainsi des solutions dont les efforts de changement sont plus limités. Il est aussi possible de noter que, bien que le schéma complet permette généralement d'obtenir une borne supérieure moyenne de meilleure qualité, l'écart entre les solutions obtenus en utilisant la PHASE 4 seule ou les phases 3 et 4 et celles obtenues grâce au schéma complet semble moins considérable lorsque $\xi > 0$. Enfin, peu importe la valeur de ξ employée, la qualité de la borne inférieure moyenne et le temps de calcul demeurent équivalents.

B3) Valeurs de Δ. Le tableau 2.15 présente les résultats obtenus pour différentes valeurs de Δ fixées par l'utilisateur. Rappelons que lorsque Δ_{opt} est la valeur de Δ correspondant à la solution optimale du problème sans contrainte de changement admissible. Ensuite, plus la valeur de Δ est petite, moins on accepte de changer la solution de référence ou, autrement dit, plus le « budget » disponible pour changer la solution de référence est limité. Pour tous les tests menés, les valeurs de α, β et ξ sont respectivement fixées à 1, 5 et 0. De plus, pour toutes les valeurs de Δ choisies, l'ensemble des instances considérées demeurent réalisables.

Les valeurs présentées au Tableau 2.15 permettent de constater que la borne supérieure moyenne semble se rapprocher de la solution optimale au fur et à mesure que la valeur de Δ diminue, c'est-à-dire que le contrôle de la solution est plus grand ou que la contrainte imposée sur le changement de la solution est plus serrée. En réduisant ainsi la valeur de Δ, l'espace de solutions réalisables peut diminuer de manière importante et, conséquemment, contribuer à l'amélioration

	Phases	$1,2\Delta_{opt}$	Δ_{opt}	$0,8\Delta_{opt}$	$0,6\Delta_{opt}$	$0,4\Delta_{opt}$	$0,2\Delta_{opt}$
$z_{BS}(\%_{moy})$	**4 et 5**	2,73%	2,54%	2,10%	1,82%	1,61%	0,45%
	3, 4 et 5	3,12%	2,42%	2,47%	1,58%	1,22%	0,45%
	3.0, 3.1, 3, 4 et 5	1,76%	1,95%	1,56%	1,54%	0,80%	0,40%
$z_{BI}(\%_{moy})$	**4 et 5**	−1,53%	−1,52%	−1,89%	−2,38%	−2,86%	−1,89%
	3, 4 et 5	−1,54%	−1,56%	−1,90%	−2,23%	−2,87%	−1,91%
	3.0, 3.1, 3, 4 et 5	−1,54%	−1,53%	−1,89%	−2,39%	−2,87%	−1,89%
$Tmps(s)$	**4 et 5**	2	2	2	3	2	1
	3, 4 et 5	16	12	13	13	13	11
	3.0, 3.1, 3, 4 et 5	99	87	93	86	77	69

Tableau 2.15 – Résultats pour différentes valeurs de Δ

de la solution finale obtenue. Il est intéressant de noter que l'application du schéma complet permet toujours d'améliorer la borne supérieure moyenne par rapport à l'utilisation de la PHASE 4 seule ou des phases 3 et 4 et que cet écart semble se réduire lorsque la valeur de Δ diminue. De plus, en ce qui concerne la borne inférieure, la valeur de Δ ne semble pas avoir d'impact significatif, du moins, il n'y a pas de tendance claire. Les temps de calcul semblent équivalents pour les deux premiers schémas algorithmiques présentés. Toutefois, lorsque l'on observe le temps de calcul pour le schéma complet, ceux-ci diminuent de façon prévisible puisque l'espace des solutions réduit avec la la valeur de Δ. Enfin, il est intéressant de constater que plus le problème est contraint, plus l'approche proposée est en mesure de réduire l'écart entre la solution obtenue et la solution optimale du problème considéré.

Les tests menés afin d'analyser les différentes valeurs de α, β, ξ et Δ ont été répétés pour le type de perturbation $M1$ afin de valider si les conclusions tirées précédemment demeurent justes. Plusieurs raisons ont justifié ce choix. Tout d'abord, en A), il a été possible de constater que le type de perturbation $M6$ donne parmi les meilleurs résultats. Les problèmes confrontés à ce type de perturbation pourraient donc être plus faciles à résoudre. Il est aussi important de rappeler que tous les tests de calibration ont été effectués pour $M6$. Il est alors possible que les paramètres soient mieux adaptés à ce type de perturbation. Le type de perturbation $M1$ semble tout indiqué pour fin de validation, puisqu'il mène à un problème où la solution de référence n'est pas nécessairement réalisable et qui semble plus difficile à résoudre (voir en A). À la suite des expérimentations, les mêmes constats ont pu être tirés. Pour cette raison, les résultats détaillés ne sont pas présentés ici.

C) Tailles de problèmes. Pour montrer que l'approche proposée permet de résoudre adéquatement des problèmes de tailles variées, différents tests ont été effectués en considérant toutes les instances des groupes G_1 à G_7 et en prenant en compte $\alpha = 1$, $\beta = 5$ et $\xi = 0$. De plus, deux valeurs de Δ ont été analysées afin de représenter les problèmes avec et sans contrainte de changement admissible. Ces résultats sont rapportés aux tableaux 2.16 et 2.17. Enfin, de la même

manière qu'en B, les résultats obtenus à partir de trois versions de l'algorithme sont présentés à titre indicatif et commentés lorsque nécessaire.

		$G1(20,10)$		$G2(30,15)$		$G3(40,20)$		$G4(50,20)$	
	Phases	$1,2\Delta_{opt}$	$0,4\Delta_{opt}$	$1,2\Delta_{opt}$	$0,4\Delta_{opt}$	$1,2\Delta_{opt}$	$0,4\Delta_{opt}$	$1,2\Delta_{opt}$	$0,4\Delta_{opt}$
$z_{BS}(\%_{moy})$	**4 et 5**	0,41%	0,14%	0,88%	1,28%	1,26%	0,20%	2,73%	1,22%
	3, 4 et 5	0,65%	0,09%	1,02%	1,71%	1,53%	0,24%	3,12%	1,12%
	3.0, 3.1, 3, 4 et 5	0,25%	0,06%	0,22%	0,74%	0,37%	0,11%	1,76%	0,99%
$z_{BI}(\%_{moy})$	**4 et 5**	−0,83%	−0,87%	−1,41%	−3,48%	−1,42%	−2,59%	−1,53%	−2,87%
	3, 4 et 5	−0,84%	−0,89%	−1,41%	−3,51%	−1,44%	−2,68%	−1,54%	−2,87%
	3.0, 3.1, 3, 4 et 5	−0,84%	−0,89%	−1,41%	−3,50%	−1,42%	−2,63%	−1,54%	−2,87%
$Tmps(s)$	**4 et 5**	1	1	1	1	3	1	4	2
	3, 4 et 5	1	1	4	3	10	4	6	13
	3.0, 3.1, 3, 4 et 5	6	5	22	17	69	42	68	77

Tableau 2.16 – Résultats pour différentes tailles de problèmes (I)

		$G5(60,30)$		$G6(75,30)$		$G7(90,30)$	
	Phases	$1,2\Delta_{opt}$	$0,4\Delta_{opt}$	$1,2\Delta_{opt}$	$0,4\Delta_{opt}$	$1,2\Delta_{opt}$	$0,4\Delta_{opt}$
$z_{BS}(\%_{moy})$	**4 et 5**	2,52%	1,68%	2,61%	1,61%	6,97%	3,42%
	3, 4 et 5	2,60%	1,49%	2,55%	1,49%	6,42%	3,45%
	3.0, 3.1, 3, 4 et 5	0,95%	1,09%	2,13%	1,47%	5,70%	3,25%
$z_{BI}(\%_{moy})$	**4 et 5**	−2,22%	−3,88%	−2,87%	−3,79%	−2,03%	−4,25%
	3, 4 et 5	−2,37%	−3,90%	−2,87%	−3,89%	−2,54%	−4,51%
	3.0, 3.1, 3, 4 et 5	−2,23%	−3,89%	−2,87%	−3,84%	−2,38%	−4,47%
$Tmps(s)$	**4 et 5**	33	3	701	13	783	18
	3, 4 et 5	35	19	364	33	850	66
	3.0, 3.1, 3, 4 et 5	443	120	4746	249	6290	403

Tableau 2.17 – Résultats pour différentes tailles de problèmes (II)

De manière générale, l'approche proposée permet de trouver de bons résultats, peu importe la taille du problème. Tout d'abord, en observant les résultats pour $\Delta = 1,2\Delta_{opt}$, il est possible de constater que, pour les instances des groupes G_1, G_2, G_3 et G_5, la borne supérieure moyenne trouvée est en-deçà de 1% de la valeur optimale. Parmi toutes les instances pour G_1, G_2, G_3 et G_5, les valeurs maximales obtenues sont respectivement de 1,03%, 1,37%, 1,10% et 2,59%. Pour les groupes G_4 et G_6, la borne supérieure moyenne est autour de 2%. Les valeurs maximales obtenues sont alors respectivement de 4,9% et 10,77%. Ces instances vont contribuer considérablement à l'augmentation de la valeur moyenne. Enfin, pour les instances du groupe G_7, la borne supérieure moyenne est d'environ 6%. Dans ce cas, l'algorithme obtient de moins bons résultats pour deux instances, $P50(23\%)$ et $P54(17\%)$, ce qui a un effet important sur la borne supérieure moyenne. En analysant la structure de ces instances, il est possible de constater que l'observation discutée en A est toujours valide ici : l'approche proposée, en utilisant le paramétrage standard, a plus de difficulté à résoudre les instances pour lesquelles le ratio $\frac{D_{tot}}{B_{tot}}$ est petit. Et cet effet semble plus marqué lorsque la taille du problème augmente. Il est donc possible de penser qu'en adaptant les paramètres aux instances considérées, une amélioration pourrait être obtenue.

En observant maintenant les résultats obtenus lorsque $\Delta = 0,4\Delta_{opt}$, il est possible de constater que l'écart moyen entre la borne supérieure déterminée et la valeur optimale pour le problème considéré est généralement plus petit. Ceci vient donc confirmer les constats émis en B. De plus, la tendance, en termes de qualité de la solution, est la même que pour $\Delta = 1,2\Delta_{opt}$. Enfin, il est possible de constater que l'application du schéma complet permet toujours d'améliorer la borne supérieure par rapport à l'utilisation de la PHASE 4 ou des phases 3 et 4. Toutefois, cet écart est plus petit pour les instances de plus petite taille puisque la phase de génération de colonnes grâce à CPLEX (PHASE 4), employée seule, permet de trouver de très bonnes solutions.
En ce qui concerne la borne inférieure, il est possible de noter que la borne inférieure moyenne se détériore généralement lorsque la taille du problème augmente. De plus, la borne inférieure moyenne semble se détériorer aussi lorsque le multiplicateur de Δ_{opt} diminue. Enfin, comme prévu, le temps de calcul augmente avec la taille du problème. Le temps de résolution peut devenir très long pour les instances des groupe G_6 et G_7. L'augmentation du temps de calcul est due principalement à la phase de recherche de la solution finale entière avec CPLEX (PHASE 5). Le temps de calcul augmente aussi lorsque $\Delta = 0,4\Delta_{opt}$, mais de manière beaucoup moins importante.

2.6 Conclusion

Ce chapitre s'est intéressé à la modélisation, à l'analyse et à la résolution d'un problème de localisation perturbé. Dans un premier temps, deux formulations ont été définies pour le problème de localisation avec capacité et affectation unique en présence de perturbations. Ces deux formulations s'inspirent respectivement des formulations proposées pour le problème de localisation sous-jacent, mais intègrent les différents mécanismes issus de la réoptimisation contrôlée. Ainsi, ces mécanismes, soit l'ajout de contraintes de changement admissible et l'introduction d'une pénalité dans la fonction objectif, permettent de tenir compte de l'impact de la modification de la solution de référence lors de la recherche d'une solution pour le problème perturbé. D'un point de vue pratique, le fait de limiter ou de contrôler les différences entre deux solutions successives permettra d'en faciliter la mise en oeuvre et de limiter les coûts d'implantation.
Dans un deuxième temps, une approche a été proposée afin de résoudre le problème étudié. La méthode développée dans ce contexte utilise une approche de génération de colonnes où le problème maître est résolu par une méthode d'ascension duale et le sous-problème est un problème de sac à dos, résolu à l'optimalité grâce à un algorithme de programmation dynamique. Le schéma de résolution, basé sur la formulation de type partitionnement, a l'avantage

d'être flexible et générique. Ainsi, il pourra être employé afin de considérer des problèmes de structure similaire lorsque confrontés à différents types de perturbations. L'approche proposée permet donc de fournir une borne inférieure de même qu'une borne supérieure déterminée à la suite de la résolution du problème maître entier, avec CPLEX, en utilisant l'ensemble de colonnes générées aux phases précédentes. Différentes techniques auraient pu être envisagées afin de déterminer la solution finale. CPLEX a été utilisé ici afin de conserver la généralité de l'approche. Néanmoins, deux phases ont été implantées afin d'améliorer la diversité de l'ensemble de colonnes généré. Ces deux mécanismes, notamment la PHASE 3.0, ont permis d'améliorer considérablement les bornes supérieures obtenues.

Différentes expérimentations ont ensuite été menées afin d'étudier le problème de localisation perturbé en analysant le compromis entre le contrôle de la solution et la qualité de la solution finale obtenue. Ainsi, différentes courbes ont été tracées, et ce, pour tous les types de perturbations considérés et pour différentes valeurs des paramètres du problème, α, β et ξ. L'analyse de ces courbes a permis de constater que, lorsque la solution de référence n'est pas réalisable, la première solution réalisable, c'est-à-dire celle qui correspond à la plus petite valeur de Δ possible, est relativement bonne et que l'amélioration potentielle est limitée. En contrepartie, maintenir le *statu quo* lorsque la solution de référence est toujours réalisable peut mener à des coûts beaucoup plus élevés que lorsqu'une réoptimisation complète est effectuée. Enfin, dans tous les cas, dès que l'on accepte de modifier un peu le plan d'opération, la qualité de la solution s'améliore dans une plus grande proportion, puis cette amélioration diminue au fur et à mesure que la valeur de Δ augmente. Les mêmes constats ont pu être faits pour différentes valeurs de α, β et ξ. Naturellement, la sélection adéquate des valeurs de ces paramètres demeurent une question importante dont la réponse pourra grandement varier en fonction du contexte d'application. Il revient donc à l'utilisateur de fixer les valeurs qu'il jugera adéquates selon le problème étudié.

Différents tests ont également été effectués afin de démontrer la flexibilité et la généralité de l'approche proposée, c'est-à-dire qu'elle fonctionne bien pour un ensemble de variantes du problème de localisation perturbé. Ainsi, il a été possible de montrer que l'approche proposée fournit de bons résultats pour tous les types de perturbations et pour toutes les tailles de problèmes considérés dans cette étude. La borne inférieure déterminée semble peu affectée par le type de perturbation ou par la taille du problème. La qualité de la borne supérieure se détériore toutefois lorsque la taille du problème augmente. Il est important de mentionner que les valeurs des paramètres standards employés pour la résolution de toutes les variantes du problème ont été fixées à partir des résultats obtenus pour des problèmes de taille moyenne. Un gain pour-

rait donc être possible en ajustant les paramètres à l'instance étudiée. Il est également possible d'observer que les paramètres α, β et ξ semblent avoir peu d'impact sur les performances générales de l'approche proposée. Toutefois, lorsque le problème est davantage contraint, l'approche semble en mesure de réduire l'écart entre la solution obtenue et la solution optimale du problème considéré, ce qui est souhaitable dans la contexte de la réoptimisation contrôlée.

Enfin, bien que l'approche proposée permette de trouver de bons résultats au problème étudié, les temps de calcul peuvent devenir importants lorsque la taille du problème augmente. D'une part, tel que mentionné précédemment, le code n'a pas été optimisé afin de limiter au maximum le temps de calcul, mais plutôt utilisé en vue d'analyses et de preuve de concept. Une optimisation du code, de même que différents paramétrages de CPLEX, pourraient être souhaitables afin d'améliorer les performances du point de vue du temps de calcul. D'autre part, il semble que la détermination d'une solution entière avec CPLEX contribue de manière importante à l'augmentation du temps de calcul. Le développement d'une méthode heuristique propre au contexte pourrait consistuer un choix judicieux afin de déterminer une bonne solution réalisable à partir de la solution de la relaxation. Les améliorations potentielles soulevées ici pourront donc être considérées lors de recherches futures. En pratique, le temps disponible à la résolution du problème dépendra fortement de la nature du problème étudié et du type de perturbation considéré.

Le problème de localisation perturbé considéré ici demeure tout à fait théorique. Son étude a permis de mieux comprendre l'impact des différents types de perturbations sur les solutions finales obtenues et de développer une approche de résolution générique et flexible qui pourra être utilisée afin de considérer des problèmes de structure similaire. L'analyse et l'adaptation de la réoptimisation contrôlée dans un contexte d'application réel constitue une extension naturelle de cette étude. La seconde partie de cette thèse s'intéressera donc à l'étude de la localisation dans le contexte de la gestion d'un service préhospitalier d'urgence, contexte où la réoptimisation contrôlée devient particulièrement pertinente.

CHAPITRE 3

DÉPLOIEMENT ET REDÉPLOIEMENT DES VÉHICULES AMBULANCIERS DANS LA GESTION D'UN SERVICE PRÉHOSPITALIER D'URGENCE

Les services préhospitaliers d'urgence (SPU) représentent une composante importante des systèmes de santé modernes. Le terme préhospitalier désigne les activités de soins cliniques et de transport ambulancier effectuées entre la réception d'un appel de détresse et la prise en charge du patient à l'hôpital. La mission principale des SPU consiste donc à répondre adéquatement aux appels de détresse en prodiguant les premiers soins aux personnes concernées et en assurant leur transport, au besoin, vers le service des urgences du centre hospitalier approprié. Dans certains cas, les SPU sont également appelés à assurer le transport des patients entre les différents centres hospitaliers. Les SPU remplissent donc un rôle important au sein des systèmes de santé modernes et leur capacité à répondre adéquatement aux appels peut avoir un impact crucial sur la santé des patients.

Ainsi, afin d'arriver à fournir un bon service à la population, les SPU doivent mobiliser un ensemble de ressources (personnel soignant, véhicules ambulanciers, centre d'appels), puis les gérer de façon efficiente. Certaines difficultés se présentent néanmoins lors de la prise de décision. Tout d'abord, les SPU sont confrontés à des demandes de service incertaines et dont la fréquence et la localisation peuvent varier selon l'heure, la journée, la semaine. De plus, les performances des SPU sont généralement mesurées en fonction du temps de réponse, défini comme le temps écoulé entre la réception d'un appel et l'arrivée du véhicule sur les lieux de l'incident. Le temps de réponse peut avoir un impact crucial sur le bien-être, voire la survie du patient. Le lien entre le temps de réponse et le succès de l'intervention demeure toutefois très difficile à déterminer, ce qui complique l'évaluation des performances réelles d'un tel système. Enfin, le déplacement des véhicules à travers un réseau routier congestionné amène une difficulté supplémentaire quant à l'évaluation du temps de déplacement et donc, du temps de réponse.

Bien que les différents SPU œuvrent dans des contextes relativement similaires, les règles et les processus déployés peuvent varier d'une ville à l'autre, d'un pays à l'autre. Toutefois, en analysant différents contextes (Gendreau *et al.*, 1997, 2001; Henderson et Mason, 2004; Andersson et Värbrand, 2007), il est possible d'identifier certaines caractéristiques communes quant à la gestion des SPU. Ainsi, dans la plupart des cas, le territoire à desservir est divisé en sous-territoires et une certaine forme de coopération existe parfois entre les véhicules affec-

tés aux différents sous-territoires. Certains SPU utilisent plusieurs types de véhicules afin de répondre aux demandes de service (premiers répondants, véhicules de soins de base (BLS), véhicules pour soins avancés (ALS), véhicules réguliers). Les nouvelles technologies, notamment les systèmes d'information géographique (SIG), sont également de plus en plus présentes, ce qui amène de nouvelles possibilités en matière de gestion. Enfin, le transport urgent et le transport inter-établissement sont gérés, dans certains cas, de façon simultanée, dans d'autres cas, de façon indépendante.

En analysant les différents SPU, il a également été possible de constater qu'ils semblent tous suivre le processus opérationnel suivant :

- Réception de l'appel ;
- Détermination de la priorité de l'appel ;
- Affectation d'un véhicule en considérant la priorité de l'appel (si aucun véhicule n'est disponible, l'appel est placé en attente) ;
- Déplacement du véhicule affecté vers le lieu de l'incident ;
- Soins au patient ;
- Transport vers un centre de traitement ou retour vers un poste d'attente (le véhicule redevient alors disponible pour l'affectation).

Les SPU classent les appels reçus en fonction d'un ensemble de priorités prédéfinies, puis appliquent les règles de gestion établies selon le cas à traiter. Lors de l'affectation d'un véhicule à un appel, le véhicule le plus proche est généralement envoyé sur les lieux de l'incident. Néanmoins, d'autres règles s'appliquent parfois lorsque, par exemple, plusieurs véhicules peuvent répondre à un appel à l'intérieur des délais prescrits.

Plusieurs questions émergent alors quant aux stratégies et aux moyens déployés afin de répondre le plus efficacement possible aux demandes de services. Combien de véhicules utiliser ? Où localiser les véhicules ? Quelle règle de répartition utiliser ? Plus concrètement, ces questions se traduisent par un ensemble de décisions à tous les niveaux de planification :

- Niveau stratégique : Localisation et construction des sites fixes (centres opérationnels ou garages), dimensionnement de la flotte de véhicules (type et nombre de véhicules), embauche du personnel.

- Niveau tactique : Localisation des postes d'attente, gestion du personnel (horaire, formation des équipes de travail), affectation des véhicules aux tâches (appels de détresse, transport inter-établissement) et aux postes d'attente.

- Niveau opérationnel : Règles de gestion (règles de répartition, politiques de redéploiement et de repositionnement, choix du centre hospitalier)

Ces décisions sont en général étroitement liées et peuvent avoir un impact considérable les unes sur les autres.

Jusqu'à maintenant, plusieurs auteurs se sont intéressés aux différentes décisions reliées à la gestion d'un service préhospitalier d'urgence et plus particulièrement aux problèmes de déploiement et de redéploiement des véhicules ambulanciers. Ce chapitre propose donc une revue des différents travaux en lien avec le déploiement et le redéploiement des véhicules d'urgence, dans le contexte particulier des SPU. Ainsi, dans un premier temps, les problèmes de déploiement et de redéploiement des véhicules ambulanciers sont brièvement décrits, puis situés par rapport aux problèmes généraux de localisation. Les approches de modélisation proposées dans la littérature, de même que les méthodes de résolution développées dans ce contexte sont ensuite abordées plus en détail. Puisque les règles de répartition, c'est-à-dire la sélection du véhicule à affecter à un appel, peuvent également avoir un impact important sur les performances du système et sur sa capacité à répondre aux demandes futures, les différents travaux à ce sujet sont également présentés brièvement. Finalement, différents commentaires reliés aux avenues de recherches potentielles viennent conclure ce chapitre.

D'autres considérations importantes sont inhérentes à la gestion des SPU, notamment quant à la prévision de la demande et au calcul adéquat des mesures de performance (Ingolfsson, 2013). Ces aspects ne seront toutefois pas abordés ici de manière systématique, mais plutôt discutés en temps opportun.

3.1 Problématique

Le déploiement et le redéploiement des véhicules ambulanciers concernent essentiellement la localisation des véhicules sur le territoire à desservir. Ces deux problèmes ont été abordés très brièvement en introduction. Dans cette section, ces deux problématiques seront présentées, puis situées par rapport aux problèmes généraux de localisation.

3.1.1 Déploiement des véhicules ambulanciers

Afin d'assurer un niveau de service adéquat à la population, les SPU utilisent un certain nombre de véhicules ambulanciers qu'ils positionnent de façon stratégique sur le territoire à desservir. Un plan de déploiement définit donc l'ensemble des sites sélectionnés pour la localisation d'un ou de plusieurs véhicules. Ainsi, entre deux affectations, les véhicules sont placés en attente aux différents sites qui constituent le plan de déploiement. Comme l'attente des véhicules ne requiert aucune installation particulière, les postes d'attente utilisés peuvent être aussi rudimentaires que le stationnement d'une station-service ou le coin d'une rue. Le problème de déploiement des véhicules ambulanciers consiste à déterminer le plan de déploiement à utiliser, c'est-à-dire à sélectionner les postes d'attente à utiliser parmi un ensemble de sites potentiels, puis à déterminer l'affectation des véhicules aux sites sélectionnés de façon à servir la population adéquatement tout en respectant un certain nombre de contraintes. Les contraintes à respecter de même que les objectifs visés peuvent néanmoins varier d'un cas à l'autre.

Dans un ouvrage paru en 1994, Swersey (1994) présente différentes problématiques liées à la gestion des services de police, d'incendie et préhospitaliers d'urgence. Il présente notamment les problèmes associés à la localisation des différents types de véhicules d'urgence. Ces problèmes possèdent tous certaines particularités qui leur sont propres. En effet, les voitures de police patrouillent géneralement les routes, ce qui n'est pas le cas pour les véhicules ambulanciers. De plus, tel que mentionné précédemment, la localisation des ambulances ne requiert aucune installation particulière, ce qui n'est pas le cas pour les services d'incendie. Ces problèmes présentent aussi diverses similitudes, notamment de par leurs objectifs. Dans tous les cas, l'objectif consiste à servir une population adéquatement. Le problème de déploiement des véhicules ambulanciers s'inscrit donc dans un cadre plus large qui regroupe l'ensemble des problèmes de localisation associés aux différents services d'urgence. Plus précisément, deux types de services d'urgence sont généralement considérés : les services avec serveurs mobiles où les serveurs se déplacent vers les clients et les services avec serveurs immobiles où les clients se déplacent vers le serveur. Bien que ces deux types de problèmes présentent des caractéristiques similaires, notamment du point de vue de la stochasticité et de la présence possible du phénomène de congestion, ils présentent également un certain nombre d'hypothèses et de caractéristiques différentes. Dans le cadre de ce chapitre, nous nous limiterons aux différents travaux en lien avec la localisation de sites d'urgence avec serveurs mobiles, puisqu'ils correspondent au problème de déploiement des véhicules ambulanciers. Le lecteur est invité à consulter l'article de Berman et Krass (2001) pour une discussion plus détaillée sur les différents modèles de localisation de sites d'urgence avec serveurs fixes.

Les problèmes de localisation liés aux services d'urgence doivent être considérés, de façon générale, comme des problèmes de localisation dans le secteur public qui incluent notamment la localisation de centres de santé, de bibliothèques et de stations de métro. Contrairement aux objectifs de minimisation des coûts ou de maximisation des profits observés dans le secteur privé, la localisation dans le secteur public vise généralement à offrir un service adéquat à la population. Ce type de localisation se caractérise également par l'incertitude reliée à la demande. En effet, dans le cas de la localisation de services d'urgence, le nombre d'appels et leur temps d'arrivée ne peuvent être prédits avec certitude. Ce phénomène d'incertitude s'observe également lors de la localisation d'une ligne de métro par exemple. De plus, dans certains cas, le volume de demandes est tel qu'il entraîne un phénomène de congestion, c'est-à-dire que les sites à localiser peuvent recevoir plus de demandes qu'ils ne peuvent en traiter. Toutes ces caractéristiques doivent donc être prises en compte lors de la résolution d'un problème de localisation dans le secteur public. Enfin, pour plus de détails au sujet des problèmes de localisation généraux, plus particulièrement ceux liés au secteur public et en contexte dynamique et stochastique, le lecteur est invité à consulter ReVelle *et al*. (1970), Marianov et Serra (2001), Berman et Krass (2001) et ReVelle et Eiselt (2005).

3.1.2 Redéploiement multi-période et dynamique des véhicules ambulanciers

Le problème de *déploiement* des véhicules ambulanciers consiste essentiellement à déterminer les différents points d'attente à utiliser pour la localisation des véhicules entre deux affectations. Le problème de *redéploiement* des véhicules ambulanciers consiste, quant à lui, à relocaliser les véhicules disponibles vers les différents points d'attente potentiels de façon à assurer, en tout temps, une couverture adéquate de la population. L'évolution du système dans le temps doit alors être considérée. Tout d'abord, l'évolution du système peut se traduire par des fluctuations de la demande dans le temps, dues notamment aux mouvements de population durant la journée. Différents plans de déploiement sont établis de façon à modifier le positionnement des véhicules entre les périodes afin de s'adapter aux changements prévus de la demande. On parlera alors de redéploiement *multi-période*. L'évolution du système peut également se traduire par une variation temporelle de l'état du système, c'est-à-dire le nombre de véhicules disponibles. Dans ce cas, les véhicules sont relocalisés lorsque l'état du système change et le justifie. Le réploiement vise alors à maintenir un niveau de service adéquat avec un nombre réduit de véhicules, certains d'entre eux étant maintenant en service (c'est-à-dire affectés à des incidents). Puisque le plan de redéploiement varie en fonction de l'état du système, on parlera plutôt de redéploiement *dynamique*.

Dans les faits, le problème de redéploiement des véhicules ambulanciers s'apparente au problème de déploiement statique. Certaines caractérisques les distinguent néanmoins (Gendreau *et al.*, 2001). Tout d'abord, le problème de déploiement est généralement résolu au niveau tactique. Le problème de redéploiement, principalement en ce qui concerne le redéploiement dynamique, est plutôt considéré au niveau opérationnel et doit, dans certains cas, être résolu en temps réel. En effet, les gestionnaires des SPU doivent généralement prendre des décisions très rapidement, voire de façon quasi-instantanée, quant à l'affectation et à la relocalisation des véhicules afin de maintenir un niveau de service adéquat. En plus des différences liées aux niveaux de planification, le problème de redéploiement considère habituellement un ensemble de contraintes supplémentaires issues de considérations pratiques visant à assurer la stabilité du système, ce qui n'est pas le cas pour le problème de déploiement statique. Des bornes sur le nombre maximal de véhicules pouvant être relocalisés ou sur la distance maximale de relocalisation peuvent être imposées. Ces contraintes visent généralement à établir un compromis entre le niveau de service offert à la population et les efforts liés au redéploiement.

3.2 Déploiement des véhicules ambulanciers

Le problème de déploiement des véhicules ambulanciers vise à déterminer la localisation des postes d'attente de façon à assurer un niveau de service adéquat à la population. Différentes approches ont été proposées afin de modéliser et de résoudre ce problème. Ainsi, dans cette section, les différentes approches de modélisation, de même que les principaux modèles développés en lien avec le problème de déploiement des véhicules ambulanciers sont abordés. Les différentes méthodes de résolution proposées dans ce contexte sont également présentées.

De manière générale, trois approches ont été proposées pour la modélisation du problème de déploiement des véhicules d'urgence : la programmation mathématique, la simulation et la théorie des files d'attente. Ces approches, de même que les principaux modèles qui y sont reliés, ont également été présentés par Goldberg (2004). Chacune de ces approches présente un certain nombre de caractéristiques qui lui sont propres. Les modèles issus de la programmation mathématique permettent la génération complète d'une solution sans nécessiter de solution initiale. La simulation et les modèles basés sur la théorie des files d'attente permettent plutôt l'évaluation d'une solution déjà établie. L'évaluation successive des différents scénarios considérés doit alors être effectuée afin d'en assurer la comparaison. Bien qu'ils ne puissent assurer la génération complète d'une solution, la simulation et les modèles basés sur la théorie des files d'attente permettent d'évaluer des solutions dans un contexte généralement plus réaliste que celui de la programmation mathématique. En effet, dans le cas de la programmation mathématique, un

grand nombre d'hypothèses simplificatrices sont souvent nécessaires afin de permettre le développement et la résolution de modèles de taille réelle dans des délais raisonnables. De toute évidence, chacune de ces approches permet de répondre à un certain nombre de besoins particuliers. Il revient au concepteur de sélectionner la méthodologie qui lui semble la plus adéquate dans le contexte auquel il s'intéresse. Certains auteurs ont également développé des modèles qui intègrent plus d'une approche de modélisation afin de bénéficier de leurs avantages respectifs.
Dans cette section, il sera question des travaux les plus importants en lien avec les trois approches mentionnées précédemment. Une attention plus particulière sera portée aux travaux en lien avec la programmation mathématique et la simulation. Nous effectuerons toutefois une brève présentation du modèle descriptif le plus utilisé, soit le modèle de l'hypercube de Larson (1974, 1975). Enfin, il est important de mentionner que cette revue vise à revoir les travaux qui sont, à notre avis, les plus pertinents, de même que les travaux qui ont été publiés plus récemment afin de constituer une synthèse représentative de l'état de l'art sur ce problème. Elle ne prétend pas couvrir tous les travaux en lien avec le déploiement des véhicules ambulanciers.

3.2.1 Programmation mathématique

Tel que mentionné précédemment, plusieurs auteurs se sont intéressés à la problématique de localisation de sites ou de véhicules d'urgence. Beaucoup de modèles basés sur la programmation mathématique ont donc été developpés au fil des années. Les modèles reliés à la localisation des véhicules d'urgence peuvent présenter différents objectifs. Certains modèles comportent des objectifs de minimisation de la distance ou du temps de réponse entre les sites sélectionnés et les zones de demande à desservir. Ces modèles considèrent alors que chaque zone de demande est couverte par le site le plus proche. Néanmoins, bien que chaque zone de demande soit couverte par le site ou le véhicule le plus proche, il est possible que celle-ci soit atteignable dans un délai de temps trop long par rapport aux objectifs généraux fixés par le SPU. Le concept de couverture a donc été introduit afin d'assurer un niveau de service acceptable à toute la population et ainsi pallier aux limites des modèles basés sur la minimisation de la distance. Une zone de demande est alors considérée couverte si elle est atteignable à l'intérieur d'une distance ou d'un délai prescrit. L'introduction du concept de couverture a donné lieu à une nouvelle famille de modèles de localisation, grandement utilisés pour la localisation des sites ou des véhicules d'urgence.

Les modèles proposés en lien avec le déploiement des véhicules ambulanciers peuvent être divisés en trois catégories :

- modèles déterministes à couverture simple ;
- modèles déterministes à couverture multiple ;
- modèles probabilistes et stochastiques.

Ces catégories suivent l'évolution chronologique des modèles proposés. En effet, au fil des années, les différents modèles ont évolué de façon à intégrer des aspects de plus en plus proches de la réalité : phénomène de congestion, incertitude de la demande, incertitude de la disponibilité des véhicules. ReVelle (1989), Marianov et ReVelle (1995) et Brotcorne *et al*. (2003) présentent un aperçu intéressant des différents modèles mathématiques appliqués au déploiement des véhicules d'urgence. Başar *et al*. (2012) proposent, quant à eux, une taxonomie des différents problèmes liés à la localisation des véhicules d'urgence.
Les modèles les plus importants en lien avec le déploiement statique des véhicules ambulanciers seront donc revus dans cette section, de même que les modèles développés plus récemment et qui n'ont pas été traités dans les revues mentionnées précédemment.
Avant d'amorcer la présentation formelle des différents modèles mathématiques, voici la notation qui sera utilisée jusqu'à la fin de ce chapitre. Certains éléments seront rappelés lors de la présentation des modèles et d'autres définitions seront apportées en cours de route, propres à chaque modèle. Les définitions suivantes demeurent toutefois communes à tous les modèles présentés. Tout d'abord, le problème de localisation des véhicules ambulanciers se définit sur un graphe $G = (V, E)$ où $V = N \cup M$, $N = \{v_1, ..., v_n\}$ et $M = \{v_{n+1}, ..., v_{n+m}\}$ représentent respectivement l'ensemble des zones de demande et des sites potentiels et où $E = \{(v_i, v_j) : v_i, v_j \in V, i < j\}$. Une zone de demande se définit comme un regroupement de population représenté par son centroïde et possédant une densité de population ou une demande donnée. Un site potentiel se définit plutôt comme un lieu physique où pourront être localisés un ou plusieurs véhicules et à partir duquel les véhicules se déplaceront pour rejoindre les différents appels de détresse. Ainsi, à chaque arête $(v_i, v_j) \in E$ est associé un temps de déplacement ou une distance d_{ij} et à chaque sommet $v_i \in N$ est associée une densité de population ou une demande a_i. Puisque la plupart des modèles utilisent le concept de couverture, M_i et M_i' sont définis comme les ensembles des sites potentiels pouvant assurer la couverture d'une zone de demande i respectivement à l'intérieur d'une distance ou d'un délai prescrit S et S', $S' > S$. L'ensemble N_j correspond à l'ensemble des zones de demande pouvant être couvertes par le site j à l'intérieur de S. Enfin, lorsqu'un nombre limité de véhicules sera considéré, ce nombre sera noté P.

3.2.1.1 Modèles déterministes à couverture simple

Toregas *et al*. (1971) ont été les premiers à formuler explicitement le problème de localisation des véhicules d'urgence en considérant le concept de couverture présenté ci-dessus. Le problème de localisation avec couverture totale (PLCT) ou *location set covering problem* (LSCP) en anglais, vise à déterminer une borne inférieure quant au nombre de véhicules à utiliser de façon à assurer la couverture de toutes les zones de demande. Ainsi, en définissant x_j, une variable binaire qui vaut 1 lorsqu'un véhicule est localisé au site j, et 0 autrement, et M_i, l'ensemble des sites potentiels pouvant assurer la couverture d'une zone de demande i à l'intérieur d'une distance ou d'un délai prescrit S, le modèle de Toregas *et al*. (1971) se formule de la façon suivante :

PLCT

$$\min \sum_{j=1}^{m} x_j \tag{3.1}$$

sous les contraintes :

$$\sum_{j \in M_i} x_j \geq 1, \ i = 1, ..., n, \tag{3.2}$$

$$x_j \in \{0, 1\}, \ j = 1, ..., m. \tag{3.3}$$

où m représente le nombre des sites potentiels et n, le nombre de zones de demande.

Concrètement, le PLCT vise à minimiser le nombre de véhicules nécessaires (3.1) de façon à garantir la couverture de toutes les zones de demande à l'intérieur de S (3.2), un seul véhicule pouvant être localisé à chaque site sélectionné (3.3). Dans l'article de Toregas *et al*. (1971), le PLCT a été appliqué à la localisation des casernes de pompiers pour la ville de New York. Le PLCT a alors été résolu par relaxation linéaire avec l'ajout de coupes. Bien qu'il ait d'abord été appliqué à la localisation des casernes de pompiers, ce modèle s'applique également, et de manière directe, à la localisation des véhicules ambulanciers.

Tel que mentionné précédemment, le PLCT vise à déterminer une borne inférieure quant au nombre de véhicules nécessaires afin d'assurer une couverture complète. Néanmoins, cette borne inférieure peut s'avérer très élevée, même irréalisable en pratique. Dans certains cas, il est donc préférable de déterminer la meilleure utilisation possible d'un nombre limité de véhicules. Ainsi, en considérant les limites pratiques associées aux solutions du PLCT, Church et ReVelle (1974) ont formulé le problème de localisation avec couverture maximale (PLCM) ou *maximal covering location problem* (MCLP) en anglais, qui vise à maximiser la population couverte en considérant un nombre donné de véhicules à localiser. En identifiant a_i, la densité

de population associée à la zone de demande i, y_i, une variable binaire qui vaut 1 si la zone de demande i est couverte par au moins un véhicule, et 0 autrement, et P, le nombre total de véhicules à localiser (les définitions de x_j et M_i demeurent les mêmes que dans le cas du PLCT), le PLCM se formule comme suit :

PLCM

$$\max \sum_{i=1}^{n} a_i y_i \tag{3.4}$$

sous les contraintes :

$$\sum_{j \in M_i} x_j \geq y_i, \ i = 1, ..., n, \tag{3.5}$$

$$\sum_{j=1}^{m} x_j = P, \tag{3.6}$$

$$x_j \in \{0, 1\}, \ j = 1, ..., m, \tag{3.7}$$

$$y_i \in \{0, 1\}, \ i = 1, ..., n. \tag{3.8}$$

Le PLCM vise donc à maximiser la population couverte (3.4) en localisant de manière optimale P véhicules (3.6). Bien que le PLCM assure une couverture maximale de la population à l'intérieur d'un délai ou d'une distance prédéfinie S, il peut être intéressant, dans certains cas, d'offrir un niveau de service minimum à l'ensemble de la population. Cela peut se traduire par l'ajout d'une contrainte qui garantit la couverture de toutes les zones de demande à l'intérieur d'une distance ou d'un délai S' moins strict, $S' > S$. À cet effet, Church et ReVelle (1974) ont proposé une contrainte supplémentaire, qu'ils appellent *mandatory closeness constraint*, très similaire à (3.2). En définissant M'_i comme l'ensemble des sites potentiels pouvant assurer la couverture d'une zone de demande i à l'intérieur de S', la contrainte proposée par Church et ReVelle se formule de la façon suivante :

$$\sum_{j \in M'_i} x_j \geq 1, \ i = 1, ..., n. \tag{3.9}$$

Aucune application précise du PLCM n'a été présentée dans l'article de Church et ReVelle (1974). Ceux-ci ont néanmoins proposé deux méthodes pour la résolution du PLCM, une méthode heuristique et une méthode de branchement, qu'ils ont appliqué par la suite à un problème test de 55 noeuds (comme ce problème a été grandement utilisé par la suite, nous y référerons comme au problème test classique). Galvão et ReVelle (1996) ont constaté que des méthodes heuristiques plus sophistiquées étaient toutefois nécessaires pour traiter des problèmes de plus grande taille. Ils ont donc proposé une heuristique lagrangienne pour la résolution du PLCM.

L'heuristique proposée se base sur la relaxation lagrangienne de la contrainte (3.5). À la suite de leurs expérimentations, Galvão et ReVelle (1996) ont pu conclure que l'heuristique proposée permettait d'obtenir de bons résultats pour des problèmes comportant jusqu'à 150 noeuds, et ce, dans des délais de temps raisonnables. Enfin, Eaton *et al*. (1985) ont appliqué le PLCM afin de déterminer la localisation des véhicules ambulanciers pour la ville d'Austin, au Texas. Dans ce cas, le modèle a été résolu grâce à la méthode heuristique de Church et ReVelle (1974). L'implantation de la solution proposée par Eaton *et al*. (1985) a mené à des bénéfices importants, notamment au niveau des coûts et du temps de réponse.

Les deux premiers modèles présentés considéraient l'utilisation d'un seul type de véhicule. Néanmoins, il n'est pas rare que des services d'urgence utilisent plus d'un type de véhicules simultanément. C'est notamment le cas pour les services d'incendie. À cet égard, Schilling *et al*. (1979) ont proposé trois modèles (*tandem equipment allocation model* (TEAM), *multiobjective tandem equipment allocation model* (MOTEAM), *facility-location, equipment-emplacement technique* (FLEET)) qui visent à maximiser la population couverte simultanément par deux types de véhicules différents. Les trois modèles proposés se distinguent principalement par le fait qu'ils acceptent ou non la localisation indépendante des différents types de véhicules. Tel que mentionné, ces modèles s'appliquent plus naturellement à la localisation dans le contexte des services d'incendie pour lesquels ils ont été conçus.

Ces modèles, principalement le PLCT et le PLCM, ont donné lieu à une quantité importante de variantes et d'extensions qui visent toujours à localiser les différents véhicules de façon à assurer un niveau de service adéquat à la population. Ainsi, bien que plutôt simples, ces modèles ont joué un rôle important dans le développement des modèles présentés dans les sous-sections suivantes. Ils ont également apporté une contribution intéressante par leur application en pratique.

3.2.1.2 Modèles déterministes à couverture multiple

Les modèles de localisation déterministes à couverture simple présentés à la section précédente présument qu'un véhicule est toujours disponible au moment où un appel d'urgence est reçu, ce qui n'est pas toujours le cas en pratique. En effet, un problème de disponibilité peut se présenter lorsque deux appels rapprochés sont placés dans une zone couverte par un seul véhicule. Le véhicule affecté à la zone de demande sert d'abord le premier appel reçu. Si le délai entre la réception du premier et du deuxième appel est trop court, le véhicule sera toujours en service et, par conséquent, non-disponible au moment où la deuxième demande de service se réalise. Les solutions déterminées grâce aux modèles de localisation à couverture simple sont généralement

peu robustes en pratique. Ainsi, les modèles à couverture multiple ont vu le jour. Ces modèles sont basés sur le principe qu'il est possible d'augmenter la probabilité pour qu'une zone de demande puisse trouver au moins un véhicule disponible dans les délais prescrits en augmentant le nombre de véhicules devant la couvrir à l'intérieur du délai prescrit. Bien que ces modèles ne soient pas probabilistes ou stochastiques en soi, ils constituent néanmoins une amélioration par rapport aux modèles précédents, puisqu'ils considèrent de façon indirecte l'aspect aléatoire des demandes.

Daskin et Stern (1981) ont été parmi les premiers à intégrer le concept de couverture multiple. En effet, le problème de couverture avec objectifs hiérarchiques (PCOH) ou *hierarchical objective set covering problem* (HOSC) en anglais, proposé par les auteurs, considère non seulement la première couverture, mais toutes les couvertures subséquentes. En définissant W, la pondération accordée au premier objectif, x_j, une variable binaire qui vaut 1 lorsqu'un véhicule est localisé au site j, et 0 autrement, s_i, le nombre de véhicules supplémentaires pouvant assurer la couverture de la zone de demande i, et N_i, l'ensemble des sites potentiels pouvant assurer la couverture d'une zone de demande i à l'intérieur d'une distance ou d'un délai prescrit S, le PCOH peut se formuler de la façon suivante :

PCOH

$$\min W \sum_{j=1}^{m} x_j - \sum_{i=1}^{n} s_i \tag{3.10}$$

sous les contraintes :

$$\sum_{j \in M_i} x_j - s_i \geq 1,\ i = 1,...,n, \tag{3.11}$$

$$x_j \in \{0,1\},\ j = 1,...,m, \tag{3.12}$$

$$s_i \geq 0,\ i = 1,...,n. \tag{3.13}$$

Le PCOH vise donc, dans un premier temps, à minimiser le nombre de véhicules nécessaires afin d'assurer la couverture de toutes les zones de demande à l'intérieur de S, puis, dans un deuxième temps, à maximiser la couverture multiple des zones de demande, c'est-à-dire le nombre de véhicules supplémentaires pouvant en assurer la couverture (3.10). Il est important de remarquer que le PCOH accorde une importance égale à tous les véhicules supplémentaires. Ceci peut avoir un effet pervers. En effet, en pratique, il peut sembler peu intéressant d'assurer la couverture d'une zone de demande par plus de deux véhicules si les deux premiers véhicules permettent de répondre largement aux demandes de service, advenant le cas que leur taux d'occupation est bas, par exemple. On peut alors penser que les troisième, quatrième et

cinquième véhicules contribueront peu à la couverture de cette zone de demande. Ainsi, même si, en théorie, le nombre de véhicules couvrant une zone de demande augmente, la couverture réelle espérée ne s'améliore plus. De plus, puisque le modèle ne considère pas la densité de population associée à chaque zone de demande, il aura tendance à regrouper les véhicules autour d'un certain nombre de zones faciles à couvrir, laissant ainsi des zones plus difficiles à desservir couvertes une seule fois. Daskin et Stern (1981) ont appliqué le PCOH à la localisation des véhicules médicaux d'urgence pour la ville d'Austin, au Texas. Le modèle a été résolu par relaxation linéaire et coupes. En comparant les résultats obtenus grâce au PLCT et au PCOH pour des instances de 33 noeuds ($m = n = 33$), les auteurs ont constaté que les solutions étaient comparables quant au nombre de véhicules nécessaires afin de garantir une couverture complète, mais que les solutions obtenues grâce au PCOH offraient une meilleure couverture multiple.

Eaton *et al*. (1986) ont tenté de pallier les faiblesses du PCOH lors de la résolution du problème de localisation des véhicules médicaux d'urgence pour la ville de Santo Domingo, en République dominicaine. Ils ont modifié la formulation de Daskin de façon à considérer la densité de population associée à chaque zone de demande. Le problème dominicain de déploiement des ambulances (PDDA) formulé par Eaton *et al*. (1986) (ou *Dominican ambulance deployment problem* (DADP) en anglais) vise donc à maximiser la population couverte plusieurs fois tout en minimisant le nombre de véhicules nécessaires afin d'assurer la couverture de toutes les zones de demande. Les véhicules supplémentaires (couverture multiple) ont toujours la même contribution dans la fonction objectif, ce qui ne permet pas d'éviter le problème d'accumulation inutile de véhicules dans des zones faciles à desservir. La formulation du PDDA est donc très similaire à la formulation du PCOH, la différence fondamentale repose sur l'ajout d'un facteur associé à la densité de population a_i dans la fonction objectif :

$$\min W \sum_{j=1}^{m} x_j - \sum_{i=1}^{n} a_i s_i. \tag{3.14}$$

Dans cette étude, étant donné la taille du problème considéré ($m = n = 214$), les auteurs ont résolu le PDDA grâce à une méthode heuristique multi-objectif.

Parallèlement à Eaton *et al*. (1986), Hogan et ReVelle (1986) ont également tenté de remédier aux problèmes associés au PCOH, en considérant non seulement la densité de population, mais également en accordant une importance hiérarchique aux différents niveaux de couverture. Les auteurs formulent donc deux modèles où seules la première et la deuxième couverture sont considérées dans la fonction objectif, ce qui permet de pallier au problème des modèles précédents quant à l'inégalité de la localisation des véhicules. En considérant maintenant u_i, une variable binaire qui vaut 1 lorsqu'une zone de demande i est couverte au moins *deux* fois, et

0 autrement, et x_j, une variable entière associée au nombre de véhicules localisés au site j, le premier modèle qu'ils ont proposé, soit le modèle de localisation avec couverture secondaire maximale (MLCSM), ou *maximal backup coverage model 1* (BACOP 1) en anglais, se formule de la façon suivante :

MLCSM

$$\max \sum_{i=1}^{n} a_i u_i \tag{3.15}$$

sous les contraintes :

$$\sum_{j \in M_i} x_j - u_i \geq 1,\ i = 1, ..., n, \tag{3.16}$$

$$\sum_{j=1}^{m} x_j = P_{min}, \tag{3.17}$$

$$0 \leq u_i \leq 1,\ i = 1, ..., n, \tag{3.18}$$

$$x_j \geq 0,\ j = 1, ..., m. \tag{3.19}$$

Le MLCSM vise à maximiser la population couverte deux fois à l'intérieur de la distance ou du délai prescrit S (3.15) en considérant le nombre minimal de véhicules nécessaires pour assurer une couverture totale (3.17). La valeur de P_{min} peut alors être déterminée par la résolution du PLCT correspondant. Plus concrètement, le MLCSM cherche la meilleure configuration possible, en termes de couverture secondaire, lorsque P_{min} véhicules sont employés. En comparant les solutions du PLCT original et de MLCSM, les auteurs ont pu constater un gain de l'ordre de $16,2\%$ quant à la couverture secondaire.

Le deuxième modèle proposé par Hogan et ReVelle (1986), soit le modèle de localisation avec couverture secondaire (MLCS), ou *backup coverage model 2* (BACOP2) en anglais, vise plutôt la maximisation simultanée de la première et de la deuxième couverture (3.20) en utilisant un nombre donné de véhicules (3.23). En considérant maintenant w, un paramètre visant à accorder une importance relative au premier et au deuxième objectif et prenant une valeur entre 0 et 1, et y_i, une variable binaire qui vaut 1 si la zone de demande i est couverte au moins une fois, et 0 autrement, le MLCS se formule de la façon suivante :

MLCS

$$\max w \sum_{i=1}^{n} a_i y_i + (1-w) \sum_{i=1}^{n} a_i u_i \tag{3.20}$$

sous les contraintes :

$$\sum_{j \in M_i} x_j - y_i - u_i \geq 0,\ i = 1, ..., n, \tag{3.21}$$

$$u_i - y_i \leq 0,\ i = 1, ..., n, \tag{3.22}$$

$$\sum_{j=1}^{m} x_j = P, \tag{3.23}$$

$$0 \leq u_i \leq 1,\ 0 \leq y_i \leq 1,\ i = 1, ..., n, \tag{3.24}$$

$$x_j \geq 0,\ j = 1, ..., m. \tag{3.25}$$

Il est important de constater que, dans le cas des modèles de localisation avec couverture secondaire, la variable x_j prend des valeurs entières, ce qui signifie que plusieurs véhicules peuvent être localisés au même site. La co-localisation des véhicules est donc possible. Hogan et ReVelle (1986) ont résolu une série d'instances tirées du problème test classique (55 noeuds) pour les deux modèles proposés par relaxation linéaire en appliquant au besoin des méthodes de branchement afin d'obtenir une solution entière. Dans le cas du MLCS, les auteurs ont constaté, à la suite de leurs expérimentations, que lorsqu'une importance plus grande est accordée à la deuxième couverture (c'est-à-dire w est plus petit), la première couverture est réduite, mais un gain considérable est observé quant à la couverture secondaire. Ils ont également noté que les modèles proposés pourraient être étendus facilement pour tenir compte de niveaux de couverture supplémentaires. À cet effet, ils ont proposé une extension de MLCS qui considère la troisième couverture. Dans tous les cas, des contraintes de type *mandatory closeness* peuvent également être ajoutées aux modèles.

Plus récemment, Gendreau *et al*. (1997) ont proposé un autre modèle, celui avec double standard (MDS) ou *double standard model* (DSM) en anglais, qui considère simultanément l'idée de double couverture et l'application de différents rayons de couverture. En fait, le MDS s'inspire de la norme émise par le *United States Emergency Medical Services (EMS) Act of 1973* qui stipule qu'une proportion α de la population doit être atteignable à l'intérieur d'un temps S, tandis que la totalité de la population doit être atteignable à l'intérieur d'un temps S', $S' > S$. Gendreau *et al*. (1997) ont donc formulé un modèle dont la solution respecte ces deux contraintes. Ainsi, en considérant maintenant p_j, la limite sur le nombre de véhicules à localiser au site j et M'_i, l'ensemble des sites potentiels pouvant assurer la couverture d'une zone de demande i à l'intérieur d'une distance ou d'un délai prescrit S' (les définitions de x_j, y_i, u_i, a_i, M_i demeurent les mêmes que pour le MLSC), le modèle de Gendreau *et al*. (1997) se formule comme suit :

MDS

$$\max \sum_{i=1}^{n} a_i u_i \tag{3.26}$$

sous les contraintes :

$$\sum_{j \in M'_i} x_j \geq 1,\ i = 1, ..., n, \tag{3.27}$$

$$\sum_{i=1}^{n} a_i y_i \geq \alpha \sum_{i=1}^{n} a_i, \tag{3.28}$$

$$\sum_{j \in M_i} x_j \geq u_i + y_i,\ i = 1, ..., n, \tag{3.29}$$

$$u_i \leq y_i,\ i = 1, ..., n, \tag{3.30}$$

$$\sum_{j=1}^{m} x_j = P, \tag{3.31}$$

$$x_j \leq p_j,\ j = 1, ..., m, \tag{3.32}$$

$$u_i, y_i \in \{0, 1\},\ i = 1, ..., n, \tag{3.33}$$

$$x_j \geq 0,\ \text{entier},\ j = 1, ..., m. \tag{3.34}$$

Le MDS vise donc à maximiser la population couverte au moins deux fois à l'intérieur de S (3.26) en assurant qu'une proportion minimale α de la population soit couverte à l'intérieur de S (3.28) et que la totalité de la population soit couverte à l'intérieur de S' (3.27). La localisation d'un nombre donné de véhicules (3.31) est considérée. Dans ce cas, la co-localisation des véhicules ambulanciers est possible, mais contrôlée par une série de paramètres prédéterminés p_j (3.32). Le modèle proposé par Gendreau *et al*. (1997) partage plusieurs caractéristiques des MLSCM et MLCS. La principale contribution du MDS repose sur l'ajout de la contrainte (3.28) visant à assurer la couverture d'une proportion α de la population à l'intérieur de S. Naturellement, le modèle peut devenir non-réalisable si le nombre de véhicules disponibles est insuffisant pour satisfaire simultanément les contraintes (3.27) et (3.28). Enfin, les auteurs proposent une méthode basée sur la recherche avec tabous pour la résolution du MDS. Pour l'ensemble des tests effectués, notamment sur les données d'Urgences-santé, une corporation responsable des SPU pour la grande région de Montréal, des solutions de bonne qualité ont pu être déterminées en quelques minutes seulement, ce qui vient confirmer la possibilité d'utiliser la méthode proposée en pratique. Les instances générées à partir des données d'Urgences-santé comportaient 2521 zones de demande, de 40 à 70 sites potentiels et de 25 à 40 véhicules à localiser.

Doerner *et al*. (2005) ont modifié le modèle proposé par Gendreau *et al*. (1997) et l'ont appliqué à la localisation des véhicules ambulanciers en Autriche. La première modification proposée repose sur la relaxation des contraintes (3.27) et (3.28) et leur intégration au sein de la fonction objectif sous la forme de termes de pénalité. La prise en compte de ces contraintes dans la

fonction objectif, plutôt que sous la forme de contraintes dures, permet alors de déterminer une solution *réalisable* en tout temps, ce qui n'était pas le cas pour le MDS. Doerner *et al*. (2005) ont également proposé l'intégration d'un troisième terme de pénalité au sein de la fonction objectif afin de limiter le nombre d'habitants affectés à un même véhicule. En effet, les auteurs ont soulevé le fait que, dans la plupart des modèles proposés, les solutions sont telles que certains véhicules sont surutilisés, c'est-à-dire qu'ils couvrent une grande proportion de la population à l'intérieur du délai prescrit, tandis que d'autres véhicules sont sous-utilisés. Une telle affectation peut devenir irréaliste lorsque la capacité des véhicules est considérée. Limiter la proportion de la population couverte par un véhicule permet d'éviter une telle situation. Afin de résoudre ce problème, l'algorithme basé sur la recherche avec tabous développé par Gendreau *et al*. (1997) a été adapté. Une méthode basée sur les colonies de fourmis a également été proposée par Doerner *et al*. (2005) Les expérimentations effectuées sur huit provinces rurales d'Autriche, comportant de 105 à 703 zones de demande, 7 à 137 localisations potentielles et de 11 à 200 véhicules ambulanciers, ont permis de montrer que les deux approches obtiennent de bons résultats. La recherche avec tabous fournit toutefois des résultats dans des délais de temps plus courts.

Laporte *et al*. (2009) ont rapporté récemment différentes applications du MDS, dont celle à la ville de Montréal au Canada et celle aux provinces autrichiennes présentée ci-dessus, de même qu'une à la Wallonie, en Belgique. Ces études ont permis de démontrer le potentiel d'utilisation de ce modèle et des algorithmes proposés pour le résoudre dans des contextes d'application réels.

Dans un autre ordre d'idée, Storbeck (1982) a proposé une formulation plus flexible, basée sur la programmation par objectifs ou *goal programming* en anglais. Le problème de localisation avec couverture maximale et multiple (PLMM) (ou *maximal-multiple location covering problem* (MMLCP) en anglais) qu'il présente vise à minimiser la population non couverte et à maximiser la couverture multiple (3.35) en considérant un nombre donné de véhicules à localiser (3.37). Ce problème considère donc simultanément plusieurs objectifs des modèles présentés précédemment. Selon ReVelle (1989), la formulation proposée par Storbeck est très versatile et figure parmi les plus intelligentes. En considérant maintenant z_i, une variable binaire qui vaut 1 si la zone de demande n'est pas couverte par au moins un véhicule, et 0 autrement (les définitions des autres variables et paramètres demeurent les mêmes que pour les modèles précédents), le modèle se formule comme suit :

PLMM

$$\min W \sum_{i=1}^{n} a_i z_i - \sum_{i=1}^{n} s_i \tag{3.35}$$

sous les contraintes :

$$\sum_{j \in M_i} x_j - s_i + z_i \geq 1,\ i = 1, ..., n, \tag{3.36}$$

$$\sum_{j=1}^{m} x_j = P, \tag{3.37}$$

$$x_j, z_i \in \{0, 1\},\ i = 1, ..., n,\ j = 1, .., m, \tag{3.38}$$

$$s_i \geq 0,\ i = 1, ..., n. \tag{3.39}$$

Storbeck a appliqué le PLMM à un problème simple de 16 noeuds ($m = n = 16$ et $P = 3$). À la suite de ses expérimentations, il a pu constater que la solution obtenue grâce au PLMM présentait le même nombre de zones de demande non couvertes que la solution obtenue grâce au PLCM, mais qu'elle offrait une amélioration importante en ce qui concerne la couverture multiple. Aucune indication n'a été donnée quant à la méthode utilisée pour résoudre le problème.

En somme, les modèles de localisation à couverture multiple viennent compléter les modèles de localisation à couverture simple en considérant maintenant le fait qu'un véhicule puisse être non disponible au moment où la demande se réalise. Ainsi, en assurant une couverture double ou multiple, la robustesse du système peut être améliorée.

3.2.1.3 Modèles probabilistes ou stochastiques

Les modèles de localisation à couverture multiple considèrent le fait que les plans de déploiement générés deviennent suffisamment robustes lorsque la double couverture ou la couverture multiple des zones de demande est assurée. Bien que ces modèles offrent une amélioration par rapport aux modèles de localisation déterministes à couverture simple, ils présentent néanmoins certaines limites. En effet, il est possible qu'en pratique la double couverture ne puisse assurer un niveau de service satisfaisant ou, à l'inverse, qu'elle ne soit tout simplement pas nécessaire. Afin de pallier aux limites des modèles à couverture multiple, certains auteurs ont décidé de considérer plus explicitement les différentes sources d'incertitude liées aux services préhospitaliers d'urgence. Les modèles de localisation probabilistes ou stochastiques ont donc été développés afin de fournir une représentation plus fidèle de la réalité. Il n'en demeure pas moins que plusieurs hypothèses simplificatrices sont toujours nécessaires au développement de tels modèles et que diverses améliorations peuvent toujours y être apportées. La façon de considérer les différentes sources d'incertitude diffère d'un modèle à l'autre donnant lieu à une

quantité importante de modèles aux hypothèses et aux objectifs variés.

La première famille de modèles probabilistes décrites sont les modèles de localisation à couverture espérée. Leur fonction objectif vise donc à maximiser la couverture espérée exprimée en fonction du taux d'occupation des véhicules où le taux d'occupation se définit comme la probabilité qu'un véhicule ne puisse être disponible pour répondre à un appel. Le fait de considérer ainsi la disponibilité des véhicules constitue une des approches possibles afin de prendre en compte l'incertitude associée à la réalisation des demandes de service. Dans certains cas, la couverture espérée s'exprime en fonction de la probabilité pour qu'un véhicule puisse atteindre une zone de demande dans les délais prescrits, ce qui permet alors de considérer des temps de déplacement stochastiques. Erkut *et al*. (2009) ont analysé le fait de considérer ou non les aspects stochastiques liés à la disponibilité des véhicules et au temps de déplacement à l'intérieur d'un modèle de localisation avec couverture maximale. L'analyse des modèles menée grâce aux données de la ville d'Edmonton, en Alberta, au Canada, ont permis de montrer que, pour un nombre donné de véhicules, les solutions déterminées grâce aux modèles de localisation avec couverture espérée obtiennent de meilleures performances par rapport aux solutions déterminées par la résolution du PLCM déterministe, et ce, jusqu'à 26 % en termes de couverture réelle. Le fait de considérer l'incertitude lors de la formulation des modèles de localisation présente donc des avantages certains.

Daskin (1982, 1983) fut un des premiers à intégrer la disponibilité réelle des véhicules dans un modèle de localisation avec couverture maximale. Il pose alors comme hypothèse que tous les véhicules présentent le même taux d'occupation q qui se définit comme le temps total nécessaire pour répondre à tous les appels divisé par le temps total de disponibilité pour tous les véhicules. Il considère également que le taux d'occupation est indépendant de la localisation et que chaque véhicule opère indépendamment, c'est-à-dire qu'il n'y a pas de coopération entre les véhicules affectés aux différentes zones de demande. Ainsi, si une zone de demande i est couverte par k véhicules, la demande espérée couverte est donnée par $E_k = a_i(1-q^k)$, où $1-q^k$ représente la probabilité pour qu'au moins un véhicule soit libre et a_i, le nombre de demandes associées à la zone de demande i. La contribution marginale du k-ième véhicule à la couverture des demandes placées en i est alors donnée par $E_k - E_{k-1} = a_i(1-q)q^{k-1}$. Ainsi, en considérant une variable binaire y_{ik} qui vaut 1 si la zone de demande i est couverte par au moins k véhicules, et 0 autrement, x_j, le nombre de véhicules localisés en j, et P, le nombre maximum de véhicules à localiser, le problème de localisation avec couverture espérée maximale (PLCEM) ou *maximum expected covering location problem* (MEXCLP) en anglais, proposé par Daskin, se formule de la façon suivante :

PLCEM

$$\max \sum_{i=1}^{n} \sum_{k=1}^{P} (1-q) q^{k-1} a_i y_{ik} \tag{3.40}$$

sous les contraintes :

$$\sum_{j \in M_i} x_j \geq \sum_{k=1}^{P} y_{ik}, \; i = 1, ..., n, \tag{3.41}$$

$$\sum_{j=1}^{m} x_j \leq P, \tag{3.42}$$

$$y_{ik} \in \{0,1\}, \; i = 1, ..., n, \; k = 1, ..., P, \tag{3.43}$$

$$x_j \geq 0, \text{ entier}, \; j = 1, ..., m. \tag{3.44}$$

Le PLCEM vise donc à maximiser la couverture espérée (3.40) en considérant la localisation d'un nombre limité de véhicules (3.42). Dans ce cas, la couverture maximale se traduit par la contribution marginale de chaque véhicule à la couverture globale. Contrairement à la formulation originale du PLCM, la co-localisation des véhicules est ici possible. De plus, il est important de noter que le modèle présente une fonction objectif concave, ce qui implique que l'inégalité $y_{i,k-1} \leq y_{ik}$ est toujours respectée. Afin de résoudre le PLCEM, Daskin (1982) a d'abord proposé l'utilisation d'une méthode de séparation et évaluation progressive. Cette approche de résolution permet de déterminer une solution optimale, mais seulement pour une valeur donnée de q. Dans un deuxième temps, Daskin (1983) a proposé une méthode heuristique basée sur des échanges simples afin de déterminer un ensemble de bonnes solutions pour la gamme complète des q. La sensibilité de la solution à la variation de q peut alors être observée. À la suite des expérimentations menées sur le problème test classique, les auteurs ont pu constater que, pour des petites valeurs de q, la solution obtenue grâce au PLCEM était équivalente, en termes de population non couverte, à celle obtenue grâce au PLCM. Pour des valeurs de q plus grandes que 0,0699, les solutions diffèrent : la couverture multiple croît alors au détriment de la première couverture. Plus récemment, Aytug et Saydam (2002) ont proposé un algorithme génétique pour la résolution du PLCEM. En comparant les résultats obtenus grâce à l'algorithme génétique proposé à ceux obtenus grâce à la méthode heuristique de Daskin (1983) pour des instances de taille variant entre 400 et 1600 noeuds ($m = n = 400, ..., 1600$), les auteurs observent que leur algorithme s'avère très intéressant en termes de qualité et d'effort de calcul lorsque la taille des problèmes traités devient plus importante. Enfin, Fujiwara *et al.* (1987) ont appliqué directement le modèle PLCEM à la localisation des véhicules ambulanciers pour la ville de Bangkok, en Thaïlande. Dans ce cas, le modèle formulé a été résolu par la méthode heuristique de Daskin (1983). Grâce aux résultats de l'expérimentation, les auteurs ont pu

conclure qu'en relocalisant les véhicules disponibles sur le territoire à desservir, un niveau de service équivalent pourrait être maintenu avec 15 véhicules par rapport aux 21 véhicules utilisés initialement, et ce, même après une augmentation de la demande.

Plusieurs variantes du problème du PLCEM ont été proposées dans la littérature. Tout d'abord, Bianchi et Church (1988) ont proposé une variante du PLCEM, MOFLEET, qui considère maintenant de façon indépendante la localisation des postes d'attente et l'allocation des véhicules. Plutôt que de considérer exclusivement la localisation d'un nombre donné de véhicules comme c'est le cas pour le PLCEM, MOFLEET propose une distinction entre le nombre de sites à localiser et le nombre de véhicules à allouer. Tout dépendant des valeurs des paramètres du modèle, le nombre de sites à localiser pourra, et même devra, être inférieur au nombre de véhicules à localiser. Daskin *et al*. (1988) ont ensuite proposé une généralisation du PLCEM qui considère différents délais prescrits pour les différents niveaux de couverture. Enfin, Repede et Bernardo (1994) ont proposé une version multi-période du PLCEM que nous discuterons plus en détail à la section 3.3.

Le PLCEM de même que les variantes présentées jusqu'ici considéraient trois hypothèses importantes : le taux d'occupation est connu et identique pour tous les véhicules, le taux d'occupation est indépendant de la localisation et chaque véhicule opère indépendamment. Malheureusement, tel que noté par Batta *et al*. (1989), ces hypothèses ne sont généralement pas respectées en pratique, ce qui peut avoir un impact non négligeable sur l'écart entre les performances réelles du système et celles prédites par le modèle. À la suite de leurs observations, Batta *et al*. (1989) ont proposé deux nouvelles variantes visant à relaxer certaines de ces hypothèses dans le but d'obtenir une estimation plus réaliste de la couverture espérée. Le premier modèle qu'ils ont proposé, le PLCEM ajusté (PLCEM-A), est très similiaire au PLCEM. En fait, le PLCEM-A se distingue du modèle original par le fait qu'il considère, à même la fonction objectif, un facteur correctif basé sur la théorie des files d'attente (Larson, 1975) qui permet la relaxation de l'hypothèse d'indépendance des véhicules. L'ajout du facteur correctif complique néanmoins le modèle final qui devient alors non linéaire. Dans ce cas, Batta *et al*. (1989) ont proposé une adaptation de la méthode heuristique de Daskin (1983) afin d'en permettre la résolution.

Le deuxième modèle proposé par Batta *et al*. (1989) se base essentiellement sur le modèle de l'hypercube de Larson (1974, 1975). Le modèle de l'hypercube de Larson sera traité plus en détail à la section 3.2.3. En fait, les auteurs proposent une procédure d'optimisation qui utilise le modèle de l'hypercube afin de mesurer la couverture espérée pour un ensemble donné de localisations. De cette façon, les trois hypothèses précédentes peuvent être relaxées. De plus, tel que mentionné par Chiyoshi *et al*. (2002), l'utilisation du modèle de l'hypercube afin d'évaluer la

couverture espérée permet de considérer la couverture des appels placés en attente, ce qui n'est pas le cas dans les modèles précédents. Le modèle de l'hypercube est donc intégré dans une méthode heuristique de descente qui vise à déterminer la localisation optimale des véhicules de façon à maximiser la couverture espérée. En observant les différents résultats obtenus pour des instances tirées du problème test classique ($m = n = 55$ et $P = 3$), les auteurs ont constaté que, bien que très similaires en termes de localisations physiques, les solutions obtenues grâce au PLCEM surestiment la couverture espérée. Celles obtenues par le PLCEM-A sous-estiment la couverture espérée pour des petites valeurs de q et la surestiment pour de grandes valeurs de q. Dans le même ordre d'idée, Galvão *et al*. (2005) ont proposé plus récemment une méthode basée sur le recuit simulé pour résoudre ce problème. De façon similaire à Batta *et al.*, la méthode heuristique de Galvão *et al*. (2005) utilise le modèle de l'hypercube afin de calculer la couverture espérée. Pour des instances variant entre 50 et 150 noeuds, le recuit simulé proposé permet une amélioration de la solution d'au plus 1% et ce, pour des temps de calcul jusqu'à 4,55 fois plus grand par rapport à la méthode de descente proposée par Batta *et al*. (1989).

Tous les modèles présentés jusqu'à maintenant considéraient des temps de déplacement déterministes. Toutefois, dans la réalité, il est fort possible que le temps de déplacement entre deux points puisse varier d'une demande d'intervention à l'autre, dû notamment à la congestion du réseau routier. Ainsi, Daskin (1987) a proposé un modèle visant à déterminer la localisation des véhicules d'urgence, l'affectation des différents véhicules aux zones de demande et la route à suivre afin d'atteindre une zone de demande à partir d'une localisation donnée en considérant des temps de déplacement stochastiques. La couverture espérée s'exprime alors en fonction de la probabilité qu'un véhicule puisse atteindre une zone de demande dans les délais prescrits, probabilité qui, dans ce cas, tient compte exclusivement de l'incertitude liée au temps de déplacement. On pose alors l'hypothèse qu'un véhicule est toujours disponible lorsqu'une demande d'intervention se réalise. Le modèle proposé par Daskin (1987) vise donc à maximiser la couverture espérée tout en minimisant le temps de réponse moyen dans le cas particulier où deux véhicules d'urgence doivent être envoyés simultanément sur le lieu d'un incident.

Goldberg *et al*. (1990b) ont proposé, puis appliqué à la ville de Tucson, en Arizona, un modèle qui possède des objectifs de couverture espérée similaires au PLCEM et à ses variantes. Le modèle vise toujours à maximiser la couverture espérée à l'intérieur d'un délai prescrit S en localisant un nombre donné de véhicules, mais considère aussi des temps de déplacement stochastiques, comme dans le cas du modèle de Daskin (1987). De plus, on y pose l'hypothèse que la répartition des véhicules est effectuée selon une liste de préférences, où k représente la position d'un site sur cette liste classée en ordre de priorité décroissant. Ainsi, la couverture es-

pérée s'exprime en fonction de la probabilité d'atteindre une zone de demande à l'intérieur du délai prescrit. Cette probabilité est déterminée en considérant la probabilité pour que le k-ième véhicule soit disponible et qu'il puisse atteindre la zone de demande en question dans le délai prescrit, de même que la probabilité pour que les véhicules localisés au $k-1$ sites plus haut dans la liste soient occupés. La probabilité pour qu'un véhicule soit occupé est indépendante de l'état du système et est calculée grâce à la théorie des files d'attente. Lorsque les localisations sont fixées, le modèle de Goldberg *et al*. (1990b) permet de calculer facilement un ensemble de mesures de performance. Il peut donc être utilisé comme un modèle descriptif ou à l'intérieur d'une approche d'optimisation. En ce sens, Goldberg et Paz (1991) ont repris le modèle proposé par Goldberg *et al*. (1990b), puis ont proposé une méthode heuristique basée sur des échanges simples afin d'en assurer la résolution.

Plus récemment, Ingolfsson *et al*. (2008) ont aussi proposé un modèle inspiré de celui de Goldberg *et al*. (1990b). Il s'en distingue toutefois puisqu'il considère, en plus des aspects stochastiques reliés à la disponbilité des véhicules et au temps de déplacement, la variabilité associée au temps écoulé entre la réception d'un appel et l'affectation d'un véhicule. En effet, les auteurs ont constaté que, dans certains cas, la variabilité du temps écoulé avant l'affectation d'un véhicule pouvait être importante. De plus, ils ont montré que les solutions obtenues diffèrent lorsque l'on considère des délais d'affectation stochastiques.

Enfin, Mandell (1988) a proposé un modèle de localisation avec couverture espérée qui considère l'utilisation de deux types de véhicules, les véhicules ALS et BLS, au sein d'un système inclusif, c'est-à-dire qu'un véhicule ALS peut assurer le même service qu'un BLS, mais pas l'inverse. Le modèle formulé par Mandell (1988) vise à maximiser le nombre espéré d'appels servis adéquatement grâce à la localisation d'un nombre donné de véhicules de type ALS et BLS. Dans le présent contexte, une zone de demande est servie adéquatement si un véhicule BLS et un véhicule ALS peuvent arriver sur les lieux de l'incident à l'intérieur d'un temps t_b et t_a, respectivement, ou encore si un véhicule ALS peut s'y rendre en un temps inférieur à t_b. Afin de formuler adéquatement le modèle, on doit pouvoir déterminer la probabilité pour qu'une zone de demande i puisse être servie adéquatement, étant donné r véhicules ALS, r' véhicules ALS et s véhicules BLS localisés de façon à pouvoir atteindre i en moins de t_a, t_b et t_b, respectivement. Ces probabilités sont obtenues en partie grâce à un modèle markovien. Enfin, bien que le modèle considère qu'un seul véhicule de chaque type puisse être localisé à un site, la co-localisation des véhicules ALS et BLS est permise.

D'une manière ou d'une autre, tous les modèles présentés jusqu'ici visaient à maximiser la couverture espérée en considérant la probabilité qu'un véhicule puisse servir ou non une zone

de demande à l'intérieur d'un délai prescrit. Une seconde approche a également été proposée afin de considérer les différentes source d'incertitudes reliées au problème de déploiement des véhicules ambulanciers : la programmation stochastique avec contraintes probabilistes (Birge et Louveaux, 2011). Contrairement aux modèles à couverture espérée, cette approche vise à considérer l'incertitude lors de la formulation des contraintes de façon à assurer la fiabilité du système, c'est-à-dire à garantir le respect de certaines contraintes avec un niveau de fiabilité donné. Dans le cas du déploiement des véhicules ambulanciers, les modèles forcent généralement le respect des contraintes de couverture. Une borne inférieure quant au nombre de véhicules nécessaires afin d'assurer une couverture adéquate est alors déterminée pour chaque zone de demande. La façon de déterminer cette borne inférieure et d'estimer le taux d'occupation des véhicules varie toutefois d'un modèle à l'autre.

En suivant l'approche de la programmation stochastique avec contraintes probabilistes, ReVelle et Hogan (1989) ont proposé le problème de localisation avec disponibilité maximale (PLDM), ou *maximum availability location problem* (MALP) en anglais, une version probabiliste du PLCM. Le PLDM vise à maximiser la population couverte avec un niveau de fiabilité donné. Le modèle garantit alors que chaque demande puisse trouver au moins un véhicule disponible à l'intérieur du délai prescrit S, et ce, avec une fiabilité α ou, plus précisément, que $1-q^k \geq \alpha$, où $k = \sum_{j \in N_i} x_j$ et $1-q^k$ représente la probabilité pour qu'au moins un véhicule soit disponible pour répondre à une demande placée en i étant donné l véhicules pour en assurer la couverture. Dans ce cas, le taux d'occupation est le même pour toutes les zones de demande et est déterminé de façon similiaire au cas du PLCEM (Daskin, 1982). La contrainte de couverture proposée peut alors se réécrire sous sa forme linéaire, $\sum_{j \in N_i} x_j \geq b$, où $b = \lceil \frac{\log(1-\alpha)}{\log q} \rceil$. Ainsi, en considérant une variable binaire y_{ik}, qui vaut 1 si le point de demande i est couvert par au moins k véhicules, une variable binaire x_j, qui vaut 1 si un véhicule est localisé au site j, a_i, le nombre de demandes associées à la zone de demande i, et P, le nombre total de véhicules à localiser, le PLDM1 se formule comme suit :

PLDM1

$$\max \sum_{i=1}^{n} a_i y_{ib} \tag{3.45}$$

sous les contraintes :

$$\sum_{k=1}^{b} y_{ik} \leq \sum_{j \in M_i} x_j, \; i = 1, ..., n, \tag{3.46}$$

$$y_{ik} \leq y_{ik-1}, \; i \in I, \; k = 2, ..., b, \tag{3.47}$$

$$\sum_{j=1}^{m} x_j = P, \tag{3.48}$$

$$x_j, y_{ik} \in \{0,1\},\ i = 1,...,n,\ j = 1,...,m,\ k = 1,...,P. \tag{3.49}$$

De manière plus concrète, le PLDM1 vise à maximiser la population servie par b véhicules (3.45) où b est le nombre de véhicules requis afin de garantir que chaque demande puisse trouver au moins un véhicule disponible à l'intérieur de S au moment où celle-ci se réalise, et ce, avec une fiabilité α. Dans ce cas, la co-localisation des véhicules n'est pas possible (3.49). La localisation indépendante de P véhicules est donc considérée (3.48).

Tel que mentionné précédemment, le PLDM1 pose l'hypothèse d'un taux d'occupation global. Afin de permettre la relaxation de cette hypothèse, ReVelle et Hogan (1989) ont proposé une deuxième version du PLDM qui considère plutôt un taux d'occupation local, c'est-à-dire pour chaque zone de demande i. Le taux d'occupation q_i s'exprime alors comme le ratio entre la durée de service des appels associés à la zone de demande i et la disponibilité des véhicules pouvant en assurer la couverture. En considérant l'expression de q_i et la probabilité pour qu'au moins un véhicule soit disponible pour répondre à une demande placée en i, une borne inférieure quant au nombre de véhicules nécessaires est déterminée pour chaque zone de demande i. Le PLDM2 est très similaire au PLDM1, à l'exception qu'il considère maintenant b_i plutôt que b. De plus, dans le cas du PLMD2, la variable x_j prend des valeurs entières plutôt que binaires, ce qui rend la co-localisation possible. Les deux versions du PLDM ont été appliquées par ReVelle et Hogan (1989) aux données de la ville de Baltimore, aux États-Unis, où $n = 207$ et $m = 31$. Dans les deux cas, le problème a été résolu par séparation et évaluation progressive. Les résultats obtenus ont permis de constater qu'une meilleure couverture est généralement obtenue en utilisant un taux d'occupation local plutôt que global, et ce, pour un nombre donné de véhicules. Plus récemment, Marianov *et al*. (2009) ont proposé pour la résolution du PLDM, une méthode heuristique basée sur une reformulation non linéaire du problème qu'ils ont nommé *heuristic concentration integer*. Les auteurs ont noté que leur méthode devient particulièrement intéressante lorsque la taille des problèmes considérés devient importante.

Dans le même ordre d'idée, ReVelle et Hogan (1988) ont formulé une version probabiliste du PLCT, soit le problème de localisation avec couverture totale probabiliste (PLCTP) ou *probabilistic location set covering problem* (PLSCP) en anglais. Le PLCTP vise toujours à minimiser le nombre de véhicules à localiser, mais en assurant maintenant que chaque demande puisse trouver au moins un véhicule disponible au moment où celle-ci est placée, et ce, avec un niveau de fiabilité donné α. De la même façon que pour le PLDM2, une contrainte probabiliste est formulée en fonction du taux d'occupation local et du niveau de fiabilité désiré. Le taux d'occupation est calculé de façon similaire au PLDM2. Les auteurs déterminent une contrainte déterministe équivalente qui garantit qu'un nombre suffisant de véhicules b_i assurent la couverture de chaque

zone de demande, où b_i est calculé à partir de q_i et α. Le PLCTP est très similaire à sa version déterministe. Il se formule de la manière suivante :

PLCTP

$$\min \sum_{j=1}^{m} x_j \tag{3.50}$$

sous les contraintes :

$$\sum_{j \in M_i} x_j \geq b_i, \; i = 1, ..., n, \tag{3.51}$$

$$x_j \geq 0, \text{ entier}, \; j = 1, ..., m. \tag{3.52}$$

Comme on peut aisément le constater, la co-localisation est acceptée dans ce cas.

ReVelle et Marianov (1991) ont également proposé une version probabiliste de FLEET (Schilling *et al*., 1979). Tout comme sa version déterministe, PROFLEET considère l'utilisation simultanée de deux types de véhicules. Le modèle vise donc à maximiser les appels pouvant être couverts simultanément par les deux types de véhicules, et ce, avec un niveau de fiabilité donné. Ainsi, de la même façon que pour le PLCTP et le PLDM, le niveau de fiabilité sert à déterminer le nombre de véhicules nécessaire pour assurer une couverture adéquate des zones de demande. Tout comme FLEET, PROFLEET s'applique plus naturellement au cas des services d'incendie, contexte pour lesquels ils ont été développés.

Les modèles présentés jusqu'à maintenant, PLCTP, PLDM et PROFLEET, considèrent tous l'indépendance du taux d'occupation des véhicules. Cette hypothèse peut, dans certains cas, engendrer un impact négatif sur l'estimation du taux d'occupation même, et donc, sur l'évaluation de la couverture. Afin de relaxer cette hypothèse et d'avoir une estimation plus fidèle des performances réelles du système, Marianov et ReVelle (1994, 1996) proposent deux nouveaux modèles, Q-PLCTP (ou Q-PLSCP en anglais) et Q-PLDM (ou Q-MALP en anglais), qui sont, en fait, des extensions des modèles pour le PLCPT et le PLDM. Les auteurs utilisent la théorie des files d'attente pour calculer la borne inférieure sur le nombre de véhicules nécessaires afin d'assurer la couverture de chaque zone de demande avec un niveau de fiabilité donné. Les modèles finaux demeurent très similaires aux modèles originaux. Dans les deux cas, les auteurs ont utilisé une méthode de séparation et évaluation progressive pour la résolution du problème test classique de 55 noeuds. Plus récemment, Harewood (2002) a également proposé une variante multi-objectifs du Q-PLDM.

Dans le même esprit, Galvão *et al*. (2005) ont proposé une autre extension du PLDM, le EMALP, qui utilise aussi la théorie des files d'attente afin d'obtenir une représentation plus fidèle de la réalité. Le problème proposé par Galvão *et al*. (2005) intègre un facteur correctif

lors de la définition des contraintes probabilistes de type (3.46). L'intégration du facteur correctif permet l'utilisation d'un taux d'occupation particulier pour chaque véhicule plutôt que pour chaque zone de demande comme c'était le cas précédemment. Elle permet également de considérer la coopération entre les véhicules. Les auteurs ont proposé une méthode heuristique basée sur le recuit simulé pour résoudre le EMALP.

Jusqu'à maintenant, la plupart des modèles présentés considéraient la disponibilité des véhicules comme principale source d'incertitude. Cela constituait la première approche présentée afin de considérer indirectement l'incertitude reliée à la réalisation de la demande. Dans leur modèle, REL-P, Ball et Lin (1993) ont plutôt décidé de considérer directement le caractère aléatoire relatif aux demandes de service. On pose alors l'hypothèse que chaque demande est générée selon une distribution de probabilité donnée. Les auteurs souhaitent alors imposer une limite supérieure quant à la probabilité r_i pour qu'aucun véhicule ne soit disponible pour répondre à l'appel au moment où celui-ci est placé, c'est-à-dire que $r_i \leq 1-\alpha$, où α représente le niveau de fiabilité recherché. De cette façon, la fiabilité du service à chaque zone de demande est assurée. Autrement dit, les auteurs souhaitent formuler une contrainte qui borne la probabilité pour que le nombre d'appels reçus à l'intérieur d'une région donnée soit plus grand que le nombre de véhicules disponibles afin de couvrir cette même région, et ce, pour une période de temps donnée. En définissant x_{jk}, une variable binaire qui vaut 1 si k véhicules sont localisés au site j, et 0 autrement, w_{jk}, les coûts fixes et les coûts variables associés à l'ouverture du site j et à la localisation de k véhicules en j, M_i, l'ensemble des sites potentiels pouvant assurer la couverture d'une zone de demande i à l'intérieur d'un délai prescrit S, N_j, l'ensemble des zones de demande couvertes par le site j à l'intérieur de S, et $D(j)$, le nombre d'appels qui se réalisent en N_j, le modèle REL-P se formule de la façon suivante :

REL-P

$$\min \sum_{j=1}^{m} \sum_{k=1}^{p_j} w_{jk} x_{jk} \tag{3.53}$$

sous les contraintes :

$$\sum_{k=1}^{p_j} x_{jk} \leq 1, j = 1, ..., m, \tag{3.54}$$

$$\sum_{j \in M_i} \sum_{1 \leq k \leq p_j} a_{jk} x_{jk} \geq b_i, i = 1, ..., n, \tag{3.55}$$

$$x_{jk} \in \{0,1\}, j = 1, ..., m, , k = 1, ..., p_j, \tag{3.56}$$

où

$$a_{jk} = -\log[P(D(j) \geq k)], \tag{3.57}$$

$$b_i = -\log(1-\alpha). \tag{3.58}$$

L'objectif de REL-P est donc de minimiser l'ensemble des coûts (3.53) de façon à garantir un service avec un niveau de fiabilité donné (3.55). La co-localisation est possible, mais à l'intérieur d'une limite p_j (3.54). Ball et Lin (1993) ont d'abord résolu le REL-P par séparation et évaluation progressive, puis ont proposé un ensemble d'inégalités valides afin d'en accélérer la résolution. L'approche proposée a été appliquée à un ensemble de problèmes test dont le problème test classique de 55 noeuds ($m = n = 55$). Borràs et Pastor (2002) proposent une variante de REL-P qui utilise un taux d'occupation propre à la localisation des véhicules plutôt qu'à la zone de demande lors du calcul de $P(D(j) \geq k)$.

Dans le même esprit que Ball et Lin (1993), Beraldi *et al*. (2004) ont proposé un modèle qui s'intéresse à la localisation et au dimensionnement d'une flotte de véhicules ambulanciers en considérant explicitement l'aspect aléatoire associé aux demandes de service. Ce modèle considère également l'affectation réelle des zones de demande aux différents véhicules. Durant la période de temps considérée, chaque véhicule peut servir uniquement une demande. De cette façon, le modèle proposé par Beraldi *et al*. (2004) s'apparente aux modèles de localisation et allocation classiques. Le modèle considère cependant, comme c'est le cas pour les modèles de couverture, qu'une localisation n'est candidate pour fournir un service que si elle est située à l'intérieur d'un temps prédéterminé S. Les auteurs ont tout d'abord présenté une version déterministe du problème servant de base à l'élaboration du modèle stochastique qui sera présenté subséquemment. En définissant y_j, une variable binaire qui vaut 1 si le site j est ouvert, et 0 autrement, x_{ij}, le nombre de véhicules localisés en j desservant la zone de demande i, a_i, les demandes de service associées à la zone de demande i, p_j, la limite sur le nombre de véhicules à localiser en j, c_{ij}, le coût d'affectation de la zone de demande i au site j, f_j, le coût d'ouverture du site j, N_j, l'ensemble des zones de demande pouvant être couvertes par j, et M_i, l'ensemble des sites pouvant assurer la couverture de i à l'intérieur de S, le modèle déterministe proposé se définit de la façon suivante :

$$\min \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} x_{ij} + \sum_{j=1}^{m} f_j y_j \tag{3.59}$$

sous les contraintes :

$$\sum_{j \in M_i} x_{ij} \geq a_i, i = 1, ..., n, \tag{3.60}$$

$$\sum_{i \in N_j} x_{ij} \leq p_j y_j, j = 1, ..., m, \tag{3.61}$$

$$x_{ij} \geq 0, \text{ entier}, i = 1,...,n, j = 1,...,m, \tag{3.62}$$

$$y_j \in \{0,1\}, j = 1,...,m. \tag{3.63}$$

Ce modèle cherche à minimiser les coûts associés à l'ouverture des sites et à l'affectation des demandes aux véhicules (3.59) de façon à ce que chaque demande a_i soit couverte à l'intérieur de S (3.60). Une limite sur le nombre de véhicules pouvant être localisés à chaque site doit également être respectée (3.61).

La fiabilité d'un tel système peut se mesurer par sa capacité à couvrir les demandes de service aléatoires. D'un point de vue mathématique, cela correspond à respecter les contraintes liées à la satisfaction de la demande (3.60) avec un niveau de fiabilité donné, α. En considérant une fiabilité globale pour tout le système, en opposition à la fiabilité locale proposée dans le cas de REL-P, la contrainte probabiliste proposée par Beraldi *et al*. (2004) peut s'exprimer par :

$$P(\sum_{j \in M_i} x_{ij} \geq \zeta_i, i = 1,...,n) \geq \alpha \tag{3.64}$$

où ζ_i représente la variable aléatoire associée aux demandes placées en i. En considérant chaque demande comme une variable aléatoire indépendante, il est possible de formuler un équivalent déterministe pour la contrainte (3.64). Le modèle stochastique proposé par Beraldi *et al*. (2004) se formule donc de façon similaire au modèle déterministe présenté précédemment, mais la contrainte (3.60) est remplacée par l'ensemble des contraintes suivantes :

$$\sum_{i=1}^{n} \sum_{k=1}^{k_i} a_{ik} z_{ik} \geq \beta \tag{3.65}$$

$$\sum_{j \in M_i} x_{ij} = l_i + \sum_{k=1}^{k_i} z_{ik}, \tag{3.66}$$

$$z_{ik} \in \{0,1\} \tag{3.67}$$

où

$$a_{ik} = \ln(F_i(l_i + k)) - \ln(F_i(l_i + k - 1)), \tag{3.68}$$

$$\beta = \ln(\alpha) - \ln(F(l)), \tag{3.69}$$

$$l_i = F_i^{-1}(\alpha) \tag{3.70}$$

et où F_i est la distribution de probabilité de ζ_i. La fiabilité du système est donc assurée par le respect des contraintes (3.65) à (3.67). Le modèle stochastique proposé par les auteurs a été résolu avec CPLEX, pour fins de validation. Bien que les modèles de Beraldi *et al*. (2004) et

de Ball et Lin (1993) aient été formulés à partir de considérations différentes, les auteurs ont constaté qu'ils arrivaient à des recommandations similaires lorsqu'appliqués au problème test classique. Les auteurs notent également que, pour la résolution de problèmes de taille importante, le développement de méthodes heuristiques pourrait s'avérer nécessaire.

Plus récemment, Alsalloum et Rand (2006) ont formulé une extension du PLCM qui intègre simultanément la couverture espérée et l'intégration de contraintes probabilistes. Plutôt que de considérer une définition strictement binaire de la couverture comme c'est le cas dans les modèles précédents, les auteurs définissent P_{ij} comme la probabilité d'atteindre une zone de demande i à partir d'une localisation j à l'intérieur d'un temps prescrit, probabilité qui dépend essentiellement du temps de déplacement entre chaque paire (i, j). L'idée de considérer les temps de déplacement lors du calcul de la probabilité pour qu'une zone de demande puisse être servie adéquatement a également été proposée par Daskin (1987) et Goldberg *et al*. (1990b). La probabilité P_{ij} intervient alors dans le calcul de la couverture espérée ce qui, selon les auteurs, permet d'améliorer l'évaluation de cette dernière, notamment lorsque le niveau d'agrégration est important. De la même façon que dans le cas du modèle de Beraldi *et al*. (2004), le modèle proposé par Alsalloum et Rand (2006) considère l'affectation réelle des zones de demande aux localisations sélectionnées en définissant y_{ij}, une variable binaire qui vaut 1 lorsque $P_{ij} \geq \rho$ où ρ est une borne prédéterminée et lorsque la localisation j est celle pour laquelle la distance par rapport à i est la plus petite, et 0 autrement. Le modèle vise donc, dans un premier temps, à sélectionner les sites à utiliser de façon à maximiser la couverture espérée et, dans un deuxième temps, à déterminer le nombre de véhicules à placer à chacune des localisations sélectionnées de façon à ce que chaque demande puisse trouver au moins un véhicule disponible. La disponibilité des véhicules est considérée par l'intermédiaire d'une contrainte probabiliste où le nombre minimum de véhicules requis pour atteindre un niveau de fiabilité adéquat est obtenu grâce à des notions issues de la théorie des files d'attente.

Ainsi, en définissant x_{jk}, une variable binaire qui vaut 1 si exactement k véhicules sont positionnés en j, et 0 autrement, λ_i, le taux d'arrivée des demandes en i, r_k, une valeur frontière du taux d'arrivée qui requiert l'ajout d'un véhicule (de k à $k+1$) déterminée par la théorie des files d'attente, et a_i, la proportion de la demande totale qui se réalise en i, les auteurs ont formulé le modèle suivant en utilisant la programmation par objectifs :

$$\min P_0 d_0^- + P_1 \sum_{j=1}^{m} d_j^+ \tag{3.71}$$

sous les contraintes :

$$\sum_{i=1}^{n}\sum_{j=1}^{m} a_i P_{ij} y_{ij} + d_0^- = 1, \tag{3.72}$$

$$\sum_{1<k<p_j} r_k x_{jk} - \sum_{i=1}^{n} \lambda_i y_{ij} - d_j^+ = 0,\ j = 1,...,m, \tag{3.73}$$

$$\sum_{j=1}^{m} y_{ij} \leq 1,\ i = 1,...,n, \tag{3.74}$$

$$\sum_{k=1}^{p_j} x_{jk} \leq 1,\ j = 1,...,m, \tag{3.75}$$

$$\sum_{j=1}^{m}\sum_{k=1}^{p_j} x_{jk} = P, \tag{3.76}$$

$$y_{ij}, x_{jk} \in \{0,1\}, i = 1,...,n, j = 1,...,m, k = 1,...,p_j. \tag{3.77}$$

Le modèle vise à minimiser la non-couverture espérée de la population de même que le nombre de véhicules à localiser (3.71). P_0 et P_1 représentent les poids associés aux objectifs tandis que d_0^- et d_j^+ représentent les déviations par rapport à ces objectifs. Ces déviations sont considérées au sein des contraintes (3.72) et (3.73). Enfin, les contraintes (3.74), (3.75) et (3.76) assurent respectivement la couverture de chaque zone de demande, le respect de la limite p_j quant au nombre de véhicules à localiser en j et la localisation d'un nombre total de sites P. Ce modèle a été appliqué à la localisation d'un nombre donné d'ambulances, variant entre 1 et 17 véhicules à travers les 92 quartiers de la ville de Riyadh, en Arabie Saoudite, et résolu de manière exacte. Aucune précision n'a toutefois été apportée quant à la méthode utilisée pour y arriver. À la suite de leurs expérimentations, les auteurs ont pu observer qu'en relocalisant les sept sites utilisés initialement, la population couverte à l'intérieur du délai prescrit passait de 74% à 85%. Deux véhicules étaient alors requis à chacun des sites. Ils ont noté que 17 sites et 24 véhicules étaient requis pour garantir que chaque demande puisse trouver au moins un véhicule disponible à l'intérieur du délai établi.

Enfin, Beraldi et Bruni (2009) ont proposé un modèle stochastique qui ne se fonde ni sur la couverture espérée, ni sur la formulation de contraintes probabilistes. Le modèle est plutôt issu de la programmation stochastique avec fonction de recours (Birge et Louveaux, 2011). À notre connaissance, il s'agit du seul modèle de ce genre appliqué au déploiement des véhicules ambulanciers. Comme le modèle proposé par Beraldi *et al.* (2004), ce modèle considère l'aspect aléatoire associé aux demandes de service de même que l'affectation réelle des véhicules aux demandes. Il considère deux étapes de décision : la première est associée à la localisation des

sites et la deuxième à l'affectation des demandes aux véhicules lorsque les demandes se réalisent. En considérant x_j, le nombre de véhicules localisés au site j, et 0 autrement, z_j, une variable binaire qui vaut 1 si le site j est utilisé, $a(w)$, le vecteur aléatoire des valeurs entières associées à la réalisation de la demande (quantité demandée), $y_{ij}(w)$, une variable binaire qui vaut 1 si la zone de demande i est affectée aux véhicules localisés en j lorsque le vecteur aléatoire $a(w)$ se réalise, c_j, le coût associé à la localisation d'un véhicule au site j, f_j, le coût associé à l'utilisation du site j, d_{ij}, la distance ou le coût pour servir une demande qui se réalise en i à partir du site j, N_j, l'ensemble des zones de demande pouvant être couvertes par le site j à l'intérieur de la distance ou du délai prescrit S, et M_i, l'ensemble des sites pouvant assurer la couverture de la zone de demande i à l'intérieur de S, le modèle proposé s'écrit comme suit :

$$\min \sum_{j=1}^{m} (c_j x_j + f_j z_j) + E_w[Q(x,z,w)] \tag{3.78}$$

$$x_j \leq p_j z_j, j = 1,...,m, \tag{3.79}$$

$$z_j \in \{0,1\}, x_j \text{ entier}, j = 1,...,m \tag{3.80}$$

où

$$Q(x,z,w) = \min_y \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij} y_{ij}(w), \tag{3.81}$$

$$\sum_{i \in N_j} a_i(w) y_{ij}(w) \leq x_j, \; j = 1,...,n, \tag{3.82}$$

$$\sum_{j \in M_i} y_{ij}(w) \geq 1, i = 1,...,n, \tag{3.83}$$

$$y_{ij}(w) \leq x_j, i = 1,...,n, \; j = 1,...,m, \tag{3.84}$$

$$y_{ij}(w) \in \{0,1\}, i = 1,...,n, \; j = 1,...,m. \tag{3.85}$$

Plus concrètement, la fonction objectif comporte un premier terme associé à la minimisation des coûts de localisation (3.78) et un second terme associé à la minimisation de la fonction de recours, c'est-à-dire à la minimisation des coûts liés à l'affectation des demandes aux différents véhicules (3.81). Les contraintes de première étape stipulent qu'un nombre maximum de véhicules peut être affecté à un site j (3.79). Les contraintes associées au recours visent, quant à elles, à ce qu'il y ait suffisamment de véhicules localisés en j pour desservir les demandes qui y sont affectées et qui se sont réalisées (3.82) et à ce qu'une demande qui se réalise soit affectée à au moins un véhicule (3.83). De plus, une demande qui se réalise en i ne peut être affectée à un véhicule localisé en j que si le site j est utilisé (3.84). Les auteurs ont également proposé

d'intégrer des contraintes probabilistes afin d'assurer le respect des contraintes (3.82) et (3.83). Naturellement, ces contraintes visent, comme c'était le cas précédemment, à assurer la fiabilité du système. Les auteurs ont proposé une approche exacte, ainsi que trois méthodes heuristiques pour résoudre le problème traité. Les auteurs ne présentent pas d'application en soi, mais plutôt une série d'instances de tailles différentes ($n = 50$ à 150 et $m = 25$ à 150) afin de tester le modèle et les différentes méthodes de résolution proposées. Le nombre de scénarios considérés pour chaque problème varie entre 10 et 40. Les auteurs ont constaté que la vitesse de résolution des méthodes heuristiques proposées augmentait significativement avec le nombre de scénarios considérés.

Comme on peut le constater au vu des différents modèles présentés dans cette section, des efforts considérables ont été déployés au fil des années afin de développer des modèles toujours plus réalistes. Toutefois, on constate que malgré des progrès certains au niveau de la modélisation de ces problèmes, plusieurs hypothèses simplificatrices sont toujours nécessaires, même pour les modèles probabilistes. En effet, il y a souvent un prix à payer entre le niveau de réalisme et la capacité à résoudre les modèles. Des questions se posent alors quant aux performances des solutions proposées par les différents modèles dans un contexte réel. Offrent-elles un niveau de couverture adéquat ? Sous-estiment-elles le nombre de véhicules requis ? La simulation peut alors constituer une alternative intéressante pour évaluer différents scénarios dans un contexte plus réaliste et justement, elle a été grandement utilisée pour traiter des problèmes tels que le déploiement des véhicules d'urgence. Différents travaux à cet effet seront recensés à la section 3.2.2.

3.2.2 Simulation

Les services préhospitaliers d'urgence font face quotidiennement à plusieurs sources d'incertitude, ce qui complexifie naturellement tout processus de décision. Considérer une ou plusieurs sources d'incertitude dans le cadre d'un modèle mathématique peut devenir extrêmement difficile à gérer, voire irréalisable. Ainsi, au fil des années, la simulation s'est avérée une avenue très intéressante pour l'analyse et l'optimisation de systèmes qui présentent plusieurs sources d'incertitude. La simulation est une approche descriptive qui permet l'évaluation de solutions déjà établies, mais dans un environnement généralement plus réaliste que dans le cas de la programmation mathématique. En effet, la programmation mathématique nécessite généralement plusieurs hypothèses simplificatrices afin d'obtenir des modèles utilisables pour traiter ce genre de problématique. Bien que plusieurs efforts aient été déployés afin d'améliorer le niveau de réalisme, notamment par l'intégration de différents éléments issus de la théorie des files d'at-

tente, la simulation s'impose parfois comme la seule avenue pour une évaluation adéquate des SPU.

Dans la littérature, trois approches ont été proposées afin d'utiliser la simulation pour analyser différentes décisions en lien avec la gestion des SPU. La première approche suggère l'utilisation de la simulation afin de développer des modèles pour l'évaluation d'un ensemble de scénarios possibles. Dans ce cas, chaque alternative est évaluée, puis la meilleure est sélectionnée. La deuxième approche vise plutôt à utiliser la simulation afin de développer des modèles permettant l'évaluation de solutions obtenues par des modèles mathématiques dans un cadre plus réaliste. De cette manière, la robustesse et les performances réelles de la solution considérée peuvent être analysées. La simulation peut également être utilisée afin de déterminer certaines mesures de performance qui ne peuvent être évaluées par les modèles mathématiques. Enfin, la troisième approche consiste à utiliser la simulation au sein même d'un processus d'optimisation, approche mieux connue sous le nom de simulation-optimisation (Andradóttir, 1998).

Dans un cas comme dans l'autre, la modélisation des différents éléments qui composent le système étudié est critique, non seulement au niveau de la validité des solutions et des résultats obtenus, mais également pour leur acceptation en pratique. En simulation, la validité se définit comme la capacité d'un modèle à représenter adéquatement un système réel (Law, 2006). Ainsi, si le modèle de simulation développé n'est pas valide aux yeux des gestionnaires d'un SPU, il y a peu de chances que les solutions proposées soient appliquées en pratique.

Grâce à l'analyse des différents modèles de simulation proposés dans la littérature, il a été possible de cibler cinq éléments principaux dont la modélisation semble particulièrement importante. Ces élements sont : la demande, le temps de service, le temps de déplacement, les règles de répartition et la priorité des appels. La modélisation de chacun de ces éléments doit donc être effectuée avec soin afin de représenter le plus fidèlement possible le système étudié. De plus, dans certains cas, différents éléments secondaires tels que le temps de maintenance des véhicules et les pauses des techniciens ambulanciers devront être modélisés afin d'assurer la validité du modèle de simulation développé. Bien entendu, les approches proposées pour la modélisation de ces éléments ont évolué au fil des années, notamment avec le raffinement des outils informatiques disponibles.

Dans cette section, nous présenterons certains modèles de simulation développés afin d'analyser les décisions reliées à la gestion des SPU et, plus particulièrement, au déploiement des véhicules ambulanciers. La présentation détaillée tous les modèles de simulation que nous avons répertoriés et qui sont en lien avec la localisation et l'affectation des véhicules ambulanciers n'est pas nécessaire pour l'objet du présent chapitre. Nous nous limiterons donc à la description de

certains modèles que nous jugeons particulièrement pertinents en portant une attention spéciale aux objectifs visés par chacun. Nous invitons le lecteur à se référer à Aboueljinane *et al*. (2013) pour une description plus étoffée des modèles de simulation développés dans le contexte de la gestion de SPU.

À la fin des années 1960, Savas (1969) fut parmi les premiers à utiliser la simulation dans le contexte de la gestion des services préhospitaliers d'urgence. Dans le système qu'il a analysé, soit le service préhospitalier du Kings County Hospital à New York, tous les véhicules ambulanciers sont localisés en attente à l'hôpital même. Afin d'évaluer la possibilité de déplacer certains véhicules vers un second poste d'attente, Savas a proposé un modèle de simulation qui permet la comparaison de différentes alternatives quant à la localisation et au nombre de véhicules déplacés en se basant sur le temps de réponse moyen, le temps écoulé entre la réception de l'appel et l'arrivée à l'hôpital, ainsi que le taux d'occupation des véhicules. Quelques années plus tard, Swoveland *et al*. (1973) ont également proposé un modèle de simulation afin d'analyser un ensemble de sites potentiels pour la localisation d'un nombre donné de véhicules pour le SPU de la ville de Vancouver, au Canada. Afin de comparer les différentes alternatives, le modèle fournit une estimation du temps de réponse, par priorité d'appels, de même qu'une estimation du taux d'utilisation des véhicules. Le modèle de simulation proposé est relativement simple. Il considère des distributions de probabilité classiques afin de modéliser la demande et les temps de déplacement. Il offre néanmoins la possibilité de considérer différentes règles de priorité, de même qu'une certaine forme de collaboration entre les véhicules affectés aux différentes zones de demande.

En 1974, Berlin et Liebman (1974) ont, quant à eux, développé un modèle de simulation afin d'évaluer différents scénarios d'affectation des véhicules en fonction d'un plan de déploiement déterminé grâce au modèle de Toregas *et al*. (1971). Dans le contexte qu'ils étudient, les localisations sont déterminées et le modèle vise plutôt à étudier l'affectation des véhicules aux différentes localisations utilisées. Le modèle de simulation développé vise également à fournir des informations quant au temps de réponse moyen et à l'utilisation des véhicules. Dans le même ordre d'idée, Lubicz et Mielczarek (1987) ont proposé, à la fin des années 1980, un modèle de simulation pour analyser la variation du nombre de véhicules à affecter à une localisation déjà déterminée à l'intérieur d'un district en fonction du temps total de service, du temps de transfert vers l'hôpital et du taux d'utilisation des véhicules. Ce modèle vise également à analyser la capacité du système à répondre à des demandes croissantes. Bien que le simulateur proposé par Lubicz et Mielczarek présente certaines similitudes avec le modèle de Berlin et Liebman, il propose une approche de modélisation plus sophistiquée des temps de déplacement.

Toujours à la fin des années 1980, Liu et Lee (1988) ont développé un modèle de simulation qui prend en compte simultanément le service préhospitalier d'urgence de même qu'un certain nombre de composantes du service des urgences de l'hôpital correspondant. En effet, tel que noté par les auteurs, le taux d'arrivée des patients aux urgences à bord de véhicules ambulanciers a un impact sur les performances du service des urgences d'un hôpital. Le taux d'occupation des lits aux urgences influence également le temps de service des véhicules ambulanciers. Les auteurs proposent donc d'analyser simultanément les deux systèmes qui sont gérés par les mêmes autorités dans le cas de Taipei, à Taiwan. Le modèle de simulation développé vise à analyser le nombre de véhicules ambulanciers nécessaires, mais également le nombre de lits requis au service des urgences. Comme la plupart des modèles précédents, le temps de réponse, le temps d'attente et le temps de service sont utilisés afin de comparer les différentes alternatives. Le modèle de Liu et Lee (1988) est intéressant, car il introduit l'idée d'intégration des différents systèmes pour une meilleure gestion d'un système de santé.

Développé dans le contexte d'Urgences-santé, corporation responsable du SPU pour la grande région de Montréal, le modèle de simulation proposé par Trudeau *et al*. (1989) vise, quant à lui, à assurer l'évaluation des différentes stratégies reliées à la gestion des SPU. Ainsi, différentes règles de répartition, plans de déploiement et stratégies de redéploiement, de même que différentes alternatives quant au nombre d'unités en service et aux horaires utilisés, pourront être évalués. Le temps de réponse espéré, la proportion des appels répondus à l'intérieur du délai prescrit, de même que la charge de travail des véhicules, sont utilisés afin de comparer les différents scénarios. Le modèle de Trudeau *et al*. (1989) est intéressant puisqu'il permet l'évaluation de différents plans de déploiement, et met à l'avant-plan l'interaction entre les différents systèmes de décision impliqués dans la gestion des SPU.

De façon plus conventionnelle, Goldberg *et al*. (1990a) ont proposé un modèle de simulation afin d'analyser et de comparer différents plans de déploiement dans le contexte de la ville de Tucson, en Arizona. Les auteurs décrivent l'utilisation de ce modèle dans le cadre de trois études. Ces études visaient respectivement à évaluer deux alternatives quant à la localisation des véhicules, à analyser l'impact de l'ajout d'un huitième véhicule et à mesurer l'impact de la relocalisation de quatre des localisations utilisées initialement. Afin de comparer les différentes alternatives, le modèle mesure le taux de succès, c'est-à-dire le pourcentage d'appels servis à l'intérieur du délai prescrit, de même que la charge de travail de chaque véhicule.

Plus récemment, Ingolfsson *et al*. (2003) ont proposé un modèle de simulation à événements discrets afin d'analyser différentes règles associées à la gestion du SPU de la ville d'Edmonton, en Alberta. Dans ce cas, l'étude visait à évaluer la possibilité d'utiliser une station unique

à partir de laquelle tous les véhicules ambulanciers amorceraient et termineraient leur quart de travail. Tel que mentionné par les auteurs, ce modèle pourrait être utilisé afin d'analyser l'augmentation du nombre de véhicules disponibles ou la modification des politiques de redéploiement en fonction de la disponibilité des véhicules et du pourcentage d'appels répondus à l'intérieur du délai prescrit.

Henderson et Mason (2004) ont proposé un modèle de simulation très complet qui vise, tout comme le modèle proposé par Trudeau *et al*. (1989), à supporter la prise de décision en permettant l'analyse de différentes stratégies liées à la gestion d'un système préhospitalier d'urgence. Dans leur article, les auteurs proposent une discussion intéressante sur l'importance de la modélisation des temps de déplacement. En effet, selon les observations des auteurs, la façon de modéliser le temps de déplacement peut avoir un impact majeur sur les performances prédites par le modèle de simulation. À cet effet, les auteurs ont développé et utilisé un modèle sophistiqué pour déterminer le temps de déplacement entre deux points en fonction de la période de la journée. Le modèle de simulation compile un certain nombre de mesures de performance, notamment en ce qui a trait au temps de réponse et au taux d'occupation des véhicules, afin de comparer les différents scénarios envisagés. Différentes applications pratiques de ce modèle de simulation ont été présentées dans Mason (2013).

Aboueljinane *et al*. (2014) ont présenté une étude de simulation menée dans le contexte de l'organisation responsable des SPU pour le département de Val-de-Marne, en France. L'objectif de cette étude consiste à évaluer différents scénarios en termes de nombre de ressources disponibles et de la localisation de ces ressources sur le territoire à desservir dans le but d'améliorer les performances du système. Afin d'effectuer cette analyse, un modèle de simulation à événements discrets est développé puis implémenté grâce à ARENA, un logiciel de simulation commercial. Les différents scénarios envisagés sont ensuite comparés en considérant deux principaux critères de performance soit la couverture, c'est-à-dire le nombre d'appels couverts à l'intérieur d'un délai prescrit, et le taux d'utilisation des ressources humaines.

Tous les modèles de simulation présentés jusqu'à maintenant ont été développés afin d'analyser, de façon plus ou moins précise, plusieurs alternatives concernant les différentes politiques de gestion des SPU. Bien que ceci ne constitue pas l'objectif premier de leur création, ces modèles permettent aussi l'évaluation de solutions déterminées par des modèles de programmation mathématique. D'autres travaux ont toutefois proposé le développement de modèles de simulation dont l'objectif premier visait plutôt à analyser les solutions déterminées par programmation mathématique. Ainsi, Uyeno et Seeberg (1984) ont proposé un modèle de simulation afin d'analyser les différents plans de déploiement déterminés grâce à un modèle de localisation classique

de type *p*-médiane dans le contexte du *British Columbia Provincial Ambulance Service*, à Vancouver. Différentes statistiques à propos du temps de réponse et de l'utilisation des véhicules ont alors été recueillies. Quelques années plus tard, Fujiwara *et al*. (1987) ont également proposé un modèle de simulation afin d'analyser les différents plans de déploiement déterminés grâce au PLCEM (Daskin, 1983) pour différents profils de demande et pour différentes valeurs de S, où S représente le délai prescrit. Celui-ci recueillait des statistiques quant aux temps de réponse et de service de même que sur l'utilisation des véhicules. Enfin, Harewood (2002) a développé un modèle de simulation afin de valider les solutions déterminées grâce à une variante multi-objectif de Q-PLDM dans le contexte d'un service d'ambulances situé sur l'île de la Bardade, dans les Antilles. Le modèle de simulation proposé se base essentiellement sur une mesure du temps de réponse, du temps de service et de la fiabilité du système afin de valider la solution proposée.

Enfin, différents travaux ont considérés la simulation au sein d'un processus de simulation-optimisation. Ainsi, Mason (2013) a présenté l'utilisation d'un modèle de simulation pour l'évaluation de solutions voisines dans un processus de recherche locale. Dans ce cas, la localisation d'un ensemble de postes d'attente pour les véhicules ambulanciers doit être déterminée. L'algorithme proposé afin de résoudre ce problème fonctionne de la manière suivante : à chaque itération, les solutions voisines (c'est-à-dire les solutions générées suite à la modification de la localisation d'un poste d'attente) sont évaluées par simulation puis la meilleure solution voisine est sélectionnée. L'algorithme itère ainsi jusqu'à ce qu'aucune amélioration ne puisse être possible. De manière similaire, Zhen *et al*. (2014) ont proposé l'utilisation de la simulation pour l'évaluation des solutions au sein d'un algorithme génétique. Ils souhaitent alors déterminer le nombre de véhicules à affecter à chaque poste d'attente en considérant un nombre donné de véhicules et un ensemble de postes d'attente dont la localisation est prédéterminée. La méthodologie proposée par Zhen *et al*. (2014) a été appliquée au contexte de la ville de Shanghai, en Chine, où 80 véhicules et 12 sites d'attente sont considérés. Finalement, Aboueljinane *et al*. (2014) ont utilisé un outil de simulation-optimisation commercial intégré au logiciel ARENA, nommé *OptQuest*, afin de déterminer les meilleurs plans de déploiement d'un nombre donné de véhicules pour un ensemble fixe de périodes, et ce, dans le contexte du département français de Val-de-Marne. *OptQuest* est un outil qui permet de définir un modèle d'optimisation dont l'objectif est composé d'une ou de plusieurs mesures de performance déterminées par simulation. Ainsi, à chaque itération du processus de recherche, les solutions sont évaluées grâce à un modèle de simulation, ce qui permet généralement une estimation plus fidèle des performances du système étudié.

Il se dégage clairement de ce qui précède que la simulation présente plusieurs possibilités quant à l'analyse de différentes stratégies de gestion dans le contexte des SPU. En effet, la simulation s'est montrée particulièrement intéressante afin d'analyser les systèmes de décision qui présentent plusieurs sources d'incertitude. La simulation permet de considérer plus facilement les sources d'incertitude les plus critiques pouvant influencer la prise de décision, ce qui assure généralement une représentation plus fidèle de la réalité. De plus, la simulation facilite l'intégration de différentes règles à caractère opérationnel telles que les règles de répartition ce qui est souvent impossible autrement. Le développement d'un modèle de simulation passe toutefois par plusieurs étapes, de la cueillette de données à l'expériementation (Law, 2006). Toutes ces étapes ne peuvent être réalisées adéquatement sans déployer certains efforts. Le développement d'un modèle de simulation peut donc nécessiter beaucoup de temps bien qu'il présente plusieurs avantages évidents.

3.2.3 Modèles descriptifs issus de la théorie des files d'attente

Plusieurs modèles descriptifs proposés en lien avec le problème de déploiement des véhicules d'urgence s'inspirent de la théorie des files d'attente. Ces modèles permettent généralement le calcul de diverses mesures de performance en fonction d'un nombre de véhicules ou d'un plan de déploiement donné. Un certain nombre d'hypothèses doivent toutefois être posées, notamment quant à la génération de la demande et à la loi de probabilité des temps de service, afin d'assurer la validité des expressions dérivées de la théorie des files d'attente. Ainsi, en posant certaines hypothèses, les SPU sont généralement modélisés comme des systèmes $M/M/n$ avec file d'attente de capacité nulle ou infinie. Dans le premier cas, les appels qui ne trouvent pas de véhicule disponible sont perdus ou répondus par un autre service. Dans le deuxième cas, ils sont placés dans une file d'attente, puis servis en suivant la règle du premier arrivé, permier servi (FIFO). L'utilisation de la théorie des files d'attente afin d'évaluer les performances d'un système et de déterminer le nombre optimal de véhicules à localiser en un site donné a été proposée par Bell et Allen (1969). L'utilisation de la théorie des files d'attente au sein de méthodologies plus sophistiquées a également été proposée par Volz (1971) et Fitzsimmons (1973).

Le modèle de l'hypercube proposé par Larson (1974) est sans contredit un des modèles descriptifs les plus utilisés afin d'évaluer les performances d'un service d'urgence. En effet, le modèle de Larson fournit, en fonction d'un ensemble de sites sélectionnés et de règles de répartition données, une série d'expressions permettant de calculer diverses mesures de performance telles que le temps de réponse, le taux d'occupation et le pourcentage des demandes servies par un véhicule donné. Pour développer son modèle, Larson considère le système étudié comme un

processus markovien en temps continu avec 2^N états possibles où chaque état correspond à une combinaison possible de véhicules libres et occupés. L'obtention des mesures de performance mentionnées précédemment se fonde sur la théorie des files d'attente et utilise la probabilité de se retrouver dans chacun des états. Le calcul de cette probabilité représente la difficulté majeure de la méthode, puisqu'elle requiert la résolution simultanée de 2^N équations linéaires, ce qui peut devenir important lorsque N, le nombre de véhicules, augmente. Cela permet néanmoins de considérer les véhicules individuellement, ce qui n'est généralement pas le cas pour les autres modèles. Cela permet également de considérer différentes règles de répartition en définissant un ordre de priorité pour les véhicules à affecter aux appels en provenance d'une zone de demande donnée. De cette façon, le modèle permet de considérer la collaboration entre les véhicules affectés à différentes régions du territoire, tout en considérant la nature stochastique associée à la demande et aux temps de service. L'hypothèse selon laquelle les demandes sont générées en suivant un processus de Poisson, de même que l'hypothèse qui stipule que le temps de service suit une loi exponentielle inverse, sont toutefois nécessaires à l'utilisation du modèle de Larson. Ce modèle pose également l'hypothèse que le temps de service ne dépend pas de l'identité du serveur, de la localisation des clients ou de l'historique du système. Enfin, le modèle de l'hypercube a été utilisé par plusieurs auteurs, notamment par Trudeau *et al*. (1989) pour le calcul des performances pour un ensemble donné de localisations, ainsi que par Batta *et al*. (1989) (voir section 3.2.1).

Tel que mentionné précédemment, la résolution exacte du modèle de l'hypercube requiert la résolution simultanée de 2^N équations linéaires, ce qui peut représenter un effort de calcul important. Afin de remédier à la situation, Larson (1975) a proposé une méthode approximative qui nécessite plutôt la résolution de N équations non linéaires, réduisant ainsi de manière considérable l'effort de calcul, notamment lorsque la taille des systèmes étudiés devient importante. Cette méthode n'utilise plus les probabilités telles qu'elles ont été définies précédemment, mais plutôt un ensemble de probabilités associées au fait que le premier véhicule disponible pour l'affectation soit le $j+1$-*ième* véhicule sélectionné aléatoirement, étant donné un ensemble de véhicules libres et occupés. En développant l'expression de la probabilité, l'auteur dérive un facteur correctif dont l'utilisation rend l'hypothèse d'indépendance des véhicules non nécessaire. Le facteur correctif proposé par Larson a été utilisé dans le cadre des méthodes proposées par Batta *et al*. (1989) et Marianov et ReVelle (1994, 1996) (voir section 3.2.1). La méthode approximative proposée par Larson calcule le taux d'occupation pour chaque véhicule grâce à une méthode itérative, puis évalue les différentes mesures de performance du modèle original.

Alanis *et al*. (2013) ont également proposé un modèle descriptif afin d'évaluer les performances d'un SPU. Toutefois, contrairement au modèle de l'hypercube, ce modèle considère le repositionnement des véhicules en fonction de l'état du système, c'est-à-dire le nombre de véhicules disponibles, à partir de *tables de positionnement* calculées *a priori*, puis appliquées en temps réel. Une *table de positionnement* définit donc, pour tous les états possibles, la localisation adéquate des véhicules disponibles. Le système ainsi décrit peut être représenté par un processus markovien. Toutefois, en considérant le repositionnement, il est possible de poser l'hypothèse selon laquelle chaque serveur est identique, ce qui n'est pas le cas pour le modèle de l'hypercube. Le nombre d'états possibles est alors réduit à $2N + 1$. Le modèle qui considère le repositionnement, proposé par Alanis *et al*. (2013), est donc plus facile à résoudre que le modèle de l'hypercube. Ainsi, à partir d'une *table de positionnement* donnée, ce modèle permet de déterminer un ensemble de mesures de performance telles que la distribution de temps de réponse et la distribution du nombre de véhicules occupés.

Que ce soit de manière exacte ou approximative, les modèles descriptifs issus de la théorie des files d'attente permettent, tout comme la simulation, l'évaluation de solutions déjà établies. La solution à évaluer devra donc être déterminée en appliquant une méthode prescriptive ou simplement en déterminant *a priori* un ensemble d'alternatives intéressantes. Afin d'assurer la génération d'une solution en soi, les modèles descriptifs issus de la théorie des files d'attente devront plutôt être intégrés au sein d'une méthodologie plus sophistiquée. En effet, les modèles descriptifs permettent généralement l'évaluation d'une solution en déployant des efforts raisonnables, principalement lorsque des méthodes approximatives sont utilisées, ce qui les rend particulièrement intéresssants pour le calcul de mesures de performance à l'intérieur d'un schéma d'optimisation.

3.2.4 Méthodes de résolution

Grâce à l'analyse des différents modèles présentés dans les trois sous-sections précédentes, il a été possible de constater que les modèles proposés pour le problème de déploiement des véhicules ambulanciers ont beaucoup évolué au fil des ans. Le même constat peut être fait en ce qui concerne les différentes méthodes de résolution proposées. Bien que la plupart de ces méthodes s'appliquent à la résolution de modèles issus de la programmation mathématique, nous avons jugé bon de placer cette section à la suite de la présentation des trois familles d'approches de modélisation, puisque nous souhaitons effectuer non seulement une brève synthèse des méthodes reliées à la programmation mathématique, mais aussi faire ressortir certaines méthodes qui proposent l'intégration de plusieurs approches.

Dans leur version originale, la plupart des modèles issus de la programmation mathématique ont été résolus de façon exacte en ayant recours à des méthodes basées sur la relaxation linéaire avec l'ajout de coupes (Toregas *et al*., 1971; Daskin et Stern, 1981) ou l'application de méthodes de branchement (Church et ReVelle, 1974; Daskin, 1983; Hogan et ReVelle, 1986; ReVelle et Hogan, 1989; Ball et Lin, 1993). Dans ces cas, les problèmes traités étaient généralement de petite ou de moyenne taille, ce qui rendait l'application des méthodes exactes envisageable. Dans la même optique, différentes méthodes heuristiques ont été proposées pour la résolution d'instances de taille raisonnable. Ces méthodes étaient principalement basées sur des échanges simples (Church et ReVelle, 1974; Eaton *et al*., 1986).

Bien entendu, les premiers modèles proposés étaient relativement simples, quoi que généralement difficiles à résoudre. Ainsi, la plupart des modèles déterministes à couverture simple, des modèles déterministes à couverture multiple, de même que les premiers modèles probabilistes, sont des modèles linéaires qui comportent un nombre raisonnable de variables et de contraintes. Plusieurs hypothèses sont toutefois nécessaires afin de permettre leur modélisation. Malheurement, tel que noté Batta *et al*. (1989), ces hypothèses ne sont généralement pas respectées en pratique, ce qui peut avoir un impact non-négligeable sur l'écart entre les performances réelles du système et les performances prédites par le modèle. Dans le but d'obtenir une meilleure estimation de la réalité, certains auteurs ont proposé différentes méthodes afin de lever ces hypothèses, donnant lieu à des modèles plus précis, mais également plus complexes. Dans la plupart des cas, ces modèles intègrent un facteur correctif issu du développement d'une méthode approximative pour la résolution du modèle de l'hypercube de Larson (1975) ou le modèle de l'hypercube lui-même (Larson, 1974) . Ainsi, lorsque le modèle de l'hypercube est utilisé pour obtenir une estimation plus précise de la couverture espérée, le développement de méthodes heuristiques afin de supporter le processus d'optimisation est nécessaire. Une méthode de descente ou une approche basée sur le recuit simulé dans laquelle la couverture espérée pour chaque solution potentielle est évaluée grâce au modèle de l'hypercube peuvent être développées afin de permettre la recherche de la solution optimale, c'est du moins ce que proposent Batta *et al*. (1989) et Galvão *et al*. (2005). De façon similaire, la simulation pourrait être utilisée afin d'évaluer les différentes solutions potentielles à l'intérieur d'un processus de recherche. Le modèle de simulation devra alors être utilisé plusieurs fois, à chaque itération du processus de recherche, ce qui peut représenter beaucoup de temps.

Au fil des ans et des applications, la taille des problèmes traités est également devenue une préoccupation importante pour les chercheurs. Premièrement, le niveau d'agrégation est directement relié à la taille des problèmes. En effet, afin d'obtenir un niveau de précision plus grand,

des problèmes moins agrégés et donc, de plus grande taille, sont souvent résolus. D'autre part, les systèmes traités sont tels que même avec un niveau d'agrégation élevé, leur taille demeure dans certains cas très grande. Naturellement, la taille des problèmes traités a un impact sur les performances des méthodes. Les méthodes exactes deviennent rapidement impraticables et les heuristiques simples mènent souvent à des résultats peu satisfaisants. La taille grandissante des problèmes à résoudre justifie le développement de méthodes de résolution plus sophistiquées. Ainsi, certains auteurs ont proposé l'utilisation d'heuristiques lagrangiennes (Galvão et ReVelle, 1996), d'algorithmes génétiques (Aytug et Saydam, 2002), de la recherche avec tabous (Gendreau *et al*., 1997) ou du recuit simulé (Galvão *et al*., 2005) afin de résoudre certains problèmes associés au déploiement statique des véhicules ambulanciers. Les méthodes ainsi proposées demeurent toutefois difficiles à comparer, puisqu'elles ont généralement été développées pour des problèmes différents et appliquées à des contextes particuliers. Elles présentent néanmoins des améliorations par rapport aux méthodes plus simples utilisés avant leur création. Enfin, il est possible de constater que peu de travaux, en lien avec les SPU, visaient l'amélioration des performances des méthodes de résolution en soi, en termes de temps de calcul. En effet, les auteurs se sont plutôt intéressés à la résolution des problèmes considérés en développant des modèles qu'ils ont validés par la suite. D'autre part, le caractère tactique du problème de déploiement des véhicules ambulanciers rend le développement de méthodes de résolution très rapide non nécessaire. Bien qu'il doive demeurer à l'intérieur d'une limite raisonnable, le temps de calcul n'est pas critique dans ce cas. La plupart des méthodes proposées permettaient de déterminer une bonne solution en quelques minutes, ce qui est tout à fait acceptable dans les circonstances, mais qui peut devenir problématique dans le cas du redéploiement dynamique. En effet, puisque le problème de redéploiement dynamique doit être résolu en temps réel, différentes stratégies devront être développées afin de réduire l'effort de calcul. Ces stratégies de modélisation et de résolution sont discutées à la section suivante.

3.3 Redéploiement multi-période et dynamique

Dans les années 1970, le redéploiement des véhicules a été considéré par Kolesar et Walker (1974) dans le contexte d'un service d'incendie. Dans le cas d'un service d'incendie, plusieurs véhicules sont généralement envoyés sur les lieux d'un incident à la suite d'un appel. Une telle affectation des camions de pompiers peut laisser un certain nombre de casernes vacantes et ainsi amener une dégradation temporaire du service offert dans les secteurs qu'elles desservent. Lorsqu'une telle situation survient, certains camions peuvent être relocalisés vers les casernes vides afin de retrouver une couverture adéquate. Dans ce contexte, Kolesar et Walker ont pro-

posé un algorithme visant à déterminer le moment opportun pour effectuer ces relocalisations, les casernes à remplir et les camions à relocaliser. Évidemment, la variation de la disponibilité des véhicules rencontrée dans le cas d'un service d'incendie se présente également dans le cas d'un SPU. Pourtant, peu d'auteurs ont abordé le problème du redéploiement dans le contexte des SPU jusqu'à présent.

Tel que mentionné précédemment, une distinction est faite dans ce chapitre entre le redéploiement multi-période et le redéploiement dynamique. Dans le cas du redéploiement multi-période, une journée de travail peut être divisée en plusieurs périodes de temps caractérisées, par exemple, par des profils de demandes différents. Les véhicules disponibles sont alors relocalisés entre les périodes en fonction de l'évolution de la demande ou des paramètres du système dans le temps. Le redéploiement dynamique vise plutôt à relocaliser les véhicules en temps réel, lorsque l'état du système, par exemple le nombre de véhicules disponibles, change ou le requiert (selon des critères prédéfinis), notamment suite à l'affectation d'un ou de plusieurs véhicules à des appels d'urgence. Dans les deux cas, le redéploiement vise à maintenir un niveau de service adéquat en tout temps. Dans cette section, nous présenterons les différents modèles qui ont été proposés jusqu'à présent afin de résoudre le problème de redéploiement multi-période et dynamique des véhicules ambulanciers.

En considérant le problème de déploiement des véhicules ambulanciers pour la ville de Louisville au Kentucky, Repede et Bernardo (1994) ont constaté que les différents modèles proposés ne considéraient pas la variation de la demande dans le temps. Pourtant, dans le cas de Louisville comme dans bien d'autres contextes, le profil de la demande varie dans le temps. Repede et Bernardo ont alors formulé ce qui semble être, à notre connaissance, le premier modèle multi-période pour le redéploiement des véhicules ambulanciers. Le problème de localisation multi-période avec couverture espérée maximale (PLCEM-MP) ou *maximal expected coverage location model with time variation* (TIMEXCLP) en anglais qu'ils proposent est, comme son nom l'indique, une variante multi-période du PLCEM présenté par Daskin (1982, 1983). Le PLCEM-MP vise toujours à maximiser la couverture espérée, mais en considérant maintenant la variation de la demande et du nombre de véhicules à localiser selon la période de temps considérée. L'ensemble des zones de demande couvertes par une localisation donnée peut également varier en fonction du temps ce qui permet de tenir compte de la congestion du réseau routier par exemple. Malheureusement, le PLCEM-MP ne tient pas compte explicitement des efforts associés au redéploiement des véhicules entre les périodes, c'est-à-dire qu'il ne cherche pas à minimiser la distance totale parcourue par les véhicules entre deux postes d'attente ou à limiter le nombre de véhicules relocalisés, ce qui fait de lui une pure extension du PLCEM. La

formulation du PLCEM-MP est très semblable à la celle du modèle original à laquelle un indice est ajouté pour tenir compte des différentes périodes de temps. Le modèle est donc résolu une seule fois et les plans de déploiement sont appliqués pour chacune des périodes considérées. Les auteurs ne présentent pas de méthode particulière pour la résolution du problème.

Plus récemment, Rajagopalan *et al*. (2008) ont proposé une variante multi-période du PLCTP (ReVelle et Hogan, 1988), le modèle dynamique de localisation avec couverture totale disponible (DLCTD) ou *dynamic available coverage location model* (DACL). Le DLCTD cherche à minimiser le nombre de véhicules à localiser de façon à ce que chaque zone de demande soit couverte avec un niveau de fiabilité donné, mais en considérant aussi plusieurs périodes de temps. Un facteur correctif est intégré lors de la formulation des contraintes probabilistes afin de garantir la fiabilité du système, ce qui n'était pas fait dans le cas du PLCTP. Comme pour le PLCEM-MP, ce modèle n'impose pas de contraintes particulières pour tenir compte du déplacement des véhicules entre les périodes faisant de lui une extension multi-période assez directe du PLCTP. Ici encore, les plans de déploiement sont calculés une seule fois, au tout début, puis appliqués au moment opportun. Rajagopalan *et al*. (2008) ont proposé un algorithme de recherche avec tabous réactif pour résoudre le problème. L'algorithme proposé fournit des résultats en quelques minutes lorsqu'appliqué aux données du *Mecklenburg County*, en Caroline du Nord, où 168 zones de demande et 8 périodes de temps ont été considérées. Saydam *et al*. (2013) ont proposé une extension du DLCTD. Cette extension vise toujours la minimisation du nombre de véhicules à localiser, mais considère également la minimisation du nombre de véhicules redéployés entre les périodes. L'algorithme développé par Rajagopalan *et al*. (2008) a été adapté puis employé afin de résoudre ce problème. Les résultats obtenus pour le *Mecklenburg County*, en Caroline du Nord, ont permis de montrer que le nombre de véhicules redéployés peut être réduit de moitié, sans une augmentation importante de la taille de la flotte requise ou une réduction de la couverture, lorsque la minimisation du nombre de véhicules redéployés est aussi considérée.

Carpentier (2006) a proposé un modèle déterministe pour le redéploiement multi-période des véhicules ambulanciers. Celui-ci, contrairement aux deux modèles probabilistes proposés précédemment, considère explicitement le déplacement des véhicules entre les périodes en intégrant dans la fonction objectif un terme qui vise à minimiser les coûts associés au déplacement des véhicules entre les périodes. De plus, de façon similaire aux modèles proposés par Beraldi *et al*. (2004) et par Beraldi et Bruni (2009), il considère non seulement la localisation des véhicules, mais également l'affectation réelle des véhicules aux zones de demande. Le nombre minimal de véhicules f_{it} à affecter à une zone de demande i durant la période t est déterminé

afin de tenir compte de l'affectation multiple des véhicules. En considérant ainsi f_{it}, plus d'une demande peuvent être affectées à un véhicule au cours de la période de temps considérée, ce qui n'était pas le cas dans les modèles de Beraldi *et al*. (2004) et de Beraldi et Bruni (2009). Des périodes de temps plus longues pourront donc être prises en compte.

En identifiant x_{ijt}, une variable binaire qui vaut 1 si la zone de demande i est couverte par un véhicule localisé au site j à la période t, et 0 autrement, z_{jlt}, une variable binaire qui vaut 1 si le véhicule l est localisé au site j à la période t, et 0 autrement, y_{jklt}, une variable binaire qui vaut 1 si le véhicule l est déplacé du site j au site k ($j \neq k$) au début de la période t, et 0 autrement, d_{ij}, la distance à parcourir pour aller de i à j, a_{it}, le nombre moyen d'appels placés en i à la période t, p_t, le nombre de véhicules disponibles à la période t, et M_i, l'ensemble des sites pouvant couvrir la zone de demande i à l'intérieur d'une distance prédéfinie S, le modèle proposé par Carpentier (2006) se formule de la façon suivante :

$$\min \sum_{i=1}^{n} \sum_{j \in M_i} \sum_{t=1}^{T} \frac{a_{it}}{f_{it}} d_{ij} x_{ij} + \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{t=1}^{T} \sum_{l=1}^{p_t} d_{jk} y_{jklt} \tag{3.86}$$

sous les contraintes :

$$\sum_{j \in M_i} x_{ijt} = f_{it},\ i = 1,...,n,\ t = 1,...,T, \tag{3.87}$$

$$\sum_{l=1}^{p_t} z_{jlt} \leq 1,\ j = 1,...,m\ t = 1,...,T, \tag{3.88}$$

$$\sum_{j=1}^{m} z_{jlt} = 1,\ l = 1,...,p_t\ \ t = 1,...,T, \tag{3.89}$$

$$\sum_{j=1}^{m} \sum_{l=1}^{p_t} z_{jlt} = p_t,\ t = 1,...,T, \tag{3.90}$$

$$x_{ijt} \leq \sum_{l=1}^{p_t} z_{jlt}, j \in N_i,\ i = 1,...,n\ t = 1,...,T, \tag{3.91}$$

$$\sum_{i=1}^{n} x_{ijt} \leq N_{max},\ j = 1,...,m,\ t = 1,...,T, \tag{3.92}$$

$$y_{jklt} \geq z_{klt} + z_{jl,t-1} - 1,\ k \in \{J | j \neq k\},\ t = 2,...,T,\ j = 1,...,m\ et\ l = 1,...,p_t, \tag{3.93}$$

$$y_{jkl1} \geq z_{kl1} + z_{jls} - 1,\ k \in \{J | j \neq k\},\ j = 1,...,m\ et\ l = 1,...,p_t, \tag{3.94}$$

$$x_{ijt}, y_{jklt}, z_{jlt} \in \{0,1\}. \tag{3.95}$$

L'objectif du modèle consiste à minimiser les coûts associés à la distance parcourue pour servir les demandes plus les coûts associés au redéploiement entre les périodes (3.86) en localisant

un nombre donné de véhicules p_t à chaque période (3.90), où p_t est tel que chaque zone de demande peut être couverte par au moins un véhicule. Chaque zone de demande doit être couverte par le nombre requis de véhicules (3.87), chaque véhicule étant affecté à un site (3.89) et pouvant couvrir un nombre limité N_{max} de zones de demande (3.92). La co-localisation des véhicules est interdite (3.88). Enfin, les contraintes (3.93) et (3.94) déterminent le déplacement des véhicules entre deux périodes consécutives. Les auteurs notent que la résolution exacte du problème devient déraisonnable pour des instances de taille supérieure à 30 localisations potentielles. Afin de remédier à cette difficulté, les auteurs ont proposé deux méthodes heuristiques pour résoudre le problème : une approche basée sur la formulation mathématique mono-période et une méthode heuristique de construction et d'amélioration locale. Les plans de redéploiement pour toutes les périodes sont déterminés une seule fois.

Başar *et al*. (2011) ont considéré le problème consistant à déterminer les sites à utiliser pour la localisation des véhicules ambulanciers en considérant un horizon de planification de plusieurs périodes, de même qu'une limite sur le nombre de sites pouvant être utilisées à chaque période. Deux rayons de couverture différents sont également pris en compte dans ce problème. Dans les faits, le modèle proposé par Başar *et al*. (2011) constitue une variante multi-période des modèles MLCSM ou BACOP en anglais (Hogan et ReVelle, 1986) et MDS ou DSM en anglais (Gendreau *et al*., 1997). Plus précisément, le modèle multi-période à couverture double (MMPCD), ou *multi-period backup double covering model* (MPBDCM) en anglais, vise à déterminer les sites à utiliser de façon à maximiser la population couverte par deux sites distincts, respectivement à l'intérieur de S et S', pour toutes les périodes de temps considérées. Ce modèle se distingue des modèles multi-période précédents d'abord parce qu'il considère le fait que lorsqu'un site est utilisé, il doit être utilisé jusqu'à la fin de l'horizon de planification, mais également parce qu'il considère des périodes de temps plus longues. Ce problème se justifie lorsque le changement de statut d'un site implique des coûts ou inconvénients relativement importants.

En considérant x_{jt}, une variable binaire qui vaut 1 si le site j est utilisé à la période t, et 0 autrement, et y_{it}, une variable binaire qui vaut 1 si la population en i est couverte deux fois respectivement à l'intérieur de S et S' à la période t, et 0 autrement, et en définissant M_i, l'ensemble des sites pouvant couvrir la zone de demande i à l'intérieur de S, M'_i, l'ensemble des sites pouvant couvrir la zone de demande i à l'intérieur de S', $S' > S$, p_t, le nombre maximal de sites utilisés à la période t et a_{it}, la population de la zone de demande i à la période t, le MMPCD se formule de la manière suivante :

MMPCD

$$\max \sum_{t=1}^{T} \sum_{i=1}^{n} a_{it} y_{it} \tag{3.96}$$

sous les contraintes :

$$\sum_{j=1}^{m} x_{jt} \leq p_t,\ t = 1,...,T, \tag{3.97}$$

$$\sum_{j \in M_i} x_{jt} - y_{it} \geq 0,\ i = 1,...,n,\ t = 1,...,T, \tag{3.98}$$

$$\sum_{j \in M_i'} x_{jt} - 2y_{it} \geq 0,\ i = 1,...,n,\ t = 1,...,T, \tag{3.99}$$

$$x_{jt} - x_{j,t-1} \geq 0,\ j = 1,...,m,\ t = 1,...,T, \tag{3.100}$$

$$x_{jt}, y_{it} \in \{0,1\},\ i = 1,...,n,\ j = 1,...,m,\ t = 1,...,T. \tag{3.101}$$

Ainsi, le MMPCD vise à maximiser la population couverte deux fois respectivement à l'intérieur de S et S' (3.96) de manière à ce que le nombre maximal de sites utilisés à chaque période soit respecté (3.97), que chaque zone de demande soit couverte adéquatement (3.98) (3.99), et enfin, qu'un site utilisé à une période donnée le demeure jusqu'à la fin de l'horizon de planification (3.100). Afin de résoudre ce problème, Başar *et al.* (2011) ont proposé un algorithme de recherche avec tabous. Les expérimentations menées sur des instances générées aléatoirement, de même que sur des instances tirées du cas de la ville d'Istanbul en Turquie, ont permis de montrer que l'algorithme proposé par les auteurs permet de fournir des bons résultats dans des délais de temps raisonnables.

Enfin, Schmid et Doerner (2010) présentent une extension multi-période du modèle avec double standard (MDS), ou *double standard model* (DSM) en anglais, proposé par Gendreau *et al.* (1997). Ce modèle, le modèle multi-période avec double standard (MMPDS), ou *multi-period double standard model* (mDSM) en anglais, se distingue de la version statique puisqu'il considère maintenant la variabilité du temps de déplacement entre les périodes, due à la congestion routière, par exemple. En effet, les auteurs ont pu constater que, dans certains contextes, le temps de déplacement peut varier de manière importante au cours d'une journée. Pourtant peu de modèles prennent en compte cet aspect, ce qui peut mener à une surestimation de la couverture globale. Afin de pallier à ce problème, les auteurs ont proposé un modèle multi-période qui considère des temps de déplacement différents pour chaque période de temps de manière à définir des régions de couverture propres à chacune de ces périodes. Le nombre de véhicules disponibles, la capacité des localisations potentielles de même que la demande demeurent toutefois constants sur tout l'horizon de planification. Le MMPDS se veut donc une extension

multi-période relativement directe du MDS où un terme de pénalité est intégré à la fonction objectif afin de limiter le nombre de véhicules déplacés entre chaque période. Le MMPDS se distingue également de la version statique puisqu'il impose une limite sur le nombre de demandes qu'un véhicule peut traiter. Cette contrainte a été introduite au modèle statique par Doerner *et al*. (2005). Une méthode de résolution basée sur la recherche à voisinage variable a été développée par Schmid et Doerner (2010) afin de résoudre le MMPDS. Les expérimentations menées sur les données de la ville de Vienne, en Autriche (3920 points de demande, entre 16 et 163 localisations potentielles, et 14 véhicules) ont permis de montrer qu'une amélioration de la couverture globale de 10 % était possible lorsque la solution du MMPDS était considérée, plutôt que la solution du modèle statique. De plus, les expérimentations ont permis de montrer que la méthode de résolution basée sur la recherche à voisinage variable permet de trouver des solutions de haute qualité, dans des délais de temps très raisonnables.

Les modèles présentés jusqu'à maintenant considéraient la variation de la demande aussi bien que du nombre de véhicules disponibles pour les différentes périodes de temps considérées. Bien qu'ils constituent une représentation améliorée de la réalité en ce qui à trait aux mouvements de la population durant une journée, ils ne permettent pas de considérer l'aspect dynamique lié au changement de l'état du système résultant de l'affectation ou de la remise en disponibilité d'un véhicule. Pour traiter cet aspect, il faudrait que les redéploiements considérés permettent de maintenir un bon niveau de service avec un nombre réduit de véhicules, certains véhicules étant alors en service.

Le premier modèle à considérer explicitement l'aspect dynamique du problème de redéploiement fut proposé par Gendreau *et al*. (2001). Le problème de redéploiement au temps t (PR-t) proposé par Gendreau *et al*. (2001) se base en fait sur le MDS proposé d'abord par les mêmes auteurs (Gendreau *et al*., 1997). Le PR-t vise donc toujours à maximiser la population couverte deux fois à l'intérieur d'un délai prescrit, mais en minimisant aussi les coûts associés au redéploiement. Afin de considérer ces coûts de redéploiement, un terme de pénalité qui prend en compte l'historique des mouvements de redéploiement des véhicules est intégré à la fonction objectif. Ce terme de pénalité vise à décourager les déplacements fréquents, les distances de redéploiement trop longues et les allers-retours entre un même poste d'attente. Ce terme de pénalité est actualisé à toutes les périodes. Ainsi, en considérant y_{jk} une variable binaire qui vaut 1 si le véhicule k est localisé en j, et 0 autrement, et M^t_{kl}, le terme de pénalité associé à la relocalisation du véhicule k de sa localisation actuelle vers la localisation j au temps t, la nouvelle fonction objectif se formule comme suit :

$$\max \sum_{i=1}^{n} a_i u_i - \sum_{j=1}^{m} \sum_{k=1}^{p} M^t_{jk} y_{jk}. \tag{3.102}$$

De plus, les contraintes (3.31) et (3.32) du MSD sont modifiées afin de considérer la variable y_{jk}, ce qui donne lieu aux contraintes suivantes :

$$\sum_{j=1}^{m} y_{jk} = 1,\ k = 1,...,p, \tag{3.103}$$

$$\sum_{k=1}^{p} y_{jk} \leq p_j,\ j = 1,...,n. \tag{3.104}$$

À toute fin pratique, le RP-t doit être résolu à chaque fois qu'un véhicule est affecté à une demande. Toutefois, dans le contexte des SPU, les décisions quant au redéploiement dynamique des véhicules doivent généralement être prises en temps réel et le temps nécessaire pour résoudre le modèle à la suite de l'affectation d'un véhicule peut s'avérer trop long. Afin de contourner cette difficulté, les auteurs proposent d'utiliser le temps disponible entre deux appels afin de déterminer un plan de relocalisation pour l'affectation possible de chaque véhicule. De cette façon, lorsque l'identité du véhicule affecté est connue, le plan de redéploiement correspondant est appliqué. Puisque le PR-t doit être résolu pour chaque affectation possible, une méthode de résolution très rapide est requise. À cet effet, les auteurs ont proposé une métaheuristique basée sur la recherche avec tabous inspirée de la méthode proposée pour la résolution du MSD (Gendreau *et al.*, 1997). Les auteurs proposent alors de résoudre en parallèle (c'est-à-dire sur plusieurs ordinateurs ou CPUs) les différents PR-t afin d'arriver à déterminer une bonne solution dans les délais disponibles. La méthode proposée a été testée avec succès sur les données fournies par Urgences-santé, à Montréal. Plus récemment, Moeini *et al.* (2013) ont proposé une extension du PR-t qui considère différentes exigences en termes de couverture, c'est-à-dire que pour certaines zones de demande, la double couverture sera comptabilisée dans la fonction objectif, alors que pour d'autres, seule la première couverture sera prise en compte. Une telle adaptation se justifie dans un contexte où l'intensité des demandes n'est pas très élevée, comme c'est le cas dans le contexte étudié, celui du département français de Val-de-Marne. Dans ce cas, de 20 à 30 appels par jour requièrent généralement l'intervention d'une des huit équipes paramédicales en service. Ainsi, puisque le volume d'appels est relativement bas, la double couverture ne sera requise que pour certaines zones de demande.

Gendreau *et al.* (2006) ont proposé un autre modèle pour la relocalisation dynamique, mais qui s'applique aux véhicules de médecins plutôt qu'aux véhicules ambulanciers. L'approche proposée est similaire à celle de Kolesar et Walker (1974), c'est-à-dire déterminer *a priori*

tous les plans de redéploiement en fonction de tous les états possibles du système. Puisque les véhicules de médecins sont peu nombreux, le nombre d'états possibles est petit, ce qui rend cette approche envisageable. Le problème de relocalisation maximale espérée (PRME), ou *maximal expected relocation problem* (MECRP) en anglais, proposé par Gendreau *et al*. (2006) vise à maximiser la couverture espérée à l'intérieur d'un délai prescrit, et ce, pour tous les états du système, où l'état du système est représenté par le nombre de véhicules disponibles. Le PRME impose également une limite sur le nombre de véhicules pouvant être relocalisés entre les états.

En identifiant par a_i, la densité de population à la zone de demande i, q_k, la probabilité d'être dans l'état k, $k = 0, ..., P$ où P est le nombre total de véhicules, x_{jk}, une variable binaire qui vaut 1 si un véhicule est localisé en j à l'état k, et 0 autrement, y_{ik}, une variable binaire qui vaut 1 si la zone de demande i est couverte par au moins un véhicule à l'état k, et 0 autrement, et u_{jk}, une variable binaire qui vaut 1 si la localisation j cesse d'être utilisée lorsque l'on passe de l'état k à l'état $k+1$, le PRME se formule comme suit :

PRME

$$\max \sum_{k=1}^{P} \sum_{i=1}^{n} a_i q_k y_{ik} \tag{3.105}$$

sous les contraintes :

$$\sum_{j \in M_i} x_{jk} \geq y_{ik},\ i = 1, ..., n,\ k = 0, ..., P, \tag{3.106}$$

$$\sum_{j=1}^{m} x_{jk} = k,\ k = 1, ..., P, \tag{3.107}$$

$$x_{jk} - x_{j,k+1} \leq u_{jk}, \tag{3.108}$$

$$\sum_{j=1}^{m} u_{jk} \leq \alpha_k,\ k = 1, ..., P-1, \tag{3.109}$$

$$x_{jk} \in \{0,1\}, u_{jk} \in \{0,1\},\ j = 1, ..., m,\ k = 1, ..., P, \tag{3.110}$$

$$y_{ik} \in \{0,1\},\ i = 1, ..., n,\ k = 1, ..., P. \tag{3.111}$$

Il est important de constater que les contraintes (3.106) et (3.107) sont essentiellement les mêmes que pour le PLCM. Les contraintes (3.108) et (3.109) visent, quant à elles, à contrôler le nombre de véhicules α_k pouvant être relocalisés. Le modèle peut être résolu une seule fois à l'avance, puis les redéploiements correspondants sont réalisés lorsque nécessaire. Le PRME a été appliqué aux données d'Urgences-santé à Montréal, au Canada, et résolu avec CPLEX.

Andersson et Värbrand (2007) ont également proposé un modèle pour le redéploiement dynamique des véhicules ambulanciers. Le modèle proposé se distingue des modèles précédents par

son utilisation de la notion de capacité de réponse, ou *preparedness* en anglais, afin de mesurer les performances du système. Les auteurs définissent la *preparedness* comme la capacité du système à répondre aux demandes présentes et futures. Ainsi, en définissant a_i, le poids associé aux demandes placées en i, K_i, le nombre de véhicules contribuant au calcul de la *preparedness* d'une zone de demande i (généralement les K_i véhicules les plus proches), t_i^k, le temps de déplacement du véhicule k vers la zone i, et γ^k, le facteur associé à la contribution du véhicule k, la *preparedness* pour la zone de demande i, ρ_i, est donnée par la formule suivante :

$$\rho_i = \frac{1}{a_i} \sum_{k=1}^{K_i} \frac{\gamma^k}{t_i^k}. \tag{3.112}$$

Le niveau de *preparedness* est vérifié régulièrement puis le redéploiement des véhicules est déclenché lorsque le niveau de *preparedness* baisse en-deçà d'une valeur prédéfinie. Ce modèle, baptisé DYNAROC par les auteurs, cherche à minimiser le temps maximal de déplacement entre les postes d'attente des véhicules relocalisés (3.113). Tout comme dans le cas du PR-t, DYNAROC considère un ensemble de contraintes pratiques. Ces contraintes visent à limiter le temps de déplacement des véhicules relocalisés (3.114) (3.115), de même que le nombre de véhicules pouvant être relocalisés (3.116). Une limite ρ_{min} est naturellement imposée quant au niveau de *preparedness* à atteindre une fois la relocalisation effectuée (3.117). En notant t_i^k, le temps de déplacement du véhicule k pour atteindre la zone i, x_i^k, une variable binaire qui vaut 1 si le véhicule k est relocalisé vers le site localisé dans la zone i, et 0 autrement, N_k, l'ensemble des zones pouvant être atteintes par le véhicule k à l'intérieur d'un délai prescrit S, P_{max}, le nombre maximal de véhicules pouvant être relocalisés, et P, le nombre total de véhicules, DYNAROC s'écrit comme suit :

$$\min z \tag{3.113}$$

sous les contraintes :

$$z \geq \sum_{i \in N_k} t_i^k x_i^k, \; k = 1, ..., P, \tag{3.114}$$

$$\sum_{i \in N_k} x_i^k \leq 1, \; k = 1, ..., P, \tag{3.115}$$

$$\sum_{k=1}^{P} \sum_{i \in N_k} x_i^k \leq P_{max}, \tag{3.116}$$

$$\frac{1}{a_i} \sum_{l=1}^{K_i} \frac{\gamma^k}{t_i^l(x_1^1, ..., x_N^P)} \geq \rho_{min}, \; i = 1, ..., n, \tag{3.117}$$

$$x_j^k \in \{0,1\},\ i = 1,...,n,\ k = 1,...,P. \tag{3.118}$$

Afin de résoudre ce modèle, les auteurs proposent une méthode heuristique simple et la testent sur des données de la ville de Stockholm, en Suède.

De façon similaire à Gendreau *et al*. (2006), Nair et Miller-Hooks (2009) ont proposé un modèle de localisation-relocalisation qui considère l'évolution de l'état du système dans le temps. Dans ce cas, l'état du système r à un instant donné est défini par le nombre de véhicules disponibles, les temps de déplacement sur le réseau routier, de même que la distribution de probabilité associée à l'arrivée des appels de détresse. Le modèle multi-objectif proposé par Nair et Miller-Hooks (2009) comporte deux objectifs visant respectivement la minimisation de la double couverture et la minimisation des coûts associés à la localisation et à la relocalisation des véhicules. En identifiant par R, l'ensemble des états possibles, $|R| = r_{max}$, u_{ir}, une variable binaire qui vaut 1 si la zone de demande i est couverte au moins deux fois à l'intérieur de délai prescrit à l'état r, p_{ir}, la probabilité pour qu'une demande soit placée à partir de la zone de demande i à l'état r, y_{jr}, une variable binaire qui vaut 1 si un véhicule est localisé en j à l'état r, et 0 autrement, c_{jr}, le coût associé à une telle localisation, $w_{jl}^{rr'}$, une variable binaire qui vaut 1 si un véhicule est relocalisé de j vers l lorsque l'état passe de r à r', et 0 autrement, $RR_{jl}^{rr'}$, le coût associé à une telle relocalisation et $PP_{rr'}$, la probabilité de passer de l'état r à l'état r', les deux objectifs se formulent comme suit :

$$Z_1 = \max \sum_{r=1}^{r_{max}} \sum_{i=1}^{n} p_{ir} u_{ir} \tag{3.119}$$

$$Z_2 = \min \sum_{r=1}^{r_{max}} \sum_{i=1}^{n} c_{ir} y_{ir} + \sum_{r=1, r \neq r'}^{r_{max}} \sum_{r'=1}^{r_{max}} \sum_{j=1}^{m} \sum_{l=1}^{m} PP_{rr'} RR_{jl}^{rr'} w_{jl}^{rr'} \tag{3.120}$$

Les contraintes du modèle sont, quant à elles, similaires aux contraintes du MDS (Gendreau *et al*., 1997), mais adaptées au contexte étudié. De la même manière que dans le cas du PRME (Gendreau *et al*., 2006), le modèle est résolu une seule fois *a priori* de manière à fournir la localisation des véhicules pour l'ensemble des états considérés. Les résultats obtenus en utilisant les données de la ville de Montréal, au Canada, ont permis de montrer qu'une amélioration des performances est possible lorsqu'une telle stratégie de redéploiement est considérée, soit de 1,3 % à 6,4 % selon le nombre de véhicules déployés.

Maxwell *et al*. (2009) ont proposé une toute nouvelle approche pour le redéploiement dynamique des véhicules ambulanciers basée sur la programmation dynamique. Dans le problème considéré par les auteurs, le redéploiement est effectué lorsqu'un véhicule se libère. Seul le véhicule libéré est impliqué dans le processus de relocalisation. On parlera aussi du problème de

repositionnement d'un véhicule à la suite de la libération. En d'autres termes, lorsqu'un véhicule se libère, le problème de redéploiement ou de repositionnement consiste à déterminer le poste d'attente où celui-ci sera localisé de façon à maximiser le nombre d'appels pouvant être atteints à l'intérieur d'un délai prescrit. Ce modèle se distingue donc des modèles de redéploiement dynamique puisqu'il considère la relocalisation d'un seul véhicule, et seulement au moment où ce dernier se libère. Cette politique a pour avantage de réduire les inconvénients causés aux techniciens ambulanciers puisqu'elle réduit la fréquence des déplacements possibles. De plus, d'un point de vue mathématique, le fait de considérer un seul véhicule lors du redéploiement permet de réduire significativement le nombre de décisions possibles.

En considérant s, l'état du système défini en fonction du temps et de l'événement courant, d'un vecteur A décrivant l'état de chaque véhicule et d'un vecteur C décrivant l'état de chaque appel, $X(s)$ l'ensemble des décisions possibles à l'état s, $c(s_k, x_k, s_{k+1})$ le coût de transition d'un état s_k à un état s_{k+1} étant donné une décision x_k, $f(s,x,w(s,x))$, une fonction de transfert qui dépend de l'état du système, de la décision effectuée et d'un élément aléatoire $w(s,x)$, α un facteur d'actualisation réel fixe avec $\alpha \in [0,1[$ et $\tau(s)$ le temps auquel le système visite l'état s, la politique qui minimise le coût total espéré actualisé étant donné un état initial s peut être déterminée en calculant la fonction valeur à partir de l'équation d'optimalité suivante :

$$J(s) = \min_{x \in X(s)} \left\{ \mathbb{E}[c(s,x,f(s,x,w(s,x))) + \alpha^{\tau(f(s,x,w(s,x)))-\tau(s)} J(f(s,x,w(s,x)))] \right\}. \qquad (3.121)$$

Il est important de rappeler que, dans ce cas, l'ensemble des décisions réalisables $X(s)$ est relativement petit, puisqu'un seul véhicule est considéré pour la relocalisation. De plus, le coût de transition permet de comptabiliser le nombre d'appels qui seront servis au-delà du délai prescrit. Ainsi $c(s_k, x_k, s_{k+1})$ vaut 1 si le prochain événement $e(s_{k+1})$ est de type *Arrivée d'un véhicule sur les lieux de l'incident*, si l'appel correspondant est urgent et si le délai prescrit est dépassé, et 0 autrement.

L'évaluation de la fonction valeur représente une tâche difficile. En effet, puisque la variable d'état est de grande dimensionnalité, le nombre de valeurs possibles pour la variable d'état est très grand. L'algorithme de la programmation dynamique classique n'est donc pas applicable directement. Pour remédier à cette situation, les auteurs proposent d'utiliser la programmation dynamique approximée, c'est-à-dire d'utiliser une approximation de la fonction valeur de la forme $J(s,r) = \sum_{p=1}^{P} r_p \phi_p(s)$, où $r = \{r_p : p = 1,...,P\}$ sont des paramètres ajustables et $\{\phi_p : p = 1,...,P\}$, des fonctions de base fixées. Le défi consiste donc à déterminer les bonnes valeurs de r et ϕ de manière à déterminer une approximation adéquate de la fonction valeur

originale. Lorsqu'une bonne approximation de ces paramètres a été trouvée, la politique optimale peut être identifiée en énumérant toutes les décisions possibles et en évaluant l'espérance associée à chacune par simulation de Monte Carlo. Les tests menés sur des instances tirées du cas de deux villes, la ville d'Edmonton au Canada et une seconde ville qui ne peut être nommée par les auteurs, ont permis de montrer que la politique obtenue grâce à la méthode proposée permet d'améliorer les performances d'un système d'environ 4 % par rapport à l'utilisation d'une politique myope, c'est-à-dire au fait de retourner un véhicule à sa base originale. De plus, les auteurs ont montré qu'il était possible d'améliorer les performances du système en augmentant la fréquence et le nombre de véhicules redéployés, à l'intérieur de certaines limites. Dans ces cas, le temps de calcul devient beaucoup plus long.

Schmid (2012) a également utilisé la programmation dynamique afin de formuler le problème de redéploiement dynamique des véhicules ambulanciers. Dans ce cas, les décisions de redéploiement sont effectuées lorsqu'un véhicule se libère, et seul le véhicule nouvellement disponible est impliqué dans les décisions de repositionnement. En effet, dans le contexte étudié par l'auteure, les déplacements entre les différents postes d'attente sont interdits par la loi. Le modèle proposé par Schmid (2012) vise donc à déterminer la localisation des véhicules une fois qu'ils se libèrent, de même que l'affectation des demandes aux véhicules disponibles de manière à minimiser le temps de réponse moyen, et ce, en prenant en compte des temps de déplacement et un volume de demandes variant en fonction du temps de même qu'en considérant l'horizon de planification fini. L'équation de Bellman associée à ce problème se formule alors de la manière suivante :

$$V_t(S_t) = \min_{x_t}(c(S_t,x_t) + \mathbb{E}\{V_{t+1}(S_{t+1}(S_t,x_t,W_{t+1}))\}), \tag{3.122}$$

où S_t représente l'état du système au temps t, c'est-à-dire l'état des demandes et l'état des véhicules, x_t, une décision effectuée au temps t, W_t, l'information rendue disponible de l'instant $t-1$ à t, et $c(S_t,x_t)$, la contribution, en termes de temps de réponse, lorsqu'une décision x_t est prise et que l'état du système est S_t. Les décisions sont effectuées en considérant une politique $X_t^\pi(S_t)$ qui fournit un vecteur de décision x_t, réalisable à l'état S_t. La politique optimale minimise donc, en considérant un facteur d'actualisation γ, la somme des temps de réponse espérés sur l'ensemble de l'horizon de planification T en utilisant :

$$\min_{\pi\in\Pi} \mathbb{E} \sum_{t=0}^{T} \gamma^t c_t(S_t, X_t^\pi(S_t)). \tag{3.123}$$

Tout comme dans le cas de Maxwell *et al.* (2009), la programmation dynamique approximée a été utilisée afin de résoudre l'équation (3.123). En comparant les résultats à ceux obtenus pour

les politiques classiques et myopes, c'est-à-dire toujours affecter le véhicule le plus proche et retourner le véhicule à son poste d'attente initial, il a été possible de constater que la politique induite par l'approche proposée permet d'améliorer les performances du système. En effet, les tests menés grâce aux données de la ville de Vienne, en Autriche, indiquent qu'en déviant des politiques de répartition et de repositionnement classiques, le temps de réponse moyen peut s'améliorer d'environ 13%.

Naoum-Sawaya et Elhedhli (2013) ont, quant à eux, considéré la programmation stochastique avec fonction de recours afin de modéliser le problème de redéploiement dynamique des véhicules ambulanciers. Dans ce cas, les décisions de première étape sont reliées au choix de la localisation des véhicules, tandis que les décisions de deuxième étape concernent l'affectation des demandes reçues aux véhicules disponibles, une fois que les demandes se réalisent. Ainsi, en identifiant c_k, le coût associé à la relocalisation du véhicule k, λ, le coût associé au fait de ne pas servir une demande dans les délais prescrits, p_s, la probabilité d'occurence du scénario $s \in S$, l'ensemble des scénarios considérés, $|S| = s_{max}$, α, le pourcentage des demandes qui devront être atteintes à l'intérieur des délais prescrits, P_j, le nombre maximal de véhicules pouvant être localisés au poste d'attente j, D_{tot}, le nombre total de demandes et $U(t)$, l'ensemble des périodes $t' \in T$ pour lesquelles un véhicule devient non-disponible puisqu'affecté à une demande et en identifiant δ_{kj}, un paramètre qui vaut 1 si la localisation du véhicule k au site j implique une relocalisation, et 0 autrement, a_{jts}, un paramètre qui vaut 1 si une demande reçue à la période t du scénario s peut être atteinte à l'intérieur des délais prescrits de la station j ,et 0 autrement, r_{kjt}, un paramètre qui vaut 1 si le véhicule k peut atteindre le site j avant le début de la période t, et 0 autrement, et γ_{ts}, un paramètre qui vaut 1 si une demande est reçue à la période t du scénario s, et 0 autrement, le problème considéré par Naoum-Sawaya et Elhedhi se formule de la manière suivante :

$$\min \sum_{k=1}^{P} \sum_{j=1}^{m} c_k \delta_{kj} y_{kj} + \lambda \sum_{s=1}^{s_{max}} p_s \sum_{t=1}^{T} (1 - \sum_{k=1}^{P} x_{kts}) \tag{3.124}$$

sous les contraintes :

$$x_{kts} - \sum_{j=1}^{m} (r_{kjt} a_{jts} y_{kj}) \leq 0,\ k = 1,...,P,\ t = 1,...,T,\ s = 1,...,s_{max}, \tag{3.125}$$

$$x_{kts} + \sum_{t' \in U(t)} x_{kt's} \leq 1,\ k = 1,...,P,\ t = 1,...,T,\ s = 1,...,s_{max}, \tag{3.126}$$

$$\sum_{k=1}^{P} x_{kts} \leq \gamma_{ts},\ t = 1,...,T,\ s = 1,...,s_{max}, \tag{3.127}$$

$$\sum_{k=1}^{P} y_{kj} \leq P_j,\ j = 1,...,m, \tag{3.128}$$

$$\sum_{k=1}^{P}\sum_{t=1}^{T}\sum_{s=1}^{s_{max}} x_{kts} \geq \alpha D_{tot}, \tag{3.129}$$

$$y_{kj} \in \{0,1\}, x_{kts} \in \{0,1\},\ k = 1,...,P,\ j = 1,...,m,\ t = 1,...,T,\ s = 1,...,s_{max}. \tag{3.130}$$

Dans ce cas, y_{kj} est une variable binaire qui vaut 1 si le véhicule k est localisé en j, et 0 autrement, et x_{kts}, une variable binaire qui vaut 1 si le véhicule k est affecté à une demande placée à la période t du scénario s, et 0 autrement. En effet, il est important de mentionner que la longueur des périodes de temps a été choisie de manière à ce qu'au plus une demande soit placée de l'instant t à l'instant $t+1$. L'objectif du modèle (3.124) consiste donc à minimiser le nombre de véhicules relocalisés, puis à minimiser le nombre de demandes qui ne peuvent être répondues à l'intérieur des délais prescrits. Un véhicule k ne pourra servir adéquatement une demande que s'il peut atteindre la demande à l'intérieur des délais prescrits (3.125). De plus, tout véhicule répondant à une demande deviendra non disponible pour une période de temps $U(t)$ (3.126). Enfin, au plus un véhicule devra être affecté à chaque demande (3.127), la capacité de chaque poste d'attente devra être respectée (3.128) et au moins une proportion α des demandes devront être répondues à l'intérieur des délais prescrits (3.129). Le modèle proposé par Naoum-Sawaya et Elhedhli (2013) a été appliqué au cas de la ville de Waterloo, au Canada, puis résolu grâce à CPLEX. Cinquante scénarios ont alors été considérés prenant en compte un horizon de planification de 2 heures fractionné en 120 périodes d'une minute. Les différents résultats obtenus ont permis de montrer que l'approche proposée permet de maintenir un niveau de service adéquat tout en minimisant le nombre de relocalisations. Le temps de calcul est également relativement court, soit en moyenne 40 secondes pour les instances considérées.

Enfin, Mason (2013) a présenté un modèle pour le problème de redéploiement dynamique des véhicules ambulanciers similaire à celui proposé par Gendreau *et al*. (2001). Ce modèle, nommé modèle de redéploiement en temps réel basé sur une couverture généralisée ou *real-time multi-view generalized-cover repositioning model* (RtMvGcRM) en anglais, a été implémenté puis implanté au sein d'un logiciel de gestion des SPU, Optima Live. Ce logiciel fournit aux SPU différentes recommandations quant à la relocalisation des véhicules ambulanciers en temps réel. Ce modèle vise donc à déterminer la localisation des véhicules disponibles de manière à maximiser le gain associé à leur localisation tout en minimisant les coûts associés à leur relocalisation. De manière similaire à Gendreau *et al*. (2001), le terme de pénalité associé aux coûts de relocalisation vise à décourager les déplacements fréquents, les distances de relocalisation trop longues et les allers-retours entre un même poste d'attente. Ainsi, en définissant V, l'ensemble

des vues considérées, c'est-à-dire les véhicules, les types d'appels (en termes de priorité) et les délais prescrits considérés dans le calcul de la couverture, $|V| = v_{max}$, d_{kjiv}, un paramètre qui vaut 1 si le véhicule k est considéré dans le calcul de la couverture et que le délai requis pour atteindre une demande placée en i à partir de j est plus petit que le délai prescrit pour la vue v et 0 autrement, et en identifiant, x_{kj}, une variable binaire valant 1 si le véhicule k est localisé en j, et 0 autrement, c_{kj}, le coût associé à la localisation du véhicule k en j, y_{iv} une variable permettant de mesurer la couverture totale de la zone de demande i pour la vue v, et $g_{iv}(y_{iv})$, le gain résultant, le RtMvGcRM se formule comme suit :

RtMvGcRM(t)

$$\max \sum_{i=1}^{n} \sum_{v=1}^{v_{max}} g_{iv}(y_{iv}) - \sum_{k=1}^{P} \sum_{j=1}^{m} c_{kj} x_{kj} \tag{3.131}$$

sous les contraintes :

$$\sum_{j=1}^{m} x_{kj} = 1,\ k = 1,...,P, \tag{3.132}$$

$$y_{iv} = \sum_{k=1}^{P} \sum_{j=1}^{m} d_{kjiv} x_{kj},\ k = 1,...,P,\ v = 1,...,v_{max}, \tag{3.133}$$

$$x_{kj} \in \{0,1\},\ k = 1,...,P,\ j = 1,...,P. \tag{3.134}$$

Le modèle présenté par Mason (2013) vise donc à maximiser le gain associé à la couverture des zones de demandes moins le coût relié à la relocalisation des véhicules (3.131) de manière à assurer que chaque véhicule soit localisé à un poste d'attente (3.132).

Dans un cas comme dans l'autre, le redéploiement multi-période et dynamique vise à maintenir un niveau de service adéquat en tout temps en considérant l'évolution du système dans le temps, ce qui n'était pas pris en compte lors du déploiement statique des véhicules ambulanciers. Nécessairement, le redéploiement est grandement influencé par la réalisation réelle de la demande. Des questions émergent alors quant aux performances réelles des plans de redéploiement établis. L'application de différents plans de déploiement permet-elle d'améliorer le niveau de service en pratique ? Les efforts de redéploiement sont-ils justifiés par un gain significatif de la couverture ? Qu'en est-il de la fréquence de redéploiement ? Comme dans le cas du déploiement statique, la simulation peut être utilisée afin d'évaluer les différentes règles de redéploiement et ainsi permettre une meilleure analyse du comportement global du système dans un contexte plus proche de la réalité. La plupart des modèles présentés dans cette section ont effectivement été analysés plus en profondeur par simulation. À la suite de leurs expérimentations, les auteurs présentés ici concluent tous que le redéploiement permet de maintenir un niveau de service adéquat et que les mécanismes mis en place permettent généralement d'y

parvenir en limitant les efforts de redéploiement. Enfin, la plupart des auteurs soulignent également la nécessité de développer des méthodes de résolution rapides afin de supporter la prise de décision en temps réel, ce qui n'était pas nécessaire dans le cas du déploiement statique.

3.4 Règles de répartition

Tel que mentionné plus tôt, les règles de répartition, c'est-à-dire la sélection du véhicule à affecter à un appel, peuvent avoir un impact important sur le temps écoulé entre la réception d'un appel et l'arrivée d'un véhicule sur les lieux de l'incident et, conséquemment, sur les performances du système. Les règles de répartition peuvent également avoir un impact sur la capacité du système à répondre aux demandes futures. En effet, à la suite de l'affectation d'un véhicule, une dégradation temporaire du service offert peut s'observer dans le secteur où celui-ci était localisé. Le redéploiement dynamique des véhicules est utilisé, dans certains cas, afin de retrouver une couverture adéquate avec un nombre réduit de véhicules, certains véhicules étant alors en service (c'est-à-dire affectés à des appels). Les règles de répartition et les stratégies de redéploiement sont donc intimement liées. Pour cette raison, nous avons jugé bon de présenter, dans cette section, différents travaux reliés aux règles de répartition. La liste des travaux présentée vise à couvrir certains travaux récents qui nous paraissent particulièrement pertinents dans le cadre de ce chapitre.

Ainsi, lorsqu'un appel de détresse est placé, un véhicule est affecté à l'appel en fonction des règles établies, et ce, le plus rapidement possible afin d'éviter des délais supplémentaires liés à la répartition. Afin d'assurer l'arrivée du véhicule sur les lieux de l'incident dans les meilleurs délais possibles, le véhicule qui se situe le plus proche de celui-ci est généralement sélectionné. Cette règle, grandement acceptée et utilisée en pratique, semble triviale pour la répartition des appels dont le niveau de priorité est élevé et qui requièrent une intervention rapide, le temps de réponse étant souvent critique pour la santé des patients. Dans ces cas, un véhicule déjà affecté à un appel de priorité inférieure peut même être réaffecté s'il peut atteindre le lieu de l'incident dans un délai de temps plus court que les véhicules disponibles. Dans la plupart des cas, la règle de répartition selon laquelle le véhicule le plus proche de l'incident est sélectionné s'applique aussi dans le cas des appels de priorité inférieure. Bien que cette règle vise à assurer une réponse dans les meilleurs délais, elle ne tient toutefois pas compte de l'impact de la non-disponibilité du véhicule affecté sur la capacité du système à répondre aux demandes futures. D'autres règles de répartition pourraient donc être utilisées dans le cas des appels de priorité inférieure afin de mieux considérer les performances futures du système, mais en assurant toujours de les servir à l'intérieur du délai prescrit. C'est du moins ce que proposent certains auteurs (Gendreau *et al.*,

2001; Carpentier, 2006; Andersson et Värbrand, 2007). Ainsi, ces études ont proposé différentes règles qui s'appliquent plus naturellement pour la répartition des appels de priorité inférieure, mais qui requièrent néanmoins une intervention rapide.

Dans un premier temps, Gendreau *et al*. (2001) ont proposé, lors du développement d'un outil de gestion intégrée pour l'affectation et le redéploiement des véhicules ambulanciers, de sélectionner, parmi les véhicules disponibles à l'intérieur du délai prescrit, le véhicule qui minimise les efforts de redéploiement suite à son affectation à l'appel. Rappelons que, dans le système de gestion proposé par les auteurs, les plans de redéploiement associés à l'affectation de tous les véhicules sont calculés entre la réception de deux appels. Ainsi, au moment où la demande se réalise, le véhicule à affecter à l'appel est sélectionné selon les règles établies et le plan de redéploiement associé est mis en place.

Dans le même ordre d'idée, Andersson et Värbrand (2007) ont proposé de sélectionner, parmi les véhicules disponibles à l'intérieur du délai prescrit, le véhicule dont la non-disponibilité cause la moins grande détérioration de la capacité de réponse, ou *preparedness* en anglais, telle que calculée en (3.112). Un algorithme simple qui sélectionne le meilleur véhicule à affecter en fonction des règles de répartition choisies a également été développé dans ce contexte.

Carpentier (2006) a proposé deux règles de répartition supplémentaires qu'il a, par la suite, analysées par simulation. La première règle proposée consiste à sélectionner, parmi les véhicules disponibles à l'intérieur du délai prescrit, le véhicule qui couvre le moins de zones de demande. La deuxième règle consiste plutôt à sélectionner, parmi les véhicules disponibles à l'intérieur du délai prescrit, le véhicule qui minimise le nombre de zones qui deviendront non couvertes suite à son affectation. La première règle vise essentiellement à maximiser le nombre de zones couvertes, et plus particulièrement, la couverture multiple des points de demande, tandis que dans le deuxième cas, le nombre de zones non couvertes est minimisé. À la suite de ses expérimentations, l'auteur a pu constater que les règles de répartition étaient généralement reliées à la qualité des plans de déploiement. Il est donc très difficile d'évaluer seules les règles de répartition. En général, les règles de répartition ne peuvent pas assurer une bonne couverture si le plan de déploiement de départ n'est pas adéquat.

Dans tous ces cas cités ci-haut, si aucun véhicule n'est disponible à l'intérieur du délai prescrit, le véhicule le plus proche est envoyé sur le lieu de l'incident.

Enfin, Schmid (2012) a, quant à elle, montré que, lorsque le niveau de priorité d'un appel le permet, le choix d'envoyer le véhicule le plus proche afin de répondre à cet appel ne constitue pas toujours la meilleure stratégie à adopter. En effet, l'adoption d'autres stratégies de répartition couplée à de meilleures règles de repositionement semblent permettre une amélioration des

performances du système. Le modèle proposé par Schmid (2012) afin de déterminer le repositionement des véhicules nouvellement libérés permet aussi d'établir les décisions de répartition en tenant compte de l'état futur du système. Ce modèle, basé sur la programmation dynamique, a été présenté plus en détail à la section 3.3.

Bien entendu, la sélection d'un véhicule à affecter pourrait être effectuée de façon similaire, afin d'équilibrer la charge de travail des véhicules ou encore de tenir compte des fins de quart ou des pauses des techniciens ambulanciers. Ainsi, lorsque plusieurs véhicules sont disponibles à l'intérieur du délai prescrit, le véhicule ayant la plus petite charge de travail ou le technicien ambulancier n'ayant pas de pause ou de fin de quart planifié dans un avenir rapproché pourrait être sélectionné. L'information nécessaire quant à la charge de travail, aux événements planifiés et à la localisation des véhicules doit être disponible en temps réel, ce qui est maintenant rendu possible grâce aux nouvelles technologies.

3.5 Conclusion

Tout au long de ce chapitre, nous nous sommes intéressés aux problèmes de déploiement et de redéploiement des véhicules ambulanciers. Le problème de *déploiement* des véhicules ambulanciers consiste essentiellement à déterminer les postes d'attente à utiliser pour la localisation des véhicules entre deux affectations. Le problème de *redéploiement* consiste, quant à lui, à relocaliser les véhicules disponibles vers les différents postes d'attente potentiels de façon à assurer, en tout temps, une couverture adéquate de la population. Différentes approches quant à la modélisation et à la résolution des problèmes de déploiement et de redéploiement ont été présentées dans ce chapitre.

Que ce soit par programmation mathématique, simulation ou au moyen de la théorie des files d'attente, plusieurs auteurs se sont intéressés à la modélisation du problème de déploiement. Les modèles proposés ont beaucoup évolué au fil des ans et des applications traitées afin de fournir une représentation plus fidèle de la réalité. Différentes stratégies ont alors été envisagées afin de mieux considérer les différentes sources d'incertitude allant de l'intégration de la couverture multiple à l'utilisation de la programmation stochastique. Essentiellement, les modèles proposés considéraient directement ou indirectement l'incertitude liée à la réalisation de la demande. La simulation a été grandement utilisée afin de comparer différentes alternatives ou de valider les solutions obtenues par programmation mathématique.

Depuis les années 1990, certains auteurs se sont également intéressés au développement de modèles en lien avec le redéploiement multi-période ou dynamique des véhicules ambulanciers. Ces modèles visaient principalement à considérer l'évolution du système dans le temps. Diffé-

rentes stratégies ont donc été mises de l'avant afin d'intégrer non seulement l'aspect évolutif du système, mais aussi afin de réduire l'effort de calcul. En effet, le problème de redéploiement, principalement en ce qui concerne le redéploiement dynamique, doit être résolu en temps réel. Il requiert donc des méthodes de résolution rapides et efficientes. Différentes règles de répartition ont également été proposées afin de mieux tenir compte de la capacité du système à répondre aux demandes futures à la suite de l'affectation de certains véhicules. Ces règles s'appliquent plus naturellement pour la répartition des appels considérés moins urgents, mais qui requièrent tout de même une intervention rapide.

Bien que des efforts constants aient été déployés au fil des ans afin de résoudre les problèmes de déploiement et de redéploiement, différentes avenues de recherche demeurent à explorer. Dans leur revue publiée en 2003, Brotcorne *et al*. anticipaient une évolution des modèles dynamiques. Ils prévoyaient également l'utilisation de la programmation stochastique avec fonction de recours dans le développement de modèles dynamiques. Depuis, quelques modèles ont été développés en lien avec le redéploiement multi-période et dynamique. De plus, la programmation stochastique avec fonction de recours a été employée par Beraldi et Bruni (2009) afin de résoudre le problème de déploiement statique. À notre avis, il ne s'agit là que d'un pas de plus vers le développement de modèles dynamiques et stochastiques plus élaborés. Cela vient tout de même confirmer les prévisions de Brotcorne *et al*. (2003) et ainsi, nous porte à croire que ces avenues de recherche sont toujours pertinentes. La taille grandissante des problèmes traités de même que l'intégration de la programmation stochastique nous laissent également penser que des efforts supplémentaires pourraient être déployés dans le futur afin de développer des méthodes de résolution plus performantes.

Dans le même ordre d'idée, la présence accrue des nouvelles technologies fait en sorte qu'une foule d'informations quant à l'état des véhicules (c'est-à-dire la localisation, la charge de travail, le temps écoulé depuis le début d'une intervention) est disponible en temps réel. Tous les modèles présentés dans ce chapitre tenaient compte des véhicules disponibles sans considérer, à notre connaissance, la possibilité qu'un véhicule se libère dans un avenir rapproché. En effet, le distribution de probabilité du temps d'intervention de même que le temps écoulé depuis le début d'une intervention et la localisation du véhicule peut nous donner une indication sur la probabilité pour qu'un véhicule se libère prochainement dans une région donnée. En intégrant cette information lors de la prise de décision, la relocalisation d'un véhicule vers une zone à couverture réduite, mais où se retrouve déjà un véhicule qui pourrait se libérer dans un délai très court, pourrait être évitée. Il semble qu'en pratique, les répartiteurs tiennent compte de cette information avant de lancer le redéploiement, mais cela ne se reflète dans aucun des modèles

présentés. La raison de cette absence est certainement la difficulté liée à l'intégration d'une telle information. Considérer cette information pourrait toutefois s'avérer utile pour formuler une nouvelle mesure de performance, par exemple. Enfin, nous pensons que la relation entre les règles de répartition et les efforts associés au redéploiement des véhicules pourrait être analysée plus finement, au moyen de la simulation, par exemple.

Les problèmes de déploiement et de redéploiement des véhicules ambulanciers s'inscrivent donc dans un processus de décision complexe auquel doit faire face un SPU. En effet, afin de fournir un service adéquat, les SPU doivent mobiliser un ensemble de ressources, puis les gérer de façon efficiente. Certaines difficultés se présentent néanmoins lors de la prise de décision, notamment quant à la prévision de la demande et à l'évaluation du taux de succès d'une intervention. Il ressort clairement de cette analyse que plusieurs travaux ont été faits afin de soutenir ce processus de décision, principalement au niveau du déploiement et du redéploiement des véhicules ambulanciers, mais à notre avis, il reste beaucoup à faire dans ce registre. Pour cette raison, la gestion des SPU présente, à nos yeux, des avenues de recherche des plus intéressantes, comme c'est d'ailleurs le cas pour la plupart des problèmes associés à la gestion des systèmes de santé. L'analyse et le développement de stratégies de gestion seront d'ailleurs abordés dans les derniers chapitres de cette thèse.

CHAPITRE 4

ÉTUDE DE STRATÉGIES DE DÉPLOIEMENT ET DE REDÉPLOIEMENT DES VÉHICULES AMBULANCIERS PAR SIMULATION

Afin d'assurer un service adéquat à la population, les services préhospitaliers d'urgence (SPU) utilisent des véhicules ambulanciers qu'ils localisent stratégiquement sur le territoire à desservir. Néanmoins, bien que la couverture du territoire ait été planifiée avec soin lors de la phase de planification statique, il est possible qu'à un certain moment dans la journée, l'ensemble de véhicules disponibles pour répondre aux appels urgents ne puisse être en mesure de servir adéquatement l'ensemble de la population. En effet, l'arrivée des appels étant difficilement prévisible et pouvant évoluer au cours d'une journée, les performances globales du système peuvent être affectées négativement si aucune action corrective n'est considérée pour rectifier la situation ou pour réagir aux différentes sources d'incertitude. L'une des actions correctives envisageables est le *redéploiement* des véhicules ambulanciers.

Tel que discuté au chapitre précédent, le problème de *déploiement* des véhicules ambulanciers consiste à déterminer les postes d'attente à utiliser pour la localisation des véhicules entre deux affectations à des appels urgents. Un plan de déploiement définit alors l'ensemble des sites sélectionnés pour la localisation d'un ou de plusieurs véhicules, de même que l'affectation des véhicules aux sites choisis. D'un point de vue pratique, le plan de déploiement associe une localisation ou un poste d'attente à chaque véhicule au début de son quart de travail. Ce plan de déploiement pourra être déterminé *a priori* ou encore de manière dynamique afin de tenir compte de l'état du système au moment où les véhicules amorcent leur quart de travail. Le poste d'attente affecté à un véhicule pourra ensuite demeurer fixe pour tout son quart de travail ou être modifié en cours de route si une stratégie de redéploiement est considérée afin de maintenir un niveau de service adéquat à tout moment. Différents types de redéploiement pourront alors être envisagés.

Plus concrètement, le problème de *redéploiement* des véhicules ambulanciers consiste à relocaliser les véhicules disponibles vers les différents postes d'attente potentiels de façon à assurer, en tout temps, une couverture adéquate de la population. L'évolution du système dans le temps est alors considérée. Elle peut se traduire par des fluctuations de la demande en raison, notamment, des mouvements de la population dans la journée. Différents plans de déploiement seront alors établis de façon à modifier le positionnement des véhicules entre les périodes afin de s'adapter aux changements prévus de la demande. On parlera alors de redéploiement *multi-période* (Başar

et al., 2011; Carpentier, 2006; Rajagopalan *et al*., 2008; Repede et Bernardo, 1994; Schmid et Doerner, 2010). L'évolution du système peut également se traduire par une variation temporelle de l'état du système, par exemple, un changement du nombre de véhicules disponibles. Le redéploiement vise alors à maintenir un niveau de service adéquat avec un nombre réduit de véhicules, certains d'entre eux étant en service (c'est-à-dire affectés à des incidents). Puisque le plan de déploiement variera en fonction de l'état du système, on parlera de redéploiement *dynamique* (Gendreau *et al*., 2001, 2006; Andersson et Värbrand, 2007; Nair et Miller-Hooks, 2009; Mason, 2013; Naoum-Sawaya et Elhedhli, 2013). Enfin, le problème de repositionnement, cas particulier du redéploiement, consiste à déterminer la localisation d'un véhicule à la suite de sa libération, une fois sa mission complétée. Traditionnellement, la règle de repositionnement consistant à retourner un véhicule à son poste d'attente initial est employée. Différentes règles de repositionnement dynamiques qui permettent de déterminer en temps réel la nouvelle localisation d'un véhicule venant d'être libéré ont toutefois été proposées récemment (Maxwell *et al*., 2009; Schmid, 2012).

Les études présentées ci-haut et décrites au *Chapitre 3* ont permis de montrer que les différentes stratégies de redéploiement proposées permettent généralement d'améliorer les performances globales du système dans leurs contextes d'application respectifs. Le redéploiement des véhicules génère toutefois des mouvements qui peuvent mener à des conséquences indésirables tant au niveau financier qu'au niveau de la gestion des ressources humaines. Une question demeure alors : est-ce que l'amélioration des performances du système à la suite d'un redéploiement justifie ces inconvénients ? Et si oui, quand et dans quelles circonstances devrait-on procéder à un redéploiement ? Malheureusement, ces questions ont été peu discutées jusqu'à maintenant. En effet, Nair et Miller-Hooks (2009) ont soulevé le fait que plusieurs recherches ont considéré le développement de stratégies de redéploiement, mais qu'aucune étude ne semble s'être intéressée à quantifier les bénéfices de telles stratégies par rapport aux stratégies plus traditionnelles de localisation statique. Ces auteurs se sont alors penchés sur l'évaluation d'une stratégie de redéploiement dynamique calculée *a priori* en prenant en compte deux objectifs, soit la maximisation de la double couverture et la minimisation des coûts de relocalisation. Les résultats obtenus grâce à la résolution d'un modèle de programmation linéaire sont présentés dans leur article, et ce, pour différent nombre de véhicules et pour différents poids accordés aux deux objectifs considérés. Dans ce cas, l'utilisation de la programmation linéaire nécessite un certain nombre d'hypothèses simplificatrices en ce qui concerne les aspects aléatoires reliés à la gestion des SPU. De plus, une seule stratégie est comparée au cas statique. Néanmoins, il s'agit d'un premier pas vers une analyse plus exhaustive de stratégies de déploiement et de redéploiement

dans un contexte plus réaliste. À notre avis, il reste encore beaucoup à faire dans ce registre.
À la lumière de ces constats, ce chapitre s'intéresse donc à l'évaluation et à l'analyse comparative de différentes stratégies de déploiement et de redéploiement. Il vise à quantifier les bénéfices des stratégies de redéploiement par rapport aux stratégies statiques plus classiques, mais aussi à comparer certaines stratégies de redéploiement en considérant différents contextes pouvant se distinguer, notamment, par leurs niveaux d'occupation ou par le profil de la demande. Afin d'analyser et de comparer les différentes stratégies de déploiement et de redéploiement dans un contexte plus réaliste, une étude de simulation est menée. En effet, l'utilisation de la simulation permet de considérer adéquatement les aspects stochastiques inhérents aux SPU, aspects qui ne peuvent être pris en compte explicitement, ou beaucoup plus difficilement, lors de la résolution de modèles mathématiques. Ainsi, un modèle de simulation flexible et générique a été développé afin de mener une telle analyse. Naturellement, ce modèle de simulation pourra servir de base pour l'analyse de différentes décisions en lien avec la gestion des SPU, et ce, à tous les niveaux de planification.
Le présent chapitre se divise donc comme suit. Les stratégies de déploiement et de redéploiement considérées dans cette étude de même que leur modélisation sont d'abord présentées. Le modèle de simulation conçu et utilisé pour l'évaluation et l'analyse des différentes stratégies est ensuite décrit brièvement, puis les résultats obtenus lors de la phase d'expérimentation sont présentés et discutés. Enfin, une réflexion sur les avenues de recherche potentielles vient clore le chapitre.
Les contributions de ce chapitre sont multiples. Dans un premier temps, ce chapitre présente la définition de quatre stratégies de gestion opérationnelle prenant en compte l'aspect dynamique du problème, puis leur modélisation à l'aide d'un cadre commun. Dans un deuxième temps, il expose la conception d'un outil de simulation générique pour l'analyse de différentes stratégies de gestion définies dans ce chapitre. Enfin, grâce aux résultats de l'analyse de simulation, il propose une analyse des bénéfices et des inconvénients liés aux stratégies dynamiques et flexibles par rapport aux stratégies statiques dans un contexte d'application pratique.

4.1 Définition des stratégies de déploiement et de redéploiement

L'étude faisant l'objet de ce chapitre s'intéresse à la comparaison et à l'analyse de différentes stratégies de gestion en lien avec le déploiement des véhicules ambulanciers sur un territoire à desservir. Une stratégie se caractérise par un ensemble de règles associées notamment au déploiement initial, au redéploiement et au repositionnement des véhicules ambulanciers à la suite de leur libération. Les différents modèles proposés jusqu'à maintenant afin de traiter chacun de

ces problèmes ont été décrits au *Chapitre 3*. Les stratégies de gestion étudiées dans ce chapitre sont, quant à elles, décrites ci-dessous et résumées au Tableau 4.1. Il est également important de mentionner que, dans tous les cas, la règle de répartition classique qui consiste à envoyer le véhicule le plus proche sur les lieux d'un incident est utilisée.

La première stratégie proposée se compose d'un ensemble de règles statiques. Ainsi, le déploiement initial d'un véhicule donné est déterminé *a priori*, puis demeure fixe pour tout son quart de travail. Le véhicule est donc dirigé vers son poste d'attente initial au début de son quart de travail, puis y retourne à chaque fois qu'il se libère ou reprend du service. La deuxième stratégie considère un déploiement initial *a priori* et une stratégie de redéploiement *multi-période*. Un plan de déploiement est alors déterminé *a priori* pour chaque période. Le poste d'attente d'un véhicule donné changera donc au cours de son quart de travail, en fonction de la période. À la suite de sa libération ou de sa remise en service, le véhicule retourne au poste d'attente qui lui est attribué selon la période. La troisième stratégie considère, quant à elle, un déploiement initial et un repositionnement dynamiques plutôt que statiques. Ainsi, le déploiement initial d'un véhicule est déterminé en considérant l'état du système au moment où le véhicule amorce son quart de travail. Le poste d'attente vers lequel un véhicule est redirigé à la suite de sa libération est également déterminé dynamiquement en considérant l'état du système. Aucun redéploiement des véhicules en attente n'est toutefois admis dans ce cas. Enfin, la quatrième stratégie proposée se compose d'un ensemble de règles dynamiques. Ainsi, le déploiement initial et la localisation d'un véhicule à la suite à de sa libération sont toujours déterminés dynamiquement, mais dans ce dernier cas de figure, le redéploiement dynamique des véhicules en attente est également permis. Certains véhicules pourront donc être relocalisés entre deux affectations à des tâches si l'état du système le justifie.

Stratégies	**Déploiement initial**	**Redéploiement**	**Repositionnement**
1	*A priori*	Aucun	Retour à localisation initiale (fixe pour tout le quart)
2	*A priori*	Multi-période	Retour à localisation initiale (en fonction de la période)
3	Dynamique	Aucun	Dynamique
4	Dynamique	Dynamique	Dynamique

Tableau 4.1 – Présentation des stratégies

4.2 Modélisation des stratégies de déploiement et de redéploiement

Les différentes stratégies de déploiement et de redéploiement présentées à la section précédente se traduisent par une série de décisions, déterminées en fonction des objectifs poursuivis et des contraintes considérées. Chaque stratégie pourra donc être représentée par un modèle ou un ensemble de règles distinctes. Dans le cadre de cette étude, toutes les stratégies analysées ont été modélisées à partir du modèle avec double standard (MDS), ou *double standard model* (DSM) en anglais, proposé par Gendreau *et al*. (1997). Plusieurs raisons ont motivé ce choix. D'une part, le MDS s'inspire de différentes règles gouvernementales employées en pratique, notamment le *United States EMS Act of 1973*. Les objectifs et les contraintes de couverture qu'il considère sont faciles à comprendre et à adapter à différents cas de figure en ce qui a trait au déploiement et au redéploiement. D'autre part, il a déjà fait l'objet de plusieurs adaptations (Gendreau *et al*., 2001; Doerner *et al*., 2005; Schmid et Doerner, 2010) et a été utilisé dans divers contextes pratiques tels que ceux mentionnés dans Laporte *et al*. (2009). Enfin, le fait de modéliser chaque stratégie à partir d'un même modèle garantit que la comparaison se fasse sur une base commune. Dans cette section, le MDS sera brièvement décrit, puis les modifications apportées au modèle afin de représenter adéquatement les stratégies proposées seront présentées. Ces différents modèles seront utilisés afin de répliquer le processus décisionnel sous-jacent aux différentes stratégies au sein du modèle de simulation.

Tel qu'il a été présenté au *Chapitre 3*, le MDS vise à intégrer l'idée de double couverture et l'utilisation de différents rayons de couverture. Il cherche à déterminer la meilleure solution qui permet de garantir qu'une proportion α de la population soit rejointe à l'intérieur d'un délai prescrit S et que la totalité de la population soit rejointe à l'intérieur d'un délai S', $S' > S$. En considérant x_j, le nombre de véhicules localisés en j, y_i, une variable binaire qui vaut 1 si la zone de demande i est couverte au moins une fois à l'intérieur de S, et 0 autrement, u_i, une variable binaire qui vaut 1 si la zone de demande i est couverte au moins deux fois à l'intérieur de S, et 0 autrement, a_i, la densité de population de la zone de demande i, p_j, la limite sur le nombre de véhicules pouvant être localisés au site j, p, le nombre total de véhicules disponibles, M_i et M'_i, les ensembles de postes d'attente potentiels pouvant assurer la couverture d'une zone de demande i respectivement à l'intérieur des délais prescrits S et S', le MDS se formule de la manière suivante :

MDS

$$\max \sum_{i=1}^{n} a_i u_i \tag{4.1}$$

sous les contraintes :

$$\sum_{j \in M'_i} x_j \geq 1,\ i = 1, ..., n, \tag{4.2}$$

$$\sum_{i=1}^{n} a_i y_i \geq \alpha \sum_{i=1}^{n} a_i, \tag{4.3}$$

$$\sum_{j \in M_i} x_j \geq u_i + y_i,\ i = 1, ..., n, \tag{4.4}$$

$$u_i \leq y_i,\ i = 1, ..., n, \tag{4.5}$$

$$\sum_{j=1}^{m} x_j = p, \tag{4.6}$$

$$x_j \leq p_j,\ j = 1, ..., m, \tag{4.7}$$

$$u_i, y_i \in \{0, 1\},\ i = 1, ..., n, \tag{4.8}$$

$$x_j \geq 0,\ \text{entier},\ j = 1, ..., m. \tag{4.9}$$

Le MDS vise donc à maximiser la population couverte au moins deux fois à l'intérieur de S (4.1) en assurant qu'une proportion minimale α de la population soit couverte à l'intérieur de S (4.3) et que la totalité de la population soit couverte à l'intérieur de S' (4.2) et cela, en considérant la localisation d'un nombre donné de véhicules (4.6). Il est important de noter que, dans certains cas, si le nombre de véhicules disponibles n'est pas suffisant, le modèle peut devenir non réalisable, c'est-à-dire que la totalité de la population ne peut être couverte à l'intérieur de S' et/ou qu'une proportion α de la population ne peut être couverte à l'intérieur de S. Il est également important de constater que le modèle fournit le nombre de véhicules à localiser à chaque poste d'attente potentiel, sans égard à l'identité des véhicules ou des équipes de travail qui y sont localisés. En effet, l'information concernant le nombre de véhicules localisés à chaque poste d'attente est généralement suffisante pour la prise de décision au niveau tactique, mais peut devenir incomplète dans le cadre de décisions à caractère opérationnel telles que le redéploiement. Dans ce cas, connaître la localisation précise d'un véhicule donné peut devenir important, notamment afin de mesurer l'impact du redéploiement, pour calculer les distances ou les temps de redéploiement, par exemple. Ainsi, quelques modifications doivent être apportées au MDS de manière à prendre en compte ces deux aspects.

Tout d'abord, afin d'assurer la faisabilité du modèle, les déviations des contraintes (4.2) et (4.3) par rapport aux bornes imposées sont pénalisées dans la fonction objectif. Pour ce faire, deux nouvelles variables sont intégrées au modèle, soit ρ_i, une variable binaire qui vaut 1 si la zone de demande i n'est pas couverte à l'intérieur de S', et 0 autrement, et δ, la population qui ne

peut être servie au moins une fois à l'intérieur de S. Une seconde modification est également apportée à la fonction objectif. En effet, plutôt que de considérer directement la maximisation de la population couverte deux fois à l'intérieur d'un délai prescrit S, on considère la maximisation de la probabilité qu'une demande d'intervention survienne dans une zone couverte deux fois à l'intérieur de S. Se faisant, le modèle pourra mieux représenter les besoins réels d'une zone de demande, de même que leur évolution dans le temps en comparaison avec la densité de population de chaque zone qui est une donnée fixe et statique dans le temps. Dans les faits, la population évoluera aussi au cours d'une journée. Cela se traduira par des besoins différents en véhicules ambulanciers au cours d'une journée. Par exemple, les besoins en véhicules ambulanciers dans un secteur résidentiel seront fort probablement différents le matin, avant qu'une partie de la population quitte pour le travail, par rapport au milieu de l'après-midi où une partie des résidents ont quitté leur demeure. Néanmoins, de tels mouvements de population sont parfois difficiles à estimer grâce aux données accessibles quant à la densité de population par zone qui incluent généralement le nombre de résidents par zone. Pour cette raison, nous avons choisi d'utiliser la probabilité qu'une demande d'intervention survienne dans une zone donnée, calculée à partir des besoins réels de chaque zone de demande.

En notant q_i, la probabilité pour qu'une demande d'intervention survienne dans la zone i et w_1, w_2 et w_3, les poids accordés respectivement aux objectifs associés à la double couverture, à la satisfaction de la contrainte de couverture totale et à la satisfaction de la contrainte de couverture partielle, la nouvelle fonction objectif se formule telle que présentée en (4.10). De plus, afin de considérer l'identité propre à chaque véhicule lors de la prise de décision, la variable x_j, telle que définie précédemment, est remplacée par la variable x_{kj}, une variable binaire qui vaut 1 si le véhicule k est localisé en j, et 0 autrement. Ainsi, en considérant cette nouvelle variable de même que les paramètres présentés ci-haut, le modèle avec double standard modifié (MDSm) se formule comme suit :

MDS*m*

$$\max w_1 \sum_{i=1}^{n} q_i u_i - w_2 \sum_{i=1}^{n} \rho_i - w_3 \delta \tag{4.10}$$

$$\sum_{j \in M_i'} \sum_{k=1}^{p} x_{kj} \geq 1 - \rho_i, \; i = 1, ..., n, \tag{4.11}$$

$$\sum_{i=1}^{n} a_i y_i \geq \alpha \sum_{i=1}^{n} a_i - \delta, \tag{4.12}$$

$$\sum_{j \in M_i} \sum_{k=1}^{p} x_{kj} \geq u_i + y_i, \; i = 1, ..., n, \tag{4.13}$$

$$u_i \leq y_i,\ i = 1,...,n, \tag{4.14}$$

$$\sum_{j=1}^{m} x_{kj} = 1,\ k = 1,...,p, \tag{4.15}$$

$$\sum_{k=1}^{p} x_{kj} \leq p_j,\ j = 1,...,m, \tag{4.16}$$

$$u_i, y_i, q_i \in \{0,1\},\ i = 1,...,n, \tag{4.17}$$

$$x_{kj} \in \{0,1\},\ j = 1,...,m,\ ,k = 1,...,p, \tag{4.18}$$

$$\delta \geq 0. \tag{4.19}$$

Le MDS ainsi modifié sera donc utilisé pour la modélisation des différentes stratégies de déploiement et de redéploiement étudiées dans le cadre de cette étude.

4.2.1 Stratégie 1 : Déploiement *a priori* sans redéploiement

Le problème de déploiement initial vise à déterminer, pour chaque véhicule, le poste d'attente qui lui sera attribué au début de son quart de travail. Ce poste d'attente demeurera ensuite fixe pour toute la durée de son quart de travail. La disponibilité des véhicules, de même que les temps de début et de fin des différents quarts de travail, devront alors être pris en compte lors de la modélisation du problème. Le modèle associé à cette stratégie se distingue donc du MDSm puisqu'il devra considérer plusieurs périodes pour intégrer adéquatement l'information concernant la disponibilité des véhicules tout au long de la journée. De cette manière, il sera aussi possible de considérer la variation de la probabilité pour qu'une demande survienne dans une zone donnée et des temps de déplacement en fonction de la période. Néanmoins, contrairement à plusieurs modèles multi-période proposés jusqu'à maintenant pour la localisation des véhicules ambulanciers, la relocalisation n'est pas permise entre les périodes : une fois le poste d'attente d'un véhicule fixé, il demeure le même pour tout son quart de travail.

Afin de prendre en compte ce contexte particulier, le MDSm doit être adapté. Pour ce faire, tous les paramètres et variables définis précédemment se verront ajouter un indice t (à l'exception de a_i et p_j, que nous considérerons indépendants de la période). Par exemple, la variable u_i devient u_i^t, une variable binaire qui vaut 1 si la zone de demande i est couverte deux fois à l'intérieur de S à la période t, et 0 autrement. Trois nouveaux paramètres sont également nécessaires afin de prendre en compte la disponibilité des véhicules soit d_k^t, un paramètre qui vaut 1 si le véhicule k est disponible à la période t, $t = 1,...,T$, et 0 autrement, t_d^k, la période de début de quart du véhicule k, t_f^k, la période de fin de quart du véhicule k. En considérant ainsi l'aspect multi-

période du problème, le modèle de déploiement avec double standard (MDDS) se reformule de la manière suivante :

MDDS

$$\max \sum_{t=1}^{T}(w_1 \sum_{i=1}^{n} q_i^t u_i^t - w_2 \sum_{i=1}^{n} \rho_i^t - w_3 \delta^t) \tag{4.20}$$

sous les contraintes :

$$\sum_{j \in M_i'^t} \sum_{k=1}^{p} x_{kj}^t \geq 1 - \rho_i^t,\ i = 1,...,n,\ t = 1,...,T, \tag{4.21}$$

$$\sum_{i=1}^{n} a_i y_i^t \geq \alpha \sum_{i=1}^{n} a_i - \delta^t,\ t = 1,...,T, \tag{4.22}$$

$$\sum_{j \in M_i^t} \sum_{k=1}^{p} x_{kj}^t \geq u_i^t + y_i^t,\ i = 1,...,n,\ t = 1,...,T, \tag{4.23}$$

$$u_i^t \leq y_i^t,\ i = 1,...,n,\ t = 1,...,T, \tag{4.24}$$

$$\sum_{j=1}^{m} x_{kj}^t = d_k^t,\ k = 1,...,p,\ t = 1,...,T, \tag{4.25}$$

$$\sum_{k=1}^{p} x_{kj}^t \leq p_j,\ j = 1,...,m,\ t = 1,...,T, \tag{4.26}$$

$$x_{kj}^t \leq x_{kj}^{t+1},\ j = 1,...,m, \forall k : t_d^k < t_f^k, t = t_d^k,...,t_f^k - 1, \tag{4.27}$$

$$x_{kj}^t \leq x_{kj}^{t+1},\ j = 1,...,m, \forall k : t_d^k > t_f^k,\ t = 0,...,t_f^k - 1 \text{ et } t = t_d^k,...,T-1, \tag{4.28}$$

$$x_{kj}^T \leq x_{kj}^0,\ j = 1,...,m, \forall k : t_d^k > t_f^k, \tag{4.29}$$

$$u_i^t, y_i^t, \rho_i^t \in \{0,1\},\ i = 1,...,n,\ t = 1,...,T, \tag{4.30}$$

$$x_{kj}^t \in \{0,1\},\ j = 1,...,m,\ ,k = 1,...,p,\ t = 1,...,T, \tag{4.31}$$

$$\delta^t \geq 0,\ t = 1,...,T. \tag{4.32}$$

Le MDDS intègre donc différentes contraintes visant, d'une part, à assurer qu'un véhicule est localisé sur le territoire seulement si ce dernier est en service (4.25) et, d'autre part, à assurer qu'une fois le poste d'attente d'un véhicule déterminé, il demeure fixe pour tout son quart de travail. Deux cas sont alors pris en compte. Dans un premier temps, la contrainte (4.27) assure que le poste d'attente d'un véhicule K demeure fixe lorsque la période de début du quart de travail est plus petite que la période de fin du quart de travail, c'est-à-dire lorsque $t_d^k < t_f^k$. Par exemple, en présumant un horizon de planification d'une journée, décomposée en périodes de 2

heures, chaque journée se répétant au cours d'une semaine, cela correspondrait à un véhicule ou une équipe paramédicale qui amorcerait son quart de travail à 8h00 pour terminer à 16h00. Les contraintes (4.28) et (4.29) assurent plutôt que le poste d'attente d'un véhicule k demeure fixe lorsque la période de début est plus grande que la période de fin, c'est-à-dire lorsque $t_d^k > t_f^k$. Ceci correspondrait, par exemple, à un véhicule ou à une équipe paramédicale qui amorcerait son quart de travail à 20h00 pour terminer à 4h00 le lendemain. Puisque les SPU offrent un service en continu, il est important de considérer ces deux cas afin de bien prendre en compte la disponibilité des véhicules en tout temps.

Le MDDS présente donc certaines similitudes avec le modèle multi-période présenté par Başar *et al*. (2011). En effet, les deux modèles considèrent qu'une fois une localisation ouverte, elle le demeurera jusqu'à la fin de l'horizon de planification. Le modèle de Başar *et al*. (2011) considère aussi des objectifs de couverture similaires à ceux poursuivis par le MDDS. Néanmoins, le problème traité par Başar *et al*. (2011) considère un problème de localisation stratégique en lien avec la construction de stations fixes. Dans le cas présent, on cherche plutôt à déterminer la localisation d'un nombre donné de véhicules ambulanciers sur le territoire à desservir pour un ensemble de périodes, en fonction de leur disponibilité.

4.2.2 Stratégie 2 : Déploiement *a priori* avec redéploiement multi-période

Le problème de redéploiement multi-période est très similaire au problème de déploiement initial présenté précédemment. Le modèle correspondant à la première stratégie permettra aussi de représenter ce problème. Quelques modifications sont toutefois nécessaires. En effet, puisque la relocalisation des véhicules entre les périodes est maintenant permise, les contraintes (4.27) à (4.29) qui visent à assurer le maintien du poste d'attente fixé en début de quart travail devront être éliminées. En retirant ce groupe de contraintes, le MDDS devient alors décomposable : un MDSm pourrait être résolu pour chaque période, en considérant la disponibilité des véhicules à chacune des périodes considérées. Néanmoins, en décomposant ainsi le problème, les inconvénients liés à la relocalisation ne pourront être pris en compte. Un grand nombre de déplacements pourraient alors être requis pour maintenir un niveau de couverture adéquat.

Afin de considérer plus explicitement les inconvénients liés au redéploiement, un modèle basé sur le MDDS, un nouveau modèle est proposé. Ce modèle est nommé le redéploiement multi-période avec double standard (MRMPDS). Contrairement au MDDS, ce modèle n'inclut pas les contraintes (4.27) à (4.29) qui visent à assurer le maintien du poste d'attente d'un véhicule. De plus, le MRMPDS poursuit maintenant deux objectifs : un objectif principal qui vise la maximisation des performances et de la satisfaction des contraintes de couverture (4.33), et

un objectif secondaire qui vise plutôt la minimisation des coûts ou des inconvénients liés à la relocalisation. Ainsi, en considérant $r^t_{kjj'}$, une variable binaire qui vaut 1 si le véhicule k est déplacé de la localisation j vers la localisation j' à la fin de la période t, et 0 autrement, et $c^t_{kjj'}$, le coût associé à la relocalisation du véhicule k de la localisation j vers la localisation j' à la fin de la période t, l'objectif secondaire associé aux coûts de relocalisation se formule de la manière suivante :

$$\min \sum_{t=1}^{T} \sum_{k=1}^{p} \sum_{j=1}^{m} \sum_{j'=1}^{m} c^t_{kjj'} r^t_{kjj'}. \tag{4.33}$$

Afin de calculer adéquatement les coûts de relocalisation encourus entre les périodes, les contraintes suivantes devront également être ajoutées au modèle :

$$r^t_{kjj'} \geq x^t_{kj} + x^{t+1}_{kj} - 1,\ j = 1,...,m,\ j' = 1,...,m,\ \forall j \neq j',\ k = 1,...,P,\ t = 0,...,T-1, \tag{4.34}$$

$$r^T_{kjj'} \geq x^T_{kj} + x^0_{kj} - 1,\ j = 1,...,m,\ j' = 1,...,m,\ \forall j \neq j',\ k = 1,...,p. \tag{4.35}$$

Le MRMPDS est donc similaire au modèle de Schmid et Doerner (2010) puisqu'il permet la relocalisation entre les périodes et considère les coûts qui y sont associés dans la fonction objectif. Il s'en distingue par le fait qu'il considère plus explicitement la disponibilité des véhicules tout au long de la journée.

4.2.3 Stratégie 3 : Déploiement et repositionnement dynamiques

Le déploiement et le repositionnement dynamiques consistent à déterminer, en temps réel, le poste d'attente d'un véhicule lorsqu'il reprend du service en début de quart de travail, après une pause ou à la suite de sa libération. Les décisions de repositionnement sont alors prises en considérant la localisation des véhicules en service et disponibles pour répondre à des appels urgents de même que l'état général du système. Dans ce cas, seul le véhicule nouvellement disponible est impliqué lors du repositionnement, le redéploiement des autres véhicules (en service et disponibles) n'étant pas admis.

Afin de maintenir des objectifs similaires aux modèles présentés ci-haut, le choix de la localisation du véhicule à la suite de sa libération ou de sa remise en service se fait en prenant en compte deux critères. Tout d'abord, si toutes les zones de demande ne peuvent être couvertes à l'intérieur de S', le véhicule nouvellement disponible est envoyé vers le poste d'attente qui permet de maximiser le nombre de zones de demande couvertes à l'intérieur de S', si la capa-

cité du poste d'attente n'est pas atteinte. Ainsi, le poste d'attente sera sélectionné de manière à pallier à la non-faisabilité du problème. Si plusieurs solutions sont équivalentes du point de vue de la faisabilité, les postes d'attente sont départagés en considérant la maximisation des performances du système, c'est-à-dire en sélectionnant le poste d'attente qui permet de maximiser la probabilité pour qu'une nouvelle demande apparaisse dans une zone couverte par au moins deux véhicules à l'intérieur de S. Enfin, si toutes les zones de demande sont couvertes à l'intérieur de S', le véhicule nouvellement disponible est dirigé vers le poste d'attente qui maximise la probabilité pour qu'une nouvelle demande apparaisse dans une zone couverte par au moins deux véhicules à l'intérieur de S, tout en respectant la capacité des postes d'attente. Dans ce cas, seule la maximisation des performances du système est considérée.

4.2.4 Stratégie 4 : Déploiement, repositionnement et redéploiement dynamiques

La dernière stratégie proposée regroupe l'ensemble des stratégies dynamiques possibles. En opposition aux stratégies précédentes, cette dernière stratégie admet le redéploiement dynamique des véhicules ambulanciers. De cette manière, si l'état du système le justifie, la localisation des véhicules en service disponibles pour répondre à des demandes, et dont le redéploiement est possible, pourra être modifiée. Dans le cas présent, on jugera que les performances du système justifient un redéploiement lorsque toutes les zones de demande ne peuvent être couvertes à l'intérieur de S'. Le redéploiement dynamique des véhicules pourrait être fréquent s'il y a peu de véhicules disponibles pour répondre aux appels à un moment précis de la journée, par exemple lorsque le système est très occupé. Toutefois, afin d'en limiter les inconvénients, une limite sur le temps minimal entre deux redéploiements successifs a été imposée. De plus, la relocalisation d'un véhicule particulier est pénalisée si le véhicule en question a été déplacé récemment, par exemple à la suite de son repositionnement.

Deux événements principaux pourront engendrer le redéploiement des véhicules ambulanciers : l'apparition d'un véhicule dans le système, à la suite de sa libération ou de sa remise en service, et la disparition d'un véhicule du système, à la suite de son affectation à un appel, au début d'une pause ou à la fin de son quart de travail. Dans le cas de l'apparition d'un véhicule, si toutes les zones de demande ne peuvent être servies à l'intérieur de S' et que le dernier redéploiement a été lancé dans un délai donné, le redéploiement dynamique des véhicules est lancé, incluant le repositionnement du véhicule nouvellement disponible dans le système. Si aucun redéploiement n'est nécessaire ou permis, le poste d'attente du véhicule nouvellement disponible est sélectionné de la même manière qu'en 4.2.3. Dans le cas de la disparition d'un véhicule, si la couverture totale à l'intérieur de S' ne peut être atteinte ou maintenue et que

la limite de temps entre deux redéploiements successifs est respectée, le redéploiement dynamique est lancé. Dans le cas contraire, la localisation des véhicules en service et disponibles pour répondre à des appels demeure la même.

Afin de modéliser adéquatement le problème de redéploiement dynamique décrit ci-haut, il importe de définir certaines variables et paramètres propres à ce problème. Ainsi, en considérant $\hat{P}$, l'ensemble de véhicules pour lesquels le redéploiement engendrera une pénalité w_4, λ_{kj}, un paramètre qui vaut 1 si le véhicule k est localisé en j avant le redéploiement et 0 autrement, et s_k, une variable binaire qui vaut 1 si le véhicule k est relocalisé (c'est-à-dire sa localisation change après le redéploiement), et 0 autrement, le modèle de redéploiement dynamique avec double standard résolu au temps t (MRDDS$_t$) s'écrit de la manière suivante :

MRDDS$_t$

$$\max w_1 \sum_{i=1}^{n} q_i u_i - w_2 \sum_{i=1}^{n} \rho_i - w_3 \delta - w_4 \sum_{k \in \hat{P}} s_k \tag{4.36}$$

$$\min \sum_{k=1}^{p} \sum_{j'=1}^{m} \sum_{j=1}^{m} \lambda_{kj'} c_{kj'j} y_{kj} \tag{4.37}$$

sous les contraintes :

$$\sum_{j \in M_i'} \sum_{k=1}^{p} x_{kj} \geq 1 - \rho_i, \ i = 1, ..., n, \tag{4.38}$$

$$\sum_{i=1}^{n} a_i y_i \geq \alpha \sum_{i=1}^{n} a_i - \delta, \tag{4.39}$$

$$\sum_{j \in M_i} \sum_{k=1}^{p} x_{kj} \geq u_i + y_i, \ i = 1, ..., n, \tag{4.40}$$

$$u_i \leq y_i, \ i = 1, ..., n, \tag{4.41}$$

$$\sum_{j=1}^{m} x_{kj} = 1, \ k = 1, ..., p, \tag{4.42}$$

$$\sum_{k=1}^{p} x_{kj} \leq p_j, \ j = 1, ..., m, \tag{4.43}$$

$$\sum_{j=1}^{m} \lambda_{kj} x_{kj} - s_k \geq 0, \ k = 1, ..., p, \tag{4.44}$$

$$u_i, y_i, \rho_i \in \{0, 1\}, \ i = 1, ..., n, \tag{4.45}$$

$$x_{kj} \in \{0, 1\}, \ j = 1, ..., m, \ , k = 1, ..., p, \tag{4.46}$$

$$s_k \in \{0, 1\}, \ k = 1, ..., p, \tag{4.47}$$

$$\delta \geq 0. \tag{4.48}$$

Le MRDDS$_t$ consiste donc en un modèle multi-objectif visant dans un premier temps, à maximiser la double couverture tout en pénalisant la non-faisabilité des contraintes (4.38) et (4.39) et la relocalisation des véhicules récemment déplacés (4.36), et, dans un deuxième temps, à minimiser les coûts associés au redéploiement (4.37). La contrainte (4.44) a été ajoutée au modèle afin de déterminer si un véhicule donné est relocalisé. Les autres contraintes demeurent très similaires à celles des autres modèles présentés jusqu'ici.

Naturellement, différents critères auraient pu être envisagés pour l'implantation du redéploiement dynamique dans un contexte d'application réel, c'est-à-dire quand lancer le redéploiement, quels véhicules considérer alors, etc. Dans le cadre de cette étude, le redéploiement dynamique a été implanté de manière à limiter le nombre et la fréquence des redéploiements en imposant une limite sur le temps minimal entre deux redéploiements successifs. Le temps minimal imposé devra toutefois être tel que le nombre de redéploiements soit suffisamment élevé afin de pouvoir mesurer l'impact réel d'une telle stratégie.

Les quatre stratégies proposées et leur modèle respectif seront implémentés au sein du modèle de simulation présenté à la section suivante, puis analysées et comparées lors de la phase d'expérimentation.

4.3 Modèle de simulation

Afin de comparer et d'analyser les différentes stratégies de déploiement et de redéploiement, un modèle de simulation a été développé. En effet, la simulation a été choisie puisqu'elle permet de considérer adéquatement les nombreux aspects incertains inhérents à la gestion des SPU. De manière plus concrète, le modèle de simulation développé comporte six composantes principales : un module *Données* qui inclut l'ensemble des paramètres nécessaires à une description adéquate du système étudié et de son état initial, un module *Demandes et variables aléatoires* qui génère l'ensemble des demandes et des variables aléatoires requises afin de mener à bien l'étude de simulation, un module *Moteur de la simulation* qui permet de gérer la simulation en soi, soit l'horloge de la simulation et la liste des événements, un module *Temps de déplacement* qui permet d'estimer le temps de déplacement entre deux localisations sur le territoire à desservir, un module *Décisions* qui permet de gérer la prise de décision en répliquant le processus décisionnel approprié au moment opportun en cours de simulation et enfin, un module *Mesures de performance* qui enregistre et compile l'information nécessaire afin d'évaluer les performances du système étudié. L'architecture du simulateur est illustrée à la Figure 4.1.

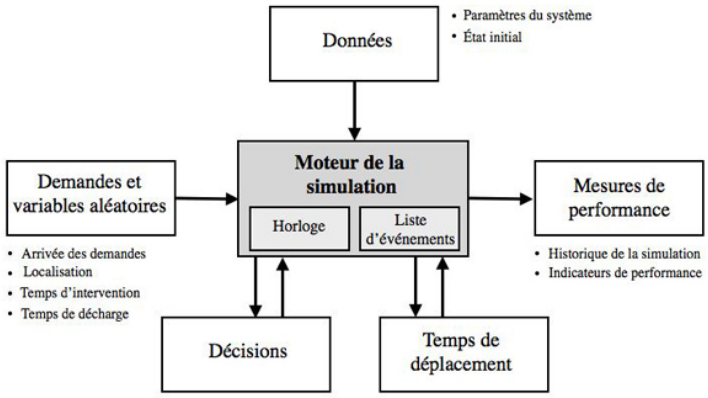


Figure 4.1 – Architecture du modèle de simulation

Dans cette section, les éléments importants qui composent le modèle de simulation développé sont brièvement présentés. Nous invitons le lecteur à se référer à Kergosien *et al*. (2015) pour une description plus détaillée de l'outil de simulation en soi, de sa conception à sa validation. Il est important de mentionner que l'outil de simulation présenté ici a été développé dans un cadre plus large où différents types de décisions pourront être pris en compte. Il est donc suffisamment flexible et générique pour permettre l'étude et l'analyse de différentes décisions, et ce, à tous les niveaux de planification et dans des contextes d'application variés.

4.3.1 Données

Afin de représenter adéquatement le contexte considéré, le modèle de simulation utilise un ensemble de données qui permettent, d'une part, de caractériser le système étudié et, d'autre part, de définir l'état initial du système. Ainsi, les données nécessaires à la simulation se classent en deux groupes : les données en lien avec les caractéristiques du système et celles décrivant son état initial. Le premier groupe de données vise donc à définir le système, et plus particulièrement, la région desservie par le SPU étudié. Ce groupe de données inclut l'information concernant le territoire à desservir et sa division en zones de demande, la liste de véhicules disponibles, l'ensemble des postes d'attente potentiels de même que l'ensemble des hôpitaux et des centres de santé de la région. Plus précisément, chaque zone de demande i se définit par son centroïde (coordonnées géographiques), sa densité de population a_i et un vecteur de probabilité. Ce vecteur définit la probabilité q_i^t pour que la prochaine demande survienne dans la zone i, à la période t. La décomposition en périodes dépendra du système étudié et demeure à la

discrétion de l'utilisateur. Les postes d'attente sont, quant à eux, définis par leurs coordonnées géographiques, de même que le nombre de véhicules p_j qui pourront y être en attente simultanément. Les hôpitaux ou établissements de santé se caractérisent aussi par leurs coordonnées géographiques. De manière générale, les données incluses dans ce premier groupe sont directement liés aux données et aux paramètres nécessaires à la formulation des modèles présentés à la section 4.2. Le deuxième groupe de données caractérise plutôt l'état initial du système, c'est-à-dire au moment où la simulation débute. On y retrouvera des données telles que la localisation initiale des véhicules et l'état des différentes ressources.

4.3.2 Demandes et variables aléatoires

Au sein d'un modèle de simulation, les événements incertains sont généralement représentés par leur distribution de probabilité. Une partie importante de la simulation est donc dédiée à l'échantillonnage de ces distributions afin d'en tirer des valeurs plausibles pour les différents événements incertains qui surviennent pendant l'exécution de la simulation. Par exemple, dans le contexte des SPU, le temps de service sur les lieux d'un incident est, en pratique, inconnu au moment où un appel est placé. Les temps de service peuvent toutefois être modélisés par une distribution de probabilité de paramètres connus. Ces distributions et paramètres pourront être déterminés, par exemple, à partir de données historiques. La valeur précise du temps de service pour une intervention donnée est alors déterminée, *a priori* ou en cours de simulation, à partir de la distribution de probabilité qui lui est associée. Dans le cas du modèle de simulation proposé, les demandes et les différentes variables aléatoires nécessaires sont générées *a priori*, puis stockées dans un fichier externe. Les valeurs ainsi générées sont ensuite utilisées au moment opportun pendant la simulation. La génération des demandes et des variables aléatoires *a priori* a été choisie ici puisqu'elle permet un meilleur contrôle et une plus grande flexibilité. Cette technique pourra faciliter la validation et la vérification du modèle de simulation, de même que son adaptation à d'autres contextes.

4.3.3 Moteur de la simulation

Le modèle de simulation proposé est basé sur la simulation à événements discrets (SED). La SED, telle que définie par Law (2006), modélise un système par son évolution dans le temps. Le système change donc de manière instantanée à différents points (instants) dans le temps, ces points correspondant aux événements qui viennent modifier l'état du système. Le système se définit par un ensemble d'entités caractérisées elles-mêmes par un ensemble d'attributs et de variables d'état dont la valeur est amenée à changer dans le temps. La gestion des événements

passe par le moteur de la simulation, une routine qui permet d'avancer l'horloge de la simulation d'un événement à un autre. Lorsqu'un événement survient, différentes procédures sont exécutées de manière à modifier l'état du système et des entités affectées par de telles décisions. De nouveaux événements peuvent également être générés. L'horloge de la simulation avance ensuite vers l'événement suivant.

4.3.3.1 Entités

Du point de vue de l'implémentation, le modèle de simulation est composé de différentes « boîtes » qui permettent de représenter les entités qui évolueront tout au long de la simulation et qui permettront de représenter le système étudié. Chaque entité se caractérise alors par un ensemble d'attributs fixes (des caractéristiques qui ne changent pas durant la simulation), de variables et d'états (des caractéristiques qui changent en cours de simulation). Plus concrètement, afin d'assurer une représentation adéquate d'un SPU, le modèle de simulation proposé comporte deux types d'entités principales : les demandes et les ambulances. D'autres entités telles que des entités *Répondants médicaux d'urgence* (RMU), *Répartiteurs* ou *Opérateurs* peuvent également être nécessaires à une représentation adéquate d'un SPU. Ces entités interviendront au tout début du processus de traitement d'une demande, afin de déterminer la priorité de l'appel et le véhicule à affecter à une demande. Dans le cas présent, le temps requis pour effectuer ces différentes tâches est agrégé en une même tâche soit le traitement de l'appel avant l'affectation de l'appel à un véhicule. Ainsi, seule l'entité *Opérateurs* est considérée. Néanmoins, si le système ou l'étude le requiert, le modèle de simulation peut facilement décomposer ces tâches et considérer de manière indépendante les différentes ressources nécessaires.

Les entités *Demandes* peuvent prendre trois états : en attente, en traitement ou terminé. Elles sont caractérisées par un grand nombre d'attributs parmi lesquels on retrouve le type de demande, le temps d'arrivée, le temps de traitement, pour n'en nommer que quelques-uns. Les entités *Ambulances* comporte également un certain nombre d'attributs (par exemple, l'heure de début et de fin de son quart de travail), de variables (par exemple, son poste d'attente) et d'états (par exemple, libre, en transit vers un poste d'attente) à travers lesquels elles évolueront au cours de la simulation. La Figure 4.2 illustre l'ensemble des états possibles pour un véhicule ambulancier, ainsi que les transitions possibles entre les états suivant la réalisation des événements. Enfin, les entités *Opérateurs* pourront prendre seulement deux états soit libre ou occupé.

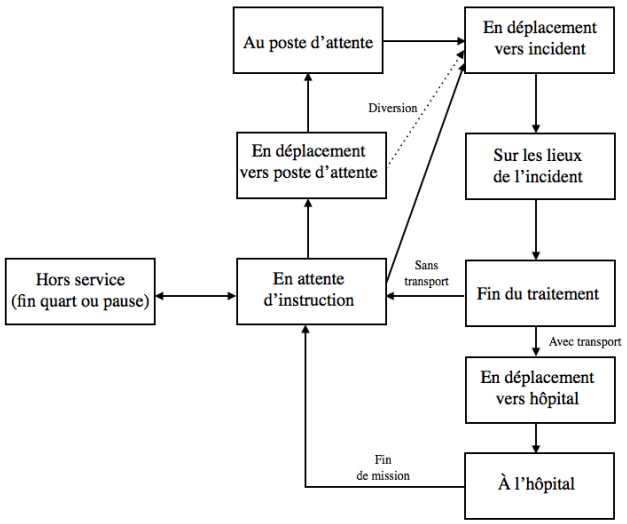


Figure 4.2 – États d'un véhicule ambulancier

4.3.3.2 Moteur de la SED

Le moteur de la simulation proposé dans le cadre de cette étude est inspiré par celui présenté par Pidd (2006) qui consiste en un algorithme en trois phases. Un tel algorithme permet à l'horloge de la simulation d'avancer de manière asynchrone d'un événement à un autre. Le moteur de la simulation fonctionne à partir d'une liste d'événement triée par ordre croissant de temps d'exécution. À chaque fois qu'un événement est exécuté, l'état du système (et éventuellement des entités) est modifié en conséquence et l'horloge de la simulation se déplace vers l'événement suivant. L'exécution d'un événement peut ne pas affecter directement l'état du système, mais plutôt générer de nouveaux événements qui seront ajoutés à la liste. Lorsqu'un nouvel événement est généré, son temps d'exécution est calculé ou déterminé à partir de variables aléatoires. Les différents événements sont classés en deux familles : les événements bornés (B) et les événements conditionnels (C). Le temps d'exécution des événements bornés peut être prédit ou simulé au moment de sa génération. Pour les événements conditionnels, leur temps d'exécution ne peut être déterminé à l'avance puisqu'il est conditionnel, c'est-à-dire qu'il va dépendre de l'état du système.

Le Tableau 4.2 présente les principaux événements possibles au cours de la simulation de même que les procédures de décisions et les changements d'état qui y sont associés. Tous les événe-

ments nécessaires au bon fonctionnement du simulateur n'y sont pas présentés. Nous avons décidé de nous limiter aux événements principaux en lien avec les changements d'état des véhicules ambulanciers et des demandes d'urgence. Naturellement, d'autres événements sont nécessaires pour gérer adéquatement les changements d'attributs et de variables tout au long de la simulation. Ces événements ne sont toutefois pas présentés ici afin d'alléger la présentation. Au tableau suivant, on identifiera D, une demande d'intervention, R, un répartiteur ou opérateur et A, un véhicule ambulancier. La variable $Transport_D$ prend la valeur 1 si la demande D requiert un transport vers un centre hospitalier et 0 autrement. Les variables $Pause_A$ et Fin_A prendront la valeur 1 si la pause ou la fin de quart de l'ambulance A est planifiée respectivement et 0 autrement, et la variable $Detour_A$ prend la valeur 1 si un détournement est planifié pour l'ambulance A et 0 autrement. Enfin, la variable globale $Liste$ est égale au nombre de demandes placées en file d'attente.

4.3.4 Temps de déplacement

En cours de simulation, l'estimation du temps de déplacement entre deux localisations à un moment donné est une tâche difficile, bien que très importante. En effet, l'estimation du temps de déplacement pourra évoluer dans le temps et variera grandement en fonction de l'origine et de la destination. Le temps de déplacement aura donc un impact important sur les résultats finaux de la simulation, mais aussi sur la crédibilité accordée à l'outil de simulation développé. Plusieurs simulateurs présentés jusqu'à maintenant utilisent des temps de déplacement ou des distances précalculés entre un ensemble de points importants sur le territoire à desservir (par exemple, centroïdes des zones de demande, postes d'attente, hôpitaux) à partir desquels le temps de déplacement est dérivé. Néanmoins, lorsque le véhicule n'est pas à une de ces localisations, par exemple lorsqu'un véhicule est détourné de sa mission, l'estimation du temps de déplacement devient plus complexe.

Événement	Type	Condition	Décisions	Changement d'état	Événements déclenchés
Début traitement de D	C	D = en attente & R = disponible		D = en traitement ; R = occupé	Affecter une ambulance à D
Affecter une ambulance à D	B		Identifier l'ambulance A à affecter	A = en déplacement vers incident ; R = disponible	Arrivée sur les lieux de l'incident
Arrivée sur les lieux de l'incident	B			A = sur les lieux de l'incident	Fin du traitement sur les lieux
Fin du traitement sur les lieux	B			A = fin du traitement	
Départ pour hôpital	C	A = fin du traitement & $Transport_D$=1		A = en déplacement vers hôpital	Arrivée à l'hôpital
Interruption demande D	C	A = fin du traitement & $Transport_D$=0		D = terminé	Fin du traitement de D
Arrivée à l'hôpital	B			A = à l'hôpital	Prise en charge du patient D
Prise en charge du patient D	B			D = terminé	Fin du traitement de D
Fin du traitement de D	B			A = en attente d'instruction	
Déplacement vers poste d'attente	C	A = en attente d'instruction & Liste=0	Déterminer le poste d'attente	A = en déplacement vers poste d'attente	Arrivée au poste d'attente
Arrivée au poste d'attente	B			A = au poste d'attente	
Détournement	C	$Détour_A$ = 1 & A = en déplacement vers poste d'attente		A = en déplacement vers incident	Arrivée sur les lieux de l'incident
Relocalisation	C	A = au poste d'attente	Relocaliser les ambulances	A = en déplacement vers poste d'attente	Arrivée au poste d'attente
Début pause A (fin de quart A)	C	(A = au poste d'attente, A=en attente d'instruction) & $Pause_A$=1 (Fin_A=1)		A = hors service	Fin pause (début de quart)
Fin pause A (début de quart A)	B			A = en attente d'instruction	

Tableau 4.2 – Événements principaux

Comme le modèle de simulation proposé permet le détournement des véhicules, c'est-à-dire de réaffecter un véhicule en déplacement vers le lieu d'un incident à un appel de priorité supérieure, il est nécessaire de pouvoir estimer où se situe le véhicule au moment de le détournement a lieu, puis de déterminer le temps de déplacement de cet endroit vers la nouvelle destination. À cet effet, différentes méthodes peuvent être utilisées, incluant des méthodes sophistiquées reliées à des systèmes d'information géographique (SIG). Le modèle de simulation proposé utilise une méthode relativement simple et générique basée sur une connaissance *a priori* de certains temps de déplacement réels ou estimés pour un ensemble de localisations importantes ou fréquemment utilisées. Anisi, en considérant M, la matrice des temps de déplacement connus entre ces localisations, $t_{a'b'}$, le temps de déplacement connu entre deux localisations a' et b' où a' et b' sont respectivement les deux localisations les plus proches de a et de b dans M, et d_{ab} et $d_{a'b'}$, les distances euclidiennes entre a et b et entre a' et b' respectivement, le temps de déplacement entre deux localisations qui ne sont pas incluses dans M, noté t_{ab}, se calcule grâce à la formule suivante :

$$t_{ab} = \frac{d_{ab} \times t_{a'b'}}{d_{a'b'}}. \tag{4.49}$$

De toute évidence, cette méthode n'est pas aussi précise qu'une méthode basée sur un SIG. Toutefois, si la matrice M contient suffisamment de points (dans le cas ici, on utilisera environ 600 points), cette méthode permettra d'approximer adéquatement les temps de déplacement en tenant compte de la présence d'obstacles ou de caractéristiques liés à l'infrastructure de transport (par exemple, des autoroutes, des ponts, des tunnels, des sens uniques) et aux conditions générales du système routier sur les itinéraires correspondant à chaque paire de points dans M (par exemple, les phénomènes de congestion). Lorsqu'un véhicule ambulancier est redirigé en cours de déplacement, on estimera qu'il se situe à une distance $\alpha \times d_{of}$ de son point d'origine o sur l'itinéraire le menant à sa destination finale f où α est le ratio entre le temps écoulé depuis le départ du véhicule vers sa destination originale et le temps de déplacement total requis pour atteindre celle-ci et où d_{of} est la distance entre o et f.

4.3.5 Décisions

L'exécution de certains événements, en cours de simulation, implique la modélisation et la réplication de processus décisionnels effectués généralement par les opérateurs ou les répartiteurs oeuvrant au sein du SPU étudié. En effet, la modélisation de ces processus vise à reproduire les pratiques et stratégies adoptées en pratique. Le module *Décisions* est donc responsable d'un

ensemble de décisions en ce qui concerne la gestion des véhicules ambulanciers, par exemple la sélection du véhicule à envoyer afin de répondre à un appel d'urgence et le choix de la localisation des véhicules disponibles à différents moments dans la journée pour n'en nommer que quelques-unes. Bien que des règles de gestion relativement simples puissent être utilisées en pratique, il est important de pouvoir intégrer facilement au sein du modèle de simulation différentes méthodes plus complexes, et éventuellement efficientes, afin de gérer l'ensemble des véhicules disponibles. Ainsi, afin de permettre une plus grande flexibilité au niveau des processus décisionnels, toutes les règles et stratégies, des plus simples aux plus sophistiquées, sont gérées à partir de routines externes. Elles peuvent donc être facilement remplacées ou modulées de manière à permettre l'analyse d'un large éventail des stratégies de gestion. Dans le cas de la présente étude, les différents modules de décisions en lien avec la localisation des véhicules sur le territoire à desservir ont été modifiés afin de considérer les stratégies de déploiement et de redéploiement propres à chaque scénario. Les différents modèles permettant de représenter les stratégies étudiées sont résolus au sein du module *Décisions*, grâce à CPLEX 12.5, sauf dans le cas de la stratégie 3 où la meilleure localisation correspondant aux critères est déterminée par énumération.

4.3.6 Mesures de performance

Le dernier module vise à supporter l'analyse des résultats issus des différents tests menés grâce à l'outil de simulation proposé. À cet effet, un historique complet de la simulation est enregistré puis utilisé afin de calculer un certain nombre de mesures de performance. Cet historique inclut tous les mouvements effectués et les demandes traitées par chaque ambulance, tous les temps associés aux différents changements d'état des entités, de même que l'information à propos des décisions effectuées en cours de simulation, par exemple le nombre de redéploiements effectués. Parmi les mesures de performance offertes à l'utilisateur, on compte le temps associé à chacune des étapes du processus mis en place pour servir un appel urgent incluant le temps de réponse (c'est-à-dire le temps entre la réception d'un appel et l'arrivée d'un véhicule ambulancier sur le lieu de l'incident), la charge de travail et le temps supplémentaire effectués par chaque équipe de techniciens ambulanciers, le nombre de fois qu'une ambulance est détournée ou redéployée, *etc*. Naturellement, selon les objectifs de l'étude effectuée, d'autres mesures de performance peuvent être calculées à partir de l'historique de la simulation. Enfin, la liste des mouvements de chaque ambulance peut également être utilisée afin de visualiser une journée de travail à partir d'une interface graphique ou pour vérifier l'outil de simulation en soi.

4.3.7 Implémentation, vérification et validation

L'implémentation, la vérification et la validation sont trois étapes importantes de toute étude de simulation. L'implémentation concerne la sélection des techniques utilisées afin d'implémenter le modèle de simulation de même que la phase de codage en soi. La vérification permet plutôt de s'assurer que le modèle de simulation implémenté est en mesure de faire ce qu'on lui demande de faire. Cette phase passe par l'inspection du code, le lancement de simulations préliminaires et la vérification de la pertinence des résultats obtenus (Altiok et Melamed, 2007). Enfin, la validation permet d'assurer que le modèle de simulation représente adéquatement le système étudié et permet d'atteindre les objectifs visés. Dans le cas qui nous intéresse, le modèle de simulation a été implémenté en utilisant un langage de programmation général (C++) afin de garantir un maximum de flexibilité. Pour de plus amples détails sur notre modèle de simulation, et plus particulièrement sa vérification et sa validation, nous référons le lecteur à un article que nous avons déjà publié à ce sujet (Kergosien *et al*., 2015).

4.4 Expérimentation

La phase d'expérimentation effectuée dans le cadre de ce projet vise à analyser et à quantifier les performances des différentes stratégies de déploiement et redéploiement proposées. Plus précisément, elle vise à analyser et à étudier le compromis entre les performances du système et les inconvénients liés à l'utilisation de stratégies de déploiement et de redéploiement dynamiques, et ce, dans différents contextes, c'est-à-dire confrontés à des niveaux d'occupation variables ou à des augmentations momentanées de la demande. Pour ce faire, un cas de base a été développé et sera utilisé tout au long de l'expérimentation. Différentes modifications seront ensuite apportées au cas de base de manière à tester les stratégies étudiées dans d'autres contextes et de valider les conclusions tirées pour différents cas de figure. Le cas de base utilisé pour la génération d'instances de même que les modifications qui y sont apportées sont présentés dans cette section. Ils sont suivis des résultats de l'étude de simulation en soi.

Afin de mener l'étude de simulation, un ensemble de données ont été générées de manière à correspondre à un service préhospitalier d'urgence fictif basé sur le contexte des villes de Montréal et de Laval (banlieue au nord de la ville de Montréal), le plus grand centre urbain de la province de Québec. En effet, tel qu'il a été mentionné dans Kergosien *et al*. (2015), le territoire à desservir a été défini à partir d'un contexte réel de manière à représenter adéquatement les difficultés liées à la gestion d'un SPU en contexte urbain telles que les zones de forte et de faible densité et la présence d'un centre-ville, pour ne nommer que ces exemples. Néanmoins,

le SPU considéré est fictif en ce sens que les décisions concernant la gestion opérationnelle et en temps réel du système, telles que les décisions de déploiement, de redéploiement et de répartition, sont basées sur un ensemble de règles considérées et acceptées dans la littérature plutôt que celles utilisées en pratique par l'organisation responsable des SPU pour la grande région de Montréal, Urgences-santé. En effet, nous n'avons pas d'information précise et officielle quant aux méthodes utilisées lors la prise de décision par Urgences-santé. Dans cette étude, les stratégies de déploiement et de redéploiement adoptées sont basées sur les modèles élaborés précédemment. De plus, nous considérerons que le véhicule le plus proche est toujours envoyé sur les lieux d'un incident. Finalement, la gestion du territoire à desservir se fait de manière totalement centralisée, c'est-à-dire que le territoire à desservir n'est pas divisé en districts. Ainsi, les répondants médicaux d'urgence (RMU), les opérateurs et les répartiteurs peuvent gérer les appels en provenance de toutes les zones de demande et toute ambulance peut être affectée à toute demande, indépendamment de la provenance de l'appel.

La Figure 4.3 illustre la région desservie par le SPU fictif considéré. Sur cette figure, chaque point représente le centroïde d'une zone de demande, la taille du point indiquant l'importance relative quant à la densité de population d'une zone donnée par rapport aux autres zones du territoire à desservir. La région considérée comporte donc 595 zones de demande, 40 postes d'attente potentiels localisés arbitrairement sur le territoire à desservir, 2 dépôts où les véhicules ambulanciers amorcent et terminent leur quart de travail et 15 hôpitaux ou centres de santé. Ce découpage du territoire est fidèle à celui utilisé par d'Urgences-santé. Un total de 150 équipes de techniciens ambulanciers travaillant chaque jour sur des quarts de travail de 8 heures, dont 125 dédiées exclusivement aux appels urgents, les autres étant affectées à d'autres tâches telles que le transport inter-établissement, ont été considérés. En prenant en compte un tel nombre de véhicules, un système présentant un niveau d'occupation élevé peut être représenté. Le nombre d'équipes en service à chaque période de la journée a été déterminé en fonction des courbes de la demande espérée, et est similaire au nombre d'équipes de travail utilisées par Urgences-santé. Le nombre d'équipes est toutefois amené à varier en fonction du scénario considéré.

De plus, puisqu'aucune donnée précise officielle concernant les demandes, les temps de déplacement et les temps d'intervention n'est disponible, un ensemble de données aléatoires ont été générées à partir d'informations provenant de plusieurs sources : rapports annuels publics d'Urgences-santé (2006), données de Statistics Canada (2011) et informations tirées de la littérature scientifique sur le domaine. Puisque la plupart de ces données se présentent sous une forme très agrégée, les paramètres du générateur de données ont été fixés de manière empirique afin d'obtenir des résultats réalistes et qui permettent de reproduire adéquatement les données

Figure 4.3 – Cartographie du territoire à desservir

agrégées collectées en termes de nombre total de demandes urgentes, de nombre de transports effectués, de nombre d'équipe de travail et du nombre de véhicules ambulanciers.

À cet effet, tel que généralement accepté dans la littérature (Ingolfsson, 2013), une distribution exponentielle a été utilisée afin de modéliser le temps entre l'arrivée de deux demandes successives. Chaque journée a été divisée en 12 périodes de deux heures, la moyenne de la distribution exponentielle associée à chaque période variant entre 1minute 30 secondes et 5 minutes en fonction de la période. De cette manière, il est possible de tenir compte adéquatement de la variation de l'intensité de la demande au cours d'une journée. Chaque fois qu'une demande est générée, elle est ensuite associée à une zone spécifique en fonction d'une distribution discrète où la probabilité pour qu'une zone de demande soit sélectionnée dépend du nombre espéré de demandes dans la zone pour la période considérée. Tel que discuté dans Aboueljinane *et al*. (2013), il s'agit là d'une des approches possibles afin de générer adéquatement les demandes.

Les temps d'intervention sur les lieux d'un incident de même que les temps de transfert à l'hôpital, incluant le temps de prise en charge du patient et les tâches administratives subséquentes, ont été modélisés grâce à des distributions Gamma(k, θ) où k correspond au paramètre de forme et θ, au paramètre d'échelle. Dans 75 % des cas, les demandes urgentes vont requérir un transport

vers un centre hospitalier. Dans ce cas, le temps d'intervention et le temps de transfert suivent respectivement une distribution Gamma(3, 5) et Gamma(8, 5). Le centre hospitalier vers lequel un patient est transporté est ensuite sélectionné aléatoirement avec un fort biais pour les centres hospitaliers les plus proches. Lorsqu'aucun transport n'est requis (dans 25 % des cas), le temps d'intervention sur les lieux de l'incident suit une distribution Gamma(3, 10). Les distributions de même que les paramètres utilisés ici sont inspirés des observations de Schmid (2012). Le nombre de RMU, d'opérateurs et de répartiteurs ont, quant à eux, été fixés de manière à ce que le temps d'attente avant le traitement d'une demande soit nul. À son tour, le temps de traitement d'un appel suit une distribution Gamma(1, 2). Enfin, les temps de déplacement d'une localisation à une autre sont générés de la manière décrite à la section 4.3.4 en utilisant la distance euclidienne entre chaque point de même qu'une matrice comportant un ensemble de valeurs réelles pour les temps de déplacement entre deux points sur le territoire à desservir.

Nous référerons donc aux instances générées à partir des données présentées ci-dessus comme aux instances de base (B). Néanmoins, afin de comparer les stratégies de déploiement et de redéploiement proposées dans d'autres contextes, le cas de base a été modifié donnant lieu à divers groupes d'instances. Dans un premier temps, trois scénarios ont été créés correspondant à différents nombres de véhicules disponibles. En faisant varier le nombre de véhicules disponibles, il est possible de représenter différentes situations où les véhicules présentent un taux d'occupation varié, ce qui permet, par le fait même, de représenter aussi des niveaux d'occupation variés. Les trois scénarios générés correspondent respectivement à une diminution de 20 % du nombre de véhicules (DV20), à une diminution de 10 % du nombre de véhicules (DV10) et à une augmentation de 10% du nombre de véhicules (AV10). Dans tous les cas, le nombre de véhicules augmente ou diminue de la même manière pour toutes les périodes. Dans un deuxième temps, six scénarios ont été générés afin de représenter différents profils d'augmentation de la demande. Ces scénarios correspondent respectivement à une augmentation de 10 % de la demande (AD10), à une augmentation de 20 % de la demande (AD20), à une augmentation de 10 % de la demande pour les périodes de jour (10h à 18h) (AJ10) , à une augmentation de 20 % de la demande pour les périodes de jour (AJ20), à une augmentation de 10 % de la demande pour les périodes de soir (18h à 2h) (AS10) et à une augmentation de 20 % de la demande pour les périodes de soir (AS20). Dans tous les cas, l'augmentation de la demande a lieu pour toutes les journées considérées dans l'étude de simulation.

La phase d'expérimentation a donc été menée en considérant les 10 groupes d'instances présentées ci-haut, en incluant le cas de base, et ce, pour les quatre stratégies décrites à la section 4.1. Afin d'analyser et de comparer les différentes stratégies de déploiement et de redéploiement

proposées, plusieurs mesures de performance ont été enregistrées tout au long des simulations. Dans ce chapitre, toutes les mesures de performance possibles ne seront pas présentées, mais plutôt celles qui nous paraissent les plus pertinentes en fonction des objectifs de cette étude. Ainsi, afin de mesurer les performances globales du système, le temps de réponse moyen (en secondes) et le pourcentage des appels servis à l'intérieur de 9 minutes (en %) sont présentés. La distance totale parcourue pour tous les véhicules (en kilomètres), la distance totale de redéploiement (en kilomètres) et le nombre de redéploiement ont été calculés de manière à évaluer les inconvénients associés aux stratégies de redéploiement. Les résultats rapportées dans la présente section sont basés sur 50 réplications, chacune étant composée de 7 jours consécutifs. Afin d'éliminer le régime transitoire correspondant au premier et au dernier jour, les résultats présentés sont déterminés en n'utilisant que les 5 jours du milieu. Pour chaque mesure, l'intervalle de confiance à 95 % est présentée aux tableaux 4.3 et 4.4. Afin de mieux illustrer les résultats rapportés dans les tableaux, les figures 4.4 à **??** présentent les courbes tracées pour le temps de réponse moyen, pour le pourcentage des appels servis à l'intérieur de 9 minutes et pour la distance totale parcourue. Dans tous les cas, la mesure a été tracée en fonction du scénario choisi, et ce, pour toutes les stratégies étudiées.

4.4.1 Variation du nombre de véhicules

Une première série de tests a été effectuée de manière à analyser les stratégies de déploiement et de redéploiement lorsque confrontées à des niveaux d'occupation variés. Tel que décrit ci-haut, afin de représenter ces niveaux d'occupation, différentes tailles de flotte de véhicules ont été utilisées. Pour le cas de base, on considérera que 125 des 150 équipes de travail disponibles pourront traiter des appels urgents, les autres équipes étant affectées aux transports non-urgents, par exemple aux transports inter-établissement. Ainsi, on considérera 101, 113 et 138 véhicules ou équipes de travail respectivement pour les cas DV20, DV10 et AV10. De cette manière, il est possible de représenter des taux d'occupation variant entre 50% et 70%.

Tout d'abord, les résultats présentés à la Figure 4.4 et au Tableau 4.3 montrent que, pour le cas considéré dans cette étude, il n'y a pas de différence notable entre les scénarios 1 et 2 au niveau de l'amélioration des performances du système. Le fait de considérer le redéploiement multi-période amène, en moyenne, une augmentation de la distance totale parcourue et, nécessairement de la distance et du nombre de redéploiement, sans impact important sur les performances du système, et ce, pour tous les scénarios considérés en termes de nombre de véhicules. Bien que plusieurs recherches aient démontré que la relocalisation multi-période améliore le niveau de service, ce n'est pas véritablement le cas ici. En effet, la stratégie 1 a été élaborée de

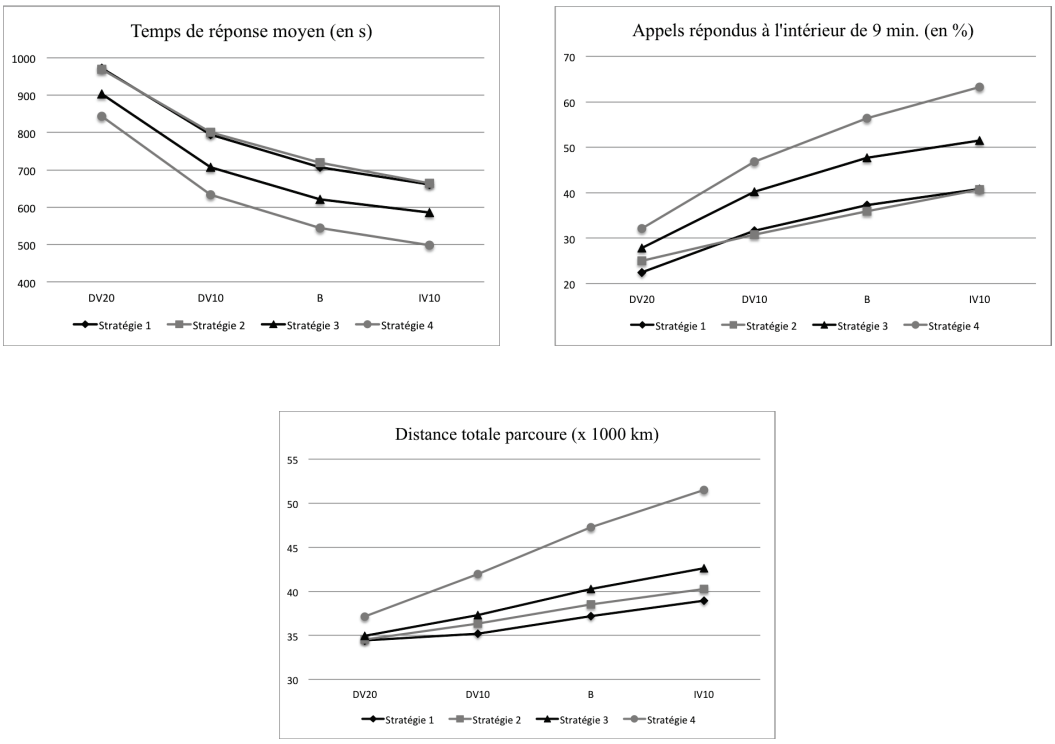


Figure 4.4 – Résultats obtenus pour différentes tailles de flotte de véhicules

manière à prendre en compte les variations du nombre de véhicules, du profil de la demande et des temps de déplacement en fonction de la période, de même que du chevauchement des quarts de travail. De cette manière, le système est en mesure de s'ajuster à chaque début de période en considérant le début du quart de travail de certains véhicules, même si le poste d'attente d'un véhicule demeure fixe pour toute la période. La stratégie 1 représente donc un hybride entre une localisation purement statique où le même nombre de véhicules est disponible à tout moment et une relocalisation multi-période. Ceci explique, du moins en partie, pourquoi les résultats obtenus pour les deux scénarios soient très similaires. La stratégie 1 est donc en mesure de s'adapter adéquatement à l'évolution du système sans pour autant avoir recours à la relocalisation.

En comparant les stratégies 3 et 4 avec la stratégie 1, il est possible d'observer que les stratégies de gestion dynamiques, que ce soit le repositionnement ou la relocalisation, contribuent à améliorer les performances du système, mais amènent aussi des coûts de relocalistion non négligeables. Plus précisément, considérer exclusivement le repositionnement dynamique (stratégie 3) permet de réduire le temps de réponse moyen (de 70 à 88 secondes en moyenne selon l'instance) de même qu'augmenter le pourcentage d'appels servis à l'intérieur de 9 minutes (de 5 à 11 % en moyenne). S'adapter un véhicule à la fois amène donc clairement des améliorations

Scénario	Mesure	Stratégie 1			Stratégie 2			Stratégie 3			Stratégie 4		
DV20	Temps de réponse (s)	973	±	10	969	±	11	903	±	11	843	±	12
	% ≤ 9 minutes	22,5	±	0,4	25,0	±	0,4	27,8	±	0,5	32,1	±	0,6
	Distance totale (km)	34 449	±	171	34 535	±	154	34 961	±	159	37 162	±	184
	Distance redép. (km)	0,0	±	0,0	177	±	20	0,0	±	0,0	4 913	±	112
	Nombre de redép.	0,0	±	0,0	102,9	±	5	0,0	±	0,0	838,4	±	18
DV10	Temps de réponse (s)	794	±	6	800	±	6	707	±	7	634	±	7
	% ≤ 9 minutes	31,7	±	0,3	30,8	±	0,4	40,2	±	0,5	46,8	±	0,5
	Distance totale (km)	35 232	±	153	36 355	±	147	37 347	±	130	41 961	±	136
	Distance redép. (km)	0,0	±	0,0	409	±	26	0,0	±	0,0	7 169	±	112
	Nombre de redép.	0,0	±	0,0	209,6	±	7.3	0,0	±	0,0	1 232,0	±	20,5
B	Temps de réponse (s)	707	±	5	720	±	4	622	±	5	545	±	4
	% ≤ 9 minutes	37,3	±	0,4	35,9	±	0,3	47,7	±	0,5	56,4	±	0,5
	Distance totale (km)	37 217	±	142	38 525	±	134	40 273	±	158	47 285	±	143
	Distance redép. (km)	0,0	±	0,0	553	±	32	0,0	±	0,0	8 101	±	84
	Nombre de redép.	0,0	±	0,0	378,5	±	9,7	0,0	±	0,0	1 391,8	±	14,0
AV10	Temps de réponse (s)	661	±	3	664	±	3	586	±	3	498	±	2
	% ≤ 9 minutes	40,8	±	0,3	40,7	±	0,4	51,5	±	0,4	63,6	±	0,3
	Distance totale (km)	38 967	±	152	40 301	±	151	42 617	±	156	51 509	±	153
	Distance redép. (km)	0,0	±	0,0	893	±	37	0,0	±	0,0	7 985	±	76
	Nombre de redép.	0,0	±	0,0	566,7	±	9,5	0,0	±	0,0	1 355,9	±	11,5

Tableau 4.3 – Résultats pour différentes tailles de flotte de véhicules

en termes de niveau de service. En contrepartie, la distance totale parcourue augmente aussi, de 512 à 3650 km selon le scénario considéré. Cela signifie donc que certains véhicules devront être envoyés relativement loin de leur position actuelle ou de leur poste d'attente précédent afin d'améliorer la capacité du système à répondre aux appels futurs. Également, plus le nombre de véhicules utilisé est grand, plus la distance totale parcourue augmente (en effet, il y a plus de véhicules), mais cette flexibilité semble aussi se concrétiser en de meilleurs résultats. Le fait de considérer à la fois le repositionnement et le redéploiement dynamique (scénario 4) améliore le niveau de service de manière encore plus importante. Selon le scénario, cette stratégie permet de réduire le temps de réponse moyen de 130 à 162 secondes, ce qui représente une amélioration deux fois plus importante que le fait de considérer exclusivement le repositionnement dynamique (stratégie 3). Le pourcentage des appels atteints à l'intérieur de 9 minutes augmente, quant à lui, de 9,59 % (DV20) à 22,48 % (AV10), ce qui correspond toujours à une amélioration d'environ du double par rapport à la stratégie 3. Néanmoins, afin d'obtenir de telles performances, les coûts de relocalisation augmentent aussi considérablement : la distance totale parcourue augmente de 2713 km à 12 542 km en moyenne selon le scénario, ce qui représente une augmentation moyenne variant entre 5 et 18 km par quart de travail. Ceci constitue une valeur trois fois plus grande que celle obtenue en considérant le repositionnement dynamique seul. De plus, comme dans le cas du stratégie 3, plus le nombre de véhicules utilisé est grand, plus la distance totale parcourue augmente. On observe toutefois que, bien que la distance totale parcourue continue d'augmenter avec le nombre de véhicules, la distance de redéploiement semble se stabiliser à un certain moment. L'augmentation de la distance totale parcourue est

donc due, entre autres, à l'augmentation du nombre de véhicules, et non une conséquence directe de l'augmentation de la distance parcourue par véhicule.

Enfin, les résultats présentés à la Figure 4.4 et au Tableau 4.3 permettent aussi de conclure que l'utilisation des stratégies de localisation et de relocalisation dynamiques en utilisant un nombre donné de véhicules peut donner des résultats aussi performants que d'augmenter le nombre de véhicules lorsqu'on n'utilise pas de stratégies dynamiques. En effet, les stratégies 3 et 4 mènent à de meilleures performances qu'une augmentation du nombre de véhicules de 10 % par rapport au cas de base. De plus, réduire les effectifs de 10 % en considérant des stratégies de gestion dynamiques est équivalent au cas de base pour lequel des stratégies statiques sont considérées. Ainsi, les stratégies dynamiques constituent des alternatives intéressantes à l'augmentation du nombre de véhicules utilisés. Néanmoins, le taux d'occupation du système est très élevé, même si les stratégies dynamiques peuvent contribuer à améliorer les performances par rapport aux stratégies statiques, elles ne permettent pas de pallier au manque de ressources. Dans ces situations, augmenter le nombre de véhicules disponibles demeure la meilleure option.

4.4.2 Augmentation de la demande

Une deuxième série de tests a été menée afin d'évaluer le comportement du système lorsque confronté à une augmentation globale ou ponctuelle de la demande. Ainsi, tel que décrit ci-haut, différents scénarios correspondant à différents profils d'augmentation de la demande, en termes d'intensité et de période, ont été évalués. Dans tous les cas, on considérera que 125 véhicules sont disponibles pour répondre aux appels urgents (*c.f.* cas de base).

Les résultats présentés au Tableau 4.4 pour différents profils de demande montrent que, tel qu'espéré, l'augmentation de la demande engendre une dégradation des performances du système, et de toute évidence, plus l'augmentation est importante, plus les performances se dégradent. Néanmoins, le système réagira mieux à une augmentation de la demande pendant le soir que pendant le jour. En effet, le système, tel que considéré dans cette étude, est moins occupé le soir, ce qui peut aider dans de telles situations. Dans les conditions étudiées, les stratégies de gestion dynamiques demeurent supérieures aux stratégies statiques en termes de niveau de service. Toutefois, on remarque que, de manière générale, les changements de la demande amènent une dégradation des performances du système dans la même mesure, quelles que soient les stratégies déployées. En effet, la dégradation du temps de réponse moyen et du pourcentage d'appels atteints à l'intérieur de 9 minutes est similaire pour toutes les stratégies. Puisque le cas de base considéré ici est déjà très chargé à en juger par les performances obtenues, cela confirme le fait que lorsque le taux d'occupation est plus élevé, aucune stratégie ne

Scénario	Mesure	Stratégie 1			Stratégie 2			Stratégie 3			Stratégie 4		
B	Temps de réponse (s)	707	±	5	720	±	4	622	±	5	545	±	4
	% ≤ 9 minutes	37,3	±	0,4	35,9	±	0,3	47,7	±	0,5	56,4	±	0,5
	Distance totale (km)	37 217	±	142	38 525	±	134	40 273	±	158	47 285	±	143
	Distance redép. (km)	0,0	±	0,0	553	±	32	0,0	±	0,0	8 101	±	84
	Nombre de redép.	0,0	±	0,0	378,5	±	9,7	0,0	±	0,0	1391,8	±	14,0
AD10	Temps de réponse (s)	778	±	6	791	±	6	697	±	7	622	±	7
	% ≤ 9 minutes	32,9	±	0,3	31,8	±	0,3	41,5	±	0,5	48,6	±	0,5
	Distance totale (km)	39 107	±	136	40 036	±	152	41 211	±	120	46 096	±	142
	Distance redép. (km)	0,0	±	0,0	408	±	87	0,0	±	0,0	7571	±	114
	Nombre de redép.	0,0	±	0,0	250,0	±	7,5	0,0	±	0,0	1294,8	±	20,7
AD20	Temps de réponse (s)	911	±	10	923	±	9	848	±	9	784	±	10
	% ≤ 9 minutes	25,2	±	0,5	25,2	±	0,4	31,6	±	0,5	36,4	±	0,6
	Distance totale (km)	41 610	±	220	42 112	±	198	42691	±	187	44692	±	192
	Distance redép. (km)	0,0	±	0,0	204,6	±	15,8	0,0	±	0,0	4709	±	123
	Nombre de redép.	0,0	±	0,0	123,8	±	6,1	0,0	±	0,0	963,0	±	22,4
AJ10	Temps de réponse (s)	749	±	6	759	±	7	667	±	6	590	±	7
	% ≤ 9 minutes	35,0	±	0,4	33,8	±	0,4	44,0	±	0,5	52,1	±	0,5
	Distance totale (km)	37 959	±	146	39 195	±	142	40 552	±	138	46 605	±	147
	Distance redép. (km)	0,0	±	0,0	564	±	29	0,0	±	0,0	7823	±	84
	Nombre de redép.	0,0	±	0,0	338,6	±	7,5	0,0	±	0,0	1350,2	±	18,0
AJ20	Temps de réponse (s)	836	±	9	843	±	9	764	±	10	690	±	9
	% ≤ 9 minutes	30,7	±	0,4	29,7	±	0,4	38,2	±	0,6	45,1	±	0,6
	Distance totale (km)	39 528	±	193	40 584	±	169	41 491	±	169	46 504	±	170
	Distance redép. (km)	0,0	±	0,0	535	±	31	0,0	±	0,0	7104	±	94
	Nombre de redép.	0,0	±	0,0	285,0	±	6,5	0,0	±	0,0	1226	±	16,4
AS10	Temps de réponse (s)	723	±	4	738	±	4	641	±	4	563	±	4
	% ≤ 9 minutes	36,0	±	0,4	34,6	±	0,3	45,4	±	0,4	54,2	±	0,5
	Distance totale (km)	37 737	±	147	38 761	±	136	40 487	±	113	46 710	±	145
	Distance redép. (km)	0,0	±	0,0	419	±	24	0,0	±	0,0	8046	±	92
	Nombre de redép.	0,0	±	0,0	327,9	±	8,2	0,0	±	0,0	1385,9	±	16,6
AS20	Temps de réponse (s)	762	±	6	776	±	6	685	±	6	609	±	6
	% ≤ 9 minutes	34,2	±	0,4	32,6	±	0,3	42,2	±	0,4	50,2	±	0,5
	Distance totale (km)	38 549	±	153	39 386	±	157	40 873	±	131	46 064	±	158
	Distance redép. (km)	0,0	±	0,0	272	±	22	0,0	±	0,0	7643	±	114
	Nombre de redép.	0,0	±	0,0	284,5	±	6,7	0,0	±	0,0	1316,2	±	21,2

Tableau 4.4 – Résultats pour différents profils de la demande

peut pallier au manque de ressources.

En observant plus attentivement les coûts de relocalisation, il est possible noter que la distance totale parcourue n'augmente pas nécessairement avec le nombre de demandes. Dans le cas d'une augmentation de la demande, la distance parcourue par les véhicules de leur poste d'attente vers les lieux de l'incident augmente, puisqu'ils ont plus de demandes à servir. La distance totale parcourue devraient augmenter conséquemment. Dans le cas des stratégies 1, 2 et 3, c'est bien le cas : la distance totale parcourue augmente avec la demande. Par contre, pour la stratégie 4, la distance totale parcourue diminue lorsque le nombre de demandes augmente. Dans les faits, une augmentation de la demande amène aussi une augmentation du taux d'occupation des véhicules. Les possibilités en termes de redéploiement sont donc plus limitées. Comme en témoignent les résultats présentés au Tableau 4.4, une réduction de la distance totale de redéploiement et une réduction du nombre de redéploiement peuvent être observées lorsque la demande augmente. Ceci mène, dans le cas de la stratégie 4, à une réduction de la distance

totale parcourue.

4.5 Conclusion

Dans ce chapitre, nous nous sommes intéressés à l'évaluation et à l'analyse comparative de différentes stratégies de déploiement et de redéploiement. Pour réaliser cette analyse, un ensemble de stratégies de gestion caractérisées par différentes règles associées au déploiement initial, au redéploiement et au repositionnement des véhicules ambulanciers à la suite de leur libération ont été définies. Dans un premier temps, les stratégies faisant l'objet de ce chapitre et les modèles sous-jacents ont été présentés de manière plus précise. Ainsi, quatre stratégies ont été proposées, soit le déploiement *a priori*, sans redéploiement (stratégie 1), le déploiement *a priori* avec redéploiement multi-période (stratégie 2), le déploiement et le redéploiement dynamiques (stratégie 3) et le déploiement, repositionnement et redéploiement dynamiques (stratégie 4). Ces stratégies de gestion ont permis de représenter adéquatement un large éventail de possibilités allant de stratégies statiques classiques aux stratégies dynamiques plus sophistiquées. Dans le cadre de cette étude, toutes les stratégies définies ont été modélisées à partir du modèle avec double standard (MDS) (Gendreau *et al*., 1997). Différentes modifications ont toutefois été apportées au MDS afin de les représenter correctement. Ces différents modèles ont été utilisés afin de répliquer le processus décisionnel sous-jacent aux différentes stratégies de gestion au sein du modèle de simulation.

Dans un deuxième temps, le modèle de simulation développé afin de comparer et d'analyser les différentes stratégies de déploiement et de redéploiement a été présenté. De manière plus concrète, le modèle de simulation développé comporte six composantes principales : un module *Données* qui inclut l'ensemble des paramètres nécessaires à une description adéquate du système étudié et de son état initial, un module *Demandes et variables aléatoires* qui génère l'ensemble des demandes et des variables aléatoires requises afin de mener à bien l'étude de simulation, un module *Moteur de la simulation* basé sur la simulation à événements discrets qui permet de gérer la simulation en soi, soit l'horloge de la simulation et la liste des événements, un module *Temps de déplacement* qui permet d'estimer le temps de déplacement entre deux localisations sur le territoire à desservir, un module *Décisions* qui permet de gérer la prise de décision en répliquant le processus décisionnel approprié au moment opportun en cours de simulation et enfin, un module *Mesures de performance* qui enregistre et compile l'information nécessaire afin d'évaluer les performances du système étudié. L'outil de simulation développé dans ce contexte est suffisamment flexible et générique pour permettre l'étude et l'analyse de différentes décisions à tous les niveaux de planification et dans des contextes d'application va-

riés. Enfin, une série de tests ont été effectués afin d'analyser et de mesurer les performances des différentes stratégies proposées et, plus précisément, de quantifier le compromis entre les performances du système et les inconvénients liés à l'utilisation de telles stratégies, et ce, dans différents contextes. Les instances utilisées lors de la phase d'expérimentation ont été générées à partir d'un cas basé sur le contexte de la géographie et de l'organisation responsable des SPU de la grande région de Montréal, Québec.

En se basant sur les résultats obtenus, il est possible d'observer que le fait de considérer l'implantation de stratégies de localisation et de relocalisation dynamiques au sein d'un SPU peut mener à des résultats similaires à une augmentation du nombre de véhicules disponibles. Étant donné les coûts très élevés associés à ce type de véhicule, les stratégies dynamiques représentent donc des alternatives fort intéressantes à l'augmentation des ressources. En considérant la possibilité d'un contexte de restrictions budgétaires dans le secteur de la santé, En considérant les compressions budgétaires présentes de nos jours, à tous les niveaux, une utilisation plus efficiente des ressources devient un enjeu majeur. Néanmoins, lorsque le taux d'occupation du système est très élevé, il est également possible de noter que, même si les stratégies dynamiques peuvent améliorer le niveau de service offert à la population, elles ne permettent pas de pallier au manque de ressources. Dans de telles situations, l'utilisation d'un plus grand nombre de véhicules demeure incontournable pour maintenir un niveau de service adéquat. D'autre part, les stratégies dynamiques mènent à de meilleures performances, mais engendrent aussi une augmentation notable de certains inconvénients potentiels ressentis par le personnel paramédical, notamment l'augmentation de la distance totale parcourue, qui peuvent devenir difficilement justifiables pour certaines organisations. À notre avis, en se basant sur les résultats obtenus et sur notre connaissance des SPU, le repositionnement dynamique peut devenir une bonne alternative à envisager pour les organisations qui ne considèrent actuellement que des stratégies statiques. Le redéploiement des véhicules, mais à l'intérieur d'une certaine limite peut aussi être très intéressant. La meilleure stratégie à adopter dépendra du contexte étudié et de l'organisation impliquée dans le processus à gérer. Cette étude a toutefois été en mesure de montrer qu'un outil d'analyse basé sur la simulation, tel que celui développé ici, permet de fournir un ensemble d'information d'une grande valeur afin de soutenir une prise de décision éclairée.

Les résultats de cette étude ont permis de confirmer que le redéploiement en temps réel des véhicules ambulanciers constitue une bonne option dans plusieurs situations, mais les inconvénients qui y sont associés peuvent être non négligeables. Le développement et l'analyse de règles et stratégies qui permettent de limiter ou de contrôler les coûts encourus lors de l'implantation d'une stratégie de redéploiement dynamique de manière plus systématique, et ainsi favoriser

l'adoption de politiques hybrides entre le repositionnement dynamique et le redéploiement dynamique complet nous paraît donc d'un grand intérêt. De cette manière, le redéploiement des véhicules ambulanciers pourra toujours être considéré, mais de façon contrôlée, tant en termes de fréquence que d'intensité. Nous aborderons ce sujet plus en détail dans les deux derniers chapitres de cette thèse. Enfin, l'analyse menée dans le cadre de cette étude pourrait être étendue à d'autres cas afin de valider si, et dans quelle mesure, les conclusions tirées ici sont les mêmes lorsque d'autres contextes d'application sont considérés, par exemple, pour des villes ayant des topographies différentes.

CHAPITRE 5

LE PROBLÈME DE REDÉPLOIEMENT ET DE PRÉAFFECTATION DANS LA GESTION EN TEMPS RÉEL D'UN SERVICE PRÉHOSPITALIER D'URGENCE

L'analyse effectuée au *Chapitre 4* a permis de montrer que des stratégies de gestion dynamiques et plus flexibles, telles que le redéploiement de véhicules ambulanciers, peuvent contribuer à l'amélioration des performances des services préhospitaliers d'urgence. En effet, dans les contextes étudiés, les stratégies de gestion dynamiques permettent d'améliorer le service à la population lorsqu'on les compare aux stratégies statiques plus traditionnelles. Elles permettent également de maintenir des performances similaires en utilisant un nombre réduit de véhicules. Ces constats sont d'autant plus vrais lorsque le système considéré est en sous-capacité. Les stratégies de gestion dynamique amènent toutefois des inconvénients liés, entre autres, aux déplacements plus fréquents des véhicules, et par conséquent, à l'augmentation du nombre de kilomètres parcourus et aux changements fréquents de postes d'attente. Elles pourront donc avoir un impact non négligeable sur les coûts d'utilisation des véhicules (c'est-à-dire les coûts reliés à la maintenance, à l'entretien et à l'essence), mais aussi sur la gestion des ressources humaines. En effet, les changements fréquents de postes d'attente, de même que les déplacements à vide peuvent être perçus négativement par les techniciens ambulanciers. La sélection d'une stratégie de redéploiement adéquate représentant un compromis intéressant entre le service à la population et les inconvénients qui y sont associés devient alors impérative. La gestion des SPU présente des enjeux importants. En effet, la capacité d'une organisation de se doter de stratégies et d'outils afin de résoudre en temps réel divers problèmes de décision peut avoir un impact considérable sur les performances de l'ensemble du système, et donc sur la santé des patients. L'élaboration de stratégies de gestion dynamiques, de même que le développement d'outils pour soutenir la prise de décision en temps réel, nous paraît donc tout à fait justifiée dans ce contexte.

Lors de la gestion d'un SPU, deux types de décisions sont généralement considérés en temps réel : la répartition des appels d'urgence et la localisation des véhicules en attente. Tout d'abord, lorsqu'un appel d'urgence est reçu, il faut déterminer quel est le véhicule qui sera envoyé sur les lieux de l'incident. C'est la répartition des appels. Traditionnellement, on considérera que le véhicule le plus proche est toujours sélectionné. Cette règle correspond à la règle employée de manière officielle par la plupart des SPU. En pratique toutefois, un répartiteur d'expérience pourrait choisir un véhicule différent lorsque la priorité de l'appel le permet. Dans cet esprit,

certains chercheurs se sont intéressés à l'étude de différentes stratégies de répartition pour les appels de priorité moindre, mais nécessitant toujours une attention immédiate. En effet, dans certains cas, il pourrait être plus intéressant d'affecter un véhicule différent de celui le plus proche de manière à maintenir un temps de réponse adéquat pour l'appel concerné, mais surtout, pour limiter une dégradation éventuelle du niveau de service global. Ainsi, des règles de répartition visant à minimiser les coûts de relocalisation (Gendreau *et al*., 2001) ou encore à minimiser la dégradation de la capacité de réponse du système (Andersson et Värbrand, 2007) ont été proposées, de même que des politiques de répartition basées sur la programmation dynamique (Schmid, 2012). Schmid (2012) a d'ailleurs montré que, pour les appels de priorité moindre, mais qui requièrent toujours une intervention immédiate, le fait de considérer d'autres politiques de répartition peut contribuer à améliorer le service à la population. Dans un deuxième temps, afin de servir la population dans des délais de temps raisonnables, les véhicules disponibles pour répondre aux appels de détresse doivent être localisés de manière adéquate sur le territoire à desservir. Tel qu'il a été discuté aux chapitres précédents, contrairement à la répartition des appels, le problème de localisation ou de déploiement des véhicules ambulanciers, et principalement la version statique de celui-ci, a été étudié de manière plus approfondie.

D'importants constats ressortent de l'analyse des différents modèles et stratégies de gestion en temps réel proposés jusqu'à maintenant. Tout d'abord, il est possible de constater que les décisions de répartition et de relocalisation sont généralement considérées indépendamment. En fait, la plupart des modèles proposés ne considèrent que le problème de localisation. Ainsi, peu importe le contexte d'application ou les objectifs visés, la règle de répartition qui consiste à envoyer le véhicule le plus proche sur les lieux d'un incident est toujours celle qui est utilisée. L'impact des décisions de répartition sur les décisions de localisation n'est donc généralement pas pris en compte explicitement. De plus, bien qu'il s'agisse d'une hypothèse réaliste, cette règle est très myope en pratique : l'impact de l'affectation d'un véhicule sur les performances futures du système n'est pas pris en compte. Ainsi, dans certains cas, envoyer le véhicule le plus proche à un appel de priorité moindre pourrait mener à une dégradation des performances globales du système, ou nécessiter un redéploiement des véhicules disponibles afin de retrouver un niveau de performance adéquat. Envoyer un véhicule différent, mais qui se situe à l'intérieur des délais acceptables, pourrait éviter de tels inconvénients. En pratique, bien que certains répartiteurs semblent le faire de façon instinctive, cela ne se traduit pas formellement dans la plupart des modèles proposés jusqu'à maintenant. Néanmoins, afin de prendre de telles décisions, il sera nécessaire de considérer de manière plus intégrée les décisions de répartition et de redéploiement.

Un deuxième constat intéressant repose sur les objectifs des modèles proposés. En effet, la plupart des modèles proposés jusqu'ici considèrent des objectifs en lien avec la couverture du territoire à desservir. Cette mesure, facile à comprendre en pratique, est en lien direct avec certaines règles gouvernementales imposées aux organismes responsables des SPU, par exemple, une proportion de la population couverte à l'intérieur d'un délai prescrit. Elle permet donc de développer des modèles réalistes. En contrepartie, elle génère souvent un grand nombre de solutions équivalentes. Afin de mesurer les performances espérées du système, Andersson et Värbrand (2007) ont proposé l'utilisation de la capacité de réponse ou *preparedness* en anglais, définie comme la capacité d'un système à servir des demandes potentielles au temps présent et dans le futur. Cette mesure permet certes de départager plus facilement les solutions potentielles, mais elle est beaucoup plus difficile à comprendre et à interpréter que les mesures de couvertures classiques. Une « bonne » valeur pour la capacité de réponse dépendra fortement du contexte. Cette mesure est donc très arbitraire. C'est probablement pour cette raison qu'elle a été si peu utilisée jusqu'à maintenant dans des contextes autres que celui considéré par Andersson et Värbrand (2007). Ainsi, d'autres mesures pourraient être envisagées en se basant sur des mesures de performance différentes qui permettraient de mieux prédire les performances du système ou anticiper les demandes futures, tout en respectant un certain nombre de critères et d'indicateurs plus traditionnels. Cela permettrait aussi de mieux départager différentes solutions. Enfin, la capacité des véhicules à répondre à des appels urgents est rarement prise en compte lors de l'élaboration des modèles de déploiement ou de redéploiement. Le fait de ne pas considérer la capacité « réelle » d'un véhicule peut mener à l'affectation d'un nombre non-réaliste de demandes à un seul véhicule et, conséquemment, à une surestimation des performances du système. De plus, cela peut engendrer un déséquilibre de la charge de travail potentielle des véhicules disponibles.

Ainsi, plusieurs questions émergent de ces différents constats. Serait-il intéressant d'intégrer, au sein d'un même modèle décisionnel, la répartition et le redéploiement/relocalisation des véhicules ambulanciers ? Lors de l'élaboration de tels modèles, serait-il utile de considérer ou de développer des mesures de performance autres que la couverture ? Si oui, lesquelles et comment les calculer ? Serait-il bénéfique d'intégrer la capacité réelle des véhicules au sein du modèle de décision ? Quel serait alors l'impact de ces différentes approches sur les solutions obtenues et sur les performances prédites du système ?

Afin d'aborder ces différentes questions, ce chapitre propose d'abord un modèle de décision pour la gestion des SPU considérant le redéploiement des véhicules, mais également une préaffectation anticipative des demandes éventuelles aux véhicules disponibles. Ce modèle vise à

minimiser le temps de réponse espéré, mesure que nous définirons dans le présent chapitre, tout en limitant les efforts liés au redéploiement (par exemple, le nombre de véhicules redéployés ou la distance de redéploiement). La capacité réelle des véhicules y sera également prise en compte. Dans le présent chapitre, le problème de redéploiement et de préaffectation des véhicules ambulanciers (PRPA) de même que ses caractéristiques principales sont d'abord décrits et discutés. La formulation de ce problème est ensuite présentée, suivie de l'analyse du modèle grâce aux résultats obtenus lors de la phase d'expérimentation. Par l'analyse et la validation de ce modèle, il sera possible de répondre aux différentes questions soulevées ci-haut et de mesurer ses avantages et ses inconvénients ici. Enfin, une discussion sur les pistes de recherche future viendra conclure le chapitre. Les contributions de ce chapitre s'inscrivent donc selon deux perspectives principales. Tout d'abord, ce chapitre présente la définition d'une nouvelle mesure de performance, de même que la formulation d'un modèle intégré pour le redéploiement et la préaffectation des véhicules ambulanciers, menant à une contribution importante au niveau théorique. Dans un deuxième temps, d'un point de vue pratique, ce chapitre fournit une analyse détaillée du modèle proposé et une discussion sur le compromis entre les performances du système en utilisant une telle stratégie et les inconvénients qui y sont associés.

5.1 Le problème de redéploiement et de préaffectation des véhicules ambulanciers (PRPA)

Le problème de redéploiement et de préaffectation des véhicules ambulanciers (PRPA) consiste à déterminer la localisation des véhicules disponibles pour répondre à des appels urgents de même qu'une liste de préaffectation pour chaque zone de demande. La localisation des véhicules pourra engendrer ou non la relocalisation des véhicules en attente vers des postes d'attente différents de ceux qu'ils occupent au moment du redéploiement. Conséquemment, des coûts de redéploiement pourront être encourus. La liste de préaffectation comportera une liste ordonnée de véhicules à affecter aux demandes apparaissant dans une zone donnée. Ainsi, si un incident survient dans une zone, le premier véhicule sur la liste sera envoyé sur les lieux de l'incident si celui-ci est disponible. Dans le cas contraire, le deuxième sera envoyé, et ainsi de suite. Si aucun véhicule n'est disponible sur la liste, le véhicule disponible le plus proche sera envoyé sur les lieux de l'incident. Si aucun véhicule n'est disponible, l'appel est placé en file d'attente ou redirigé vers un autre service. Le PRPA vise alors à déterminer la localisation des véhicules et les listes de préaffectation des zones de manière à maximiser les performances du système tout en limitant les inconvénients liés au redéploiement. La Figure 5.1 représente une solution réalisable pour une instance du PRPA où trois véhicules sont disponibles pour répondre aux appels et où deux véhicules sont inclus sur chaque liste de préaffectation.

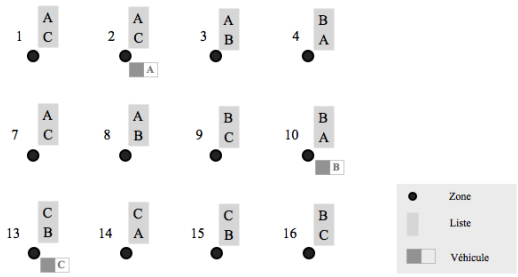


Figure 5.1 – Solution réalisable - Problème comportant trois véhicules et des listes de préaffectation de taille 2

Trois caractéristiques principales distinguent le PRPA des problèmes étudiés jusqu'à présent. Tout d'abord, le PRPA considère, en plus des décisions de redéploiement, l'établissement de listes de préaffectation ordonnées pour chaque zone de demande. Ces listes de préaffectation permettront de guider les décisions de répartition, principalement pour les demandes de priorité moindre, mais nécessitant toujours une intervention immédiate. Au meilleur de notre connaissance, seuls Goldberg *et al*. (1990b) ont considéré l'hypothèse d'une répartition des véhicules effectuée selon une liste de préférence dans le contexte du problème de localisation statique. De plus, l'établissement des listes de préaffectation permettra de considérer plus explicitement la capacité des véhicules. Cette idée a été introduite par Doerner *et al*. (2005) afin de limiter le nombre d'habitants affectés à un même véhicule lors de la formulation d'une version modifiée du MDS (Gendreau *et al*., 1997). Elle a été adaptée ici au contexte considéré dans cette étude. Enfin, le temps de réponse espéré, tel que défini dans cette section, sera utilisé afin de prédire les performances du système et de sélectionner la meilleure solution possible.

5.1.1 Les listes de préaffectation et la capacité des véhicules

Le PRPA considère explicitement l'établissement de listes de préaffectation pour chaque zone de demande. Le fait de considérer ainsi une liste de préaffectation permettra, lorsque le niveau

de priorité de l'appel le permet, d'envoyer sur les lieux d'un incident un véhicule différent du véhicule le plus proche, si les performances globales et/ou les contraintes du problème le justifient. L'affectation des zones de demande aux véhicules disponibles selon une liste ordonnée permettra aussi de prendre en compte plus explicitement leur capacité. Dans le contexte étudié, la capacité des véhicules est définie comme le nombre d'interventions qu'un véhicule peut servir sur une période de temps donnée ou comme le nombre maximal d'interventions qui pourront lui être affectés sur la période considérée. Le nombre espéré d'interventions servies par un véhicule dépend, quant à lui, du taux d'occupation, c'est-à-dire la probabilité pour qu'un véhicule soit occupé au moment où un appel est placé, du nombre espéré d'interventions dans une zone pour la période considérée et de l'affectation des zones de demande aux véhicules. Le nombre espéré d'interventions servies par un véhicule est donc limité par sa capacité.

5.1.2 Le temps de réponse pour prédire les performances du système

Afin de mesurer et de prédire les performances du système, le temps de réponse est considéré ici. Tel que discuté au *Chapitre 3*, le temps de réponse est défini comme le temps écoulé entre la réception d'un appel et l'arrivée d'un véhicule sur les lieux de l'incident. Bien que peu utilisée explicitement dans les modèles mathématiques formulés jusqu'à maintenant pour la localisation des véhicules ambulanciers, cette mesure reflète pourtant une préoccupation claire des organisations offrant des SPU. En effet, plusieurs d'entre elles, notamment Urgences-santé (2013), utilisent maintenant le temps de réponse moyen afin d'évaluer les performances de leur service plutôt que le pourcentage d'appels servis à l'intérieur d'un délai prédéfini utilisé précédemment. Cette mesure offre donc, à notre avis, une bonne évaluation des performances espérées des SPU et une différenciation adéquate des solutions potentielles, tout en ayant une signification concrète tant pour les utilisateurs du système que leurs gestionnaires.

Le temps de réponse pour une demande dépend du véhicule affecté à cet appel et de sa localisation. Afin de mesurer et de prédire le temps de réponse, il est donc nécessaire de connaître à la fois la localisation des véhicules et les décisions de répartition, c'est-à-dire quel véhicule sera affecté à un appel. Un autre facteur important qui viendra influencer la sélection du véhicule pour servir une demande, et par conséquent, le temps de réponse, est le taux d'occupation des véhicules. En effet, le premier véhicule sur la liste sera affecté à un appel survenant dans une zone donnée que s'il est libre au moment où l'appel est reçu, sinon le deuxième sera envoyé s'il est libre, et ainsi de suite. La distance entre chaque véhicule disponible et la zone de demande concernée, le taux d'occupation des véhicules de même que la liste de préaffectation pour la zone en question permettront de calculer le temps de réponse espéré. Le temps de réponse es-

péré suit la même logique que la couverture espérée telle que définie par Daskin (1983) pour le MECRP. La contribution de chaque véhicule dans le calcul du temps de réponse espéré dépendra de son taux d'occupation. Ainsi, considérer le fait que le véhicule le plus proche n'est pas toujours disponible lors du calcul du temps de réponse espéré permet, à notre avis, d'offrir une meilleure estimation. Dans le cas où la capacité des véhicules est illimitée, le fait de considérer le temps de réponse espéré comme une mesure de la performance aura pour effet d'ordonner les véhicules dans les listes de préaffectation selon un ordre de distance croissant. Par contre, lorsque la capacité du système est limitée, les listes de préaffectation s'ajustent pour tenir compte de la capacité « réelle » des véhicules à servir des demandes. Néanmoins, dans tous les cas, l'impact de l'affectation éventuelle des véhicules aux appels sur les décisions de localisation/relocalisation est considéré.

Les objectifs en lien avec le temps de réponse pourront se décliner sous différentes formes, par exemple la minimisation du temps de réponse moyen ou la minimisation du temps de réponse pour la zone de demande la moins bien servie. Dans le premier cas, on cherchera à garantir un bon niveau de performance global, tandis que dans le deuxième, on cherchera plutôt l'équité à travers les zones de demande. D'autres objectifs pourront également être pris en compte de manière à limiter les coûts ou les inconvénients liés au redéploiement. Des objectifs tels que la minimisation de la distance ou du temps total de redéploiement ou encore la minimisation de la distance ou du temps maximal de redéploiement, pour n'en nommer que quelques-uns, pourront alors être considérés. Le choix du ou des objectifs considérés et/ou la pondération qui leur est accordée dépendra alors du contexte d'application.

5.2 Modélisation du PRPA

Le problème de redéploiement et de préaffectation des véhicules ambulanciers (PRPA) se définit sur un graphe $G = (V,E)$ où $V = I \cup J$, $I = \{v_1, ..., v_n\}$ et $J = \{v_{n+1}, ..., v_{n+m}\}$ représentent respectivement l'ensemble des zones de demande et des postes d'attente potentiels, et où $E = \{(v_i, v_j) : v_i, v_j \in V\}$ est l'ensemble des arêtes reliant les sommets du graphe. Une zone de demande se définit comme un regroupement de population représenté par son centroïde et possédant une densité de population ou une demande donnée. Un site ou poste d'attente potentiel se définit plutôt comme un lieu physique où pourront être localisés un ou plusieurs véhicules et à partir duquel les véhicules se déplaceront pour rejoindre les différents appels de détresse. Ainsi, à chaque arête $(v_i, v_j) \in E$ est associé un temps de déplacement t_{ij} et à chaque sommet $v_i \in I$ est associée une densité de population ou une demande d_i. De plus, à chaque sommet $v_j \in J$ est associé un nombre maximal de véhicules p_j pouvant y être placés en attente.

On définira également K, l'ensemble des véhicules disponibles, et Z, l'ensemble des positions considérées sur la liste de préaffectation, $|Z|$ représentant la taille de la liste de préaffectation. Ainsi, à chaque véhicule $k \in K$ sont associés un paramètre λ_j^k qui vaut 1 si le véhicule k est localisé en $v_j \in J$ avant la relocalisation, et 0 autrement, un paramètre θ^k qui vaut 1 si le véhicule k est considéré dans le calcul des coûts de relocalisation, et 0 autrement, et une pénalité $M_{j_1 j_2}^k$ associée au déplacement du véhicule k de v_{j_1}en v_{j_2}. On considérera également q, le taux d'occupation des véhicules, et W, la capacité ou le nombre maximal d'interventions qu'un véhicule peut servir sur une période donnée. Enfin, ξ_s et ξ_r représentent les poids ou l'importance relative accordée respectivement aux objectifs liés au service à la population et aux inconvénients associés au redéploiement des véhicules lors de la prise de décision.
Les hypothèses suivantes sont également prises en compte lors de la formulation du PRPA :

1. Les décisions de répartition sont prises en fonction des listes de préaffectation établies pour chaque zone de demande, et ce, pour tous les appels dont le niveau de priorité le permet. Pour les autres, le véhicule le plus proche est envoyé sur les lieux de l'incident.

2. Si aucun véhicule n'est disponible sur la liste de préaffectation d'une zone, le véhicule disponible le plus proche est envoyé pour répondre à l'appel.

3. Si aucun véhicule n'est disponible pour répondre à l'appel, ce dernier est placé en file d'attente ou redirigé vers un autre service. Le temps de réponse est alors T, une valeur arbitraire que l'on fixera.

4. Les décisions de localisation et de préaffectation sont prises pour un horizon de planification donné, qui dépendra du contexte.

5. Le nombre de véhicules disponibles pour l'horizon de planification considéré est connu.

6. Tous les véhicules ont la même capacité en termes de charge de travail.

7. Tous les véhicules ont le même taux d'occupation.

8. Le nombre espéré d'interventions pour une zone pour la période considérée est connu.

Avant d'aller plus loin, il importe de discuter brièvement les hypothèses *6* et *7*. À notre avis, l'hypothèse *6* stipulant que chaque véhicule dispose de la même capacité à servir des interventions semble raisonnable comme en pratique, chaque équipe de travail possède la même capacité, une équipe étant associée à chaque véhicule. L'hypothèse *7* mérite toutefois d'être discutée davantage. En effet, un taux d'occupation global se justifie ici de deux manières. Tout

d'abord, le modèle proposé pour la localisation/relocalisation et l'établissement des listes de préaffectation considère explicitement le fait que chaque véhicule a une capacité limitée à servir des demandes. L'ajout d'une telle contrainte vise à limiter le nombre de demandes potentielles affectées à un véhicule et éventuellement à mieux répartir les demandes prévues entre les véhicules. Cela devrait mener à un meilleur équilibre de la charge de travail, et donc, justifier l'hypothèse selon laquelle le taux d'occupation est le même pour tous les véhicules. De plus, cette hypothèse a été posée dans plusieurs travaux précédents, notamment les travaux de Daskin (1982, 1983). Différentes méthodes ont également été proposées afin de relaxer cette hypothèse, se basant sur des facteurs correctifs ou des méthodes basées sur la théorie de files d'attente (Batta *et al.*, 1989). Les modèles qui en découlent sont donc, en général, plus difficiles à traiter. Dans le cas présent, nous souhaitons analyser plus en détail l'impact des listes de préaffectation et de la capacité des véhicules sur les solutions obtenues et les performances espérées du système. Nous pensons donc que cette hypothèse est tout à fait raisonnable pour une première analyse.

5.2.1 Décisions

Le PRPA considère deux types de décisions : des décisions de localisation qui attribuent un poste d'attente à chaque véhicule disponible (c'est-à-dire en attente, nouvellement libéré, en début de quart de travail, en retour de pause) et des décisions de préaffectation, qui affectent une liste ordonnée de véhicules à chaque zone de demande. Afin de considérer adéquatement ces différentes décisions, trois groupes de variables ont été définis soit : $x^k_{j_1 j_2}$, une variable binaire qui vaut 1 si le véhicule k se déplace du site $v_{j_1} \in J$ vers le site $v_{j_2} \in J$ à la suite du redéploiement/relocalisation, et 0 autrement (lorsque $j_1 = j_2$, aucune relocalisation n'est requise) ; w_i^{zk} un variable binaire qui vaut 1 si le véhicule k est le z-ième appelé sur la liste de préaffectation de la zone de demande située en $v_i \in I$, et 0 autrement ; et y_{ij}^{zk}, une variable binaire qui vaut 1 si le véhicule k, localisé en $v_j \in J$ à la suite de la relocalisation, est à la z-ième position sur la liste de préaffectation de la zone de demande située en $v_i \in I$, et 0 autrement.

5.2.2 Objectifs

Deux grandes familles d'objectifs sont prises en compte lors de la formulation du PRPA. La première famille d'objectif vise la maximisation des performances du système, celles-ci étant définies en fonction du temps de réponse espéré. En considérant les listes de préaffectation, le calcul du temps de réponse espéré pour une zone de demande donnée se divise en trois parties :

1. La contribution des véhicules sur la liste de préaffectation de la zone de demande en question.

2. La contribution des véhicules disponibles pour répondre aux appels de détresse, mais qui ne se retrouvent pas sur la liste de la zone de demande concernée.

3. Le temps de réponse associé aux appels en provenance de la zone de demande, mais qui ont dû être placés en file d'attente ou référés à un autre service puisqu'aucun véhicule n'était disponible pour répondre à un appel au moment où il a été reçu.

En définissant K_Z^i, l'ensemble des véhicules sur la liste de préaffectation de la zone $v_i \in I$, classés selon la liste de préaffectation où z_k est la position du véhicule k sur la liste, et K_A^i, l'ensemble des véhicules qui ne se retrouvent pas sur la liste de préaffectation de la zone i, classés en ordre croissant de distance t_{ki} où a_k est la position du véhicule k sur la liste ordonnée, le temps de réponse espéré pour les appels en provenance de la zone de demande $v_i \in I$ se formule de la manière suivante :

$$\sum_{k \in K_z^i} (1-q) q^{z_k - 1} t_{ki} + \sum_{k \in K_a^i} (1-q) q^{|Z| + a_k - 1} t_{ki} + q^{|K|} T. \tag{5.1}$$

Lors de la formulation du PRPA, seule la contribution des véhicules sur les listes de préaffectation (1) est considérée. Les termes (2) et (3) pourront être calculés ensuite en fonction des décisions et des caractéristiques du système.

Le temps de réponse espéré peut être utilisé afin de formuler différents objectifs. La minimisation du temps de réponse moyen ou total permettra de garantir de bonnes performances globales. En utilisant une telle mesure, il est possible que certaines zones soient très bien servies alors que d'autres soient servies très pauvrement. Afin d'améliorer l'équité entre les différentes zones de demande, d'autres objectifs pourraient être considérés tels que la minimisation du temps de réponse espéré pour la zone de demande la moins bien servie. La deuxième famille d'objectifs que nous considérerons s'intéresse plutôt à la minimisation des inconvénients liés au redéploiement. Ainsi, le nombre de véhicules redéployés, la distance/temps total ou maximal de redéploiement ou le mouvement de véhicules particuliers pourront être minimisés. Au sein du PRPA, ces deux familles d'objectifs sont considérées pour la formulation d'un objectif combiné où ξ_s et ξ_r sont respectivement les poids accordés aux objectifs en lien avec le service à la population et aux inconvénients de redéploiement. En faisant varier les valeurs de ξ_s et ξ_r, il sera possible de représenter différentes situations où une importance plus ou moins grande est accordée aux coûts de redéploiement.

Le Tableau 5.1 résume les objectifs possibles pour le PRPA, de même que leur formulation et les contraintes qui devront être ajoutées afin de les considérer adéquatement. La formulation du problème présentée à la sous-section suivante considère la minimisation du temps de réponse total pour répondre à tous les appels, puis la minimisation des temps totaux nécessaires au redéploiement. Puisque ces objectifs sont d'abord considérés lors de la phase d'expérimentation, ils ont été choisis afin de présenter la formulation initiale du modèle. D'autres objectifs en lien avec le service à la population sont aussi évalués lors de la phase d'expérimentation afin d'illustrer les différentes solutions obtenues en fonction des objectifs choisis.

5.2.3 Formulation du PRPA

$$\min \xi_s \sum_{i\in I}\sum_{z\in Z}\sum_{k\in K}\sum_{j\in J}(1-q)q^{z-1}d_i t_{ji} y_{ij}^{zk} + \xi_r \sum_{k\in K}\sum_{j_1\in J}\sum_{j_2\in J}\theta^k t_{j_1 j_2} x_{j_1 j_2}^k \tag{5.2}$$

sous les contraintes :

$$\sum_{j_1\in J}\sum_{j_2\in J}\lambda_{kj_1} x_{j_1 j_2}^k = 1,\ \forall k \in K, \tag{5.3}$$

$$\sum_{j_1\in J}\sum_{j_2\in J}(1-\lambda_{kj_1}) x_{j_1 j_2}^k = 0,\ \forall k \in K, \tag{5.4}$$

$$\sum_{k\in K}\sum_{j_1\in J} x_{j_1 j_2}^k \leq p_j,\ \forall j_2 \in J, \tag{5.5}$$

$$\sum_{z\in Z}\sum_{i\in I}(1-q)q^{z-1}d_i w_i^{zk} \leq W,\ \forall k \in K, \tag{5.6}$$

$$\sum_{z\in Z} w_i^{zk} \leq 1,\ \forall k \in K,\ \forall i \in I, \tag{5.7}$$

$$\sum_{k\in K} w_i^{zk} = 1,\ \forall z \in Z,\ \forall i \in I, \tag{5.8}$$

$$y_{ij_2}^{zk} \leq \sum_{j_1\in J} x_{j_1 j_2}^k,\ \forall z \in Z,\ \forall i \in I,\ \forall k \in K,\ \forall j_2 \in J, \tag{5.9}$$

$$w_i^{zk} = \sum_{j\in J} y_{ij}^{zk}, \forall i \in I, \forall z \in Z, \forall k \in K. \tag{5.10}$$

$$x_{j_1 j_2}^k \in \{0,1\},\ \forall j_1 \in J,\ \forall j_2 \in J,\ \forall k \in K, \tag{5.11}$$

$$w_i^{zk} \in \{0,1\},\ \forall z \in Z,\ \forall i \in I,\ \forall k \in K, \tag{5.12}$$

$$y_{ij}^{zk} \in \{0,1\},\ \forall z \in Z,\ \forall i \in I,\ \forall k \in K,\ \forall j \in J. \tag{5.13}$$

La fonction objectif du PRPA présentée ci-haut consiste à minimiser les temps de réponse totaux, puis les temps nécessaires au redéploiement (5.2). Les contraintes (5.3) et (5.4) visent à assurer que chaque véhicule disponible soit localisé à un poste d'attente, tandis que les contraintes (5.5) garantissent que le nombre maximal de véhicules pouvant être localisés à un poste d'attente donné soit respecté. Les contraintes (5.6) assurent, quant à elles, la satisfaction de la capacité des véhicules à traiter des demandes ou, autrement dit, assurent que le nombre espéré de demandes affectées à un véhicule soit inférieur ou égal à sa capacité. Le nombre espéré de demandes affectées à un véhicule dépendra du taux d'occupation des véhicules et de sa présence sur les listes de préaffectation des zones de demande du territoire à desservir. Ainsi, si le premier véhicule sur la liste de préaffectation de la zone est libre au moment où un appel est placé, soit dans une proportion de $1-q$ des cas, il sera envoyé sur les lieux de l'incident. Si le premier véhicule est occupé, soit la proportion des cas restants q, le deuxième véhicule sur la liste de préaffectation de la zone sera envoyé sur les lieux de l'incident, s'il est disponible, et ainsi de suite. Le modèle considère aussi qu'un véhicule ne pourra occuper plus d'une position sur la liste de préaffectation d'une zone de demande donnée (5.7) et qu'exactement un véhicule devra se retrouver à chaque position de la liste de préaffectation d'une zone de demande (5.8). Enfin, les contraintes (5.9) et (5.10) assurent le lien entre les variables de décision, tandis que les contraintes (5.11) à (5.13) assurent leur intégralité.

Il est possible d'observer que si la valeur de ξ_r est fixée à 0, la position actuelle des véhicules ne viendra pas influencer les décisions. En effet, dans ce cas, les coûts de relocalisation ne sont pas pris en compte. Plus la valeur de ξ_r prendra de l'importance par rapport à la valeur de ξ_s, plus la solution sera influencée par la position actuelle des véhicules, et donc, par les coûts de relocalisation.

Objectifs de performance	**Objectifs associés aux coûts/inconvénients**
Minimiser le temps de réponse maximal (T_{max})	*Minimiser le nombre de redéploiement*
$\min T_{max}$	$\min \sum_{k\in K}\sum_{j_1\in J}\sum_{j_2\in J, j_2\neq j_1} \theta^k x^k_{j_1 j_2}$
Ajouter la contrainte au modèle :	*Minimiser la distance ou le temps*
$\sum_{z\in Z}\sum_{k\in K}\sum_{j\in J}(1-q)q^{z-1}t_{ji}y^{zk}_{ij} \leq T_{max},\ \forall i \in I$	*de relocalisation maximum (D_{max})*
	$\min D_{max}$
	Ajouter la contrainte au modèle :
	$\sum_{j_1\in J}\sum_{j_2\in J} t_{j_1 j_2} x^k_{j_1 j_2} \leq D_{max},\ \forall k \in K$
	Minimiser le mouvement pondéré des véhicules
	$\min \sum_{j_1\in J}\sum_{j_2\in J, j_2\neq j_1} M^k_{j_1 j_2} x^k_{j_1 j_2}$

Tableau 5.1 – Objectifs possibles pour le PRPA

5.3 Expérimentation

Une série d'expérimentations a été effectuée afin d'analyser le modèle proposé pour le redéploiement et la préaffectation des véhicules ambulanciers. Plus précisément, les résultats obtenus lors de la phase d'expérimentation permettront d'analyser l'impact des décisions de préaffectation sur les décisions de relocalisation, de même que celui de considérer la capacité des véhicules sur les solutions obtenues et les performances prédites du système, et ce, pour un ensemble varié de paramètres.

Deux groupes d'instances ont été générés afin de valider et d'analyser le modèle. Le premier groupe d'instances a été généré aléatoirement à partir d'un territoire structuré comportant un nombre réduit de zones de demandes. Ce groupe d'instances permettra de valider la cohérence des solutions fournies par le modèle, de même que d'amener des pistes de réponse aux questions soulevées en introduction. Le deuxième groupe d'instances a été généré de façon à refléter un cas réel, c'est-à-dire le même que celui considéré au *Chapitre* 4. Il sera alors possible de vérifier comment un tel modèle se comporte lorsque confronté à une situation plus réaliste.

Les instances appartenant au premier groupe ont été générées aléatoirement ou déterminées arbitrairement de manière à représenter un problème structuré de taille réduite. Les paramètres permettant de décrire un tel contexte se définissent comme suit. Tout d'abord, le territoire à desservir, comportant 24 zones de demande, a été défini sur un rectangle 16 km $\times$ 20 km (voir Figure 5.2). À chaque zone de demande est associée une demande d_i qui varie selon deux cas. Dans le premier cas, *Cas U*, la demande associée à chaque zone suit une distribution uniforme de borne inférieure 1 et de borne supérieure 100. Ce cas permettra de représenter un contexte urbain avec des zones à plus forte et à plus faible densité de population. Dans le deuxième cas, *Cas R*, la demande est constante pour toutes les zones de demande, soit $d_i = 50$, ce qui permettra de représenter un contexte rural où la demande est répartie plus uniformément sur le territoire à desservir. Chaque zone de demande comporte un poste d'attente potentiel de capacité 1. Le nombre de véhicules disponibles pour répondre aux appels variant entre 3 et 4 et leur capacité est fixée à 425 interventions. Cette valeur a été déterminée en tenant compte du profil de la demande générée. Le taux d'occupation global considéré est de $0,4$. La matrice de temps de déplacement a été définie à partir des distances euclidiennes entre deux localisations et d'une vitesse constante. Le temps de déplacement entre deux points est indiqué sur le Figure 5.2 en secondes.

Les instances, telles que décrites ci-haut, constituent le cas de base des instances générées aléatoirement. Tout au long de l'expérimentation, différents paramètres sont ensuite modifiés de manière à effectuer une analyse détaillée de l'impact de la valeur de la capacité des véhicules,

de la taille des listes de préaffectation et du taux d'occupation sur les solutions obtenues et les performances espérées du système. L'utilisation de différentes fonctions objectif sera également évaluée.

Les instances pseudo-réelles ont été générées à partir du contexte développé dans le cadre du *Chapitre 4*, de même que d'une analyse rigoureuse des résultats obtenus par simulation. Les instances pseudo-réelles se divisent en deux groupes correspondant différentes tailles de problèmes : réduite (30 zones de demande) et moyenne (149 zones de demande). Ces instances ont été définies à partir d'un sous-ensemble du territoire des villes de Montréal et de Laval. Tel qu'il a été présenté au *Chapitre 4*, ce territoire a été défini de manière à représenter adéquatement les difficultés liées à la gestion d'un SPU en contexte urbain, ce qui n'était pas le cas pour les instances aléatoires. Ces instances visent donc à analyser le modèle dans un contexte plus réaliste. Le territoire utilisé pour la génération des instances est représenté à la Figure 5.3. Chaque point sur cette figure représente le centroïde d'une zone de demande, la taille du point indiquant l'importance relative d'une densité de population par rapport aux autres zones du territoire à desservir.

À chaque zone de demande correspond une densité de population de même qu'une demande espérée. La demande espérée pour une zone de demande est définie comme le nombre d'interventions prévues ou contre lesquelles on souhaite se prémunir pour une période donnée. Naturellement, la demande prévue et la capacité des véhicules sont influencées par la longueur de la période choisie. Dans le cas de cette étude, la demande prévue pour une période donnée a été tirée à partir des données simulées. Ainsi, pour chaque cas de figure (réduit, moyen et réel), une liste d'appels a été générée tel que décrit au *Chapitre 4*. La liste des appels générés a ensuite été analysée plus en détail afin de déterminer le profil de demande pour chaque zone considérée. Une journée a alors été divisée en 4 périodes de 6 heures, correspondant respectivement aux périodes du matin (6 h à 12 h), de l'après-midi (12 h à 18 h), du soir (18 h à 0 h) et de la nuit (0 h à 6 h). En effet, il a été noté que l'intensité des demandes variait généralement pour chacune de ces périodes. Cela mènera à la définition de 4 scénarios différents en termes de profil de demande. Lorsque disponibles, des données historiques pourront également être analysées afin de déterminer le nombre d'interventions prévues par période pour chaque zone de demande.

L'ensemble des postes d'attente potentiels où les véhicules pourront être placés varie, quant à lui, de 30 à 48 postes d'attente répartis sur le territoire à desservir, incluant les postes d'attente choisis arbitrairement, les centres hospitaliers et les centres opérationnels. Pour l'instance de petite taille, chaque zone de demande inclut un poste d'attente potentiel. Pour l'instance

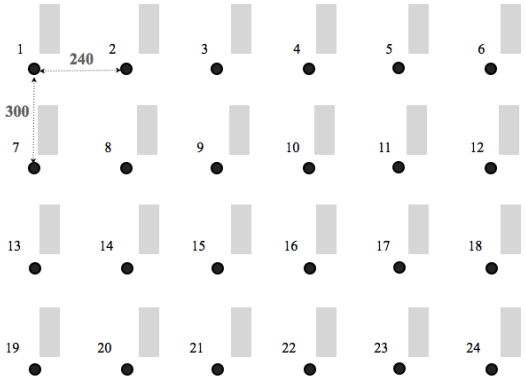


Figure 5.2 – Territoire à desservir - Instances aléatoires

Figure 5.3 – Territoire à desservir - Instances pseudo-réelles

de moyenne taille, 48 zones sont sélectionnés de manière aléatoire pour devenir candidate à recevoir un véhicule en attente. Il est important de constater que, pour les instances de taille moyenne, le nombre de localisations potentielles est plus petit que le nombre de zones de demande. Ces valeurs, fixées arbitrairement, ont été choisies afin de limiter le temps de calcul tout en demeurant tout à fait réaliste. La capacité de ces localisations a été fixée à 1. La matrice comportant les temps de déplacement entre les différentes localisations sur le territoire à desservir a été extraite de la matrice comportant les temps « réels » entre les différents centroïdes des zones.

Selon la taille du problème, le nombre de véhicules ou d'équipes de travail disponibles variera de 2 à 14 pour la période considérée, correspondant à différents scénarios allant de 7,5 à 20 zones par véhicule. Lorsqu'il sera pertinent de le faire, la localisation d'un véhicule avant la relocalistion sera déterminée de trois manières : soit aléatoirement, soit à partir d'un sous-ensemble des postes d'attente optimaux ou quasi-optimaux pour un nombre de véhicules plus élevé, soit une combinaison des deux. La capacité des véhicules a été fixée à 5 interventions par période, en considérant un taux d'occupation de 0,5. Ces deux valeurs ont été déterminées à la suite de l'analyse des données générées et des résultats obtenus par simulation.

Le Tableau 5.2 résume l'ensemble des paramètres importants pour chaque groupe d'instances. Les valeurs entre crochets correspondent aux valeurs utilisées afin d'analyser la sensibilité du modèle par rapport aux différents paramètres. L'analyse du modèle proposé pour le redéploiement et la préaffectation des véhicules ambulanciers comporte deux étapes. Dans un premier temps, le modèle est analysé plus en détail, grâce aux instances de taille réduite (U, R et R30). Ainsi, il sera possible d'évaluer :

- L'impact des listes de préaffectation sur les décisions de localisation ;
- L'impact de la capacité du système sur les décisions de localisation et de préaffectation ;
- La sensibilité par rapport à la valeur du taux d'occupation ;
- La sensibilité par rapport au profil de la demande ;
- L'impact de la fonction objectif sur les décisions de localisation.

Dans les quatre premiers cas, les coûts associés aux efforts de relocalisation ne sont pas pris en compte, c'est-à-dire $\xi_s = 1$ et $\xi_r = 0$. Les poids attribués à chaque famille d'objectifs sont ensuite analysés plus explicitement. Dans un deuxième temps, les résultats principaux obtenus pour les instances de taille réduite sont validés et illustrés grâce aux instances de taille réelle. Dans tous les cas, les résultats ont été obtenus à l'aide de CPLEX *12.5* en imposant un temps

limite de 28 800 secondes. Les tests ont été lancés à partir d'une machine Windows avec huit cœurs et 15 Go de RAM, installée sur un serveur équipé d'un processeur AMD Opteron 6328.

5.3.1 Résultats et analyse pour les instances de taille réduite (*U*, *R* et *R30*)

Afin de valider et d'analyser plus en détail les différentes caractéristiques du modèle, confronté à des situations variées en termes de territoire et de profil de demande, différents tests ont été menés grâce aux instances de taille réduite. Les résultats obtenus dans ce cas sont discutés dans la présente section. Chaque tableau de résultats présente d'abord les caractéristiques principales de l'instance et du scénario considéré, puis reporte les localisations obtenues à la suite de la résolution du modèle, le temps de réponse correspondant (TR), le temps de calcul (Temps) et l'écart entre la solution trouvée et la borne inférieure fournie par CPLEX (*GAP*). Afin d'assurer une comparaison adéquate des scénarios considérés, les trois composantes du temps de réponse telles que présentées à la section 6.2.2, ont aussi été calculées puis reportées dans les tableaux. À moins d'indication contraire, TR^1 est le temps de réponse total correspondant à la contribution des véhicules sur les listes de préaffectation, TR^{1+2}, le temps de réponse total correspondant à tous les véhicules disponibles, incluant les véhicules qui ne figurent pas sur les listes de préaffectation alors classés en fonction de la distance pour chaque zone de demande, et TR^{1+2+3}, le temps de réponse total pour le système en considérant la probabilité pour qu'un appel puisse être placé en file d'attente ou redirigé vers un autre service. Dans ce cas, $T = 360$ pour U et R et $T = 220$ pour $R30$. Pour les instances aléatoires, la valeur de T a été déterminée de manière arbitraire, tandis que pour l'instance pseudo-réelle, elle a été extraite des résultats des études de simulation présentés au *Chapitre* 4. Tous les temps de réponse sont reportés en secondes. Pour TR^{1+2+3}, le temps de réponse moyen par intervention est également donné, entre parenthèses, en secondes par intervention. Les temps de calcul et les GAP sont reportés respectivement en

	Instances			
	U	**R**	**R30**	**R149**
$\|I\|$	24		30	149
$\|J\|$	24		30	48
$\|K\|$	3 ou 4		2, 3 ou 4	Entre 8 et 14
$\|Z\|$	2 [1, 3, 4]			2 [1]
d_i	UNIF[1,100]	50	N, AM, PM, S	PM
q	0.4 [0.2, 0.3, 0.5, 0.6]		0.5 [0.3, 0.4, 0.6, 0.7]	0.5
W	425 [400, 500, 2400]		5 [2, 3, 4, 1000]	5
t_{ij}	Distance euclidienne et vitesse constante.		À partir des temps de déplacements « réels ».	

Tableau 5.2 – Paramètres par groupe d'instances

secondes et en pourcentage. Lorsque d'autres mesures ou caractéristiques sont utilisées, elles seront d'abord présentées dans la sous-section correspondante.

5.3.1.1 Impact des listes de préaffectation sur les décisions de localisation

Les résultats de l'expérimentation menée en considérant des listes de préaffectation de taille variée sont présentés ici, et ce, pour toutes les instances de taille réduite (U, R et R30). En analysant les résultats obtenus, il est possible d'évaluer dans quelle mesure le fait de considérer plusieurs véhicules au sein des listes de préaffectation a un impact sur les décisions de localisation, et par conséquent, sur les performances espérées du système. Puisque qu'exactement un véhicule devra être présent à chaque position sur la liste de préaffectation d'une zone donnée et qu'un véhicule ne peut occuper plus d'une position sur une liste, seuls les cas réalisables sont considérés, c'est-à-dire celles où $|K| \geq |Z|$.

	|Z|	**|K|**	**Localisation**	**TR**^{1} [s]	**TR**$^{1+2}$ [s]	**TR**$^{1+2+3}$ [s (s/int)]	**Temps** [s]	***GAP*** [%]
	4	4	[9, 11, 14, 16]	505 212	505 212	530 467 (397)	28 800	5,97
	3	3	[10, 14, 16]	531 012	531 012	595 801 (446)	6642	0,00
	3	4	[8, 11, 14, 16]	463 374	505 207	530 463 (397)	28 800	5,28
U	2	3	[11, 14, 16]	427 872	532 471	597 353 (447)	132	0,00
	2	4	[8, 11, 14, 17]	366 018	507 621	532 978 (399)	276	0,00
	1	3	[5, 14, 17]	242 640	573 534	641 044 (479)	144	0,00
	1	4	[5, 8, 16, 19]	211 428	546 160	572 564 (428)	143	0,00
	4	4	[8, 11, 14, 17]	476 652	476 652	501 176 (375)	28 800	18,27
	3	3	[8, 11, 16]	495 906	495 906	558 448 (418)	28 800	10,08
	3	4	[8, 11, 14, 17]	426 642	466 501	490 765 (367)	28 800	12,99
R	2	3	[8, 11, 15]	402 192	495 907	558 450 (417)	302	0,00
	2	4	[8, 11, 14, 17]	338 400	466 502	490 767 (367)	793	0,00
	1	3	[8, 11, 22]	227 880	515 102	578 873 (433)	2	0,00
	1	4	[2, 5, 14, 17]	200 160	539 759	565 899 (423)	2	0,00
	4	4	[2, 7, 11, 20]	929,64	929,64	1822 (271)	28 800	9,70
	3	3	[2, 7, 11]	914,14	914,14	1955 (291)	9963	0,00
	3	4	[2, 7, 11, 20]	813,61	929,63	1822 (271)	28 800	8,76
	2	2	[2, 11]	844,69	844,69	2168 (322)	27	0,00
R30	2	3	[2, 7, 20]	697,05	929,11	1972 (293)	460	0,00
	2	4	[2, 4, 11, 20]	608,57	929,38	1822 (271)	1731	0,00
	1	2	[2, 20]	423,29	885,22	2218 (330)	2	0,00
	1	3	[2, 6, 20]	350,76	948,26	1994 (297)	2	0,00
	1	4	[2, 6, 20, 30]	310,24	995,25	1892 (281)	2	0,00

Tableau 5.3 – Impact des listes de préaffectation

Les résultats présentés au Tableau 5.3 permettent d'émettre trois constats principaux. Tout d'abord, il est possible de constater que, pour tous les cas étudiés, les listes de préaffectation ont un impact important sur les décisions de localisation. Ce constat est particulièrement vrai lorsque l'on inclut deux véhicules sur les listes de préaffectation plutôt qu'un seul. L'impact devient moins important lorsque trois ou quatre véhicules sont considérés : l'ajout d'un véhicule supplémentaire sur la liste de préaffectation a une contribution moins importante au sein de la fonction objectif ce qui engendre une moins grande variation en termes de localisation des

véhicules. Ce constat confirme par le fait même la pertinence de la notion de double couverture proposée par plusieurs auteurs (Hogan et ReVelle, 1986; Gendreau *et al*., 1997), tout en permettant une distinction plus fine entre les différentes solutions. Il est également possible d'observer que les performances prédites du système s'améliorent aussi lorsque l'on considère les listes de préaffectation. Selon l'instance considérée, une amélioration allant jusqu'à 56 secondes du temps de réponse par intervention peut être notée, lorsque deux véhicules sont présents sur les listes plutôt qu'un seul. Cette réduction est toutefois moins notable lorsque l'on considère trois ou quatre véhicules sur les listes de préaffectation. Enfin, les temps de calcul augmentent de manière considérable avec la taille des listes. Pour cette raison, pour l'ensemble des tests subséquents, la taille des listes de préaffectation a été fixée à 2, ce qui représente à notre avis, un compromis intéressant entre l'amélioration des décisions et le temps de calcul.

5.3.1.2 Impact de la capacité du système sur les décisions de localisation et de préaffectation

La capacité du système varie en fonction de deux paramètres : le nombre de véhicules disponibles et la capacité des véhicules. Différentes combinaisons des valeurs de $|K|$ et de W ont donc été prises en compte afin de définir différents scénarios en termes de capacité du système. Ces scénarios permettront d'analyser l'impact de la capacité sur les décisions de localisation et de préaffectation, et ce, pour les trois cas de figure U, R et R30. Dans tous les cas, $W = 2400$ représente un système sans contrainte de capacité.

En analysant les résultats présentés au Tableau 5.4, il est possible de constater que, pour tous les cas étudiés, la capacité du système n'a pas d'impact réel sur les décisions de localisation lorsque l'on compare les localisations obtenues en considérant un même nombre de véhicules, mais pour différentes valeurs de la capacité. Lorsque le nombre de véhicules varie, mais que la capacité demeure la même, les décisions de localisation changent, c'est-à-dire les localisations obtenues en considérant $|K|$ véhicules n'est pas un sous-ensemble des localisations obtenues pour $|K| + 1$ véhicules. Cette observation se vérifie aussi au Tableau 5.3. Dans tous les cas, les performances s'améliorent lorsque le nombre de véhicules augmente. Enfin, bien que les localisations ne semblent pas influencées par la valeur de la capacité, il est possible de noter que lorsque la capacité des véhicules est limitée, les listes de préaffectation s'ajustent afin de prendre en compte la capacité « réelle » des véhicules à traiter des demandes urgentes, ce qui permet d'obtenir une meilleure estimation des performances du système.

	W	|K|	Localisation	TR^1 [s]	TR^{1+2} [s]	TR^{1+2+3} [s (s/int)]	Temps [s]	*GAP* [%]
	2400	3	[11, 14, 16]	427 512	532 111	596 970 (446)	277	0,00
	2400	4	[8, 11, 14, 17]	366 018	507 621	532 938 (399)	985	0,00
	500	3	[11, 14, 16]	427 512	532 111	596 970 (446)	213	0,00
U	500	4	[8, 11, 14, 17]	366 018	507 621	532 938 (399)	1148	0,00
	425	3	[11, 14, 16]	427 872	532 471	597 353 (447)	136	0,00
	425	4	[8, 11, 14, 17]	366 018	507 621	532 938 (399)	277	0,00
	400	3	[11, 14, 16]	430 554	534 790	599 821 (449)	273	0,00
	400	4	[8, 11, 14, 17]	366 018	507 621	532 938 (399)	1544	0,00
	2400	3	[10, 14, 17]	402 192	495 907	558 450 (418)	3601	0,00
	2400	4	[8, 11, 14, 17]	338 400	466 502	490 767 (367)	7018	0,00
	500	3	[10, 14, 17]	402 192	495 907	558 450 (418)	653	0,00
R	500	4	[8, 11, 14, 17]	338 400	466 502	490 767 (367)	2146	0,00
	425	3	[8, 11, 15]	402 192	495 907	558 450 (418)	302	0,00
	425	4	[8, 11, 14, 17]	338 400	466 502	490 767 (367)	793	0,00
	400	3	[10, 14, 17]	402 336	495 994	558 542 (418)	771	0,00
	400	4	[8, 11, 14, 17]	338 400	466 502	490 767 (367)	983	9,70
	2400	2	[2, 11]	844,69	844,69	2168 (322)	25	0,00
	2400	3	[2, 7, 20]	697,05	929,11	1972 (293)	462	0,00
	2400	4	[2, 4, 11, 20]	608,57	929,38	1822 (271)	1790	0,00
	5	2	[2, 11]	844,69	844,69	2168 (322)	27	0,00
	5	3	[2, 7, 20]	697,05	929,11	1972 (293)	460	0,00
	5	4	[2, 4, 11, 20]	608,57	929,38	1822 (271)	1731	0,00
R30	4	2	[2, 11]	844,69	844,69	2168 (322)	25	0,00
	4	3	[2, 7, 20]	697,05	929,11	1972 (293)	464	0,00
	4	4	[2, 4, 11, 20]	608,57	929,38	1822 (271)	1731	0,00
	3	2	[2, 11]	844,69	844,69	2168 (322)	29	0,00
	3	3	[2, 7, 20]	697,05	929,11	1972 (293)	2074	0,00
	3	4	[2, 4, 11, 20]	608,57	929,38	1822 (271)	793	0,00
	2	2	Non réalisable	-	-	-	-	-
	2	3	[2, 7, 20]	697,05	929,11	1972 (293)	336	0,00
	2	4	[2, 4, 11, 20]	608,57	929,38	1822 (271)	1146	0,00

Tableau 5.4 – Impact de la capacité du système

5.3.1.3 Sensibilité par rapport à la valeur du taux d'occupation

Le taux d'occupation est un paramètre intrinsèque du système et est considéré ici comme étant le même pour tous les véhicules. En pratique, il peut être difficile d'estimer avec exactitude la valeur de q. De plus, l'hypothèse d'un taux d'occupation global ne se vérifie pas toujours. Le taux d'occupation influence deux composantes principales du modèle : la fonction objectif et la contrainte de capacité. Mais qu'en est-il des décisions de localisation optimales ? Afin d'analyser la sensibilité des décisions de localisation par rapport au taux d'occupation, une série d'expérimentations a été menée en considérant différentes valeurs de q, et ce, pour les instances U, R et R30. Les résultats obtenus sont présentés au Tableau 5.5.

Les résultats obtenus permettent d'observer que le taux d'occupation n'influence pas fortement les décisions de localisation ou de relocalisation. Sa valeur influencera toutefois le temps de réponse associé à une solution donnée puisque q intervient dans le calcul de la fonction objectif. Ainsi, plus le taux d'occupation est grand, plus le temps de réponse est grand pour un nombre fixe de véhicules. En effet, pour de grandes valeurs de q, la probabilité que le premier véhicule

	q	|K|	Localisation	TR^1 [s]	TR^{1+2} [s]	TR^{1+2+3} [s (s/int)]	Temps [s]	*GAP* [%]
	0,6	3	[11, 14, 16]	343 716	530 614	749 192 (560)	551	0,00
	0,6	4	[8, 11, 14, 17]	294 552	540 019	672 384 (503)	1756	0,00
	0,5	3	[11, 14, 16]	392 952	529 148	655 457 (490)	373	0,00
	0,5	4	[8, 11, 14, 17]	336 606	535 334	598 875 (448)	837	0,00
U	0,4	3	[11, 14, 16]	427 872	532 471	597 353 (447)	132	0,00
	0,4	4	[8, 14, 15, 17]	366 018	507 621	532 938 (399)	276	0,00
	0,3	3	[11, 14, 16]	452 190	519 995	547 030 (409)	138	0,00
	0,3	4	[8, 11, 14, 17]	382 806	468 500	476 194 (356)	149	0,00
	0,2	3	Non réalisable	-	-	-	-	-
	0,2	4	[8, 11, 14, 17]	386 952	426 806	428 259 (320)	253	0,00
	0,6	3	[7, 10, 14]	323 712	464 285	672 215 (503)	13 559	0,00
	0,6	4	[8, 11, 14, 17]	271 680	545 760	670 518 (502)	5436	0,00
	0,5	3	[6, 8, 11]	369 900	491 925	612 255 (458)	3828	0,00
	0,5	4	[8, 11, 14, 17]	310 800	490 575	550 740 (412)	2626	0,00
R	0,4	3	[8, 11, 15]	402 192	495 907	557 516 (417)	302	0,00
	0,4	4	[8, 11, 14, 17]	338 400	466 502	491 146 (367)	793	9,70
	0,3	3	[9, 14, 17]	420 714	482 172	508 163 (380)	312	0,00
	0,3	4	[8, 11, 14, 17]	354 480	432 008	439 805 (329)	633	0,00
	0,2	3	[9, 14, 17]	425 184	456 403	464 104 (347)	93	0,00
	0,2	4	[8, 11, 14, 17]	359 040	395 098	396 638 (297)	110	0,00
	0,7	2	[2, 7]	604,96	604,96	2283 (340)	49	0,00
	0,7	3	[2, 7, 20]	497,93	770,83	2129 (317)	1285	0,00
	0,7	4	[2, 4, 11, 20]	430,66	863,40	1981 (295)	9892	0,00
	0,6	2	[2, 11]	741,81	741,81	2168 (322)	41	0,00
	0,6	3	[2, 7, 20]	610,77	878,10	1972 (293)	1718	0,00
	0,6	4	[2, 4, 11, 20]	530,54	927,27	1822 (271)	9376	0,00
R30	0,5	2	[2, 11]	844,69	844,69	2040 (303)	27	0,00
	0,5	3	[2, 7, 20]	697,05	929,11	1832 (272)	460	0,00
	0,5	4	[2, 4, 11, 20]	608,57	929,38	1695 (252)	1731	0,00
	0,4	2	[2, 11]	914,54	914,54	1912 (284)	16	0,00
	0,4	3	[2, 7, 20]	756,75	934,97	1713 (255)	153	0,00
	0,4	4	[2, 4, 11, 20]	664,77	893,04	1594 (237)	1683	0,00
	0,3	2	[2, 11]	951,36	951,36	2368 (352)	3	0,00
	0,3	3	[2, 7, 20]	789,89	906,84	2285 (340)	43	0,00
	0,3	4	[2, 6, 11, 20]	696,42	832,66	2168 (323)	551	0,00

Tableau 5.5 – Impact du taux d'occupation

sur la liste de préaffectation d'une zone de demande soit libre pour répondre à un appel est plus petite, ce qui a pour effet d'accroître le temps de réponse total. Inversement, un taux d'occupation plus petit correspond à une plus grande probabilité de trouver le premier véhicule sur la liste libre au moment où un appel est placé et, conséquemment, à un meilleur temps de réponse. Toutefois, du point de vue de la contrainte de capacité, cela signifie aussi qu'un véhicule aura à répondre plus fréquemment aux demandes en provenance des zones dont il est le premier sur la liste. La contrainte de capacité est donc plus « serrée » lorsque les valeurs de q sont plus petites, ce qui justifie le fait que le problème n'est pas réalisable dans le cas U où $q = 0.2$ et $|K| = 3$. La valeur de q viendra affecter indirectement la contrainte de capacité : les valeurs de q plus petites correspondant à des capacités moindres.

5.3.1.4 Sensibilité par rapport au profil de la demande

De manière plus générale, les décisions de localisation et de relocalisation sont influencées d'une part, par les caractéristiques géographiques du système (t_{ij}), et d'autre part, par l'intensité des demandes placées dans une zone donnée (d_i). Afin d'analyser la sensibilité des solutions par rapport au profil de la demande, une série d'expérimentations ont été menées en considérant différentes périodes au profil de demande varié, et ce, pour R30. Jusqu'à présent, tous les résultats présentés considéraient la même période, soit la période de l'après-midi, c'est-à-dire de 12 h 00 à 18 h 00 (PM). Le Tableau 5.6 reporte donc les résultats pour les autres périodes de la journée.

	Période	**|K|**	**Localisation**	**TR**^{1} [s]	**TR**$^{1+2}$ [s]	**TR**$^{1+2+3}$ [s (s/int)]	**Temps** [s]	***GAP*** [%]
	N	2	[2, 11]	423,31	423,31	1115 (315)	19	0,00
	N	3	[2, 7, 20]	345,66	463,73	1011 (285)	102	0,00
	N	4	[2, 4, 11, 20]	298,83	461,02	930 (262)	1248	0,00
	AM	2	[2, 11]	676,11	676,11	1747 (320)	35	0,00
	AM	3	[2, 7, 20]	556,83	738,97	1583 (290)	317	0,00
R30	AM	4	[2, 4, 11, 20]	484,35	738,78	1462 (263)	2411	0,00
	PM	2	[2, 11]	844,69	844,69	2167 (322)	27	0,00
	PM	3	[2, 7, 20]	697,05	929,11	1972 (293)	460	0,00
	PM	4	[2, 4, 11, 20]	608,57	929,38	1822 (271)	1731	0,00
	E	2	[2, 11]	732,77	732,77	1920 (316)	28	0,00
	E	3	[2, 7, 20]	602,48	806,07	1744 (287)	346	0,00
	E	4	[2, 4, 11, 20]	523,29	801,69	1605 (265)	1900	0,00

Tableau 5.6 – Impact du profil de la demande

En observant les résultats présentés au Tableau 5.6, il est possible de constater que, pour R30, la période de la journée n'influence pas les décisions de localisation et de relocalisation. La demande évolue donc au fil de la journée (elle diminue ou augmente) de manière similaire pour la plupart des zones de demande, ce qui ne justifie pas la modification des postes d'attente optimaux. Il est fort possible que les décisions de localisation ou de relocalisation puissent varier dans un contexte où les probabilités d'occurence varient beaucoup au cours de la journée. Ce pourrait être le cas, par exemple, lorsqu'il y a des déplacements importants de population. Néanmoins, à la vue des résultats présentés au Tableau 5.6, ce phénomène ne semble pas marqué dans le cas étudié.

5.3.1.5 Analyse de différentes variantes de la fonction objectif

Les résultats présentés jusqu'à maintenant ont été déterminés en ne considérant que la minimisation du temps de réponse total (ou de manière équivalente, du temps de réponse moyen). Aucune autre mesure de la qualité du service n'a été évaluée et aucun coût de relocalisation n'a été pris en compte. De cette manière, il a été possible d'analyser et de mieux comprendre les

caractéristiques du modèle, dans un même cadre de comparaison. Afin de compléter l'analyse du modèle proposé, il importe maintenant d'aller un peu plus loin dans l'analyse de la fonction objectif. C'est ce que nous ferons dans la présente sous-section.

Dans un premier temps, les objectifs en lien avec la qualité du service sont analysés ici. Tel que discuté précédemment, la minimisation du temps de réponse total permet d'assurer un bon service global, mais peut aussi mener à des zones de demande pauvrement desservies. Afin d'améliorer l'équité entre les zones de demande, il est possible de considérer, au sein de la fonction objectif, la minimisation du temps de réponse maximal, plutôt que celle du temps de réponse global. Cela aura pour effet d'améliorer le service pour la zone la moins bien servie, au détriment du temps de réponse total, mais dans quelle mesure ? Différents tests ont donc été menés afin de répondre à cette question et d'évaluer les différences entre les solutions obtenues en considérant ces deux objectifs, et ce, pour U, R et R30. De plus, une mesure alternative, soit la somme du temps de réponse moyen et du temps de réponse maximal, a également été évaluée. Cette nouvelle fonction objectif permettra d'obtenir un compromis entre les performances globales du système et ses pires performances. Le Tableau 5.7 présente les résultats obtenus en considérant les trois fonctions objectif : temps de réponse moyen (TR_{moy}), temps de réponse maximal (TR_{max}) ou la somme des deux (TR_{m+m}). Pour chaque scénario considéré, la valeur de la fonction objectif est présentée (en gras), puis les autres mesures, calculées en fonction de la solution trouvée, sont reportés. Dans tous les cas, les coûts associés aux effort de relocalisation ne sont pas considérés, c'est-à-dire que $\xi_r = 0$.

	Obj.	**|K|**	**Localisation**	$\mathbf{TR}^1_{moy}$ [s/int]	$\mathbf{TR}^1_{max}$ [s/int]	$\mathbf{TR}^1_{m+m}$ [s/int]	**Temps** [s]	***GAP*** [%]
U	TR_{moy}	3	[11, 14, 16]	**319,8**	613,2	933,6	132	0,00
	TR_{moy}	4	[8, 11, 14, 17]	**273,6**	385,9	659,7	276	0,00
	TR_{max}	3	[4, 13, 22]	409,2	**532,8**	942,0	792	0,00
	TR_{max}	4	[5, 8, 14, 23]	316,2	**385,9**	700,2	1851	0,00
	TR_{m+m}	3	[10, 13, 16]	328,8	547,2	**876,0**	28 880	2,53
	TR_{m+m}	4	[8, 11, 14, 17]	273,6	385,	**659,7**	730	0,00
R	TR_{moy}	3	[8, 11, 15]	**335,4**	576,0	911,2	302	0,00
	TR_{moy}	4	[8, 11, 14 , 17]	**282,0**	385,9	667,9	793	0,00
	TR_{max}	3	[3, 8, 20]	411,36	**532,8**	944,2	3619	0,00
	TR_{max}	4	[2, 5, 20, 23]	340,6	**385,9**	726,5	272	0,00
	TR_{m+m}	3	[7, 10, 16]	344,7	547,2	**891,9**	4364	0,00
	TR_{m+m}	4	[8, 11, 14, 17]	282,0	385,9	**667,9**	1246	0,00
R30	TR_{moy}	2	[2, 11]	**125,7**	261,0	386,7	27	0,00
	TR_{moy}	3	[2, 7, 20]	**103,7**	288,8	345,4	460	0,00
	TR_{moy}	4	[2, 4, 11, 20]	**90,5**	229,0	319,5	1731	0,00
	TR_{max}	2	[7, 30]	156,6	**212,3**	368,9	23	0,00
	TR_{max}	3	[1, 3, 9]	146,6	**171,0**	317,6	792	0,00
	TR_{max}	4	[7, 9, 28, 29]	121,6	**153,0**	274,6	3604	0,00
	TR_{m+m}	2	[7, 30]	152,1	212,3	**363,3**	37	0,00
	TR_{m+m}	3	[4, 10, 29]	122,7	172,8	**295,5**	1064	0,00
	TR_{m+m}	4	[2, 7, 20, 29]	96,6	159,0	**255,6**	11 665	0,00

Tableau 5.7 – Impact du choix de la fonction objectif

Les résultats présentés au Tableau 5.7 confirment que le choix de la fonction objectif influence fortement les localisations optimales, et par le fait même, les performances globales et maximales obtenues. Naturellement, le choix de minimiser les performances moyennes amène une dégradation des pires performances, et inversement. Lorsque l'on considère la minimisation du temps de réponse moyen, le temps de réponse maximal correspondant dépasse de 0 % à 79 % le temps de réponse maximal optimal. Lorsque le temps de réponse maximal est plutôt optimisé, les performances moyennes se dégradent de 20 % à 34 % selon l'instance et le scénario considéré, ce qui demeure non négligeable. Le fait de considérer une combinaison des deux mesures permet d'obtenir des solutions intermédiaires qui se rapprochent davantage des deux objectifs : on parle alors d'une dégradation de 0 % à 21 % pour le temps de réponse moyen et de 0 % à 5 % pour le temps de réponse maximal.

Dans un deuxième temps, l'impact des coûts liés aux efforts de relocalisation sur les décisions de localisation sont analysés plus en détail. Pour ce faire, différents scénarios ont été crées en considérant un ensemble de valeurs variées pour ξ_s et ξ_r. Pour chaque cas, les valeurs de ξ_s sont d'abord fixées à 1, 0,75, 0,5, 0,25 et 0, puis les valeurs de ξ_r sont fixées de sorte que $\xi_s + \xi_r = 1$. À la suite d'une expérimentation préliminaire, nous avons choisi d'inclure aussi le cas $\xi_s = 0,875$ et $\xi_r = 0,125$ afin de fournir une meilleure analyse du compromis entre le temps de réponse et le temps total de relocalisation. Cette analyse est effectuée ici pour l'instance R30, en prenant en compte différents plans de localisation initiale, c'est-à-dire avant le redéploiement. En effet, la localisation initiale des véhicules aura un impact important sur les solutions finales. Afin de représenter un éventail varié de situations, pour chaque scénario en termes de nombre de véhicules, différents plans de localisation initiale ont été pris en compte. Ces plans de localisation ont été déterminés de trois manières : en sélectionnant la localisation initiale des véhicules parmi les localisation optimales dans le cas où le nombre de véhicules disponibles est égal à $|K|+1$, en sélectionnant les localisations initiales de manière aléatoire, ou grâce à une combinaison des deux. Dans tous les cas, les localisations appartenant à l'ensemble des localisations initiales pour $|K|+1$ sont identifiées en gras. Les tableaux 5.8, 5.9 et 5.10 présentent les résultats obtenus respectivement dans les cas où 2, 3 ou 4 véhicules sont disponibles pour répondre aux appels d'urgence, et ce, pour différentes valeurs de ξ_s et ξ_r. Dans chaque tableau, la localisation initiale des véhicules, les valeurs de ξ_s et ξ_r employées de même que les solutions finales obtenues en termes de localisation finale des véhicules et de la valeur de la fonction objectif correspondante sont reportées. Afin de fournir une base de comparaison commune et d'étudier de manière plus concrète le compromis entre le gain en performance et les efforts de relocalisation, les valeurs du temps total de réponse (TR^1) et du temps total de

relocalisation (TL) correspondant à chacune des solutions sont aussi présentées. La Figure 5.4 présente différents graphiques où le temps de réponse total a été tracé en fonction du temps total de relocalisation. Ces tableaux et graphiques sont utilisés afin d'analyser et de discuter les résultats obtenus.

Les résultats obtenus pour différentes valeurs de ξ_s et ξ_r permettent d'abord de conclure que, clairement, le poids accordé à chacun des objectifs a un impact sur les décisions de localisation et de relocalisation. On observe toutefois que, pour les scénarios considérés, les efforts de relocalisation prennent rapidement le dessus sur le gain en performance au fur et à mesure que la valeur de ξ_s diminue, c'est-à-dire que rapidement, le gain en performance ne justifie plus les efforts de relocalisation. Nous avons d'ailleurs choisi d'inclure à la présentation des résultats le cas $\xi_s = 0.875$ et $\xi_r = 0.125$ de manière à fournir une meilleure appréciation du compromis entre le temps de réponse et le temps total de relocalisation. En effet, dans presque tous les cas, lorsque $\xi_s \leq 0.5$, les efforts de relocalisation ne justifient pas les gains en performance. La localisation initiale des véhicules avant le redéploiement a aussi un impact sur les décisions de relocalisation. Plus les localisations initiales sont proches des localisations optimales pour $|K|+1$, moins le gain en performance ne justifie la modification des postes d'attente des véhicules en attente. En effet, on observe que lorsque toutes les localisations initiales appartiennent à l'ensemble des localisations optimales dans le cas où $|K|+1$ véhicules sont disponibles (en gras dans le tableau), la solution courante, sans relocalisation, demeure de bonne qualité. On parle alors d'un écart de 1,4 % par rapport à la solution optimale dans le cas où 2 véhicules sont disponibles, de 2,3 % pour 3 véhicules et de 8,7 % pour 4 véhicules. Néanmoins, plus le nombre de véhicules est grand et donc, plus le système dispose d'une certaine flexibilité, plus la stratégie de « ne rien faire » est perdante. Les gains potentiels à considérer la relocalisation sont plus importants plus le nombre de véhicules est grand, et ce, peu importe les localisations initiales. D'un point de vue pratique, dans le contexte étudié, il ne devient vraiment intéressant de modifier la localisation des véhicules en attente que si le plan de localisation actuel s'écarte davantage du plan optimal pour $|K|+1$ véhicules. Autrement dit, à la suite de l'affectation d'un véhicule, il ne devient intéressant de revoir la localisation des véhicules restants que si leur localisation est considérablement différentes des localisations optimales pour $|K|+1$ véhicules.

En observant maintenant plus en détail la Figure 5.4, il est possible de voir clairement que dès que l'on accepte de modifier un peu le plan de localisation actuel, la qualité de la solution s'améliore dans une plus grande proportion. Il est donc intéressant de modifier la localisation des véhicules (surtout si on s'écarte beaucoup du plan optimal dans le cas où $|K|+1$ véhicules sont disponibles) afin de déterminer une bonne solution, mais les efforts supplémentaires en

termes de temps de relocalisation ne justifient pas toujours le gain en performance par la suite. Ces observations viennent donc confirmer les constats émis au *Chapitre 2*, dans un contexte d'application bien précis.

Loc. initiale	ξ_s	ξ_r	**Localisation finale**	**Obj,** [s]	**TR**[1] [s]	**TL** [s]	**Temps** [s]	***GAP*** [%]
	1	0	[2, 11]	844,69	844,69	137	35	0,00
	0,875	0,125	[2, 7]	749,39	856,44	0	19	0,00
[**2**, **7**]	0,75	0,25	[2, 7]	642,33	856,44	0	211	0,00
	0,5	0,5	[2, 7]	428,22	856,44	0	3	0,00
	0,25	0,75	[2, 7]	214,11	856,44	0	3	0,00
	0	1	[2, 7]	0,00	856,44	0	3	0,00
	1	0	[2, 11]	844,69	844,69	298	35	0,00
	0,875	0,125	[7, 11]	773,36	871,84	84	28	0,00
[**7**, 19]	0,75	0,25	[7, 11]	674,88	871,84	84	15	0,00
	0,5	0,5	[7, 19]	464,68	929,37	0	4	0,00
	0,25	0,75	[7, 19]	232,34	929,37	0	3	0,00
	0	1	[7, 19]	0,00	929,37	0	3	0,00
	1	0	[2, 11]	844,69	844,69	432	35	0,00
	0,875	0,125	[2, 11]	793,1	844,69	432	26	0,00
[15, 25]	0,75	0,25	[2, 11]	741,52	844,69	432	28	0,00
	0,5	0,5	[15, 25]	558,4	1116,8	0	6	0,00
	0,25	0,75	[15, 25]	279,2	1116,8	0	3	0,00
	0	1	[15, 25]	0,00	1116,8	0	3	0,00

Tableau 5.8 – Impact des coûts de relocalisation - 2 véhicules

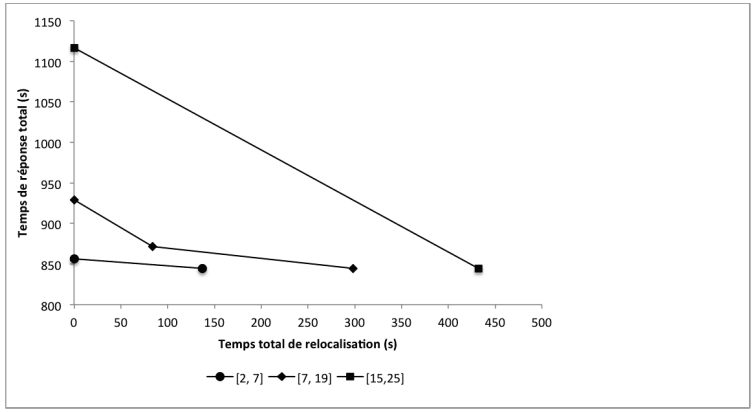
Loc. initiale	ξ_s	ξ_r	**Localisation finale**	**Obj.** [s]	**TR**[1] [s]	**TL** [s]	**Temps** [s]	***GAP*** [%]
	1	0	[2, 7, 20]	697,05	697,05	70	460	0,00
	0,875	0,125	[2, 7, 20]	618,67	697,05	70	76	0,00
[**2**, **4**	0,75	0,25	[2, 4, 20]	534,79	713,06	0	10	0,00
20]	0,5	0,5	[2, 4, 20]	356,53	713,06	0	3	0,00
	0,25	0,75	[2, 4, 20]	178,26	713,06	0	3	0,00
	0	1	[2, 4, 20]	0,00	713,06	0	3	0,00
	1	0	[2, 7, 20]	697,05	697,05	389	460	0,00
	0,875	0,125	[2, 7, 20]	658,54	697,05	389	362	0,00
[**4**, **11**,	0,75	0,25	[4, 11, 21]	580,22	761,5	0	10	0,00
21]	0,5	0,5	[4, 11, 21]	386,81	761,5	0	3	0,00
	0,25	0,75	[4, 11, 21]	193,41	761,5	0	3	0,00
	0	1	[4, 11, 21]	0,00	761,5	0	3	0,00
	1	0	[2, 7, 20]	697,05	697,05	510	460	0,00
	0,875	0,125	[2, 7, 11]	671,17	705,19	433	241	0,00
[1, 3,	0,75	0,25	[1, 2, 11]	616,49	775,32	140	89	0,00
11]	0,5	0,5	[1, 3, 11]	423,15	846,29	0	3	0,00
	0,25	0,75	[1, 3, 11]	211,57	846,29	0	3	0,00
	0	1	[1, 3, 11]	0	846,29	0	3	0,00
	1	0	[2, 7, 20]	697,05	697,05	310	460	0,00
	0,875	0,125	[2, 7, 20]	648,67	697,05	310	121	0,00
[3, 6,	0,75	0,25	[3, 7, 20]	590,95	731,27	60	31	0,00
22]	0,5	0,5	[3, 6, 22]	402,6	805,2	0	3	0,00
	0,25	0,75	[3, 6, 22]	201,3	805,2	0	3	0,00
	0	1	[3, 6, 22]	0	805,2	0	3	0,00

Tableau 5.9 – Impact des coûts de relocalisation - 3 véhicules

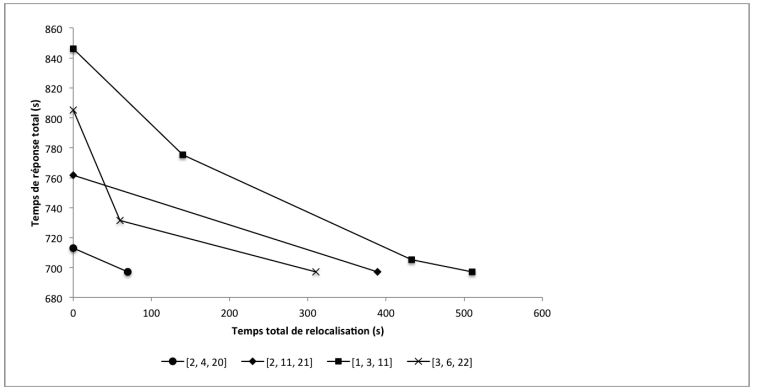
Loc. initiale	ξ_s	ξ_r	**Localisation finale**	**Obj.** [s]	**TR**[1] [s]	**TL** [s]	**Temps** [s]	***GAP*** [%]
[**2**, **4**, **11**, **28**]	1	0	[2, 4, 11, 20]	608,57	608,57	345	1731	0,00
	0,875	0,125	[2, 7, 11, 28]	564,19	634,79	70	359	0,00
	0,75	0,25	[2, 7, 11, 28]	493,59	634,79	70	16	0,00
	0,5	0,5	[2, 4, 11, 28]	333,4	666,8	0	3	0,00
	0,25	0,75	[2, 4, 11, 28]	166,7	666,8	0	3	0,00
	0	1	[2, 4, 11, 28]	0	666,8	0	3	0,00
[**2**, **6**, **28**, 30]	1	0	[2, 4, 11, 20]	608,57	608,57	683	1731	0,00
	0,875	0,125	[2, 7, 20, 28]	597,69	626,65	395	644	0,00
	0,75	0,25	[2, 6, 20, 28]	563,90	640,18	335	66	0,00
	0,5	0,5	[2, 6, 28, 30]	400,62	801,24	0	5	0,00
	0,25	0,75	[2, 6, 28, 30]	200,31	801,24	0	3	0,00
	0	1	[2, 6, 28, 30]	0,00	801,24	0	3	0,00
[3, 10, **11**, **20**]	1	0	[2, 4, 11, 20]	608,57	608,57	321	1731	0,00
	0,875	0,125	[2, 4, 11, 20]	572,62	608,57	321	394	0,00
	0,75	0,25	[3, 4, 11, 20]	526,19	638,92	188	25	0,00
	0,5	0,5	[3, 10, 11, 20]	365,96	731,92	0	3	0,00
	0,25	0,75	[3, 10, 11, 20]	182,98	731,92	0	3	0,00
	0	1	[3, 10, 11, 20]	0,00	731,92	0	3	0,00
[**6**, 13, 16, 24]	1	0	[2, 4, 11, 20]	608,57	608,57	459	1731	0,00
	0,875	0,125	[2, 6, 11, 24]	583,58	618,95	247	170	0,00
	0,75	0,25	[2, 6, 11, 24]	535,5	618,95	247	55	0,00
	0,5	0,5	[6, 11, 16, 24]	390,58	740,87	31	4	0,00
	0,25	0,75	[6, 13, 16, 24]	195,41	773,06	0	3	0,00
	0	1	[6, 13, 16, 24]	0	773,06	0	3	0,00
[9, 15, 27, 30]	1	0	[2, 4, 11, 20]	608,57	608,57	626	1731	0,00
	0,875	0,125	[2, 11, 20, 30]	601	630,29	421	295	0,00
	0,75	0,25	[2, 9, 11, 30]	562,05	655,07	283	27	0,00
	0,5	0,5	[9, 15, 27, 30]	392,82	785,63	0	3	0,00
	0,25	0,75	[9, 15, 27, 30]	196,41	785,63	0	3	0,00
	0	1	[9, 15, 27, 30]	0	785,63	0	3	0,00

Tableau 5.10 – Impact des coûts de relocalisation - 4 véhicules

(a) 2 véhicules



(b) 3 véhicules



(c) 4 véhicules

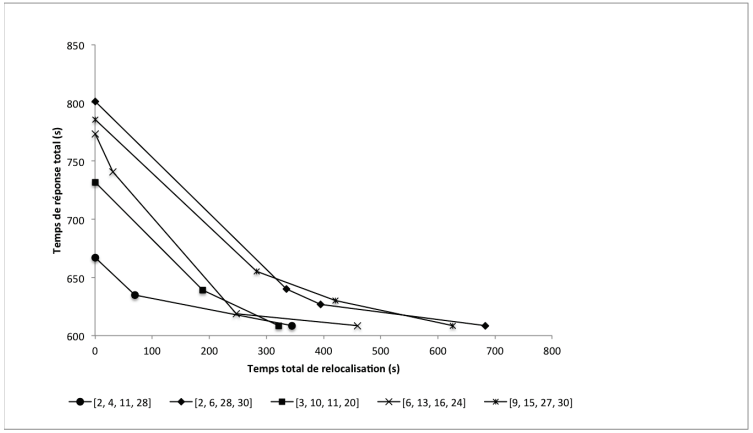


Figure 5.4 – Temps de réponse en fonction du temps de relocalisation

5.3.2 Validation et illustration pour les instances de taille moyenne

L'analyse effectuée pour les instances de taille réduite a permis de montrer que le fait de considérer plus d'un véhicule dans les listes de préaffectation influence les décisions de localisation et conséquemment, les performances du système. La capacité des véhicules influence aussi les décisions, mais dans les cas étudiés ici, ce sont principalement les décisions de préaffectation qui sont touchées par la capacité du système. De cette manière, les performances prédites du système se rapprochent davantage des performances réelles du système que lorsque la capacité des véhicules n'est pas tenue en compte. Enfin, la fonction objectif influence clairement les décisions, et en utilisant les mêmes mesures de la qualité du service, les poids attribués au service à la population et aux efforts de relocalisation influencent aussi les décisions. Maintenant, est-ce que les principaux constats résumés ici se vérifient lorsque l'on considère des instances de taille plus réaliste ? La présente section vise donc à valider ces observations grâce aux instances de taille réelle. Pour ce faire, les instances R149 ont été considérées.

|K|	$\|Z\|$	ξ_s	ξ_r	**TR**[1] [s]	**Temps** [s]	***GAP*** [%]
8	1	1	0	6429,2	9	0,00
	2	1	0	6069,6	28800	16,36
	2	0,5	0,5	6573,6	28800	12,04
10	1	1	0	5664,6	30	0,00
	2	1	0	5784,1	28800	16,17
	2	0,5	0,5	6134,4	28800	11,36
12	1	1	0	5361,8	14	0,00
	2	1	0	5460,1	28800	17,15
	2	0,5	0,5	5456,7	28800	8,93
14	1	1	0	5073,5	16	0,00
	2	1	0	5298,1	28800	19,90
	2	0,5	0,5	5549,3	28800	8,29

Tableau 5.11 – Résultats obtenus pour l'instance R149

(a) 8 véhicules

(b) 10 véhicules

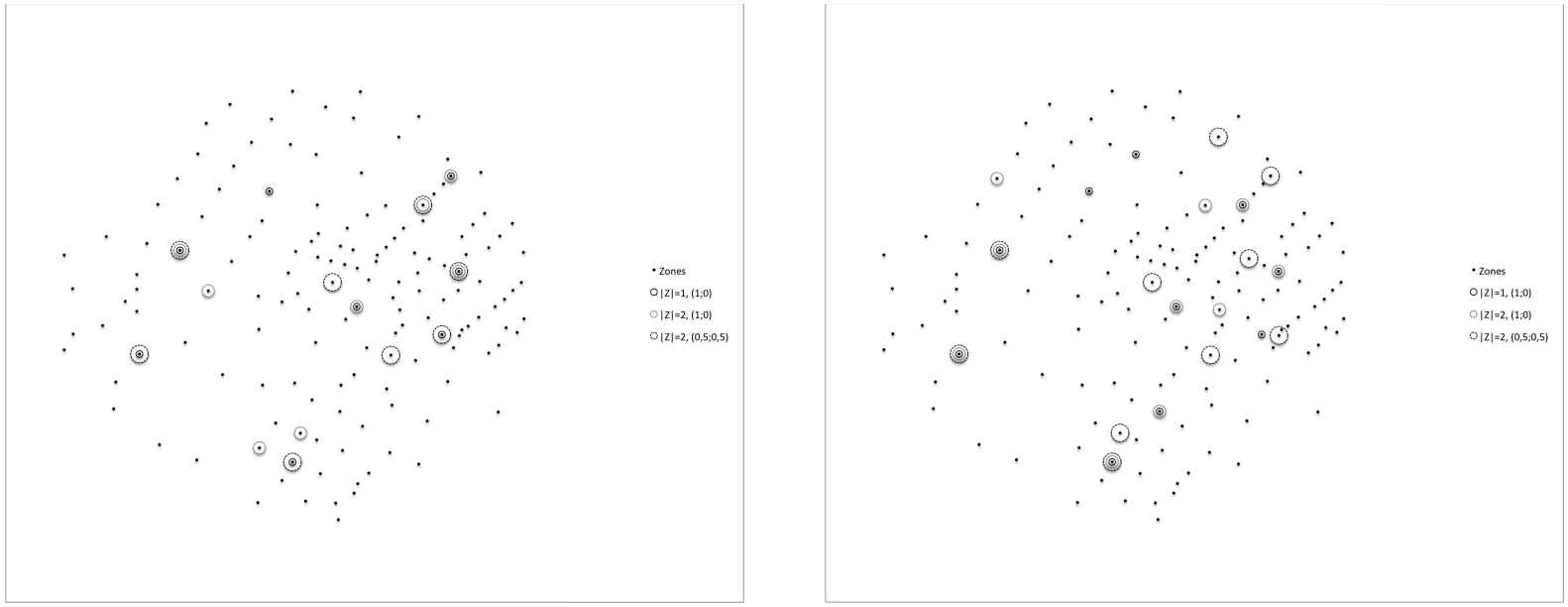
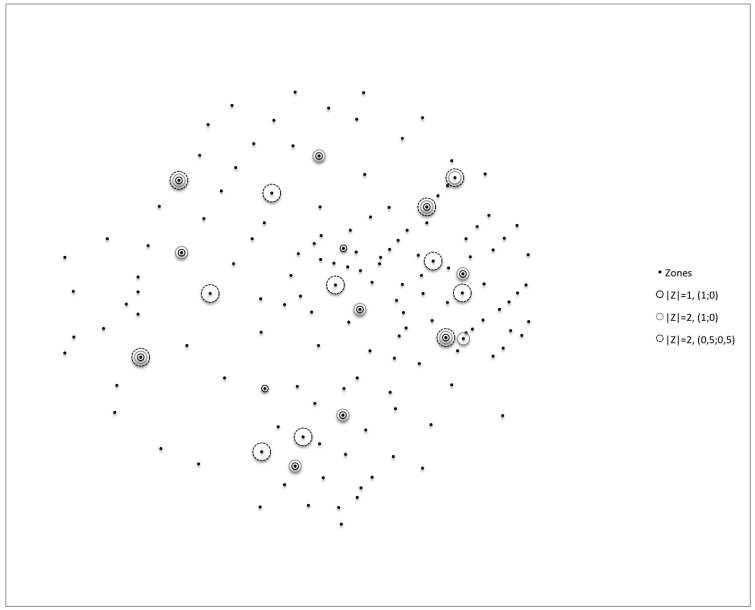


Figure 5.5 – Résultats obtenues pour l'instance R149 (I)

(c) 12 véhicules



(d) 14 véhicules

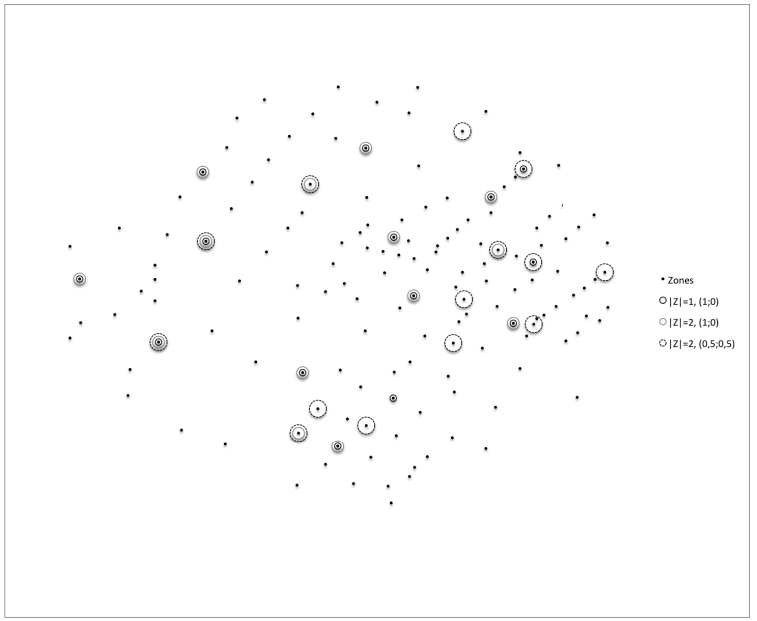


Figure 5.6 – Résultats obtenues pour l'instance R149 (II)

Les figures 5.5 et 5.6 présentent les résultats obtenus pour l'instance R149 pour quatre cas différents en termes de nombre de véhicules : 8, 10, 12 et 14 véhicules. Dans tous les cas de figures, trois situations ont été testées afin de valider les constats observés précédemment : $|Z| = 1$, $\xi_s = 1$ et $\xi_r = 0$; $|Z| = 2$, $\xi_s = 1$ et $\xi_r = 0$; $|Z| = 2$, $\xi_s = 0,5$ et $\xi_r = 0,5$. Les figures 5.5 et 5.6 montrent que les solutions obtenues semblent toujours différentes que l'on considère un ou deux véhicules sur les listes de préaffectation. De plus, le fait d'attribuer un poids différent aux objectifs, en conservant les mêmes tailles de liste, semble aussi mener à des solutions différentes en termes de localisation. Enfin, on peut observer que certaines localisations sont intéressantes peu importe les paramètres sélectionnés.

Si on analyse maintenant les résultats reportés au Tableau 5.11 pour les mêmes scénarios, il est possible d'observer que le fait de considérer deux véhicules sur les listes de préaffectation mène à des solutions différentes aussi en termes de la valeur de la fonction objectif. Afin d'assurer une meilleure comparaison des scénarios, nous avons calculé, pour $|Z| = 1$, la valeur optimale de la fonction objectif lorsque les listes de préaffectation incluent 2 véhicules, mais lorsque les localisations optimales pour $|Z| = 1$ sont maintenues. C'est cette valeur qui est reportée au Tableau 5.11 sous TR[1]. Tout d'abord, dans le cas où 8 véhicules sont disponibles pour répondre aux appels d'urgence et en considérant $\xi_s = 1$, il est possible d'observer que la solution obtenue pour $|Z| = 2$ est de meilleure qualité (on note une amélioration d'environ 7%) que celle obtenue pour $|Z| = 1$, et ce, malgré l'écart par rapport à la borne inférieure fournie par CPLEX (GAP) de 16,36%. Dans les cas où 10, 12 et 14 véhicules sont disponibles pour répondre aux appels et en considérant toujours $\xi_s = 1$, les solutions trouvées en imposant $|Z| = 1$ sont de meilleure qualité que celles déterminées en imposant $|Z| = 2$. En effet, dans ces cas, comme le GAP demeure grand pour $|Z| = 2$, la solution trouvée en posant $|Z| = 1$ demeure de meilleure qualité que celle déterminée en imposant $|Z| = 2$ et un temps limite de 28 800 secondes. La valeur de la fonction objectif pour 8 véhicules confirme toutefois que si on arrive à trouver une solution de bonne qualité, c'est-à-dire si on arrive à réduire le *GAP*, il peut être possible de faire un gain considérable. Le développement d'une approche de résolution qui permettra de trouver une solution de bonne qualité dans des délais de temps raisonnables est donc tout à fait justifié dans ce cas.

Les tests effectués jusqu'à maintenant pour l'instance R149 semblent bien valider les observations réalisées pour les instances de taille réduite, bien qu'il soit difficile de tirer des conclusions hors de tout doute. En effet, la plupart des instances n'ont pu être résolues à l'optimalité. Les difficultés liées à la résolution des ces instances justifient le développement d'une approche pour la résolution de taille d'instances réalistes, sujet qui sera abordé au prochain chapitre.

5.4 Conclusion

Dans ce chapitre, il a été question du développement et de l'analyse d'un modèle de décision pour la gestion des SPU considérant le redéploiement des véhicules, mais également une préaffectation anticipative des demandes éventuelles aux véhicules disponibles. Ce modèle vise à minimiser le temps de réponse espéré tout en limitant les efforts liés au redéploiement. Dans ce cas, le problème de redéploiement et de préaffectation des véhicules ambulanciers (PRPA) a été formulé afin de déterminer la localisation des véhicules disponibles pour répondre à des appels urgents de même que la liste de préaffectation pour chaque zone de demande. La liste de préaffectation d'une zone comporte une liste ordonnée de véhicules à affecter aux demandes placées dans la zone en question. Le PRPA tel que proposé dans ce chapitre se distingue donc des problèmes étudiés jusqu'à présent par trois aspects principaux. Tout d'abord, il intègre, en plus des décisions de relocalisation, l'établissement d'une liste ordonnée de préaffectation pour chaque zone de demande. De plus, il considère plus explicitement la capacité « réelle » des véhicules. Enfin, le temps de réponse espéré est formulé et utilisé afin d'évaluer les performances anticipées du système et sélectionner la meilleure solution possible.

Les résultats obtenus ont permis de montrer que le fait de considérer plus d'un véhicule dans les listes de préaffectation influence les décisions de localisation et, conséquemment, les performances du système. La capacité des véhicules influence aussi les décisions, mais dans les cas étudiés ici, ce sont principalement les décisions de préaffectation qui sont touchées par la capacité du système. De cette manière, les performances prédites du système se rapprochent davantage des performances réelles du système. Enfin, la fonction objectif influence clairement les décisions et, en utilisant les mêmes mesures de la qualité du service, les poids attribués aux objectifs liés au service à la population et aux efforts de relocalisation influencent aussi les décisions. Ces constats se vérifient principalement pour les instances de taille réduite bien que les résultats semblent aussi se vérifier pour les instances de taille plus réelle. Néanmoins, il est difficile de tirer des conclusions dans ce cas, puisque toutes les instances n'ont pu être résolues à l'optimalité dans les délais de temps imposés. L'expérimentation a donc permis de constater que le modèle proposé et les outils utilisés dans leur forme actuelle ne permettent pas de résoudre toutes les instances étudiées dans des délais de temps raisonnables. Ceci justifie hors de tout doute le développement d'une approche de résolution qui permette de déterminer de manière efficiente de bonnes solutions au problème. Le développement d'une telle méthode est abordée au *Chapitre 6*.

Le modèle proposé dans ce chapitre permet de déterminer la localisation des véhicules ambulanciers et une liste de préaffectation des zones de demande aux véhicules. De cette manière, il

peut être possible de mieux se préparer pour répondre aux demandes éventuelles. Néanmoins, l'anticipation des demandes se fait de manière implicite. Maintenant, il serait intéressant de considérer plus explicitement l'évolution et les performances futures du système. Dans ce cas, non seulement l'arrivée des demandes futures, mais aussi la libération prévue des véhicules, pourraient être pris en compte de manière à mieux guider la relocalisation des véhicules. À notre avis, cela pourrait contribuer à l'amélioration des performances globales pour l'ensemble du système, au temps présent et dans le futur. Dans le cadre de cette thèse, nous n'irons pas plus loin dans le développement du modèle en soi, ceci pourra faire l'objet de recherches futures. Au prochain chapitre, nous aborderons plutôt le développement d'approche de résolution afin de pouvoir considérer le modèle pour la prise de décision en pratique.

CHAPITRE 6

UNE APPROCHE MATHEURISTIQUE POUR LA RÉSOLUTION DU PROBLÈME DE REDÉPLOIEMENT ET DE PRÉAFFECTATION DES VÉHICULES AMBULANCIERS (PRPA)

Le problème de redéploiement et de préaffectation des véhicules ambulanciers a été défini au chapitre précédent afin de déterminer la localisation des véhicules disponibles à la suite d'un redéploiement, de même qu'un ensemble de listes de préaffectation pour chaque zone de demande. Il a alors été montré que le fait de considérer conjointement la relocalisation et la préaffectation a un impact réel sur la prise de décision et permet de mieux prédire les performances du système. De plus, le fait de prendre en compte la capacité des véhicules au sein du modèle influence également les décisions, mais dans ce cas, ce sont principalement les décisions de préaffectation qui sont touchées. Considérer le PRPA afin de supporter la prise de décision présente donc un potentiel clair. Néanmoins, l'expérimentation menée au *Chapitre 5* a aussi permis de soulever certains défis quant à sa résolution. D'une part, les temps de résolution peuvent devenir importants lorsque les listes de préaffectation comportent plus d'un véhicule. D'autre part, le modèle et les outils utilisés dans leur forme actuelle ne permettent pas de résoudre adéquatement toutes les instances de taille réelle. Dans ce contexte, le développement d'approches et de méthodes de résolution capables de fournir des solutions de bonne qualité à des problèmes de taille réelle dans des délais de temps raisonnables devient inévitable.

Jusqu'à présent, différentes méthodes ont été considérées pour la résolution de problèmes similaires au PRPA, c'est-à-dire différentes versions des problèmes de déploiement et de redéploiement. Ces méthodes de résolution ont été discutées au *Chapitre 3*. À la suite de l'analyse de la littérature existante, il a été possible de constater que certains auteurs ont eu recours à des méthodes exactes pour résoudre les problèmes étudiés. Des méthodes basées sur la relaxation linéaire avec l'ajout de coupes (Toregas *et al*., 1971; Daskin et Stern, 1981) ou l'application de méthodes de branchement (Church et ReVelle, 1974; Daskin, 1983; Hogan et ReVelle, 1986; ReVelle et Hogan, 1989; Ball et Lin, 1993) ont été utilisées. Les méthodes de résolution exactes deviennent toutefois rapidement impraticables lorsque la taille du problème ou le niveau de complexité du modèle s'accroît. C'est ce que nous avons constaté aussi au *Chapitre 5*. Certains auteurs ont donc opté pour le développement de méthodes heuristiques ou de métaheuristiques, propres au contexte étudié, pour assurer la résolution d'instances de taille réelle dans des délais de temps raisonnables. La plupart des métaheuristiques proposées dans ces cas sont basées

sur les principes de la recherche par voisinage, par exemple des algorithmes de recherche avec tabous (Gendreau *et al.*, 2001; Rajagopalan *et al.*, 2008; Başar *et al.*, 2011; Saydam *et al.*, 2013) ou des algorithmes de recherche à voisinage variable (Schmid et Doerner, 2010). Bien que différentes méthodes de résolution aient été proposées pour résoudre différentes versions du problème de redéploiement, aucune d'entre elles ne considère explicitement l'affectation des zones de demande aux véhicules au moyen de listes de préaffectation de taille variée. Aucune des approches existantes n'est donc directement applicable à la résolution du PRPA.

Le présent chapitre abordera donc le développement d'une première approche pour la résolution du PRPA. Pour ce faire, une approche de type matheuristique, composée de différents sous-problèmes et tirant profit d'une décomposition du territoire en sous-régions, est proposée ici. Chaque sous-problème correspond alors à un modèle de programmation linéaire, issu du modèle global tel qu'il a été défini au chapitre précédent. Le développement d'une approche matheuristique basée sur la décomposition du territoire se justifie par deux aspects principaux. Tout d'abord, une décomposition du problème basée sur une division du territoire en sous-régions est tout à fait naturelle dans le contexte de la gestion d'un SPU. En effet, certaines organisations, telles qu'Urgences-santé, divisent déjà le territoire à desservir en sous-régions, chaque sous-région étant gérée de manière indépendante [1]. De plus, une telle méthodologie permet de tirer profit de la puissance des méthodes de résolution exactes disponibles, telles que celles employées au chapitre précédent.

Ce chapitre poursuit donc deux objectifs principaux. Dans un premier temps, il vise à présenter une première approche pour la résolution du PRPA qui permettra de traiter des instances de taille réelle. Cette méthode pourra alors être utilisée en pratique afin de supporter la prise de décision à différents niveaux, mais elle pourra aussi être employée afin de valider les observations effectuées précédemment. Dans un deuxième temps, ce chapitre discute de manière claire l'utilisation d'un tel modèle en pratique, c'est-à-dire comme outil d'aide à la décision. La contribution principale de ce chapitre est méthodologique par le développement d'une approche de type matheuristique et d'un outil d'aide à la décision pour le déploiement/redéploiement et la préaffectation des véhicules ambulanciers. Le présent chapitre s'organise comme suit. Tout d'abord, l'approche matheuristique proposée est décrite en portant une attention particulière à la modélisation des différents sous-problèmes qui la composent. Les résultats obtenus pour un ensemble d'instances tirées d'un cas d'application réel sont ensuite présentés, suivis d'une discussion au sujet de l'utilisation du PRPA comme outil d'aide à la décision. Enfin, une conclusion, de même que différentes perspectives de recherche, viendront clore le chapitre.

[1] www.emergensys.net

6.1 Description de l'approche matheuristique

L'approche proposée ici est basée sur une décomposition du territoire à desservir en sous-régions. Dans le contexte de la gestion des SPU, une décomposition du territoire en sous-régions est tout à fait naturelle. En effet, tel que discuté en introduction, certaines organisations divisent le territoire à desservir en sous-régions. Chaque sous-région est ensuite gérée indépendamment, c'est-à-dire que chaque sous-région dispose de sa propre flotte de véhicules, les véhicules ne répondant principalement qu'aux appels en provenance de la sous-région à laquelle ils sont affectés. L'intervention de véhicules en provenance des autres sous-régions est admise lorsque nécessaire, mais de manière contrôlée. L'idée de diviser le territoire en sous-régions est donc cohérente avec la pratique. Le nombre de sous-régions considérées, de même que la division du territoire, deviennent alors des données du problème, c'est-à-dire que l'utilisateur a le contrôle sur la division du territoire en sous-régions. La décomposition du territoire n'est pas seulement une stratégie de réduction du temps de calcul au sein de l'algorithme de résolution, mais sous-tend plutôt une stratégie de gestion. Néanmoins, le nombre de sous-régions choisi influencera nécessairement la taille des sous-problèmes et, conséquemment, la qualité de la solution et le temps nécessaire à la résolution de chaque sous-problème.
La Figure 6.1 présente la structure de l'approche matheuristique proposée. Plus formellement, cette approche comporte trois phases principales :

- **Phase 1** : Affectation des véhicules disponibles aux sous-régions ;
- **Phase 2** : Relocalisation des véhicules et établissement des listes de préaffectation à l'intérieur des sous-régions ;
- **Phase 3** : Mise à jour des listes de préaffectation.

La première phase de l'algorithme vise à déterminer la localisation des véhicules disponibles de manière agrégée, c'est-à-dire au niveau des sous-régions, puis à déterminer l'affectation des appels en provenance des différentes sous-régions aux véhicules disponibles. Puisque cette phase prend en compte une version agrégée du problème, on considérera qu'un véhicule est toujours disponible lorsqu'un appel est placé. En effet, comme plusieurs véhicules sont localisés dans une même région (si la demande le justifie), la capacité globale des véhicules affectés à une sous-région permettra de servir adéquatement les demandes : les véhicules disponibles pourront compenser pour la non-disponibilité de certains. L'affectation plus détaillée des véhicules aux zones de demande appartenant à une même sous-région sera plutôt effectuée à la phase 2. En ce sens, la deuxième phase consiste à déterminer la localisation des véhicules à l'intérieur de

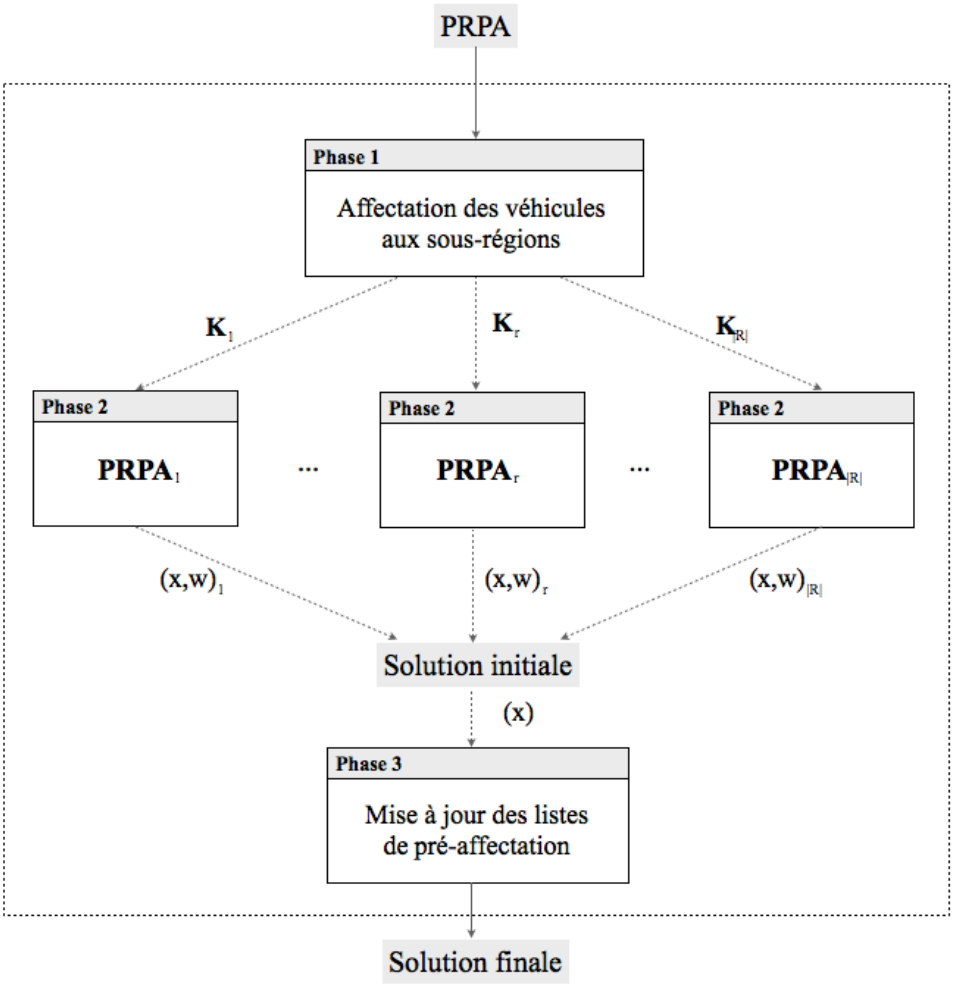


Figure 6.1 – Schéma de l'approche matheuristique

la sous-région considérée, puis à constituer la liste de préaffectation de chaque zone de demande appartenant à la sous-région concernée. Contrairement à la phase 1, la phase 2 considérera la non-disponibilité des véhicules en prenant explicitement en compte le taux d'occupation des véhicules lors de la formulation de la fonction objectif et des contraintes de capacité, comme dans le modèle original proposé pour le PRPA. Enfin, la troisième phase consiste à revoir les listes de préaffectation en considérant les localisations déterminées à la phase 2 comme étant fixes. En effet, toute approche de décomposition du territoire peut amener d'importants problèmes aux frontières des sous-régions. Afin de pallier à cette difficulté, une troisième phase a été ajoutée de manière à déterminer le meilleur plan de préaffectation global en fonction des décisions de localisations choisies à la phase 2. La troisième phase ne sera lancée que si les règles de gestion sont telles qu'un véhicule affecté à une sous-région peut répondre à un appel reçu à l'extérieur de sa sous-région si c'est nécessaire de le faire.

Chacune des trois phases présentées ci-haut correspond à un modèle de programmation mathématique linéaire issu du modèle global. Chaque sous-problème est alors résolu de manière optimale ou encore en imposant un temps limite pour la résolution des différents sous-problèmes, si le temps disponible à leur résolution ne permet pas l'obtention d'une garantie d'optimalité. Dans le meilleur des cas, tous les sous-problèmes pourront être résolus à l'optimalité : l'écart entre la solution trouvée et la solution optimale du problème global ne sera alors qu'une conséquence de la décomposition du problème.

La première version de l'approche, telle que décrite ici, consiste à effectuer chacune des trois phases de manière séquentielle, c'est-à-dire qu'il n'y a pas de boucle de rétroaction entre les différentes phases. L'objectif de ce chapitre étant de présenter une première approche pour la résolution de PRPA lorsque confronté à des instances de taille réelle et permettre la validation finale du modèle, l'amélioration de l'approche matheuristique en soi ne sera abordée que dans le cadre de recherches futures. Il est probable, selon nous, que l'approche matheuristique puisse bénéficier considérablement de l'introduction de différents mécanismes de rétroaction et de modification de la solution, issus de la recherche locale par exemple. Les pistes d'améliorations potentielles à privilégier seront discutées brièvement à la suite de la présentation des résultats expérimentaux. Dans cette section, nous décrirons plutôt les différentes phases de l'approche matheuristique proposée plus en détail. À moins d'indication contraire, la description des modèles associés à chacune des phases considère la même notation que celle employée au chapitre précédent. De plus, puisque l'ensemble des positions sur la liste de préaffectation et les poids attribués aux différents objectifs peuvent varier, on définira Z^2 et Z^3, l'ensemble des positions sur les listes pour les phases 2 et 3. On notera aussi R, l'ensemble des sous-régions.

6.1.1 Phase 1 : Affectation des véhicules disponibles aux sous-régions

La phase 1 consiste à déterminer la localisation des véhicules disponibles de même que l'affectation des véhicules disponibles aux demandes en provenance des différentes sous-régions. Cette phase cherche alors à déterminer le meilleur plan de redéploiement de manière à minimiser à la fois le temps total nécessaire pour répondre à tous les appels et le temps total nécessaire au redéploiement des véhicules entre les sous-régions. On considérera qu'un véhicule pourra répondre aux appels en provenance de la sous-région à laquelle il est affecté, mais pourra aussi répondre à ceux en provenance d'autres sous-régions si l'état du système le justifie. De cette manière, les échanges entre les sous-régions pourront être considérés. Pour chaque sous-région $r \in R$, la demande d_r correspond à la somme des demandes pour l'ensemble des zones appartenant à la sous-région concernée, notée I_r, chaque demande étant placée à partir du centroïde de la sous-région. Le centroïde d'une sous-région correspond, quant à lui, au centroïde de la zone pour laquelle la distance euclidienne entre le centre géométrique de la sous-région et le centroïde de la zone en question est la plus petite. Le centroïde de la sous-région représente aussi l'unique poste d'attente de la sous-région, un poste d'attente dont la capacité est égale à la somme des capacités de tous les postes d'attente appartenant à la sous-région concernée, noté J_r. La localisation initiale d'un véhicule, c'est-à-dire sa localisation avant le redéploiement, correspond alors au centroïde de la sous-région où il se retrouve. Enfin, la matrice des temps de déplacement entre les centroïdes des sous-régions est déterminée à partir de la matrice globale comportant tous les temps de déplacement du centroïde d'une zone vers les centroïdes des autres zones.

Ainsi, en définissant, $x^k_{r_1 r_2}$ une variable binaire qui vaut 1 si le véhicule k se déplace de la région r_1 à la région r_2 à la suite de la relocalisation, et 0 autrement ; w^k_r, la proportion des demandes de la région r affectées au véhicule k et $y^k_{r_1 r_2}$, la proportion des demandes de la région r_1 affectées au véhicule k, localisé en r_2 à la suite de la relocalisation, le modèle associé à la phase 1 se formule de la manière suivante :

$$\min \xi_s \sum_{k \in K} \sum_{r_1 \in R} \sum_{r_2 \in R} d_{r_1} t_{r_2 r_1} y^k_{r_1 r_2} + \xi_r \sum_{k \in K} \sum_{r_1 \in R} \sum_{r_2 \in R} \theta^k t_{r_1 r_2} x^k_{r_1 r_2} \tag{6.1}$$

sous les contraintes :

$$\sum_{r_1 \in R} \sum_{r_2 \in R} \lambda_{k r_1} x^k_{r_1 r_2} = 1, \ \forall k \in K, \tag{6.2}$$

$$\sum_{r_1 \in R} \sum_{r_2 \in R} (1 - \lambda_{k r_1}) x^k_{r_1 r_2} = 0, \ \forall k \in K, \tag{6.3}$$

$$\sum_{k\in K}\sum_{r_1\in R} x^k_{r_1 r_2} \leq p_r,\ \forall r_2 \in R, \tag{6.4}$$

$$\sum_{r\in R} a_r w^k_r \leq W,\ \forall k \in K, \tag{6.5}$$

$$\sum_{k\in K} w^k_r \geq 1,\ \forall r \in R, \tag{6.6}$$

$$y^k_{r_1 r_2} \leq \sum_{r_3\in R} x^k_{r_3 r_2},\ \forall r_1 \in R,\ \forall r_2 \in R,\ \forall k \in K, \tag{6.7}$$

$$w^k_{r_1} = \sum_{r_2\in R} y^k_{r_1 r_2}, \forall r_1 \in R, \forall k \in K, \tag{6.8}$$

$$x^k_{r_1 r_2} \in \{0,1\},\ \forall r_1 \in R,\ \forall r_2 \in R,\ \forall k \in K, \tag{6.9}$$

$$w^k_r \geq 0,\ \forall r \in R,\ \forall k \in K, \tag{6.10}$$

$$y^k_{r_1 r_2} \geq 0,\ \forall r_1 \in R,\ \forall r_2 \in R,\ \forall k \in K, \tag{6.11}$$

$$w^k_r \leq 1,\ \forall r \in R,\ \forall k \in K, \tag{6.12}$$

$$y^k_{r_1 r_2} \leq 1,\ \forall r_1 \in R,\ \forall r_2 \in R,\ \forall k \in K. \tag{6.13}$$

Tel que discuté précédemment, la fonction objectif minimise le temps total nécessaire afin de servir toutes les demandes de même que le temps total nécessaire au redéploiement des véhicules entre les sous-régions. La fonction objectif, telle que formulée en (6.15), peut engendrer un nombre considérable de solutions équivalentes, principalement lorsque les coûts de relocalisation ne sont pas considérés, *i.e.* $\xi_r = 0$. Afin de mieux départager les solutions, un objectif secondaire a été introduit. Cet objectif vise à répartir le plus équitablement possible les véhicules entre les sous-régions, c'est-à-dire à minimiser la différence entre le nombre minimal et le nombre maximal de véhicules affectés à une sous-région, tout en assurant que chaque demande soit servie par un véhicule disponible. D'autres objectifs secondaires auraient pu être considérés afin de mieux prendre en compte l'intensité et la dispersion des demandes au sein d'une même sous-région, deux facteurs qui peuvent influencer le temps de réponse pour une sous-région donnée. Néanmoins, à la suite de divers tests préliminaires, le second objectif choisi semblait fournir de bons résultats. Cette approche constitue donc, à notre avis, un bon premier pas pour la résolution de la phase 1, principalement si une attention particulière est portée à la décomposition du territoire de façon à obtenir des sous-régions les plus uniformes possibles en termes de demandes. Les contraintes du modèle assurent, quant à elles, la localisation adéquate des véhicules (6.2) (6.3), de même que le respect de la capacité des postes d'attente (6.4) et des véhicules disponibles (6.5). Le modèle garantit aussi que chaque demande est couverte

par au moins un véhicule (6.6). Enfin, les contraintes (6.8) et (6.9) assurent le lien entre les variables de décisions, tandis que les contraintes (6.10) à (6.13) assurent leur intégralité ou le respect de leurs bornes. Le modèle associé à la phase 1 s'apparente aux modèles de localisation et d'affectation (ReVelle et Eiselt, 2005; Current *et al.*, 2001), mais où les coûts associés à la relocalisation, définis ici comme le temps total de relocalisation, sont considérés explicitement au sein de la fonction objectif.

6.1.2 Phase 2 : Relocalisation des véhicules et établissement des listes de préaffectation à l'intérieur des sous-régions

La phase 2 correspond essentiellement à la résolution du PRPA tel qu'il a été défini au chapitre précédent. Un PRPA est alors résolu pour chaque sous-région : $|R|$ PRPA sont résolus de manière à fournir une borne supérieure pour le problème global. Lorsque $|R| = 1$, le problème correspond au problème global. La solution déterminée à la phase 1 permet de définir chacun des $|R|$ problèmes à résoudre à la phase 2. L'ensemble des véhicules localisés à chaque sous-région r, noté K_r, est alors extrait de la solution obtenue à la suite de la phase 1. La localisation initiale des véhicules appartenant à K_r à l'intérieur de la sous-région est, quant à elle, déterminée de la manière suivante. Tout d'abord, si la sous-région du véhicule avant le redéploiement est la même que celle choisie à la suite de la phase 1, la localisation initiale du véhicule demeure la même à l'intérieur de sa sous-région. Cependant, si la solution obtenue à la suite de la phase 1 comporte un changement de sous-région, la localisation initiale du véhicule (pour fins de résolution de la phase 2) change. Sa localisation initiale est alors choisie aléatoirement parmi l'ensemble des postes d'attente libres appartenant à la dite sous-région. Néanmoins, les coûts associés à la relocalisation du véhicule à l'intérieur de la sous-région ne sont pas comptabilisés puisqu'ils ont déjà été encourus à la phase 1. La valeur de θ_k est alors fixée à 0.

Pour chaque sous-problème $r \in R$, on considérera l'ensemble des zones de demande et des postes d'attente appartenant à la sous-région concernée, notés respectivement I_r et J_r. La demande des zones et la capacité des postes d'attente correspondent alors à leur valeur originale. Enfin, la matrice des temps de déplacement entre les centroïdes des zones de demande appartenant à I_r est déterminée à partir de la matrice globale des temps de déplacement entre les centroïdes des zones. De manière similaire à la formulation originale, le PRPA$_r$ considère $x^k_{j_1 j_2}$, une variable binaire qui vaut 1 si le véhicule k se déplace du site $v_{j_1} \in J_r$ vers le site $v_{j_2} \in J_r$ à la suite du redéploiement/relocalisation, et 0 autrement ; w_i^{zk} un variable binaire qui vaut 1 si le véhicule k est le z-ième appelé sur la liste de préaffectation de la zone de demande située en $v_i \in I_r$, et 0 autrement ; et y_{ij}^{zk}, une variable binaire qui vaut 1 si le véhicule k, localisé en $v_j \in J_r$

à la suite du redéploiement, est à la z-ième position sur la liste de préaffectation de la zone de demande située en $v_i \in I_r$, et 0 autrement. Les autres paramètres demeurent les mêmes que ceux définis au chapitre précédent. Ainsi, pour chaque sous-région $r \in R$, le problème suivant est résolu :

PRPA$_r$

$$\min \xi_s \sum_{i \in I_r} \sum_{z \in Z^2} \sum_{k \in K_r} \sum_{j \in J_r} (1-q) q^{z-1} d_i t_{ji} y_{ij}^{zk} + \xi_r \sum_{k \in K_r} \sum_{j_1 \in J_r} \sum_{j_2 \in J_r} \theta^k t_{j_1 j_2} x_{j_1 j_2}^k \tag{6.14}$$

sous les contraintes :

$$\sum_{j_1 \in J_r} \sum_{j_2 \in J_r} \lambda_{k j_1} x_{j_1 j_2}^k = 1, \ \forall k \in K_r, \tag{6.15}$$

$$\sum_{j_1 \in J_r} \sum_{j_2 \in J_r} (1 - \lambda_{k j_1}) x_{j_1 j_2}^k = 0, \ \forall k \in K_r, \tag{6.16}$$

$$\sum_{k \in K_r} \sum_{j_1 \in J_r} x_{j_1 j_2}^k \leq p_j, \ \forall j_2 \in J_r, \tag{6.17}$$

$$\sum_{z \in Z^2} \sum_{i \in I_r} (1-q) q^{z-1} d_i w_i^{zk} \leq W, \ \forall k \in K_r, \tag{6.18}$$

$$\sum_{z \in Z^2} w_i^{zk} \leq 1, \ \forall k \in K_r, \ \forall i \in I_r, \tag{6.19}$$

$$\sum_{k \in K_r} w_i^{zk} = 1, \ \forall z \in Z^2, \ \forall i \in I_r, \tag{6.20}$$

$$y_{ij_2}^{zk} \leq \sum_{j_1 \in J_r} x_{j_1 j_2}^k, \ \forall z \in Z^2, \ \forall i \in I_r, \ \forall k \in K_r, \ \forall j_2 \in J_r, \tag{6.21}$$

$$w_i^{zk} = \sum_{j \in J} y_{ij}^{zk}, \forall i \in I_r, \forall z \in Z^2, \forall k \in K_r, \tag{6.22}$$

$$x_{j_1 j_2}^k \in \{0,1\}, \ \forall j_1 \in J_r, \ \forall j_2 \in J_r, \ \forall k \in K_r, \tag{6.23}$$

$$w_i^{zk} \in \{0,1\}, \ \forall z \in Z^2, \ \forall i \in I_r, \ \forall k \in K_r, \tag{6.24}$$

$$y_{ij}^{zk} \in \{0,1\}, \ \forall z \in Z^2, \ \forall i \in I_r, \ \forall k \in K_r, \ \forall j \in J_r. \tag{6.25}$$

Le PRPA$_r$ minimise le temps de réponse espéré total, de même que le temps total de redéploiement pour les zones de demande et les véhicules à l'intérieur de la sous-région concernée. Comme pour le problème original, les contraintes (6.15) et (6.16) assurent la localisation adéquate des véhicules disponibles, tandis que les contraintes (6.17) garantissent le respect de la capacité des postes d'attente. Les contraintes (6.18) assurent, quant à elles, la satisfaction de la capacité des véhicules à traiter des demandes. En plus des contraintes liantes et d'intégralité

(6.21) à (6.25), le modèle impose aussi un ensemble de contraintes stipulant qu'un véhicule ne peut occuper plus d'une position sur la liste de préaffectation d'une zone de demande donnée (6.19) et qu'exactement un véhicule doit se retrouver à chaque position sur la liste de préaffectation d'une zone de demande (6.20).

Pour chaque sous-région, la résolution du PRPA permet de déterminer la localisation précise des véhicules disponibles à l'intérieur de la sous-région, de même que les listes de préaffectation pour toutes les zones de demande appartenant à la sous-région considérée, identifiées par $(x,w)_r$ sur la Figure 6.1. La combinaison des solutions pour toutes les sous-régions $r \in R$ permettra de fournir une borne supérieure initiale pour le problème étudié. Néanmoins, la décomposition de territoires en sous-régions peut générer des problèmes aux frontières. En effet, il pourrait devenir plus avantageux d'affecter une zone de demande à un véhicule hors de sa sous-région, si ce véhicule peut répondre plus rapidement à une demande donnée. La phase 3 a été conçue en ce sens. Les listes de préaffectation pourront alors être revues de manière globale, en fixant les localisations choisies à la suite de la phase 2, identifiées par (x) sur la Figure 6.1. Cette phase pourra, voire devra, être lancée si les échanges de véhicules sont permis entre les sous-régions, c'est-à-dire qu'un véhicule peut servir un appel à l'extérieur de sa sous-région, ce qui n'est pas toujours le cas en pratique. Si ces échanges ne sont pas permis, seules les phases 1 et 2 sont pertinentes.

6.1.3 Phase 3 : Mise à jour des listes de préaffectation

Si le contexte d'application le permet, la phase 3 consiste à revoir globalement les listes de préaffectation en maintenant les décisions de localisation et de relocalisation déterminées à la suite de la phase 2. Dans ce cas, toutes les zones de demande sont considérées simultanément de manière à améliorer la qualité de la borne supérieure initiale. La demande pour chaque zone demeure égale à sa demande originale. Lors de la formulation du modèle, comme la localisation des véhicules est connue, t_{ki} est défini (plutôt que t_{ji}) comme le temps nécessaire au véhicule k pour atteindre une zone de demande localisée en v_i. La matrice des temps de déplacement entre les véhicules et les zones de demandes est toujours déterminée à partir de la matrice globale des temps de déplacement entre les centroïdes des zones de demande.

En considérant w_i^{zk}, une variable binaire qui vaut 1 si le véhicule k est le z-ième appelé sur la liste de préaffectation de la zone de demande située en $v_i \in I$, et 0 autrement, le modèle correspondant à la phase 3 se formule comme suit :

$$\min \sum_{i \in I} \sum_{z \in Z_3} \sum_{k \in K} \sum_{j \in J_r} (1-q) q^{z-1} d_i t_{ki} w_i^{zk} \tag{6.26}$$

sous les contraintes :

$$\sum_{z \in Z^3} \sum_{i \in I} (1-q) q^{z-1} d_i w_i^{zk} \leq W, \ \forall k \in K, \tag{6.27}$$

$$\sum_{z \in Z^3} w_i^{zk} \leq 1, \ \forall k \in K, \ \forall i \in I, \tag{6.28}$$

$$\sum_{k \in K_r} w_i^{zk} = 1, \ \forall z \in Z^3, \ \forall i \in I_r, \tag{6.29}$$

$$w_i^{zk} \in \{0,1\}, \ \forall z \in Z^3, \ \forall i \in I_r, \ \forall k \in K_r. \tag{6.30}$$

De manière similaire à la phase 2, la fonction objectif (6.26) vise la minimisation du temps de réponse espéré total de sorte que la capacité des véhicules à servir des demandes urgentes soit respectée (6.27), qu'un véhicule ne se retrouve qu'à au plus une position sur une liste de préaffectation (6.28) et qu'exactement un véhicule se retrouve à chaque position sur la liste de préaffectation d'une zone donnée (6.29). Enfin, l'intégralité des variables de décision est garantie grâce aux contraintes (6.30).

6.2 Expérimentation

Afin d'analyser et de valider l'approche de résolution proposée, une série d'expérimentations a été menée. L'expérimentation poursuit trois objectifs principaux. Dans un premier temps, elle vise à valider la méthode en soi, c'est-à-dire à mesurer l'impact de la décomposition et de la limite imposée sur le temps de calcul pour la résolution des différents sous-problèmes sur la qualité de la solution finale. Dans un deuxième temps, elle cherche à démontrer que la méthode proposée permet bien de résoudre des problèmes de taille réelle et, conséquemment, grâce aux résultats obtenus, à confirmer les principales observations issues de l'analyse effectuée au chapitre précédent. Enfin, elle permet de soulever différentes pistes d'améliorations potentielles qui pourront être explorées plus en détail dans le cadre de recherches futures.

Afin de mener l'expérimentation, les instances pseudo-réelles considérées au chapitre précédent ont été récupérées ici. Ces instances se divisent maintenant en trois groupes correspondant à trois différentes tailles de problèmes : R30 (30 zones de demande), R149 (149 zones de demande) et R595 (595 zones de demande). Quelle que soit la taille du problème, ces instances ont été définies à partir du territoire ou d'un sous-ensemble du territoire des villes de Montréal et de Laval. Ces instances permettront d'analyser le comportement de la méthodologie proposée dans un contexte d'application réaliste. Dans tous les cas, la demande, la capacité des véhicules de même que l'ensemble des postes d'attention potentiels ont été déterminés tel que décrit au chapitre précédent. Le nombre de véhicules a, quant à lui, été fixé de manière à obtenir un

ratio zones par véhicule d'environ 12, sauf dans le cas R30 où un plus petit ratio a été choisi de façon à obtenir une solution réalisable pour tous les scénarios de décomposition étudiés. À moins d'indication contraire, 4, 12 et 50 véhicules sont utilisés respectivement pour R30, R149 et R595. Les données et les paramètres les plus importants sont résumés au Tableau 6.1.

	Instances		
	R30	**R149**	**R595**
$\|I\|$	30	149	595
$\|J\|$	30	48	218
$\|R\|$	2	[2 ; 3 ; 4]	15 [5 ; 10]
$\|K\|$	4	12	50 [40 ; 60]
$\|Z^2\|$	[1 ; 2]		2 [1 ; 3]
$\|Z^3\|$	2		2 [3]
d_i	PM		
q	0,5		
W	5		
t_{ij}	À partir des temps de déplacement « réels »		

Tableau 6.1 – Paramètres par groupe d'instances

Pour valider adéquatement la méthode de résolution proposée, chaque instance a été décomposée en sous-régions. Tel qu'indiqué au Tableau 6.1, l'instance R30 a été divisée en 2 sous-régions, l'instance R149 en 2, 3 et 4 sous-régions, tandis que l'instance R595 a été divisée en 5, 10 et 15 sous-régions. Chaque décomposition a été établie de manière arbitraire, mais de manière à obtenir une division relativement uniforme en termes de zones par sous-région. Pour les instances R149 et R595, cette division du territoire en sous-régions correspond à des ratios variant entre 37,5 à 120 zones par sous-région. Ces différents scénarios ont été générés de manière à obtenir des sous-problèmes de taille raisonnable et pouvant être résolus adéquatement grâce aux outils en place. Considérant la taille réduite de l'instance R30, le ratio zones par sous-région est beaucoup plus petit. On parle alors de 15 zones par sous-région. Pour l'instance R30, deux divisions différentes du territoire sont évaluées. Les figures 6.2 à 6.4 illustrent la décomposition du territoire en sous-régions, et ce, pour chaque instance. Chaque point sur le territoire à desservir représente le centroïde d'une zone, les zones appartenant à différentes sous-régions étant identifiées par des points de forme différente.

a) 2 sous-régions (R30-I)

b) 2 sous-régions (R30-II)

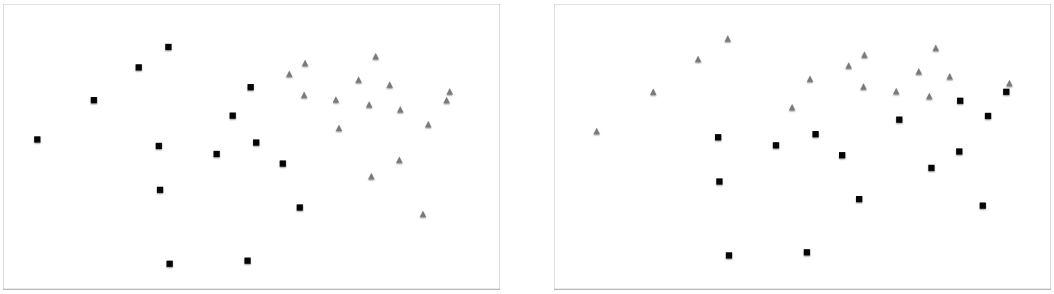
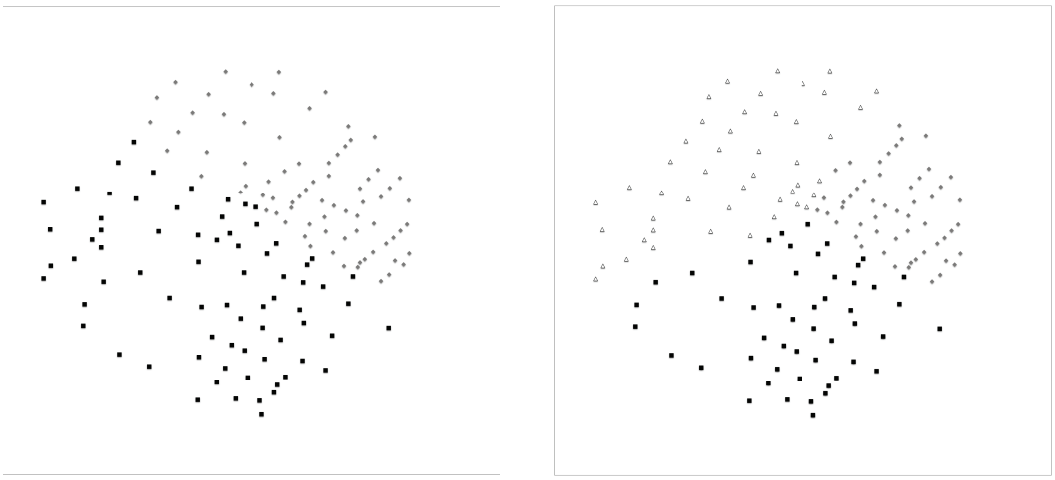


Figure 6.2 – Instance R30

a) 2 sous-régions (R149-2)

b) 3 sous-régions (R149-3)



c) 4 sous-régions (R149-4)

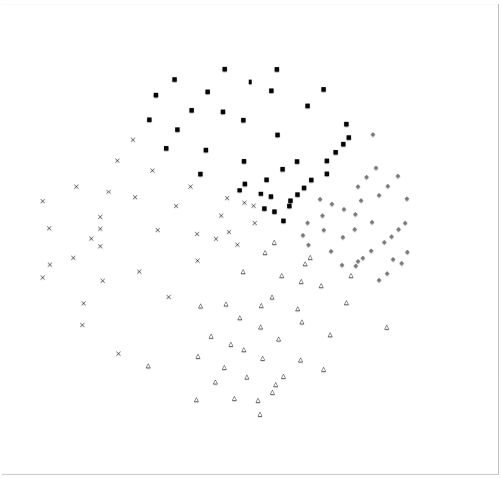
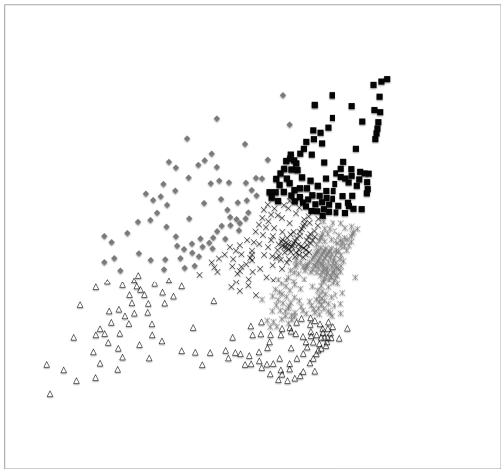
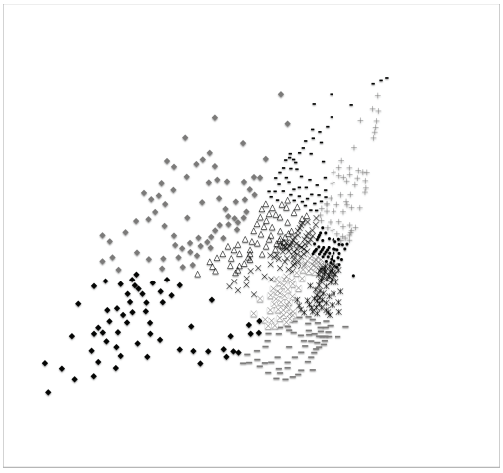


Figure 6.3 – Instance R149

a) 5 sous-régions (R595-5)



b) 10 sous-régions (R595-10)



c) 15 sous-régions (R595-15)

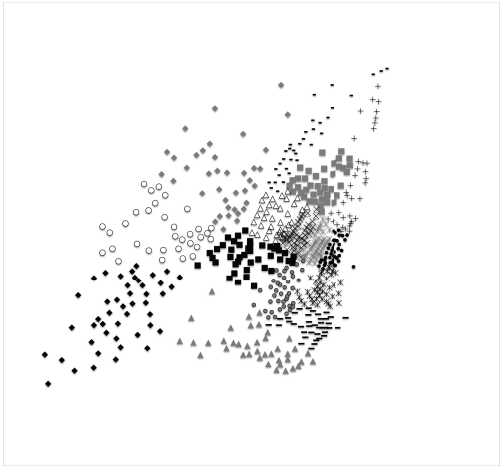


Figure 6.4 – Instance R595

Dans cette sous-section, les résultats obtenus pour les différentes expérimentations menées sont présentés et discutés. Dans tous les cas, les résultats obtenus pour chaque sous-problème ont été déterminés au moyen de CPLEX 12.5. Les tests ont été lancés à partir d'une machine Windows avec huit cœurs et 15 Go de RAM, installée sur un serveur équipé d'un processeur AMD Opteron 6328.

6.2.1 Impact de la décomposition sur la valeur de la solution finale

La décomposition du territoire en sous-régions peut avoir un impact important sur la prise de décision. En effet, le nombre de sous-régions considérées, de même que la composition de chaque sous-région, viendront influencer à la fois les décisions de localisation et de préaffectation. Une question se pose alors : qu'en est-il de la qualité des solutions obtenues grâce à l'approche matheuristique proposée quand on les compare aux solutions optimales pour le problème original ? Autrement dit, quelle est la perte de qualité des résultats engendrée par la décomposition ? Afin de répondre à cette importante question, une série d'expérimentations a été menée à partir des instances R30, R149 et R595. Pour l'instance R30, la solution optimale a été déterminée au chapitre précédent. Pour l'instance R149, seule une borne supérieure a pu être déterminée. Enfin, aucune solution n'est connue pour l'instance R595. En effet, la méthode proposée au chapitre précédent n'a pas permis d'obtenir une solution réalisable pour l'instance R595. Différents niveaux de décomposition ont tout de même été testés afin d'analyser leur impact sur la qualité de la solution finale. Une décomposition du territoire en 5, 10 et 15 sous-régions a alors été étudiée. La présente section présente donc une analyse plus détaillée de l'impact de la décomposition sur la qualité de la solution et la perte d'optimalité qui en découle.

Les résultats présentés aux tableaux 6.2 à 6.4 ont été obtenus au moyen de l'approche matheuristique proposée. Chaque sous-problème a alors été résolu au moyen de CPLEX 12.5, en imposant un temps limite de 3600 secondes. Chacun des tableaux de résultats présente d'abord les paramètres importants utilisés soit le nombre de zones considérées ($|R|$), le taille des listes considérées aux phases 2 et 3 ($|Z^2|$, $|Z^3|$), puis l'ensemble des phases lancées afin de déterminer la solution finale (Phases). La meilleure solution trouvée au *Chapitre* 5 ($TR^\star$) est ensuite reportée. Enfin, différentes mesures associées à la solution déterminée par l'approche matheuristique sont présentées. Ces mesures sont : la valeur de la solution finale (TR), le temps moyen de calcul pour l'ensemble des sous-problèmes de la phase 2 ($Temps^2_{moy}$), le temps total de résolution ($Temps_{tot}$), l'écart moyen entre la solution obtenue et la borne inférieure fournie par CPLEX pour chaque sous-problème de la phase 2 (GAP^2), l'écart maximal entre la solution obtenue et la borne inférieure fournie par CPLEX pour chaque sous-problème de la phase 2 (GAP_{max}) et

le nombre de sous-problème de la phase 2 qui n'ont pu être résolu à l'optimalité à l'intérieur du temps limite (Nb^2_{tot}). Ces différentes mesures sont analysées puis discutées dans la présente sous-section. Il est important de mentionner que le temps de calcul pour la phase 3 n'a pas été reporté explicitement puisqu'il est négligeable par rapport au temps total de calcul. De plus, dans tous les cas, le problème de la phase 3 a pu être résolu à l'optimalité.

|R|	(|Z^2|, |Z^3|)	**Phases**	**TR*** [s]	**TR** [s]	**Temps**^{2} [s]	**Temps**$_{tot}$ [s]	**GAP**^{2} [%]	**GAP**$_{max}$ [%]	Nb^2_{s-opt} [-]
2-I	(2,2)	1+2	608,6	778,4	0,12	1,9	0	0	0
	(2,2)	1+2+3		659,7	0,11	2,1	0	0	0
2-II	(2,2)	1+2	608,6	631,4	0,12	2,0	0	0	0
	(2,2)	1+2+3		**608,6**	0,12	2,1	0	0	0

Tableau 6.2 – Résultats obtenus pour R30

|R|	(|Z^2|, |Z^3|)	**Phases**	**TR*** [s]	**TR** [s]	**Temps**^{2} [s]	**Temps**$_{tot}$ [s]	**GAP**^{2} [%]	**GAP**$_{max}$ [%]	Nb^2_{s-opt} [-]
2	(2,2)	1+2	5361,8	5563,3	3600	7200	9,82	12,09	2
	(2,2)	1+2+3		5320,3	3600	7200	9,82	12,09	2
3	(2,2)	1+2	5361,8	5724,9	2610	7830	2,96	7,932	2
	(2,2)	1+2+3		5460,5	2610	7831	2,96	7,932	2
4	(2,2)	1+2	5361,8	5871,4	203	813	0	0	0
	(2,2)	1+2+3		5461,0	203	814	0	0	0

Tableau 6.3 – Résultats obtenus pour R149

|R|	(|Z^2|, |Z^3|)	**Phases**	**TR*** [s]	**TR** [s]	**Temps**^{2} [s]	**Temps**$_{tot}$ [s]	**GAP**^{2} [%]	**GAP**$_{max}$ [%]	Nb^2_{s-opt} [-]
5	(2,2)	1+2	-	37569,1	3600	18026	33,12	65,15	5
	(2,2)	1+2+3	-	28381,3	3600	18028	33,12	65,15	5
10	(2,2)	1+2	-	31165,1	3562	35615	11,22	20,57	9
	(2,2)	1+2+3	-	28097,0	3562	35617	11,22	20,57	9
15	(2,2)	1+2	-	31161,7	1174	17613	1,01	9,45	3
	(2,2)	1+2+3	-	27929,0	1174	17616	1,01	9,45	3

Tableau 6.4 – Résultats obtenus pour R595

Dans un premier temps, il est possible d'observer que le nombre de sous-régions considéré a un impact important sur la solution obtenue. En effet, dans le cas R149, les meilleures solutions ont été obtenues en considérant un plus petit nombre de sous-régions. Ainsi, plus la division du territoire est grossière, plus on se rapproche du problème global. On observe toutefois le phénomène inverse lorsque R595 est résolu : plus $|R|$ est grand, meilleure est la solution. Dans tous les cas, considérer un plus petit nombre de sous-régions mène aussi à des sous-problèmes de plus grande taille et, par conséquent, plus difficiles à résoudre. En effet, dans le cas R149, lorsque deux sous-régions sont prises en compte, deux instances n'ont pu être résolues à l'optimalité en imposant un temps limite de 3600 secondes. Le GAP moyen et maximal est alors de

9,82 % et 12,09 % respectivement. Le phénomène est encore plus important pour R595. Dans le cas où $|R| = 5$, aucune instance n'a pu être résolue à l'optimalité : le GAP moyen et maximal demeure relativement important, soit respectivement 33,12 % et 65,15 %. Ceci peut donc expliquer, du moins partiellement, la dégradation de la solution en considérant un nombre plus petit de sous-régions, ce qui est contre-intuitif *a priori*. La décomposition du problème influence donc clairement la qualité de la solution finale, bien que le comportement soit difficilement prévisible.

Naturellement, le fait de décomposer le problème en sous-problèmes mène à une perte d'optimalité. Lorsque tous les sous-problèmes sont résolus à l'optimalité, la perte en optimalité observée, c'est-à-dire l'écart entre la solution obtenue grâce à l'approche matheuristique complète et la solution optimale pour le problème original, est directement liée à la décomposition. Lorsque tous les sous-problèmes de la phase 2 sont résolus à l'optimalité, soit pour R30-I, R30-II et R149-4, la perte en optimalité varie entre 0 % et 9,27 %. La solution optimale du problème global a pu être déterminée pour R30-II. Pour l'instance R149, même dans les cas où les sous-problèmes ne sont pas tous résolus à l'optimalité, la solution fournie grâce à l'approche matheuristique, en lançant les 3 phases, permet même d'améliorer la borne supérieure déterminée au chapitre précédent de 0,8 %. Dans le pire des cas, l'écart par rapport à la borne supérieure trouvée au chapitre précédent est de 3,95 %.

Enfin, l'expérimentation a permis de constater que la phase 3 contribue à améliorer considérablement la valeur de la solution finale. En effet, selon l'instance étudiée, une amélioration allant jusqu'à 15 % par rapport à la solution obtenue à la suite de la phase 2 peut être notée. L'impact de la phase 3 semble être encore plus marqué lorsque le nombre de sous-srégions augmente. En effet, plus il y a de frontières, plus il devient avantageux de revoir les listes. Ainsi, lorsqu'il est possible de le faire en pratique, revoir les listes de préaffectation présente un intérêt clair.

6.2.2 Impact de la limite imposée sur le temps de calcul pour la résolution des sous-problèmes de la phase 2

En pratique, le temps de calcul peut devenir un élément crucial lors de la mise en place d'une stratégie de gestion dynamique, telle que celle basée sur le PRPA. Les résultats présentés à la sous-section précédente ont permis de constater que, selon le niveau de décomposition et les instances traitées, le temps de résolution demeure important. De plus, le temps limite imposé pour la résolution des sous-problèmes de la phase 2 est tel qu'il n'est pas toujours possible de les résoudre à l'optimalité. Un GAP important demeure pour certaines instances, malgré un temps de résolution relativement long.

Une option rendue possible par l'approche de décomposition proposée afin d'améliorer le temps de calcul est la parallélisation. Grâce à cette technique, chaque sous-problème de la phase 2 pourrait alors être résolu en parallèle menant ainsi à une réduction importante du temps total de résolution. Ce dernier pourra alors être réduit d'un facteur $|R|$. Néanmoins, malgré la parallélisation, si le temps nécessaire à la résolution d'un sous-problème de la phase 2 est long, le temps de résolution total en est tout autant. Une alternative intéressante afin de réduire le temps de calcul total grâce aux méthodes déjà en place consiste à réduire la limite imposée sur le temps de calcul pour la résolution des sous-problèmes de la phase 2. L'impact de cette limite sur la qualité de la solution est analysé à la présente section. Trois valeurs pour le temps limite sont alors évaluées : 300 secondes, 900 secondes et 3600 secondes (*c.f.* sous-section précédente). Puisque les résultats obtenus à la sous-section précédente ont permis de montrer que les meilleures solutions pour l'instance de taille réelle (R595) ont été obtenues en décomposant le problème en 15 sous-régions, ce scénario est considéré ici. Le Tableau 6.5 présente les résultats obtenus pour différentes limites sur le temps de calcul. De manière similaire aux tableaux présentés à la sous-section précédente, le Tableau 6.5 reporte d'abord les caractéristiques principales de l'instance soit le temps limite imposé pour la résolution de chaque problème de la phase 2 (T^2_{lim}) et les valeurs choisies pour $|Z^2|$ et $|Z^3|$. Le tableau présente ensuite les valeurs du temps de réponse, de même qu'un ensemble de mesures en lien avec le temps de résolution et l'écart entre la solution obtenue et la solution optimale pour chaque sous-problème. Ces mesures ont été définies précédemment.

T^2_{lim}	($\|Z^2\|$, $\|Z^3\|$)	**Phases**	**TR** [s]	**Temps**^{2} [s]	**Temps**$_{tot}$ [s]	**GAP**^{2} [%]	**GAP**$_{max}$ [%]	Nb^2_{s-opt} [-]
300	(2,2)	1+2	31277,5	252	3772	5,17	18,99	11
	(2,2)	1+2+3	27929,0	252	3774	5,17	18,99	11
	(3,3)	1+2	43032,7	294	4409	18,35	42,06	14
	(3,3)	1+2+3	37140,4	294	4410	18,35	42,06	14
900	(2,2)	1+2	31185,9	572	8573	2,21	15,62	5
	(2,2)	1+2+3	27858,8	572	8575	2,21	15,62	5
	(3,3)	1+2	42306,8	828	12412	13,31	32,89	13
	(3,3)	1+2+3	37100,7	828	12414	13,31	32,89	13
3600	(2,2)	1+2	31161,7	1174	17613	1,01	9,45	3
	(2,2)	1+2+3	27929,0	1174	17615	1,01	9,45	3
	(3,3)	1+2	42217,2	3019	45285	8,82	30,54	12
	(3,3)	1+2+3	37100,7	3019	45286	8,82	30,54	12

Tableau 6.5 – Résultats obtenus pour différents limites imposées sur le temps de calcul

En observant les résultats présentés au Tableau 6.5, il est possible de constater qu'en effet, la valeur de la solution obtenue à la suite de la phase 2, c'est-à-dire sans mise à jour des listes de préaffectation, diminue au fur et à mesure que le temps imposé augmente. Le GAP moyen et maximal, de même que le nombre de sous-problèmes ne pouvant être résolus à l'optimalité, di-

minuent aussi lorsque la limite imposée sur le temps augmente. Il est alors possible d'observer que le GAP moyen diminue de 5,17 % à 1,01 % lorsque le temps limite passe de 300 à 3600 secondes pour$|Z|^2 = 2$ et de 18,35 % à 8,82 % pour $|Z|^2 = 3$. En contrepartie, la valeur de la solution ne s'améliore que de 0,4 % pour $|Z|^2 = 2$ et de 1,89 % pour $|Z|^2 = 3$. En comparant la valeur des solutions obtenues à la suite de la phase 3 pour les trois cas de figures, on remarque que la solution finale ne s'améliore pas nécessairement avec le temps. En effet, les meilleures solutions sont obtenues en important un temps limite de 900 secondes ($T^2_{lim} = 900$). Les solutions obtenues pour $T^2_{lim} = 300$ et $T^2_{lim} = 3600$ sont, quant à elles, équivalentes. La phase 3 permet donc de pallier, d'une certaine manière, à la sous-optimalité des sous-problèmes de la phase 2.

Tel que discuté précédemment, les solutions obtenues peuvent être fortement influencées par la décomposition du territoire en sous-régions. De plus, même si les sous-problèmes sont résolus de manière optimale, un GAP dû à la décomposition demeure toujours. Ce GAP est difficilement prévisible. Les efforts supplémentaires, en termes de temps total de calcul pour la résolution de la phase 2, ne sont donc pas forcément justifiés. C'est du moins ce que le cas étudié ici nous permet d'affirmer. Lorsque le temps disponible pour la résolution du problème est plus grand, il serait plus avantageux à notre avis d'introduire différents mécanismes d'améliorations de la solution, basés par exemple sur la recherche locale, de manière à améliorer la qualité de la solution globale plutôt que d'augmenter le temps de calcul pour la résolution des problèmes de phase 2. Nous discuterons de cet aspect à la fin de la présente sous-section. En considérant les résultats obtenus ici, dans ce qui suit, un temps de 900 secondes est imposé pour la résolution des sous-problèmes de la phase 2.

6.2.3 Validation de la méthode pour différentes valeurs de paramètres

Les deux premières séries d'expérimentations menées ont permis d'analyser la méthode de résolution en soi, c'est-à-dire la perte d'optimalité due à la décomposition et l'impact du temps de résolution sur la qualité de la solution finale. Dans cette sous-section, nous présentons plutôt une série de tests visant, dans un premier temps, à montrer que la méthode permet de fournir une solution à un problème de taille réelle, soit l'instance R595, pour un ensemble varié de paramètres, et, dans un deuxième temps, à valider les constats émis au *Chapitre* 5. Dans ce cas, deux types de paramètres sont évalués : la taille des listes utilisées lors de la résolution des sous-problèmes des phases 2 et 3 ($|Z^2|$ et $|Z^3|$) et les poids attribués dans la fonction objectif au service et aux efforts de redéploiement (ξ_s et ξ_r). La taille des listes, de même que les poids dans la fonction objectif, sont des paramètres contrôlés par l'utilisateur et issus de décisions

de gestion en opposition aux caractéristiques du système tels que q et W, d'où l'intérêt de leur analyse. De plus, afin d'enrichir l'expérimentation, l'impact des listes de préaffectation est évalué pour 40, 50 et 60 véhicules. Les résultats obtenus pour l'instance de taille réelle (R595) divisée en 15 sous-régions et en imposant un temps limite 900 secondes lors de la résolution des sous-problèmes de la phase 2 sont présentés aux tableaux 6.6 et 6.7. Au Tableau 6.6, on présentera d'abord le nombre de véhicules utilisé ($|K|$), puis les tailles des listes ($|Z^2|$ et $|Z^3|$). Au Tableau 6.7, les poids ξ_s et ξ_r seront plutôt reportés. Dans ce cas, $|Z^2|$ et $|Z^3|$ sont toujours égaux à 2 et la localisation initiale des véhicules est déterminée aléatoirement. Enfin, dans les deux cas, seuls les résultats obtenus pour l'algorithme complet, c'est-à-dire en considérant les 3 phases, sont reportés. Les mesures utilisées pour l'analyse sont les mêmes que celles présentées précédemment. La valeur du temps total de relocalisation total (TL) associée à la solution déterminée est également présentée au Tableau 6.7.

|K|	($\|Z^2\|$, $\|Z^3\|$)	**TR** [s]	**Temps**^{2} [s]	**Temps**$_{tot}$ [s]	**GAP**^{2} [%]	**GAP**$_{max}$ [%]	Nb^2_{s-opt} [-]
40	(1,2)	31260,8	0,16	2,2	0	0	0
	(2,2)	31608,7	440,33	6605,0	2,28	11,63	5
	(1,3)	40626,8	0,16	2,2	0	0	0
	(2,3)	40791,4	440,33	6605,0	2,28	11,63	5
	(3,3)	Non réalisable	-	-	-	-	-
50	(1,2)	28424,5	0,13	2,0	0	0	0
	(2,2)	27858,8	571,50	8572,5	2,21	15,62	5
	(1,3)	37032,5	0,13	2,0	0	0	0
	(2,3)	36296,8	571,50	8572,5	2,21	15,62	5
	(3,3)	37100,7	827,46	12411,9	13,31	32,89	13
60	(1,2)	26330,0	0,15	2,3	0	0	0
	(2,2)	25971,8	779,94	11699,2	5,28	14,85	10
	(1,3)	34394,8	0,15	2,3	0	0	0
	(2,3)	33967,1	779,94	11699,2	5,28	14,85	10
	(3,3)	34132,8	900,00	13501,2	18,35	36,30	15

Tableau 6.6 – Résultats obtenus pour différents nombres de véhicules

Tout d'abord, les résultats présentés au Tableau 6.6 permettent d'observer que l'approche matheuristique proposée permet d'obtenir une solution de bonne qualité pour un problème de taille réelle, soit l'instance R595, ce qui n'était pas le cas au *Chapitre* 5. Bien qu'il ne soit pas possible d'évaluer la valeur de la solution en termes d'optimalité globale, c'est-à-dire l'écart entre la solution obtenue et la solution optimale du problème global, l'écart moyen entre la solution optimale et la solution obtenue pour chaque sous-problème de la phase 2 se situe entre 2,21 % et 5,28 % pour $|Z^2| = 2$, selon la valeur de $|K|$. Cet écart nous paraît raisonnable étant donné les résultats obtenus au chapitre précédent, la taille du problème et le temps limite imposé de 900 secondes. L'approche matheuristique permet donc de résoudre de manière adéquate des instances de taille réelle.

De plus, les résultats obtenus en utilisant $|K| = 50$ (valeur fixée pour tous les tests précédents)

permettent de constater que le fait de considérer des combinaisons variées de $(|Z^2|, |Z^3|)$ assure l'obtention de solutions différentes. Les efforts de calcul sont aussi très différents que l'on fixe le nombre de véhicules sur les listes préaffectation lors de la résolution de la phase 2 à 1 ou à 2. Considérer $|Z^2| = 1$, puis revoir les listes de préaffectation à la phase 3 en utilisant $|Z^3| = 2$ ou $|Z^3| = 3$, selon le cas, permet de fournir une bonne borne supérieure au problème original, c'est-à-dire lorsque $|Z^2| = 2$ ou $|Z^2| = 3$, dans des délais de temps très courts. On parle alors d'un temps total de résolution d'environ 2 secondes. Néanmoins, le fait de considérer $|Z^2| = 2$ améliore la solution obtenue. On parle alors d'une amélioration d'environ 2 % dans le cas où $|Z^2| = 2$ et $|Z^3| = 2$ par rapport au cas où $|Z^2| = 1$ et $|Z^3| = 2$, et ce, malgré un GAP local moyen de 2.21 %. Ceci vient donc confirmer les constats émis au chapitre précédent : le fait de considérer plus d'un véhicule sur les listes de préaffectation peut avoir un impact important sur les décisions de localisation (et conséquemment la valeur de la solution finale). Une amélioration de 2% est également notable lorsque l'on compare le cas $|Z^2| = 1$ et $|Z^3| = 3$ au cas $|Z^2| = 2$ et $|Z^3| = 3$. Toutefois, la solution trouvée grâce à $|Z^2| = 3$ et $|Z^3| = 3$ ne s'améliore pas par rapport aux deux autres. Ceci est probablement dû à la difficulté à résoudre les sous-problèmes de la phase 2. En effet, dans ce cas, le GAP local est encore très grand. On parle de 13,31% en moyenne. Ces observations se confirment aussi lorsque $|K| = 60$. Par contre, lorsque $|K| = 40$, c'est-à-dire lorsque le problème est contraint davantage en termes de capacité du système, la borne inférieure trouvée en considérant $|Z^2| = 1$ est de meilleure qualité que lorsque $|Z^2| = 2$ ou $|Z^2| = 3$.

ξ_s	ξ_r	$(\|Z^2\|, \|Z^3\|)$	**TR** [s]	**TL** [s]	**Temps**^{2} [s]	**Temps**$_{tot}$ [s]	**GAP**^{2} [%]	**GAP**$_{max}$ [%]	Nb^2_{s-opt} [-]
1	0	(2,2)	27858,8	19693	571,5	8573	2,21	15,62	5
0,75	0,25	(2,2)	28491,1	13148	362,3	5674	0,99	8,27	3
0,5	0,5	(2,2)	31412,1	5745	79,6	2098	0,27	3,99	1
0,25	0,75	(2,2)	32041,2	2635	28,2	1145	0	0	0
0	1	(2,2)	35352,7	0	0,1	434	0	0	0

Tableau 6.7 – Résultats obtenus pour différents poids dans la fonction objectif

En observant les résultats présentés au Tableau 6.7, il est possible de constater que le fait de fixer différentes valeurs pour les poids dans la fonction objectif engendre des solutions variées, menant à différents temps de réponse espérés et différents temps de relocalisation. Naturellement, plus le poids attribué aux efforts de relocalisation augmente par rapport au poids attribué au service à la population, plus le temps total de relocalisation diminue. Conséquemment, le temps de réponse espéré se dégrade. En observant la Figure 6.5 où le temps de réponse total a été tracé en fonction du temps total de relocalisation, on remarque aussi que, dès que l'on accepte de modifier un peu le plan de localisation actuel, la qualité de la solution s'améliore dans

une plus grande proportion. Il est donc intéressant de changer la localisation des véhicules disponibles de manière à déterminer une bonne solution. Néanmoins, les efforts supplémentaires liés à la relocalisation ne justifient pas toujours le gain en performance. Ces constats confirment donc les observations discutées au chapitre précédent.

6.2.4 Pistes d'améliorations potentielles

L'approche matheuristique proposée dans ce chapitre est composée de trois phases. Chacune des phases correspond à un sous-problème, issu du problème global, et chaque sous-problème est résolu de manière séquentielle. Aucune boucle de rétroaction ou mécanisme d'amélioration de la solution globale n'a été implantée dans la première version de l'approche. Néanmoins, à notre avis, il pourrait être intéressant d'envisager l'implantation de tels mécanismes. L'expérimentation effectuée dans ce chapitre, de même qu'une série de tests préliminaires, ont permis d'identifier certaines pistes d'améliorations potentielles. Bien que ces améliorations ne sont pas implantées dans le cadre de cette thèse, nous pensons tout de même qu'une discussion à ce sujet s'impose de façon à mettre en lumière le potentiel de la méthode proposée. Dans la présente section, trois pistes d'améliorations potentielles sont discutées. Ces pistes d'améliorations sont en lien avec : l'affectation des véhicules aux sous-régions ; l'amélioration de la résolution pour les sous-problèmes de la phase 2 ; et l'analyse de la division du territoire en sous-régions.

6.2.4.1 Affectation des véhicules aux sous-régions

L'affectation des véhicules aux différentes sous-régions définies a un impact important sur la qualité de la solution obtenue. Par exemple, le fait d'affecter 2, 3 ou 4 véhicules à une sous-région affecte les décisions de localisation et la qualité de la solution déterminée localement pour la sous-région concernée. Cela aura aussi un impact sur la qualité de la solution globale.

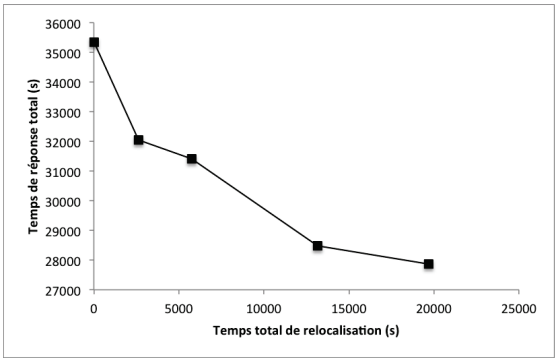


Figure 6.5 – Temps de réponse en fonction du temps de relocalisation

L'analyse effectuée au *Chapitre 5* a d'ailleurs permis de montrer que le nombre de véhicules influence le temps de réponse espéré et que cette relation n'est pas nécessairement linéaire. Des tests préliminaires ont montré que le plan d'affectation initial obtenu à la suite de la phase 1 influence aussi la qualité de la solution finale. En effet, des tests réalisés en utilisant des plans différents de ceux générés par modèle de la phase 1 ont permis d'obtenir des solutions variées. Bien que la phase 1, telle qu'elle est modélisée actuellement, permet d'offrir, dans la plupart des cas testés, la meilleure solution, dans certains cas, de meilleures solutions sont obtenues grâce à d'autres plans d'affectation. Explorer différents plans d'affectation des véhicules aux sous-régions nous paraît donc particulièrement intéressant. Dans cette optique, deux avenues pourraient être empruntées afin d'améliorer la présente méthodologie : (1) l'amélioration du modèle de la phase 1 et l'intégration de boucles de rétroaction ; et (2) l'amélioration de la solution en impliquant le déplacement d'un nombre limité de véhicules.

Afin d'améliorer les solutions fournies à la suite de la phase 1, la relation entre le nombre de véhicules affectés à une sous-région donnée et la qualité de la solution déterminée localement pourrait être mieux évaluée. En effet, tel que mentionné précédemment, il est difficile d'établir à la base l'impact d'affecter x véhicules plutôt que y à une sous-région particulière. Pour ce faire, une des options possibles réside en l'introduction de boucles de rétroaction entre les phases 2 et 1. La résolution des différents sous-problèmes de la phase 2 permettra d'obtenir une meilleure estimation de la relation entre le nombre de véhicules affectés à une sous-région et la qualité de la solution finale. Cette information pourra ensuite être prise en compte afin d'adapter le modèle défini à la phase 1, qui sera résolu à nouveau, et ainsi de suite. L'algorithme pourrait alors itérer entre les phases 1 et 2 jusqu'à ce qu'un critère d'arrêt soit atteint, après quoi, la phase 3 pourra être lancée.

Dans le même ordre d'idée, afin d'améliorer la qualité de la solution finale, des déplacements de véhicules d'une sous-région à une autre pourraient être envisagés à la suite de la phase 2. Un véhicule pourrait, par exemple, être déplacé d'une sous-région où le temps de réponse est très bas par rapport aux autres vers une sous-région moins bien desservie. La phase 1 servira alors à déterminer une bonne solution initiale, puis différentes versions des sous-problèmes de la phase 2 seront résolus de manière successive jusqu'à ce qu'un critère d'arrêt soit atteint. La phase 3 pourra être lancée ensuite. Chaque version des sous-problèmes de la phase 2 différera par l'affectation des véhicules aux différentes sous-régions. Chaque plan d'affectation successif sera déterminé en appliquant différents mécanismes issus de la recherche locale, par exemple la recherche avec tabous ou la recherche à voisinage variable. Dans ce cas, la modification des plans d'affectation sera implantée entre deux résolutions successives de la phase 2. Une autre

alternative pourrait aussi être envisagée afin de modifier la localisation d'un ou de plusieurs véhicules. En effet, le modèle de la phase 3 pourrait être modifié de manière à permettre la relocalisation d'un nombre limité de véhicules lors de sa résolution plutôt que de considérer toutes les localisations comme étant fixes. De cette façon, il sera possible d'évaluer différents voisinages de taille contrôlée, par la résolution successive de la phase 3.

6.2.4.2 Amélioration de la résolution pour les sous-problèmes de la phase 2

Tout au long de l'expérimentation, il a été possible d'observer que la résolution des sous-problèmes appartenant à la phase 2 peut représenter des défis importants. En effet, dans sa forme actuelle, même en utilisant la parallélisation, le temps total de calcul peut devenir considérable. Pour des décisions *a priori*, de nature tactique voire opérationnelle, la méthodologie proposée permet de fournir des solutions dans des délais raisonnables. Par contre, le temps disponible pour la résolution du problème peut devenir insuffisant lorsque des décisions doivent être prises en temps réel. Le développement d'une approche de résolution propre au sous-problème de la phase 2 pourrait alors être envisagé afin de réduire le temps de calcul total. Ainsi, plutôt que de résoudre chaque sous-problème de la phase 2 de manière exacte grâce à un solveur commercial, il pourrait être résolu de manière heuristique. Différentes avenues pourraient être empruntées, par exemple, la recherche avec tabous, la recherche à voisinage variable, le *local branching*. En améliorant la résolution des sous-problèmes de la phase 2, il serait possible de considérer une division plus grossière du territoire, c'est-à-dire d'utiliser un plus petit nombre de sous-régions. Ceci pourrait contribuer à améliorer la solution globalement, au prix de la perte d'une garantie d'optimalité locale.

6.2.4.3 Analyse de la division du territoire en sous-régions

La division du territoire à desservir en sous-régions est un autre élément pouvant influencer de manière importante la qualité de la solution finale obtenue. En effet, l'analyse effectuée pour R30-I et R30-II a bien montré que les deux décompositions sélectionnées permettaient d'obtenir des résultats différents. Dans sa forme actuelle, l'approche proposée utilise une décomposition du territoire en sous-régions contrôlée par l'utilisateur, c'est-à-dire que la décomposition en sous-régions est une donnée du problème. À notre avis, analyser plus en profondeur comment les différents patrons de découpage du territoire peuvent affecter la qualité de la solution et valider si une tendance peut être observée à ce niveau présente un intérêt clair. Autrement dit, il serait intéressant de voir si certains patrons de découpage permettent d'obtenir de meilleures solutions que d'autres. La décomposition du territoire en sous-régions est un problème en soi

qui pourrait être abordée lors de recherches futures, grâce à la méthodologie proposée dans ce chapitre. À terme, une analyse et des recommandations en termes de décomposition du territoire pourraient guider l'utilisateur vers l'obtention de meilleures solutions.

6.3 Le PRPA comme outil d'aide à la décision

Le PRPA défini et formulé au chapitre précédent présente des avantages certains, notamment par son habileté à prendre en compte la capacité des véhicules et à prédire les performances espérées du système. Son utilisation en pratique afin de supporter la prise de décision s'avère donc très intéressante. Il a toutefois été noté lors de l'expérimentation menée au chapitre précédent que des difficultés peuvent être rencontrées lors de la résolution d'instances de taille réelle grâce aux outils mis en place. Afin de pallier à ce problème, une approche matheuristique basée sur une décomposition du problème en sous-régions a été développée afin de fournir une solution à des instances de taille réelle et ainsi permettre son utilisation en pratique. Malgré les pistes d'améliorations potentielles soulevées précédemment, l'approche matheuristique proposée dans ce chapitre s'est avérée une méthode efficace pour résoudre le problème étudié.
Afin de faciliter son utilisation en pratique, un prototype logiciel a été développé. Ce prototype dispose d'une interface graphique qui permet à un utilisateur de modifier facilement les différents paramètres importants et de lancer aisément la résolution du problème. Cet outil a d'ailleurs été utilisé pour lancer les différents tests effectués dans le cadre de ce chapitre. Dans la présente sous-section, le prototype logiciel développé est brièvement décrit, puis une discussion sur son application potentielle en pratique sera présentée.

6.3.1 Présentation du prototype logiciel

Le prototype logiciel développé permet l'utilisation de l'approche matheuristique proposée dans un environnement de travail convivial. Il assure la génération et la résolution automatique des différents sous-problèmes à partir du problème global. Il permet aussi le paramétrage du problème à même son interface. Enfin, il offre à l'utilisateur la possibilité d'effectuer un lancement simultané ou indépendant des trois phases de la méthode, ce qui lui confère un certain niveau de flexibilité. Le prototype logiciel comporte trois composantes principales : la gestion des données ; l'interface et la routine interne ; et la résolution des sous-problèmes. La Figure 6.6 présente l'architecture de l'outil développé de même qu'un aperçu de son interface.
Afin d'utiliser adéquatement l'approche proposée, l'utilisateur doit remplir un fichier global comportant toutes les données correspondant aux caractéristiques du problème : les zones de demande, leur sous-région, leur demande et leurs coordonnées géographiques ; les postes d'attente

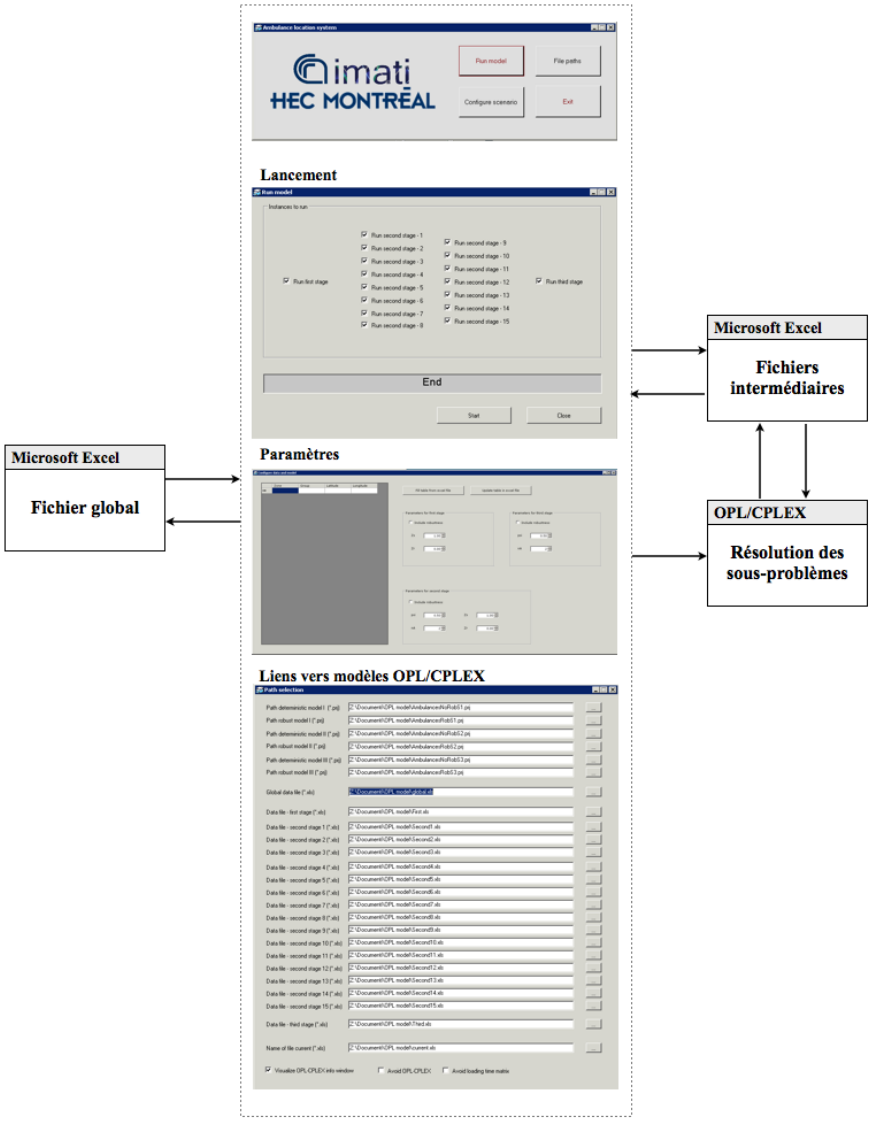


Figure 6.6 – Architecture du prototype logiciel

potentiels et leur capacité ; les véhicules disponibles et leur capacité ; et la matrice des temps de déplacement entre les centroïdes des zones. Ce fichier Excel sera ensuite lu, puis utilisé afin de générer un ensemble de fichiers intermédiaires nécessaires à la résolution des sous-problèmes des phases 1, 2 et 3. En effet, la routine interne du logiciel gère le processus d'agrégation et de désagrégation des données tout au long de l'exécution de l'algorithme. L'interface en soi et la routine interne ont été codés en Visual Basic. L'interface permet à l'utilisateur de modifier au sein de la fenêtre *Paramètres* certaines caractéristiques du système telles que le taux d'occupation q, de même que les paramètres de gestion tels que la taille des listes et les poids attribués aux objectifs pour les différentes phases de l'approche matheurisitque. Des situations variées pourront alors être représentées aisément. L'interface permet aussi de lancer simultanément ou indépendamment chacune des phases de l'approche via la fenêtre *Lancement*. La fenêtre *Liens vers modèles OPL-CPLEX* permet de fournir à l'outil les bons liens vers les modèles nécessaires à la résolution de chaque sous-problème. En fonction des paramètres choisis et du fichier global, la routine interne générera automatiquement les fichiers correspondant à chaque sous-problème, puis appellera OPL-CPLEX pour leur résolution. Les solutions obtenues sont ensuite stockées dans les fichiers intermédiaires, un fichier par sous-problème, puis utilisées au besoin pour la définition des autres sous-problèmes. Enfin, la solution finale est reportée dans le fichier global.

6.3.2 Décisions considérées grâce à l'outil d'aide à la décision

Bien qu'ils aient été développés pour assurer la résolution d'un problème de redéploiement et de préaffectation des véhicules ambulanciers aux demandes urgentes, l'approche de décomposition et l'outil en soi permettent d'aller plus loin et d'analyser différents types de décision, et ce, à tous les niveaux décisionnels : stratégique, tactique et opérationnel (voire temps réel). Parmi les décisions à caractère stratégique notons le dimensionnement de la flotte et le découpage du territoire en sous-régions, parmi les décisions à caractère tactique, l'affectation des véhicules aux sous-régions, la sélection d'un ensemble de postes d'attente potentiels et la détermination d'un plan de déploiement, et enfin, parmi les décisions à caractère opérationnel, la relocalisation des véhicules et l'établissement de listes de préaffectation. Dans cette sous-section, les différentes analyses possibles grâce à l'outil développé sont brièvement discutées.

6.3.2.1 Dimensionnement de la flotte

Le problème de dimensionnement consiste à déterminer le nombre de véhicules à utiliser afin d'assurer un service adéquat à la population. Dans ce contexte, l'outil développé pourrait être

utilisé afin d'analyser différents scénarios en termes de nombre de véhicules utilisés. Afin de considérer différentes tailles de flotte, le nombre de véhicules disponibles devra être modifié dans le fichier global. Comme le problème est considéré au niveau stratégique, la prise en compte des efforts de relocalisation peut devenir non-pertinente. La valeur de ξ_r pourra alors être fixée à 0. La phase 1 affectera alors les véhicules disponibles aux différentes sous-régions, puis la relation entre le nombre de véhicules dans chaque sous-région et les performances espérées du système sera évaluée plus en détail aux phases 2 et 3. Des valeurs de $|Z^2| = |Z^3|$ égale à 2 ou 3 pourraient alors être considérées de manière à fournir une analyse plus précise des performances espérées du système.

6.3.2.2 Découpage du territoire

Tel que discuté précédemment, pour certaines organisations, le découpage du territoire en sous-régions représente un enjeu important. Ceci est particulièrement pertinent pour les organisations qui utilisent une gestion indépendante ou partiellement indépendante des sous-régions. En effet, dans ces cas, le découpage du territoire peut avoir un impact sur les performances du système et l'utilisation des ressources. Le découpage du territoire est une décision qui pourrait aussi être analysée grâce à l'outil proposé. Il a déjà été mentionné que le découpage du territoire est une donnée du problème, incluse dans le fichier global. En faisant varier le nombre de sous-régions et leur composition au sein du fichier global, il sera possible d'évaluer différents scénarios quant au découpage du territoire. Des paramètres tels que ceux mentionnés pour le dimensionnement de la flotte pourront être utilisés ici. Enfin, la phase 3 ne sera lancée que s'il est permis de servir des demandes en provenance d'une sous-région grâce aux véhicules affectés à d'autres sous-régions.

6.3.2.3 Affectation des véhicules aux sous-régions

Lorsque le nombre de véhicules disponibles est fixé, la phase 1 consiste à déterminer l'affectation des véhicules aux différentes sous-régions et, conséquemment, le nombre de véhicules qui seront localisés à chaque sous-région. En omettant la phase 1, c'est-à-dire en la déconnectant du modèle qui permet actuellement de le résoudre, il est possible de tester différents plans d'affectation des véhicules aux sous-régions. Les plans d'affectation testés devront être déterminés manuellement ou grâce à une autre méthode. Les phases 2 et 3 serviront alors à évaluer les plans d'affectation choisis. Vu le caractère tactique de la décision d'affectation, des paramètres similaires à ceux considérés précédemment, par exemple $\xi_r = 0$ et $|Z^2| = |Z^3| = 2$, pourraient être considérés.

6.3.2.4 Sélection d'un sous-ensemble de postes d'attente potentiels ou d'un plan de déploiement

Au chapitre précédent, il a été possible de constater que différentes valeurs des paramètres permettaient souvent d'obtenir des solutions variées. Néanmoins, bien que les solutions puissent être différentes, certains postes d'attente se démarquent et semblent constituer de bonnes alternatives dans plusieurs cas de figure. La résolution du problème original, mais en considérant un ensemble de paramètres variés (par exemple, $|K|$, $|R|$, $|Z|^2$, q), peut donc mener à une analyse intéressante afin d'identifier un sous-ensemble de postes d'attente qui semblent particulièrement intéressants pour la localisation des véhicules ambulanciers. Le nombre de localisations choisies à la suite de l'analyse des plans de déploiement obtenus pour différentes valeurs des paramètres dépendra du contexte d'application. Lorsque les décisions de localisation sont revues en temps réel, il pourrait être intéressant d'identifier *a priori* un sous-ensemble de « bons » postes d'attente. Cela pourra contribuer à réduire le temps de calcul de manière importante. Dans le cas présenté ci-haut, 80 postes d'attente potentiels pourraient, par exemple, être sélectionnés parmi les 218 proposés au départ. Ces 80 postes d'attente pourraient ensuite être utilisés lors de la résolution du problème de redéploiement et de préaffectation lorsque considéré en temps réel. De la même manière, l'outil pourrait être utilisé afin de déterminer le meilleur plan de déploiement pour un nombre donné de véhicules. Lorsque tous les véhicules sont libres pour répondre à des appels urgents, on trouvera un véhicule par poste d'attente. Lorsque certains véhicules deviennent non disponibles, le redéploiement des véhicules, lorsque permis, pourrait se limiter aux postes d'attente appartenant au plan de déploiement original. Des paramètres similaires à ceux proposés pour l'analyse des autres décisions à caractère stratégique et tactique pourront être considérés ici.

6.3.2.5 Redéploiement des véhicules

Le redéploiement des véhicules consiste à déterminer vers quels postes d'attentes diriger ou rediriger les véhicules disponibles pour répondre aux appels urgents. Le PRPA a été formulé afin de prendre en compte à la fois les décisions de relocalisation et les décisions de préaffectation. Il a alors été montré que le fait de considérer les listes de préaffectation peut avoir un impact important sur les décisions de localisation/relocalisation. Dans ce cas, la valeur de ξ_r pourrait, voire devrait, prendre une valeur non nulle. Cette valeur dépendra fortement du contexte d'application et du compromis possible entre le service offert à la population et les efforts de relocalisation. Une analyse *a priori* pourrait être justifiée afin des fixer des valeurs adéquates pour ξ_s et ξ_r.

En pratique, le redéploiement des véhicules pourra être lancé dans différentes circonstances : de manière périodique ou encore à la suite de l'affectation d'un véhicule à un appel. Dans un premier temps, un redéploiement des véhicules peut être lancé à certains moments fixés dans la journée, à chaque changement de périodes par exemple. Les périodes pourront alors correspondre au début et au fin de quart des véhicules. Si toutes les données du problème sont connues à l'avance, pour toutes les périodes considérées, l'outil proposé pourra être utilisé de manière à déterminer le redéploiement des véhicules à chaque début de période en considérant le plan de localisation de la période précédente pour les véhicules déjà en service. La demande associée à chacune des zones et la capacité des véhicules seront alors fixées selon la longueur de la période considérée. Dans le cas d'un redéploiement mutli-période résolu *a priori*, deux véhicules pourront être inclus au sein des listes de préaffectation pour les phases 2 et 3, puisque le temps nécessaire à la résolution est généralement disponible. Une valeur plus grande pour $|Z^3|$ pourra toutefois être fixée si nécessaire. Si les données du problème sont amenées à changer de manière dynamique, l'outil pourra être utilisé de façon similaire, mais en temps réel plutôt qu'*a priori*. Les mêmes paramètres pourront être utilisés, mais en imposant un temps réduit pour la résolution des sous-problèmes de la phase 2 de façon à limiter le temps de calcul.

Le redéploiement des véhicules pourra aussi être lancé lorsque l'état du système le justifie, par exemple, lorsque les performances espérées du système se dégradent en deçà d'un seuil critique. Les performances espérées du système pourront alors être calculées à la suite de l'affectation d'un véhicule, un moment où les performances espérées peuvent se dégrader de manière plus importante. Le redéploiement des véhicules disponibles et en attente pour répondre aux appels d'urgence sera alors lancé afin de retrouver un niveau de service adéquat grâce aux véhicules disponibles, tout en limitant les efforts de relocalisation. À la limite, le redéploiement des véhicules pourra, voire devra, être lancé à chaque affectation d'un véhicule de façon à assurer ou à maintenir un bon niveau de service grâce aux véhicules disponibles. Des paramètres similaires à ceux discutés précédemment pourront être considérés, mais adaptés à un horizon de planification plus court. Certaines limites se posent toutefois quant à la résolution du problème en temps réel, lorsque le temps disponible pour la résolution est court. En effet, selon les paramètres fixés et la taille des problèmes, le temps de résolution demeurent relativement long en comparaison au temp généralement disponible. À cet effet, certaines pistes ont soulevées afin d'accélérer le processus de résolution. Néanmoins, en fonction des mécanismes déjà en place, deux options sont possibles. Tout d'abord, afin de réduire le temps de résolution total, il est possible de réduire le temps limite pour la résolution des sous-problèmes de la phase 2. En effet, l'expérimentation a permis d'observer que, même en réduisant le temps limite pour la résolu-

tion des sous-problèmes de la phase 2, il est possible, grâce à la phase 3 en partie, d'obtenir de bons résultats. Une autre option consiste à utiliser $|Z^2| = 1$, puis $|Z^3| = 2$ ou 3. Cela permettra d'obtenir une bonne borne supérieure au problème dans des délais de temps très courts. Il s'agit donc d'une bonne alternative à considérer lorsque le temps disponible à la résolution est réduit.

6.3.2.6 Génération et mise à jour des listes de préaffectation

La phase 3 de l'approche permet de revoir les listes de préaffectation en fonction des décisions de localisation déterminées à la suite de la phase 2. En pratique, la phase 3 pourrait aussi être utilisée afin de revoir les listes de préaffectation en temps réel, lorsque l'état du système le justifie, mais ne justifie pas nécessairement le redéploiement des véhicules disponibles et en attente. En effet, le modèle correspondant à la phase 3 peut être résolu très rapidement, généralement en moins d'une seconde, ce qui permet de revoir les listes de préaffectation, en fonction des véhicules disponibles pour répondre aux appels urgents, aussi souvent que nécessaire. La phase 3 pourrait donc être lancée à chaque fois qu'un véhicule est affecté à un appel ou qu'il se libère, ou encore lorsque les performances espérées du système se dégrade en deçà d'un seuil limite. Néanmoins, le fait de revoir seulement les listes de préaffectation ne garantit pas toujours de regagner un niveau de performance acceptable, mais pourra justement aider à identifier les situations où le redéploiement des véhicules deviendra nécessaire. Lors de la résolution de la phase 3, des paramètres similaires à ceux employés pour le problème de redéploiement et de préaffectation correspondant pourront toujours être utilisés.

6.4 Conclusion

Dans ce chapitre, une première approche pour la résolution du problème de redéploiement et de préaffectation des véhicules ambulanciers (PRPA) a été proposée. L'approche matheuristique présentée ici tire profit d'une décomposition du territoire en sous-régions. Cette approche se divise en trois phases : l'affectation des véhicules disponibles aux sous-régions ; la relocalisation des véhicules et l'établissement des listes de préaffectation à l'intérieur des sous-régions ; et la mise à jour des listes de préaffectation. Chacune des trois phases correspond à un modèle de programmation linéaire, formulé dans le présent chapitre et issu d'un modèle global. Chaque sous-problème est résolu grâce à un solveur commercial en imposant un temps limite pour la résolution des différents sous-problèmes. Dans le meilleur des cas, tous les sous-problèmes pourront être résolus à l'optimalité : l'écart entre la solution trouvée et la solution optimale au problème global n'étant qu'une conséquence de la décomposition du problème.

Les résultats obtenus grâce à une série d'instances pseudo-réelles ont permis de montrer que la décomposition du problème influence la qualité de la solution obtenue et entraîne une perte d'optimalité. La perte d'optimalité semble toutefois raisonnable. Pour les instances étudiées, cette perte d'optimalité se chiffre entre 0 % et 9,27 %. Pour les instances de plus grande taille, l'approche proposée permet d'améliorer la meilleure borne obtenue au chapitre précédent, à l'intérieur de délais raisonnables. L'expérimentation a aussi permis de montrer que l'approche matheuristique est efficace pour résoudre des instances de taille réelle, ce qui n'était pas possible grâce au modèle et aux outils mis en place au chapitre précédent. Ainsi, deux constats importants ont pu être validés grâce à l'instance de taille réelle. Tout d'abord, le fait de considérer plus d'un véhicule sur les listes de préaffectation peut avoir un impact important sur les décisions de localisation et, conséquemment, sur la valeur de la solution finale. Dans un deuxième temps, le fait de considérer différents poids au sein de la fonction objectif affecte les solutions finales. On constate alors que dès que l'on accepte de modifier un peu le plan de localisation actuel, la qualité de la solution s'améliore dans une plus grande proportion. Les efforts de relocalisation sont moins justifiés par la suite. Enfin, l'expérimentation a permis d'identifier différentes pistes d'améliorations potentielles qui pourront être abordées dans le cadre de recherches futures.

Bien que l'approche et le prototype logiciel présentés dans ce chapitre se soient avérés efficaces, ils ne constituent qu'une première approche pour la résolution du PRPA et son application en pratique. Un certain nombre de pistes d'améliorations potentielles ont déjà été identifiéss et discutées dans ce chapitre. Une des avenues qui nous semble prioritaire repose sur l'intégration de mécanismes empruntés aux métaheuristiques classiques telles que la recherche avec tabous ou la recherche à voisinage variable. En effet, à notre avis, c'est grâce à ce type de mécanismes que l'approche matheuristique proposée pourra démontrer le potentiel de l'intégration de la programmation mathématique et des métaheuristiques. De plus, les différents modèles, tels qu'ils ont été définis ici, ne considèrent qu'implicitement l'incertitude en lien avec l'arrivée des demandes et la non-disponibilité des véhicules. Il pourrait alors être intéressant d'aller un peu plus loin en ce sens et d'intégrer plus explicitement, au sein des modèles, cette notion d'incertitude inhérentes aux SPU, en utilisant, par exemple, la programmation robuste. C'est du moins une des avenues que nous comptons explorer dans le futur.

CONCLUSION

La présente thèse s'est intéressée au développement et à l'analyse de stratégies de gestion afin de réagir et de s'adapter, dans un cadre clair et formel, aux réalisations possibles de l'incertitude en cours d'opération et à l'évolution du système dans le temps. Dans un premier temps, les stratégies de gestion réactives ont été étudiées dans le contexte d'un problème général de localisation, en présence de perturbations. Le concept de réoptimisation contrôlée a alors été défini, puis utilisé pour la formulation du problème de localisation perturbé. Plus concrètement, la réoptimisation contrôlée a été définie de manière à limiter ou à contrôler la différence entre deux solutions successives afin d'en faciliter la mise en oeuvre et de limiter les coûts d'implantation. Une méthode de résolution, utilisant une approche de génération de colonnes, a ensuite été développée afin de déterminer une solution au problème étudié. Cette méthode, flexible et générique, pourra également être utilisée pour résoudre différents problèmes dont la structure est similaire à celui étudié ici. Les résultats obtenus à la suite de l'analyse du problème de localisation perturbé ont permis de montrer que le fait de maintenir les décisions initialement établies lorsque l'état du système évolue ou qu'un événement imprévisible survient n'est pas sans conséquence. En effet, cela peut mener à une augmentation importante des coûts lorsqu'on les compare à ceux d'une réoptimisation complète. Ensuite, dès que l'on accepte de modifier un peu le plan d'opération, la qualité de la solution s'améliore significativement, puis cette amélioration continue d'augmenter, mais de moins en moins au fur et à mesure que le plan d'opération s'écarte du plan initial. Revoir la solution en cours d'opération peut donc mener à des bénéfices importants, mais des efforts non négligeables sont souvent nécessaires afin de les rendre possibles. Les différents mécanismes issus de la réoptimisation contrôlée permettent toutefois de les limiter, dans une certaine mesure.

Dans un deuxième temps, les stratégies de gestion dynamiques, appliquées en cours d'opération, ont été étudiées dans le contexte particulier de la gestion des services préhospitaliers d'urgence (SPU). Dans le cas de la gestion des SPU, l'utilisation efficiente des véhicules ambulanciers disponibles, de même que leur localisation sur le territoire à desservir, représentent des aspects critiques. Les SPU opèrent dans un environnement en constante évolution (c'est-à-dire fortement dynamique) étant donné la nature incertaine de la demande, des temps d'intervention et de la disponibilité des ressources, ce qui complexifie le prise de décision. La capacité d'une telle organisation à répondre adéquatement aux appels de détresse est donc en lien étroit avec sa capacité à anticiper puis à réagir aux différentes sources d'incertitude.

Dans le contexte des SPU, différentes stratégies de gestion opérationnelle prenant en compte l'aspect dynamique du problème ont été définies, puis modélisées dans un cadre commun. Ces stratégies ont ensuite été analysées par simulation. Les résultats obtenus ont permis de quantifier les bénéfices des stratégies de gestion plus dynamiques et flexibles par rapport aux stratégies statiques, et ce, dans divers contextes. Il a alors été possible d'observer que le fait de considérer l'implantation de stratégies de localisation et de relocalisation dynamiques au sein d'un SPU peut devenir équivalent à une augmentation du nombre de véhicules disponibles. Elles mènent aussi à de meilleures performances lorsqu'on les compare aux stratégies statiques. Les stratégies de gestion dynamiques génèrent toutefois une augmentation notable des inconvénients ressentis par le personnel paramédical, notamment une augmentation de la distance totale parcourue à vide. La meilleure stratégie à adopter dépendra alors du contexte étudié et de l'organisation impliquée dans le processus à gérer. L'étude menée ici a montré que les outils de simulation, tels que celui développé dans le contexte de cette thèse, permettent de fournir un ensemble d'informations d'une grande valeur afin de soutenir une prise de décision éclairée. Elle permet aussi de confirmer l'intérêt des stratégies dynamiques, telles que le redéploiement, et ce, malgré les inconvénients qui peuvent y être associés.

Pour donner suite à ce projet, un modèle de décision pour la gestion des SPU qui considère, notamment, le redéploiement dynamique des véhicules ambulanciers a été proposé. Ce modèle se distingue des modèles proposés précédemment puisqu'il considère, en plus des décisions de relocalisation, une préaffectation anticipative des demandes éventuelles sous la forme de listes. Il considère aussi plus explicitement la capacité « réelle » des véhicules à servir des demandes. Au sein de ce modèle, les performances espérées de l'ensemble du système sont calculées, en partie, grâce au temps de réponse espéré, une mesure définie dans cette thèse. Afin de limiter les coûts associés au redéploiement des véhicules, le temps total de redéploiement est également pris en compte dans la fonction objectif. L'analyse du modèle proposé a permis de constater que le fait de considérer plus d'un véhicule dans les listes de préaffectation influence les décisions de localisation et, conséquemment, les performances du système. La capacité des véhicules influence aussi les décisions, mais dans les cas étudiés ici, ce sont principalement les décisions de préaffectation qui sont touchées par la capacité du système. De cette manière, les performances prédites du système se rapprochent davantage de ses performances réelles. Enfin, la fonction objectif et les poids attribués aux différents objectifs influencent clairement les décisions. L'expérimentation a aussi permis de constater que l'outil utilisé pour résoudre le modèle dans sa forme actuelle, soit un solveur commercial, ne permet pas de traiter toutes les instances étudiées dans des délais de temps raisonnables. Une approche de type matheuristique

a donc été conçue afin de résoudre des instances de tailles réelles dans des délais raisonnables.

L'approche matheuristique proposée pour la résolution du problème de redéploiement et de préaffectation tire profit d'une décomposition du territoire à desservir en sous-régions. Cette approche se divise en trois phases soit l'affectation des véhicules disponibles aux sous-régions, la relocalisation des véhicules et l'établissement des listes de préaffectation à l'intérieur des sous-régions, et la mise à jour des listes de préaffectation. Chacune des trois phases correspond à un modèle de programmation linéaire, issu du modèle de décision global. Les résultats obtenus grâce à une série d'instances pseudo-réelles ont permis de constater que la décomposition du problème influence la qualité de la solution obtenue et entraîne une perte d'optimalité. Cette perte d'optimalité semble toutefois tout à fait raisonnable dans le contexte, soit de l'ordre de 0 % à 10 %. L'expérimentation a aussi permis de montrer que l'approche matheuristique s'avère un outil efficace pour la résolution d'instances de taille réelle.

Il ressort clairement de cette thèse que les stratégies de gestion dynamiques et flexibles présentent des avantages considérables. Elles permettent de prendre ou de modifier des décisions en cours d'opération, en fonction de l'état du système. Ainsi, un meilleur niveau de service pourra être maintenu en tout temps. Néanmoins, ces stratégies amènent aussi des inconvénients et des efforts supplémentaires qui peuvent prendre différentes formes : impact économique, inconvénients perçus par les employés, etc. Il est donc crucial pour les organisations de se doter d'un cadre clair et formel afin de prendre les meilleures décisions possibles en temps réel, tout en limitant les impacts négatifs sur le système dans son ensemble. La reoptimisation contrôlée offre un cadre intéressant pour prendre de telles décisions. La mise en oeuvre en pratique de stratégies de gestion dynamiques présente également des défis importants tels que la nécessité de disposer de technologies de communications adéquates, le traitement de l'information en temps réel, la nécessité d'un temps de réaction court et d'une prise de décision rapide, pour ne nommer que quelques exemples. Des analyses telles que celles menées dans le cadre de cette thèse permettent toutefois de démontrer le potentiel de ces stratégies dans un contexte particulier et de justifier leur implantation en pratique, le cas échéant. Cette implantation est d'autant plus justifiée dans des contextes où des vies peuvent être en jeu et où chaque seconde compte, tel que ceux des SPU.

Cette thèse a amorcé une discussion importante sur les pratiques de gestion permettant à une organisation de réagir et de s'adapter aux réalisations possibles de l'incertitude et à l'évolution du système dans le temps. L'intérêt d'aller plus loin dans le développement de stratégies et de méthodes afin de soutenir la prise de décision dynamiquement est clair, et cette thèse n'a que commencé à aborder le sujet. Plusieurs avenues de recherche pourront être empruntées

à la suite de cette thèse. Des pistes de recherche futures ont déjà été soulevées à la fin de chaque chapitre. De manière générale, les méthodologies développées pourront être menées plus loin, puis appliquées à d'autres contextes où elles seront justifiées. À terme, les méthodes développées dans le cadre de cette thèse pourront, à mon avis, par leur implantation en pratique, contribuer à l'amélioration du service à la population. Néanmoins, il reste encore beaucoup à faire dans ce registre.

BIBLIOGRAPHIE

Aboolian, R., T. Cui et Z.-J. M. Shen. 2013, «An efficient approach for solving reliable facility location models», *INFORMS Journal on Computing*, vol. 25, p. 720–729.

Aboueljinane, L., E. Sahin et Z. Jemai. 2013, «A review of simulation models applied to emergency medical service operations», *Computers & Industrial Engineering*, vol. 66, p. 734–750.

Aboueljinane, L., E. Sahin, Z. Jemai et J. Marty. 2014, «A simulation study to improve the performance of an emergency medical service : Application to the French Val-de-Marne department», *Simulation Modelling Practice and Theory*, vol. 47, p. 46–59.

Acuna-Agost, R., P. Michelon, D. Feillet et S. Gueye. 2011, «SAPI : Statistical analysis of propagation of incidents. A new approach for rescheduling trains after disruptions.», *European Journal of Operational Research*, vol. 215, p. 227–243.

Agar, M. et S. Salhi. 1998, «Lagrangean heuristics applied to a variety of large capacitated plant location problems.», *Journal of Operational Research Society*, vol. 49, p. 1072–1084.

Agra, A., M. Christiansen, R. Figueiredo, L. M. Hvattum, M. Poss et C. Requejo. 2013, «The robust vehicle routing problem with time windows», *Computers & Operations Research*, vol. 40, p. 856–866.

Ahuja, R., J. B. Orlin, S. Pallottino et M. P. Scutellà. 2004, «A multi-exchange heuristic for the single-source capacitated facility location problem», *Management Science*, vol. 50, p. 749–760.

Alanis, R., A. Ingolfsson et B. Kolfal. 2013, «A Markov chain model for EMS system with repositioning», *Production and Operations Management*, vol. 22, p. 216–231.

Albers, S. 2003, «Online algorithms : A survey», *Mathematical Programming*, vol. 97, p. 3–26.

Alem, D. J. et R. Morabito. 2012, «Production planning in furniture setting via robust optimization», *Computers & Operations Research*, vol. 39, p. 139–150.

Alsalloum, O. I. et G. K. Rand. 2006, «Extensions to emergency vehicle location model», *Computers & Operations Research*, vol. 33, p. 2725–2743.

Altay, N. et W. G. Green. 2006, «OR/MS research in disaster operations management», *European Journal of Operational Research*, vol. 175, p. 475–493.

Altiok, T. et B. Melamed. 2007, *Simulation Modeling and Analysis with Arena*, Academic Press, Amsterdam.

Anaya-Arenas, A. M., J. Renaud et A. Ruiz. 2012, «Relief distribution networks : A systematic review», Document de travail CIRRELT-2012-55, CIRRELT.

Andersson, T. et P. Värbrand. 2007, «Decision support tools for ambulance dispatch and relocation», *Journal of the Operational Research Society*, vol. 58, p. 195–201.

Andradóttir, S. 1998, «Simulation Optimization», dans *Handbook of simulation*, édité par J. Banks, Wiley, New York, N.Y., p. 307–334.

Audet, I. 2009, «Quatre Montréalais sur dix boudent le vaccin», *La Presse*.

Averbakh, I. et O. Berman. 2000, «Minmax regret median location on network under uncertainty», *INFORMS Journal on Computing*, vol. 12, p. 104–110.

Aytug, H. et C. Saydam. 2002, «Solving large-scale maximum expected covering location problem by genetic algorithms : A comparative study», *European Journal of Operational Research*, vol. 141, p. 480–494.

Başar, A., B. Çatay et T. Ünlüyurt. 2011, «A multi-period double coverage approach for locating the emergency medical service stations in Istanbul», *Journal of the Operational Research Society*, vol. 62, p. 627–637.

Başar, A., B. Çatay et T. Ünlüyurt. 2012, «A taxonomy for emergency service station location problem», *Operations Research Letters*, vol. 6, p. 1147–1160.

Balcik, B. et B. M. Beamon. 2008, «Facility location in humanitarian relief», *International Journal of Logistics : Research and Applications*, vol. 2, p. 101–121.

Ball, M. O., C. Barnhart, G. L. Nemhauser et A. R. Odoni. 2007, «Air transportation : Irreguler operations and control», dans *Handbooks in Operations Research and Management Science*, vol. 14, édité par C. Barnhart et G. Laporte, North-Holland, Amsterdam, p. 1–67.

Ball, M. O. et L. F. Lin. 1993, «A reliability model applied to emergency service vehicle location», *Operations Research*, vol. 41, p. 18–36.

Barceló, J. et J. Casanovas. 1984, «A heuristic lagrangean algorithm for the capacitated plant location problem», *European Journal of Operational Research*, vol. 15, p. 212–226.

Bard, J., G. Yu et M. F. Arguello. 2001, «Optimizing aircraft routing in response to groundings and delays», *IIE Transactions on Operations Engineering*, vol. 33, p. 931–947.

Baron, O., J. Milner et H. Naseraldin. 2011, «Facility location : A robust optimization approach», *Production and Operations Management*, vol. 20, p. 772–785.

Batta, R., J. M. Dolan et N. N. Krishnamurty. 1989, «The maximal expected covering location problem : Revisited», *Transportation Science*, vol. 23, p. 277–287.

Beale, E. 1955, «On minimizing a convex function subject to linear inequalities», *Journal of the Royal Statistical Society Serie B*, vol. 17, p. 173–184.

Beasley, J. E. 1993, «Lagrangean heuristics for location problems», *European Journal of Operational Research*, vol. 65, p. 383–399.

Beaudry, A., G. Laporte, T. Melo et S. Nickel. 2009, «Dynamic transportation of patients in hospitals», *OR Spectrum*, vol. 32, p. 77–107.

Begen, M. 2011, «Stochastic dynamic programming models and applications», dans *Wiley Encyclopedia of Operations Research and Management Science*, édité par J. J. Cochran, P. Keskinocak, J. P. Kharoufeh et J. C. Smith, John Wiley & Sons, Hoboken, N.J.

Bélanger, V., A. Ruiz et P. Soriano. 2012, «Déploiement et redéploiement des véhicules ambulanciers dans la gestion des services préhospitaliers d'urgence», *INFOR*, vol. 50, p. 1–30.

Bell, C. N. et D. Allen. 1969, «Optimal planning of the emergency ambulance services», *Socio-Economic Planning Sciences*, vol. 3, p. 95–101.

Bellman, R. 1957, *Dynamic programming*, Princeton University Press, Princeton, N.J.

Ben-Tal, A., L. E. Ghaoui et A. Nemirovski. 2009, *Robust Optimization*, Princeton University Press, Princeton, N.J.

Ben-Tal, A., A. Goryashko, E. Guslitzer et A. Nemirovski. 2004, «Adjustable robust solution of uncertain linear programs», *Mathematical Programming Serie A*, vol. 99, p. 351–376.

Ben-Tal, A. et A. Nemirovski. 1998, «Robust convex optimization», *Mathematics of Operations Research*, vol. 23, p. 769–805.

Ben-Tal, A. et A. Nemirovski. 1999, «Robust solutions to uncertain programs», *Operations Research Letters*, vol. 25, p. 1–13.

Ben-Tal, A. et A. Nemirovski. 2000, «Robust solutions of linear programing problems contaminated with uncertain data», *Mathematical Programming*, vol. 88, p. 411–424.

Ben-Tal, A. et A. Nemirovski. 2002, «Robust optimization - methodology and applications», *Mathematical Programming*, vol. 92, p. 453–480.

Beraldi, P. et M. E. Bruni. 2009, «A probabilistic model applied to emergency service vehicle location», *European Journal of Operational Research*, vol. 196, p. 323–331.

Beraldi, P., M. E. Bruni et D. Conforti. 2004, «Designing robust emergency medical service via stochastic programming», *European Journal of Operational Research*, vol. 158, p. 183–193.

Berbeglia, G., J.-F. Cordeau et G. Laporte. 2010, «Dynamic pickup and delivery problems», *European Journal of Operational Research*, vol. 202, p. 8–15.

Berlin, G. N. et J. C. Liebman. 1974, «Mathematical analysis of emergency ambulance location», *Socio-Economic Planning Sciences*, vol. 8, p. 323–328.

Berman, O. et D. Krass. 2001, «Facility location with stochastic demands and congestion», dans *Facility Location. A Survey of Applications and Theory*, édité par Z. Drezner et H. W. Hamacher, Springer, New York, N.Y., p. 329–372.

Berman, O., D. Krass et M. B. C. Menezes. 2007, «Facility reliability issues in network p-median problems : Strategic centralization and co-location effects», *Operations Research*, vol. 55, p. 332–350.

Bertsekas, D. P. 1987, *Dynamic programming : deterministic and stochastic models*, Prentice-Hall, Englewood Cliffs, N.J.

Bertsekas, D. P. 2012, *Dynamic programming and optimal control*, Athena Scientific, Belmont, MA.

Bertsekas, D. P. et J. N. Tsitsiklis. 1996, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA.

Bertsimas, D., D. B. Brown et C. Caramanis. 2011, «Theory and applications of robust optimization», *SIAM Review*, vol. 53, p. 464–501.

Bertsimas, D. et M. Sim. 2004, «The price of robustness», *Operations Research*, vol. 52, p. 35–53.

Beyer, H.-G. et B. Sendhoff. 2007, «Robust optimization - A comprehensive survey», *Computational Methods Applied to Mechanical Engineering*, vol. 196, p. 3190–3218.

Bianchi, G. et R. L. Church. 1988, «A hybrid fleet model for emergency medical service system design», *Social Sciences & Medicine*, vol. 26, p. 163–171.

Birge, J. R. 1997, «Stochastic programming : Computation and Applications», *INFORMS Journal on Computing*, vol. 9, p. 111–133.

Birge, J. R. et F. V. Louveaux. 2011, *Introduction to Stochastic Programming*, Springer, New York, N.Y.

Bohle, C., S. Maturana et J. Vera. 2010, «A robust optimization approach to wine grape harvesting scheduling», *European Journal of Operational Research*, vol. 200, p. 245–252.

Boloori Arabani, A. et R. Z. Farahani. 2012, «Facility location dynamics : An overview of classifications and applications», *Computers & Operations Research*, vol. 62, p. 408–420.

Borràs, F. et J. T. Pastor. 2002, «The ex-post evalulation of the minimum location reliability : An enhanced probabilistic location set covering model», *Annals of Operations Research*, vol. 111, p. 51–74.

Boschetti, M. A., A. Mingozzi et S. Ricciardelli. 2008, «A dual ascent procedure for the set partitioning problem», *Discrete Optimization*, vol. 5, p. 735–747.

Brotcorne, L., G. Laporte et F. Semet. 2003, «Ambulance location and relocation models», *European Journal of Operational Research*, vol. 147, p. 451–463.

Cameron, D. 2009, «Le tiers de la population est vacciné», *Cyberpresse*. URL `www.cyberpresse.ca`.

Carpentier, G. 2006, *La conception et la gestion d'un réseau de service ambulancier*, mémoire de maîtrise, Université Laval.

Carrizosa, E. et S. Nickel. 2003, «Robust facility location», *Mathematical methods in Operations Research*, vol. 58, p. 331–349.

Ceselli, A., F. Liberatore et G. Righini. 2009, «A computational evaluation of a general branch-and-price framework for capacitated location problems», *Annals of Operations Research*, vol. 167, p. 209–251.

Charnes, A. et W. W. Cooper. 1959, «Chance-constrained programming», *Management Science*, vol. 5, p. 73–79.

Chen, B. et C.-S. Lin. 1998, «Minmax-regret robust 1-median location on a tree», *Networks*, vol. 31, p. 93–103.

Chen, C. et C. Ting. 2008, «Combining lagrangian heuristic and ant colony system to solve the single source capacitated location problem», *Transportation Research Part E*, vol. 44, p. 1099–1122.

Chen, K. et T. Xiao. 2009, «Demand disruption and coordination of the supply chain with a dominant retailer», *European Journal of Operational Research*, vol. 197, p. 225–234.

Chiyoshi, F. Y., R. D. Galvão et R. Morabito. 2002, «A note on solution to the maximal expected covering location problem», *Computers & Operations Research*, vol. 30, p. 87–96.

Church, R. L. et C. S. ReVelle. 1974, «The maximal covering location problem», *Papers of Regional Science Association*, vol. 32, p. 101–118.

Clausen, J., J. Hansen, J. Larsen et A. Larsen. 2001, «Disruption management», *OR/MS Today*, vol. 28, p. 40–43.

Clausen, J., A. Larsen, J. Larsen et N. J. Rezanova. 2010, «Disruption management in the airline industry - concepts, models and methods», *Computers & Operations Research*, vol. 37, p. 809–821.

Concannon, K., M. Elder, K. Hindle, J. Tremble et S. Tse. 2007, *Simulation Modeling with Simul8*, Visual Thinking International, Missisauga, ON.

Contreras, I. et J. A. Diaz. 2008, «Scatter search for the single source capacitated facility location problem», *Annals of Operations Research*, vol. 157, p. 73–89.

Cordeau, J.-F. et G. Laporte. 2007, «The dial-a-ride problem : Models and algorithms», *Annals of Operations Research*, vol. 153, p. 29–46.

Cortinhal, M. J. et M. E. Captivo. 2003, «Upper and lower bound for the single source capacitated location problem», *European Journal of Operational Research*, vol. 151, p. 333–351.

Crainic, T., X. Fu, M. Gendreau, W. Rei et S. W. Wallace. 2011, «Progressive hedging-based metaheuristics for stochastic network design», *Networks*, vol. 58, p. 114–124.

Cui, T., Y. Ouyang et Z.-J. M. Shen. 2010, «Reliable facility location under the risk of disruptions», *Operations Research*, vol. 58, p. 998–1011.

Current, J., M. S. Daskin et D. A. Schilling. 2001, «Discrete network location models», dans *Facility Location. Applications and Theory*, édité par Z. Drezner et H. W. Hamacher, Springer, Berlin, p. 81–108.

Daganzo, C. F. et G. F. Newell. 1986, «Configuration of physical distribution networks», *Networks*, vol. 16, p. 113–132.

Dantzig, G. 1955, «Linear programming under uncertainty», *Management Science*, vol. 1, p. 197–206.

Dantzig, G. B. et J. H. Ramser. 1959, «The truck dispatching problem», *Management Science*, vol. 6, p. 80–91.

Daskin, M. S. 1982, «Application of an expected covering model to emergency medical service design», *Decision Sciences*, vol. 13, p. 416–439.

Daskin, M. S. 1983, «A maximum expected location problem : Formulation, properties and heuristic solution», *Transportation Science*, vol. 17, p. 416–439.

Daskin, M. S. 1987, «Location, dispatching, and routing models for emergency services with stochastic travel times», dans *Spatial Analysis and Location-Allocation Models*, édité par A. Ghosh et G. Rushton, Van Nostrand Reinhold, New York, N.Y., p. 224–265.

Daskin, M. S. 2008, «What you should know about location modeling», *Naval Research Logistics*, vol. 55, p. 283–294.

Daskin, M. S., K. Hogan et C. S. ReVelle. 1988, «Integration of multiple, excess, backup and expected covering models», *Environment and Planning B*, vol. 15, p. 15–35.

Daskin, M. S. et E. H. Stern. 1981, «A hierarchical objective set covering model for emergency medical service vehicle deployment», *Transportation Science*, vol. 15, p. 137–152.

Delmaire, H., J. A. Diaz, E. Fernandez et M. Ortega. 1999, «Reactive GRASP and Tabu Search based heuristic for the single source capacitated plant location problem», *INFOR*, vol. 37, p. 194–225.

Denardo, E. V. 1982, *Dynamic programming : models and applications*, Prentice-Hall, Englewood Cliffs, N. J.

Diaz, J. A. et E. Fernandez. 2002, «A branch-and-price algorithm for the single source capacitated plant location problem», *Journal of the Operational Research Society*, vol. 53, p. 728–740.

Doerner, K. F., W. J. Gutjahr, R. F. Hartl, M. Karall et M. Reimann. 2005, «Heuristic solution of an extended double-coverage ambulance location problem for austria», *Central European Journal of Operations Research*, vol. 13, p. 325–340.

Dreyfus, S. 2002, «Richard Bellman on the birth of dynamic programming», *Operations Research*, vol. 50, p. 48–51.

Drezner, Z. 1987, «Heuristic solution methods for two location problems with unreliable facilities», *Journal of the Operational Research Society*, vol. 38, p. 509–514.

Drezner, Z. et H. W. Hamacher. 2001, *Facility Location. Applications and Theory*, Springer, Berlin.

Dror, M., G. Laporte et P. Trudeau. 1989, «Vehicule routing with stochastic demands : Properties and solution frameworks», *Transportation Science*, vol. 23, p. 166–176.

Durbin, M. et K. Hoffman. 2008, «The dance of the thirty-ton trucks : Dispatching and scheduling in a dynamic environment», *Operations Research*, vol. 56, p. 3–19.

Eaton, D. J., M. S. Daskin, D. Simmons, B. Bulloch et G. Jansma. 1985, «Determining emergency medical deployment in Austin, Texas», *Interfaces*, vol. 15, p. 96–108.

Eaton, D. J., H. M. U. Sanchez, R. R. Lantigua et J. Morgan. 1986, «Determining ambulance deployment in Santo Domingo, Dominican Republic», *Journal of the Operational Research Society*, vol. 37, p. 113–126.

Eiselt, H. A., M. Gendreau et G. Laporte. 1992, «Location of facilities on a network subject to a single-edge failure», *Networks*, vol. 22, p. 231–246.

Erera, A. L., J. C. Morales et M. Savelsbergh. 2009, «Robust optimization for empty repositioning problems», *Operations Research*, vol. 57, p. 468–483.

Erkut, E., A. Ingolfsson, T. Sim et G. Erdogan. 2009, «Computational comparison of five maximal covering models for locating ambulances», *Geographical Analysis*, vol. 41, p. 43–65.

Fiat, A. et G. J. Woeginger. 1998, *Online Algorithms : The State of the Art*, Springer-Verlag, Berlin.

Filar, J., P. Manyem et K. White. 2001, «How airlines and airports recover from schedule perturbations : A survey», *Annals of Operations Research*, vol. 108, p. 315–333.

Fitzsimmons, J. A. 1973, «A methodology for emergency ambulance deployment», *Management Science*, vol. 19, p. 627–636.

Fu, M. C., F. W. Glover et J. April. 2005, «Simulation Optimization : A review, new developments and applications», dans *Proceedings of the 2005 Winter Simulation Conference*, édité par M. E. Kuhl, N. M. Steiger, F. B. Armstrong et J. A. Joines, p. 83–95.

Fujiwara, O., T. Makjamroen et K. K. Gupta. 1987, «Ambulance deployment analysis : A study case of Bangkok», *European Journal of Operational Research*, vol. 31, p. 9–18.

Gabrel, V., C. Murat et A. Thiele. 2014, «Recent advances in robust optimization : An overview», *European Journal of Operational Research*, vol. 235, p. 471–483.

Galindo, G. et R. Batta. 2013, «Review of recent development in OR/MS research in disaster operations management», *European Journal of Operational Research*, vol. 230, p. 201–211.

Galvão, R. D., F. Y. Chiyoshi et R. Morabito. 2005, «Towards unified formulations and extensions of two classical probabilistic location model», *Computers & Operations Research*, vol. 32, p. 15–33.

Galvão, R. D. et C. S. ReVelle. 1996, «A lagrangean heuristic for the maximal covering location problem», *European Journal of Operational Research*, vol. 88, p. 114–123.

Garey, M. R. et D. S. Johnson. 1979, *Computers and Intractability : A guide to the Theory of NP-Completeness*, W. H Freeman and Co., San Francisco, CA.

Gautam, N. 2012, *Analysis of queues : methods and applications*, Taylor & Francis, Boca Raton, FL.

Gendreau, M., O. Jabali et W. Rei. 2015, «Stochastic vehicle routing problem», dans *Vehicle Routing : Problems, methods and applications*, édité par D. Vigo et P. Toth, SIAM, Philadelphie, PA, p. 213–240.

Gendreau, M., G. Laporte et F. Semet. 1997, «Solving an ambulance location model by tabu search», *Location Science*, vol. 5, p. 75–88.

Gendreau, M., G. Laporte et F. Semet. 2001, «A dynamic model and parallel tabu search heuristic for real-time ambulance relocation», *Parallel Computing*, vol. 27, p. 1641–1653.

Gendreau, M., G. Laporte et F. Semet. 2006, «The maximal expected relocation problem for emergency vehicles», *Journal of the Operational Research Society*, vol. 57, p. 22–28.

Gilmore, P. et R. E. Gomory. 1961, «A linear programming approach to the cutting stock problem», *Operations Research*, vol. 9, p. 849–859.

Goldberg, J. 2004, «Operations research models for the deployment of emergency services vehicle», *EMS Management Journal*, vol. 1, p. 20–39.

Goldberg, J., R. Dietrich, J. M. Chen et M. G. Mitwasi. 1990a, «A simulation model for evaluating a set of emergency vehicle base locations : Development, validation and usage», *Socio-Economic Planning Sciences*, vol. 24, p. 124–141.

Goldberg, J., R. Dietrich, J. M. Chen, M. G. Mitwasi, T. Valenzuela et E. Criss. 1990b, «Validating and applying a model for locating emergency medical services in Tucson, AZ», *European Journal of Operational Research*, vol. 49, p. 308–324.

Goldberg, J. et L. Paz. 1991, «Locating emergency vehicle bases when service time depends on call location», *Transportation Science*, vol. 25, p. 264–280.

Golden, B., S. Raghavan et E. Wasil. 2008, *The vehicle routing problem : Latest advances and new challenges*, Springer, London, New York, N.Y.

Gregory, C., K. Darby-Dowman et G. Mitra. 2011, «Robust optimization and portfolio selection : The cost of robustness», *European Journal of Operational Research*, vol. 212, p. 417–428.

Gross, D. 2008, *Fundamentals of queueing theory*, Wiley & Sons, Hoboken, N.J.

Hanne, T., T. Melo et S. Nickel. 2009, «Bringing robustness to patient flow management through optimized patient transport in hospitals», *Interfaces*, vol. 39, p. 241–255.

Harewood, S. I. 2002, «Emergency ambulance deployment in Barbados : a multi-objective approach», *Journal of the Operational Research Society*, vol. 53, p. 185–192.

Harrell, C., B. K. Ghosh et R. Bowden. 2011, *Simulation Using Promodel*, McGraw-Hill, New York, N.Y.

Henderson, S. G. et A. J. Mason. 2004, «Ambulance service planning : simulation and data visualisation», dans *Operations Research and Health care*, édité par M. L. Brandeau, F. Sainfort et W. Pierskalla, Springer, New York, N.Y., p. 77–102.

Hindi, K. et K. Pienkosz. 1999, «Efficient solution of large scale, single source, capacitated plant location problem», *Journal of the Operational Research Society*, vol. 50, p. 268–274.

Hogan, K. et C. S. ReVelle. 1986, «Concepts and application of backup coverage», *Management Science*, vol. 34, p. 1434–1444.

Holmberg, K., M. Rönnqvist et D. Yuan. 1999, «An exact algorithm for the capacitated facility location problems with single sourcing», *European Journal of Operational Research*, vol. 113, p. 544–559.

Howard, R. A. 1960, *Dynamic programming and Markov processes*, The M.I.T Press, Cambridge, MA.

Ingolfsson, A. 2013, «EMS planning and management», dans *Operations Research and Healthcare policy*, édité par G. S. Zaric, Springer, New York, N.Y., p. 105–128.

Ingolfsson, A., S. Budge et E. Erkut. 2008, «Optimal ambulance location with random delays and travel times», *Health Care Management Science*, vol. 11, p. 262–274.

Ingolfsson, A., E. Erkut et S. Budge. 2003, «Simulation of single start station for Edmonton EMS», *Journal of the Operational Research Society*, vol. 54, p. 736–746.

Jespersen-Groth, J., D. Potthoff, J. Clausen, D. Huisman, L. Kroon, G. Maróti et M. N. Nielsen. 2009, «Disruption management in passenger railway transportation», dans *Robust and Online Large-Scale Optimization*, édité par R. Ahuja, R. Möhring et C. Zaroliagis, Springer, Berlin, p. 399–421.

Kall, P. 1982, «Stochastic programming», *European Journal of Operational Research*, vol. 10, p. 125–130.

Kall, P. et S. W. Wallace. 1995, *Stochastic programming*, Wiley, New York, N.Y.

Kelton, W. D., R. P. Sadowski et D. T. Sturrock. 2006, *Simulation with Arena*, McGraw-Hill, New York, N.Y.

Kergosien, Y., V. Bélanger, P. Soriano, M. Gendreau et A. Ruiz. 2015, «A generic and flexible simulation-based analysis tool for EMS management», *International Joural of Production Research, sous presse*.

Kergosien, Y., C. Lenté, D. Piton et J.-C. Billaut. 2011, «A tabu search heuristic fot the dynamic transportation of patiens between care units», *European Journal of Operational Research*, vol. 214, p. 442–452.

Kingman, J. 2009, «The first erlang century - and the next», *Queueing system*, vol. 63, p. 3–12.

Kleywegt, A. J., A. Shapiro et T. Hommem-De-Mello. 2001, «The sample average approximation method for stochastic discrete optimization», *SIAM Journal on Optimization*, vol. 12, p. 479–502.

Klibi, W., S. Ichoua et A. Martel. 2013, «Prepositioning emergency supplies to support disaster relief : A stochastic programming approach», Document de travail CIRRELT-2013-19, CIRRELT.

Klibi, W., A. Martel et A. Guitouni. 2010, «The design of robust value-creating supply chain networks : A critical review», *European Journal of Operational Research*, vol. 203, p. 283–293.

Klincewick, J. et H. Luss. 1986, «A lagrangean relaxation heuristic for the capacitated facility location with single-source constraints», *Journal of Operational Research Society*, vol. 37, p. 495–500.

Kohl, N., A. Larsen, J. Larsen, A. Ross et S. Tiourine. 2007, «Airline disruption management - perspective, experiences and outlook», *Journal of Air Transport Management*, vol. 13, p. 149–162.

Kolesar, P. J. et W. E. Walker. 1974, «An algorithm for the dynamic relocation of fire companies», *Operations Research*, vol. 22, p. 249–274.

Kouvelis, P. et G. Yu. 1997, *Robust Discrete Optimization and its application*, Kluwer Academic Publishers, Boston, MA.

Labbé, M., J.-F. Thisse et R. E. Wendell. 1991, «Sensitivity analysis in minisum facility location problems», *Operations Research*, vol. 39, p. 961–969.

Laporte, G. et F. Louveaux. 1993, «Integer L-shape method for stochastic integer program with complete recourse», *Operations Research Letters*, vol. 13, p. 133–142.

Laporte, G., F. Louveaux et H. Mercure. 1989, «Models and exact solutions for a class of stochastic location-routing problems», *European Journal of Operational Research*, vol. 39, p. 71–78.

Laporte, G., F. V. Louveaux, F. Semet et A. Thirion. 2009, «Applications of the double standard model for ambulance location», dans *Innovations in Distribution Logistics*, édité par L. Bertazzi, M. G. Speranza et J. van Nunen, Springer, Berlin, p. 235–249.

Laporte, G., S. Nickel et F. S. de Gama, éd.. 2015, *Location Science*, Springer, New York, N.Y.

Larson, R. C. 1974, «A hypercube queuing model for facility location and redistricting in urban emergency services», *Computers & Operations Research*, vol. 1, p. 67–85.

Larson, R. C. 1975, «Approximating the performance of urban emergency service systems», *Operations Research*, vol. 23, p. 845–868.

Lavery, E., M. Beaverstock, A. Greenwood et W. Nordgren. 2011, *Applied Simulation : Modeling and Analysis Using FlexSim*, FlexSim Software Products Inc., Orem, UT.

Law, A. M. 2006, *Simulation Modeling & Analysis*, McGraw-Hill, Boston, MA.

Li, X. et Y. Ouyang. 2010, «A continuum approximation approach to reliable facility location design under correlated probabilistic disruptions», *Transportation Research Part B*, vol. 44, p. 535–548.

Li, Z. et M. G. Ierapetritou. 2008, «Robust optimization for process scheduling under uncertainty», *Industrial & engineering chemistry research*, vol. 47, p. 4148–4157.

Lim, M., M. S. Daskin, A. Bassamboo et S. Chopra. 2010, «A facility reliability model : Formulation, properties, and algorithm», *Naval Research Logistics*, vol. 57, p. 58–70.

Lim, M., M. S. Daskin, A. Bassamboo et S. Chopra. 2013, «Facility location decisions with random disruptions and imperfect estimation», *Manufacturing and Service Operations Management*, vol. 15, p. 239–249.

Listes, O. et R. Dekker. 2005, «A stochastic approach to a case study for product recovery network design», *European Journal of Operational Research*, vol. 160, p. 268–287.

Liu, M. S. et J. T. Lee. 1988, «A simulation model of a hospital emergency call system using SLAMII», *Simulation*, vol. 51, p. 216–221.

Lokketangen, A. et D. L. Woodruff. 1996, «Progressive hedging and tabu search applied to mixed integer (0,1) multi-stage programming», *Journal of Heuristics*, vol. 2, p. 111–128.

Louveaux, F. V. 1986, «Discrete stochastic location models», *Annals of Operations Research*, vol. 6, p. 23–34.

Lubicz, M. et B. Mielczarek. 1987, «Simulation modelling of emergency medical services», *European Journal of Operational Research*, vol. 29, p. 178–185.

Lund, K., O. Madsen et M. M. Solomon. 1996, «Vehicle routing problems with varying degrees of dynamism», cahier de recherche, Institute of Mathematical Modeling, Technical University of Denmark.

Mahdian, M., N. Hamid et A. Saberi. 2012, «Online optimization with uncertain information», *ACM Transactions on Algorithms*, vol. 8, p. 1–29.

Mandell, M. B. 1988, «Covering models for two-tiered emergency medical services system», *Location Science*, vol. 6, p. 355–368.

Manno, I. 1999, *Introduction to Monte-Carlo method*, Akadémiai Kiadó, Budapest.

Marianov, V., M. Mizumori et C. S. ReVelle. 2009, «The heuristic concentration-integer and its application to a class of location problems», *Computers & Operations Research*, vol. 36, p. 1406–1422.

Marianov, V. et C. S. ReVelle. 1994, «The queuing probabilistic location set covering problem and some extensions», *Socio-Economic Planning Sciences*, vol. 28, p. 167–178.

Marianov, V. et C. S. ReVelle. 1995, «Siting emergency services», dans *Facility Location. A survey of Applications and Methods*, édité par Z. Drezner, Springer, New York, N.Y., p. 119–223.

Marianov, V. et C. S. ReVelle. 1996, «The queuing maximal availability location problem : A model for the siting of emergency vehicles», *European Journal of Operational Research*, vol. 93, p. 110–120.

Marianov, V. et D. Serra. 2001, «Location problems in public sector», dans *Facility Location. Applications and Theory*, édité par Z. Drezner et H. W. Hamacher, Springer, Berlin, p. 119–130.

Mason, A. J. 2013, «Simulation and real-time optimised relocation for improving ambulance operations», dans *Handbook of Healthcare Operations : Methods and Applications*, édité par B. Denton, Springer, New York, N.Y., p. 289–317.

Maxwell, M. S., M. Restepo, S. G. Henderson et H. Topaloglu. 2009, «Approximate dynamic programming for ambulance redeployment», *INFORMS Journal on Computing*, vol. 22, p. 266–281.

Melachrinoudis, E. et M. E. Helander. 1996, «A single facility location problem on a tree with unreliable edges», *Networks*, vol. 27, p. 219–237.

Mitrović-Minić, S., R. Krishnamurtia et G. Laporte. 2004, «Double-horizon based heuristics for the dynamic pickup and delivery problem with time windows», *Transportation Research Part B*, vol. 38, p. 669–685.

Moeini, M., Z. Jemai et E. Sahin. 2013, «An integer programming model for the dynamic location and relocation of emergency vehicles : A case study», dans *Proceedings of the 12th International Symposium on Operational Research (SOR'2013)*, édité par T. Csendes, L. L. Stim et J. Zerovnik, Slovenia, p. 343–350.

Morales, J. C. 2006, *Planning robust freight transportation operations*, thèse de doctorat, Georgia Institute of Technology.

Mu, Q., Z. Fu, J. Lysgaard et R. Eglese. 2011, «Disruption management of the vehicle routing problem with vehicle breakdown», *Journal of the Operational Research Society*, vol. 62, p. 742–749.

Mulvey, J., R. Vanderbei et S. Zenios. 1995, «Robust optimization of large-scale systems», *Operations Research*, vol. 43, p. 264–281.

Nair, R. et E. Miller-Hooks. 2009, «Evaluation of relocation strategies for emergency medical service vehicles», *Journal of the Transportation Research Board*, vol. 2137, p. 63–73.

Naoum-Sawaya, J. et S. Elhedhli. 2013, «A stochastic optimization model for real-time ambulance redeployment», *Computers & Operations Research*, vol. 40, p. 1972–1978.

Narayanaswami, S. et N. Rangaraj. 2011, «Scheduling and rescheduling of railway operations : A review and expository analysis», *Technology Operation Management*, vol. 2, p. 102–122.

Narayanaswami, S. et N. Rangaraj. 2013, «Modelling disruptions and resolving conflicts optimally in a railway schedule», *Computers & Industrial Engineering*, vol. 64, p. 469–481.

Neebe, A. et M. Rao. 1983, «An algorithm for the fixed-charge assigning users to sources problem», *Journal of the Operational Research Society*, vol. 34, p. 1107–1113.

Nel, L. D. et C. J. Colbourn. 1990, «Locating a broadcast facility in an unreliable network», *INFOR*, vol. 28, p. 363–379.

Nielsen, L. K., L. Kroon et G. Maróti. 2012, «A rolling horizon approach for disruption management of railway rolling stock», *European Journal of Operational Research*, vol. 220, p. 496–509.

Novoa, C. et R. Storer. 2009, «An approximate dynamic programming approach for the vehicle routing problem with stochastic demands», *European Journal of Operational Research*, vol. 196, p. 509–515.

Noyan, N. 2012, «Risk-averse two-stage stochastic programming with an application to disaster management», *Computers & Operations Research*, vol. 39, p. 541–559.

Owen, S. H. et M. S. Daskin. 1998, «Strategic facility location : a review», *European Journal of Operational Research*, vol. 111, p. 423–447.

Pidd, M. 2006, *Computer simulation in management science*, John Wiley & Sons, New York, N.Y.

Pidd, M. 2014, «The ways forward : A personal view on system dynamics and discrete-event simulation», dans *Discrete-Event Simulation and System Dynamics for Management Decision Making*, édité par S. Brailsford, L. Churilov et B. Dangerfield, John Wiley & Sons, New York, N.Y., p. 318–336.

Pillac, V., M. Gendreau, C. Guéret et A. L. Medaglia. 2013, «A review of dynamic vehicle routing problems», *European Journal of Operational Research*, vol. 225, p. 1–11.

Pirkul, H. 1987, «Efficient algorithms for the capacitated concentrator problem», *Computers & Operations Research*, vol. 14, p. 197–208.

Pisinger, D. 1997, «A minimal algorithm for the 0-1 knapsack problem», *Operations Research*, vol. 45, p. 758–767.

Powell, W. B. 20011, *Approximate dynamic programming : Solving the Curses of dimensionality.*, Wiley-Interscience, Hoboken, N.J.

Puterman, M. L. 1994, *Markov decision processes : discrete stochastic dynamic programming*, Wiley-Interscience, New York, N.Y., Toronto.

Qi, X., J. Bard et G. Yu. 2004, «Supply chain coordination with demand disruptions», *Omega*, vol. 32, p. 301–312.

Qi, X., J. F. Bard et G. Yu. 2006, «Disruption management for machine scheduling : The case of SPT schedules», *International Journal of Production Economics*, vol. 103, p. 166–184.

Rajagopalan, H. K., C. Saydam et J. Xiao. 2008, «A multiperiod set covering location model for dynamic redeployment of ambulances», *Computers & Operations Research*, vol. 35, p. 814–826.

Ravi, R. et A. Sinha. 2004, «Hedging uncertainty : approximation algorithms for stochastic optimization problems», *Lecture Notes in Computer Science*, vol. 3064, p. 101–115.

Rawls, C. G. et M. A. Turnquist. 2011, «Pre-positioning planning for emergency response with service quality constraints», *OR Spectrum*, vol. 33, p. 481–498.

Repede, J. F. et J. J. Bernardo. 1994, «Developping and validating a decision support system for location emergency medical vehicles in Louisville, Kentucky», *European Journal of Operational Research*, vol. 75, p. 567–581.

ReVelle, C. S. 1989, «Review, extension and prediction in emergency services siting models», *European Journal of Operational Research*, vol. 40, p. 58–69.

ReVelle, C. S. et H. A. Eiselt. 2005, «Location analysis : a synthesis and survey», *European Journal of Operational Research*, vol. 165, p. 1–19.

ReVelle, C. S., H. A. Eiselt et M. S. Daskin. 2008, «A bibliography of some fundamental problem categories in discrete location science», *European Journal of Operational Research*, vol. 184, p. 817–848.

ReVelle, C. S. et K. Hogan. 1988, «A reliability constrained siting model with local estimates of busy fractions», *Environment and Planning B*, vol. 15, p. 143–152.

ReVelle, C. S. et K. Hogan. 1989, «The maximum availability location problem», *Transportation Science*, vol. 23, p. 192–200.

ReVelle, C. S. et V. Marianov. 1991, «A probabilistic FLEET model with individual reliability requirements», *European Journal of Operational Research*, vol. 53, p. 93–105.

ReVelle, C. S., D. Marks et J. C. Liebman. 1970, «An analysis of private and public sector location models», *Management Science*, vol. 16, p. 692–707.

Rockafellar, R. T. et R. J. B. Wets. 1991, «Scenarios and policy aggregation in optimization under uncertainty», *Mathematics of Operations Research*, vol. 16, p. 119–147.

Rönnqvist, M., S. Tragantalerngsak et J. Holt. 1999, «A repeated matching heuristic for the single-source capacitated facility location problem», *European Journal of Operational Research*, vol. 116, p. 51–68.

Ross, S. M. 2013, *Simulation*, Academic Press, Amsterdam.

Rubinstein, R. Y. 1981, *Simulation and Monte Carlo method*, Wiley, New York, N.Y., Toronto.

Sahinidis, N. V. 2004, «Optimization under uncertainty : state-of-the art and opportunities», *Computers and Chemical Engineering*, vol. 28, p. 971–983.

Savas, E. S. 1969, «Simulation and cost-effictiveness analysis of New York's emergency ambulance service», *Management Science*, vol. 15, p. 602–627.

Saydam, C., H. K. Rajagopalan, E. Sharer et K. Lawrimore-Belanger. 2013, «The dynamic redeployment coverage location model», *Health Systems*, vol. 2, p. 103–119.

Schilling, D. A., D. J. Elzinga, J. Cohon, R. L. Church et C. S. ReVelle. 1979, «The TEAM/FLEET models for simultaneous facility and equipment setting», *Transportation Science*, vol. 13, p. 163–175.

Schmid, V. 2012, «Solving the dynamic ambulance relocation problem and dispatching problem using approximate dynamic programming», *European Journal of Operational Research*, vol. 219, p. 611–621.

Schmid, V. et K. F. Doerner. 2010, «Ambulance location and relocation problems with time-dependent travel times», *European Journal of Operational Research*, vol. 207, p. 1293–1303.

Secomandi, N. 2000, «Comparing neuro-dynamic programming algorithms fort the vehicle routing problem with stochastic demands», *Computers & Operations Research*, vol. 27, p. 1201–1225.

Secomandi, N. 2003, «Analysis of a rollout approach to sequencing problems with stochastic routing applications», *Journal of Heuristics*, vol. 9, p. 321–352.

Shen, Z.-J. M., R. L. Zhan et J. Zhang. 2011, «The reliable facility location problem : Formulations, heuristics, and approximation algorithms», *INFORMS Journal on Computing*, vol. 23, p. 470–482.

Sniedovich, M. 2010, *Dynamic programming : Foundations and principles*, CRC Press, Boca Raton, FL.

Snyder, L. 2006, «Facility location under uncertainty», *IIE Transactions*, vol. 38, p. 537–554.

Snyder, L. V., Z. Atan, P. Peng, Y. Rong, A. J. Schmitt et B. Sinsoysal. 2014, «OR/MS models for supply chain disruptions : A review», http ://ssrn.com/abstract=1989882.

Snyder, L. V. et M. S. Daskin. 2005, «Reliability models for facility location : The expected failure cost case», *Transportation Science*, vol. 39, p. 400–416.

Snyder, L. V., M. P. Scaparra, M. S. Daskin et R. C. Church. 2006, «Planning for disruptions in supply chain networks», *INFORMS Tutorials in Operations Research*, p. 235–257.

Soyester, A. L. 1973, «Convex programming with set-inclusive constraints and applications to inexact linear programming», *Operations Research*, vol. 21, p. 1154–1157.

Sridharan, R. 1993, «A lagrangian heuristic for the capacitated plant location problem with single source constraints», *European Journal of Operational Research*, vol. 66, p. 305–312.

Statistics Canada. 2011, «2011 census profile», URL `https://www12.statcan.gc.ca/census-recensement/2011/`.

Storbeck, J. 1982, «Slack, natural slack and location covering», *Socio-Economic Planning Sciences*, vol. 16, p. 99–105.

Sundarapandian, V. 2009, *Probability, Statistics and Queueing theory*, PHI Learning, Delhi.

Sungur, I., F. Ordónez et M. Dessouky. 2008, «A robust optimization approach for the capacitated vehicle routing problem with demand uncertainty», *IIE Transactions*, vol. 40, p. 509–523.

Swersey, J. A. 1994, «The deployment of police, fire, and emergency medical units», dans *Operations Research and The Public Sector*, édité par S. Pollock, M. Rothkopf et A. Barnett, Elsevier B. V., Amsterdam, p. 151–200.

Swoveland, C., D. Uyeno, I. Vertinsky et R. Vikson. 1973, «A simulation model-based methodology for optimization of ambulance service policies», *Socio-Economic Planning Sciences*, vol. 7, p. 697–703.

Tekin, W. et I. Sabuncuoglu. 2004, «Simulation Optimization : A comprehensive review on theory and applications», *IIE Transactions*, vol. 36, p. 1067–1081.

The Toronto Star. 2009, «Quebec begins H1N1 vaccinations after school outbreak», *The Toronto Star*.

Tocher, K. D. 1962, *The Art of Simulation*, English Universities Press, London.

Toregas, C., R. Swain, C. S. ReVelle et L. Bergman. 1971, «The location of emergency service facilities», *Operations Research*, vol. 19, p. 1363–1373.

Toth, P. et D. Vigo. 2015, *The vehicle routing problem*, SIAM, Philadelphie, PA.

Trudeau, P., J.-M. Rousseau, J. A. Ferland et J. Choquette. 1989, «An operations research approach for the planning and operation of an ambulance service», *INFOR*, vol. 27, p. 95–113.

Urgences-santé. 2006, «Rapport annuel», cahier de recherche. URL `http://www.urgences-sante.qc.ca/`.

Urgences-santé. 2013, «Rapport annuel», cahier de recherche. URL `http://www.urgences-sante.qc.ca/`.

Uyeno, D. et C. Seeberg. 1984, «A practical methodology for ambulance location», *Simulation*, vol. 43, p. 79–87.

Van Hentenryck, P. et R. Bent. 2006, *Online Stochastic Combinatorial Optimization*, MIT Press, Cambridge, MA.

Volz, R. A. 1971, «Optimum ambulance location in semi-rural areas», *Transportation Science*, vol. 5, p. 193–203.

Weaver, J. R. et R. L. Church. 1983, «Computational procedures for location problems on stochastic networks», *Transportation Science*, vol. 17, p. 168–180.

White, D. J. 1985, «Real applications of markov decision processes», *Interfaces*, vol. 15, p. 73–83.

White, D. J. 1988, «Further real applications of markov decision processes», *Interfaces*, vol. 18, p. 55–61.

White, D. J. 1993, «A survey of applications of markov decision processes», *Journal of the Operational Research Society*, vol. 44, p. 1073–1096.

Xia, Y., M.-H. Yang, B. Golany, S. M. Gilbert et G. Yu. 2004, «Real-time disruption management in a two-stage production and inventory system», *IIE Transactions*, vol. 36, p. 111–125.

Xiao, T. J., X. Qi et G. Yu. 2007, «Coordination of supply chain after demand disruptions when retailers compete», *International Journal of Production Economics*, vol. 109, p. 162–179.

Xiao, T. J. et X. T. Qi. 2008, «Price competition, cost and demand disruptions and coordination of a supply chain with one manufacturer and two competing retailers», *Omega*, vol. 36, p. 741–753.

Xiao, T. J., G. Yu, Z. H. Sheng et Y. S. Xia. 2005, «Coordination of a supply chain with one-manufacturer and two-retailers under demand promotion and disruption management decisions», *Annals of Operations Research*, vol. 135, p. 87–109.

Yang, J., X. Qi et G. Yu. 2005, «Disruption management in production planning», *Naval Research Logistics*, vol. 52, p. 420–442.

Yang, Z., F. Chu et H. Chen. 2012, «A cut-and-solve based algorithm for the single-source capacitated facility location problem», *European Journal of Operational Research*, vol. 221, p. 521–532.

Yu, G., M. F. Arguello, M. Song, S. McCowan et A. White. 2003, «A new era for crew recovery at Continental Airlines», *Interfaces*, vol. 33, p. 5–22.

Yu, G. et X. Qi. 2004, *Disruption Management : framework, models and applications*, World Scientific., River Edge, N. J.

Zhen, L., K. Wang, H. Hu et D. Chang. 2014, «A simulation optimization framework for ambulance deployment and relocation problems», *Computers & Industrial Engineering*, vol. 72, p. 12–23.