Modélisation et analyse des données pour l'aide à la décision

Ahmed Bounekkar

Université Claude Bernard Lyon 1

Université de Lyon

N° d'ordre : 0472014

Année : 2014

Habilitation à diriger les recherches

Spécialité : Informatique

Modélisation et analyse des données pour l'aide à la décision

Ahmed BOUNEKKAR

Soutenue publiquement le 8 décembre 2014, devant le jury composé de

Marc BUI	Professeur EPHE, Université Paris 8	Rapporteur
Bernard DOUSSET	Professeur IRIT Université Toulouse 3	Rapporteur
Mohand Saïd HACID	Professeur LIRIS Université Lyon 1	Examinateur
Michel LAMURE	Professeur SIS, Université Lyon 1	Examinateur
Mohamed MOSBAH	Professeur LABRI Université de Bordeaux	Rapporteur
Gabriela SALZANO	Maître de conférences HDR Université Paris-EST	Examinateur

Résumé

Mes travaux de recherche après la thèse se situent essentiellement dans le domaine de la structuration et de la fouille des données (analyse des données spatiales, clustering,…) pour la conception de méthodologies d'aide à la décision. Ces travaux de recherche peuvent être résumés comme suit :

1. Le concept de contiguïté, simple ou généralisé, utilisé à partir de distances dans mon travail de thèse, se passe très bien de cette dernière notion et peut en fait être posé simplement dès que l'on dispose d'une notion de relation de « voisinage » sans faire intervenir des notions de métriques. Ainsi, nous avons travaillé sur la mise au point d'algorithmes qui, partant d'un graphe de contiguïté entre des objets sur lesquels sont mesurées des variables, permettent de structurer l'ensemble des données. Nous nous sommes intéressés également à la mise au point d'algorithmes prétopologiques s'intéressant plus généralement à la gestion des données. Dans ce même contexte, nous étions amenés à généraliser le modèle de l'auto-régression spatiale en intégrant des variables exogènes susceptibles d'expliquer le phénomène étudié en chaque point de l'espace. Par rapport au contexte de l'imagerie numérique, le gain en volume de données à traiter est immense. En revanche, le modèle est formellement plus complexe notamment la non régularité des données, ce qui débouche sur d'autres problèmes algorithmiques à résoudre. D'autre part, la nécessité d'appliquer la régression logistique à des données spatiales, nous a amenée à proposer un modèle de régression logistique spatiale (SLR). D'une manière générale, ces travaux s'insèrent dans l'analyse des processus de décision et ont été appliqués dans divers domaines, notamment en santé et en environnement.

2. La classification automatique est l'une des méthodes d'analyse des données les plus utilisées dans divers domaines. Elle a pour objectif d'identifier des groupes d'individus (ou d'éléments) ayant un comportement homogène vis-à-vis d'un certain nombre de caractéristiques. Les applications sont très nombreuses et concernent de nombreux domaines. Dans la plupart des travaux cités ci-dessous on propose des méthodes de classification applicables à l'analyse des données manquantes et des donnes mixtes (décrites par des caractères quantitatifs et qualitatifs à la fois). C'est un des points forts de nos méthodes car, dans des situations concrètes, les données ne sont pas toujours représentables dans un espace métrique. Nous avons apporté plusieurs contributions dans ce domaine, notamment pour les données non métriques :

   La méthode de classification ascendante hiérarchique ne vise généralement pas à optimiser un critère global portant sur la hiérarchie indicée obtenue sur l'ultramétrique. Nous avons proposé une méthode de recherche de l'ultramétrique la plus proche de la dissimilarité définie sur l'ensemble des individus en utilisant des techniques d'optimisation combinatoire. Cette méthode permet d'optimiser l'homogénéité des clusters obtenus.
   Dans le domaine de classification automatique des données mixtes (qualitatives et quantitatives), nous avons proposé une méthode basée sur la notion d'agrégation d'opinions. Elle consiste à associer à chaque variable une fonction de classement qui va jouer le rôle d'un juge. Ce dernier va classer les individus selon ses propres critères. En rassemblant les classements de toutes les variables, pour tous les couples d'individus, on cherche à construire un classement collectif sur les individus (meilleure agrégation possible). La recherche de la partition résultat est basée sur la maximisation de la fonction de concordance globale.
   Lorsqu'on est amené à faire des opérations de classification, de comparaison ou de structuration des données, on est amené à choisir une mesure de proximité entre les individus ou les éléments étudiés. Les résultats obtenus sont souvent impactés par le choix de la mesure de proximité adoptée. Ces mesures de proximité sont caractérisées par des propriétés mathématiques précises. Sont-elles, pour autant, toutes équivalentes ? Peuvent-elles être utilisées dans la pratique de manière indifférenciée ? Autrement dit, est-ce que, par exemple, la mesure de proximité entre individus plongés dans un espace multidimensionnel, influence-t-elle le résultat des opérations comme la classification en groupes ou la recherche des k-plus-proches voisins? Après la présentation de l'approche d'équivalence topologique, nous avons proposé une classification permettant de fournir des groupes «équivalents» de mesures de proximité.
   La classification des données hétérogènes répond à un besoin réel car les données ne sont pas toujours numériques et structurées. Nous avons proposé dans ce sens des méthodes de classification basées sur le concept de prétopologie. Une des méthodes proposées consiste à structurer l'espace des données en utilisant la fonction d'adhérence qui mène à définir les fermés et les fermés minimaux. Ceci nous permet d'extraire des germes qui serviront par la suite dans les méthodes de classification par réallocation. L'intérêt principal de ce travail est de déterminer une structure forte et de fournir un nombre de classes de manière « naturelle ». Nous avons ensuite utilisé les fermés minimaux pour obtenir la classification prétopologique encadrant la structure prétopologique.
   Une autre méthode de proposée consiste à pouvoir classer des objets sur la base d'évaluations seulement qualitatives en se dotant de la possibilité de n'utiliser qu'un sous ensemble d'évaluations.

3. Les métaheuristiques hybrides évolutionnaires pour l'aide à la décision multi-objectifs.

La majorité des problèmes d'aide à la décision dans le monde réel concerne plusieurs objectifs contradictoires, les solutions forment un ensemble de compromis appelé l'ensemble l'optimum de Pareto. Récemment, des algorithmes basés sur les métaheuristiques hybrides évolutionnaires ont prouvé leur efficacité dans le traitement de ce type de problèmes. Nous nous sommes intéressés à l'élaboration d'une nouvelle méthodologie en intégrant diverses stratégies de recherches, en exploitant les avantages et en évitant les inconvénients des stratégies originelles. Nous avons proposé plusieurs méthodes de résolution qui ont donné lieu à des publications de haut niveau en informatique. Parmi ces méthodes on peut citer : HMEH, HMEH2 et HESSA.

4. Aujourd'hui, les réseaux complexes concernent différentes catégories de réseaux : sociaux, technologiques, de connaissances,…Une analyse des réseaux complexes et les graphes adaptés aux réseaux sociaux nous a amenée à proposer un modèle de réseaux stochastiques. Ce modèle présente un nouveau concept basé sur le couplage de deux théories sous-jacentes, la prétopologie et les ensembles aléatoires. Par la généralisation de graphes aléatoires, ce modèle permet de prendre en compte par exemple, dans la gestion des crises sanitaires, des évènements imprévus et néanmoins très fréquents. Des modèles de propagation de grippe ont été mis en proposés.

5. Certains travaux liés à des projets de recherche ou à des thèses sont cités brièvement ici :

   Plusieurs travaux ont été menés pour modéliser le phénomène de diffusion de grippe dans le cadre de projets (préfecture du Vaucluse, Europe) ou de sujets de thèse. Parmi les modèles proposés, on peut citer une nouvelle méthode de modélisation de la propagation des maladies infectieuses SEIR-SW basé sur le modèle d'épidémie SEIR et le réseau social petit monde.
   Dans le cadre du projet européen FLURESP nous avons contribué à redéfinir les principaux scénarios d'une pandémie humaine au niveau européen, à décrire des stratégies de réponses possibles et d'évaluer ces stratégies d'intervention dans un cadre d'analyse multicritères et des analyses coût-efficacité.
   Dans le cadre d'un projet interdisciplinaire du CNRS ayant pour objectif d'élaborer un système d'alerte à la pollution de l'air à Ouagadougou. Nous avons mis en place un modèle spatial de propagation de la pollution en utilisant des méthodes d'analyse spatiale (Régression, autocorrélation) basées sur la contiguïté. Etude des corrélations entre les lieux de pollution et des pathologies (maladies de la peau, des yeux, respiratoires,..) dans la région de Ouagadougou.
   Dans le cadre des projets thématiques de recherche de la région Rhône-Alpes, nous avons élaboré de règles prédictives des durées interventions dans les blocs opératoires en fonction des paramètres ayant une influence significative sur les durées. Nous avons également mis en place une méthodologie d'évaluation des performances dans les établissements hospitaliers. Nous avons également étudié les facteurs influençant le taux d'occupation des salles dans les blocs opératoires et fourni une déclinaison et une analyse des indicateurs stratégiques dans le contexte de regroupement de plateaux médicotechniques.
   Nous avons proposé une méthodologie rigoureuse s'appuyant sur un cadre conceptuel approprié dans le but de tester la validité transculturelle des instruments de Qualité de Vie. Ce travail est motivé par l'internationalisation de la recherche clinique et le besoin grandissant d'instruments subjectifs pouvant être utilisés de manière transculturelle, la recherche transculturelle prend un caractère essentiel dans le domaine de la qualité de vie. Dans ce cadre, si les processus de traduction sont particulièrement bien étudiés, les méthodes statistiques permettant d'évaluer de manière quantitative la validité transculturelle des questionnaires sont encore assez peu connues.
   Dans le cadre de la perspective de réalisation des objectifs du millénaire pour le développement, nous avons mené une étude pour la conception d'un système d'information sanitaire et la mise en place d'une base d'indicateurs statistiques en santé publique pour le mali. Ce système d'information est capable de gérer, d'acquérir d'évaluer et de rendre visible et lisible les nombreux projets et activités qui donnent un aspect d'appartenance au secteur informel comme l'immense majorité des efforts des pays en voie de développement.

Table des matières

Introduction

Ce mémoire présente une partie non exhaustive de mes travaux de recherches durant ma carrière après la thèse de doctorat. Le début de mes travaux est dans la continuité de mon doctorat et porte sur la modélisation et l'analyse des données spatiales jusqu'en 2008. Parallèlement, je me suis intéressé à la thématique de classification automatique des données (clustering). Par ailleurs, depuis 2009, je m'intéresse à l'optimisation multi-objectifs pour l'aide à la décision. Les thématiques de recherches de mon ancien laboratoire LASS m'avaient conduit à me lancer dans la recherche de modélisation mathématique et informatique appliquée à des problématiques en santé. Dans ce contexte, j'ai travaillé sur Plusieurs projets de recherche en santé (Région, CNRS, Europe,..). Ces travaux sont présentés ici dans un ordre thématique que chronologique.

Les travaux de recherche après la thèse se situent essentiellement dans le domaine de la structuration et de la fouille des données (analyse des données spatiales, clustering,…) pour la conception de méthodologies d'aide à la décision. Ces travaux de recherche peuvent être résumés comme suit :

Le chapitre 1 porte sur la problématique de classification automatique. C'est l'une des catégories de méthodes d'analyse des données les plus utilisées car elle répond à un besoin réel des utilisateurs et elle présente un vaste champ d'applications. Elle a pour objectif d'identifier des groupes d'individus (ou d'éléments) ayant un comportement homogène vis-à-vis d'un certain nombre de caractéristiques. Les applications sont très nombreuses et concernent tous les domaines. Dans certains travaux, on propose des méthodes de classification applicables à l'analyse des données hétérogènes, c'est-à-dire les données décrites par des caractères quantitatifs et qualitatifs à la fois. C'est un des points forts de nos méthodes car, dans des situations concrètes, les données ne sont pas toujours représentables dans un espace métrique. Nous avons apporté plusieurs contributions dans ce domaine, notamment pour les données non métriques : elles portent sur les méthodes de CAH, les problèmes des données mixtes, le problème de choix de mesures de proximité et enfin les situations à données hétérogènes.

La méthode de classification ascendante hiérarchique ne vise généralement pas à optimiser un critère global portant sur la hiérarchie indicée obtenue sur l'ultramétrique. Nous avons proposé une méthode de recherche de l'ultramétrique la plus proche de la dissimilarité définie sur l'ensemble des individus en utilisant des techniques d'optimisation combinatoire. Cette méthode permet d'optimiser l'homogénéité des clusters obtenus

Dans le domaine de classification automatique des données mixtes (qualitatives et quantitatives), nous avons proposé une méthode basée sur la notion d'agrégation d'opinions. Elle consiste à associer à chaque variable une fonction de classement qui va jouer le rôle d'un juge. Ce dernier va classer les individus selon ses propres critères. En rassemblant les classements de toutes les variables, pour tous les couples d'individus, on cherche à construire un classement collectif sur les individus (meilleure agrégation possible). La recherche de la partition résultat est basée sur la maximisation de la fonction de concordance globale.

Lorsqu'on est amené à faire des opérations de classification, de comparaison ou de structuration des données, on est amené à choisir une mesure de proximité entre les individus ou les éléments étudiés. Les résultats obtenus sont souvent impactés par le choix de la mesure de proximité adoptée. Ces mesures de proximité sont caractérisées par des propriétés mathématiques précises. Sont-elles, pour autant, toutes équivalentes ? Peuvent-elles être utilisées dans la pratique de manière indifférenciée ? Autrement dit, est-ce que, par exemple, la mesure de proximité entre individus plongés dans un espace multidimensionnel, influence-t-elle le résultat des opérations comme la classification en groupes ou la recherche des k-plus-proches voisins? Après la présentation de l'approche d'équivalence topologique, nous avons proposé une classification permettant de fournir des groupes «équivalents» de mesures de proximité.

La classification des données hétérogènes répond à un besoin réel car les données ne sont pas toujours numériques et structurées. Nous avons proposé dans ce sens des méthodes de classification basées sur le concept de prétopologie. Une des méthodes proposées consiste à structurer l'espace des données en utilisant la fonction d'adhérence qui mène à définir les fermés et les fermés minimaux. Ceci nous permet d'extraire des germes qui serviront par la suite dans les méthodes de classification par réallocation. L'intérêt principal de ce travail est de déterminer une structure forte et de fournir un nombre de classes de manière « naturelle ». Nous avons ensuite utilisé les fermés minimaux pour obtenir la classification prétopologique encadrant la structure prétopologique. Une autre méthode de proposée consiste à pouvoir classer des objets sur la base d'évaluations seulement qualitatives en se dotant de la possibilité de n'utiliser qu'un sous ensemble d'évaluations. Ce travail n'a pas été développé dans ce mémoire.

Le chapitre 2 concerne l'analyse des données spatiales. Le concept de contiguïté, simple ou généralisé, utilisé à partir de distances dans mon travail de thèse de doctorat, se passe très bien de cette dernière notion et peut en fait être posé simplement dès que l'on dispose d'une notion de relation de « voisinage » sans faire intervenir des notions de métriques. Il s'agit de travailler non plus avec de simple espaces métriques, mais bien avec des espaces plus généraux, tels que les espaces prétopologiques qui ont l'avantage d'être utilisables sur des objets non quantitatifs. Ainsi, nous avons travaillé à la mise au point d'algorithmes qui, en partant d'un graphe de contiguïté entre des objets sur lesquels sont mesurées des variables, permettent de structurer l'ensemble des données. Nous nous sommes intéressés également à la mise au point d'algorithmes prétopologiques s'intéressant plus généralement à la gestion des données. Dans ce même contexte, nous étions amenés à généraliser le modèle de l'auto-régression spatiale en intégrant des variables exogènes susceptibles d'expliquer le phénomène étudié en chaque point de l'espace. Par rapport au contexte de l'imagerie, le gain en volume de données à traiter est immense. En revanche, le modèle est formellement plus complexe, ce qui débouche sur d'autres problèmes algorithmiques à résoudre. D'autre part, la généralisation de la notion de régression logistique à des données spatiales, nous a amenée à proposer un modèle de régression logistique spatiale (SLR). D'une manière générale, ces travaux s'insèrent dans l'analyse des processus de décision et ont été appliqués dans divers domaines du secteur de santé.

Le chapitre 3 porte sur les métaheuristiques hybrides évolutionnaires pour l'aide à la décision multi-objectifs. La majorité des problèmes d'aide à la décision dans le monde réel concerne plusieurs objectifs contradictoires, les solutions forment un ensemble de compromis qu'on peut assimiler à un ensemble « d'optima de Pareto ». Récemment, des algorithmes basés sur les métaheuristiques hybrides évolutionnaires ont prouvé leur efficacité dans le traitement de ce type de problèmes. Nous nous sommes intéressés à l'élaboration d'une nouvelle méthodologie en intégrant diverses stratégies de recherches, en exploitant les avantages et en évitant les inconvénients des stratégies originelles. Nous avons proposé plusieurs méthodes de résolution qui ont donné lieu à des résultats très satisfaisants. Parmi ces méthodes on peut citer : HMEH, HMEH2 et HESSA.

Le chapitre 4 est une synthèse de quelques travaux de modélisation pour l'aide à la décision en santé. Aujourd'hui, les réseaux complexes concernent différentes catégories de réseaux : sociaux, technologiques, de connaissances,…Une analyse des réseaux complexes et les graphes adaptés aux

réseaux sociaux nous a amenée à proposer un modèle de réseaux stochastiques. Ce modèle présente un nouveau concept basé sur le couplage de deux théories sous-jacentes, la prétopologie et les ensembles aléatoires. Par la généralisation des graphes aléatoires, ce modèle permet de prendre en compte par exemple, dans la gestion des crises sanitaires, des évènements imprévus et néanmoins très fréquents. Plusieurs travaux ont été menés pour modéliser le phénomène de diffusion de grippe dans le cadre de projets (préfecture du Vaucluse, Europe) ou de sujets de thèse à l'encadrement desquelles j'ai contribué. Parmi les modèles proposés, on peut citer les travaux récents sur la modélisation de la propagation des maladies infectieuses SEIR-SW basé sur le modèle d'épidémie SEIR et le réseau social « petit monde ».

Dans chacun de ces chapitres, après l'exposé de la problématique, un bref état de l'art des méthodes existantes est dressé. Ils sont suivis des contributions que nous avons apportées à la problématique étudiée.

Le chapitre 5 est une synthèse de ma carrière professionnelle. Les différentes responsabilités occupées, les encadrements, les publications et les projets de recherche sont mentionnés.

Chapitre 1

Les méthodes de classification

Sommaire

1.1 Introduction

La notion de classification est essentielle dans de nombreux domaines, elle permet aux scientifiques de mettre de l'ordre dans les connaissances qu'ils ont sur le monde. Ainsi, par exemple, depuis longtemps des chercheurs et des scientifiques ont essayé de classifier les espèces animales. C'est une des raisons pour laquelle les spécialistes en mathématiques appliquées, d'abord seuls, puis avec l'aide des outils et méthodes informatiques ont travaillé dans ce domaine, produisant ainsi un grand nombre de méthodes et algorithmes de classification. Face à ces classifications, les scientifiques sont souvent incapables de désigner la meilleure d'entre elles, car chacune présente un intérêt par rapport aux autres et à la tâche considérée. La classification est une forme d'abstraction, puisque l'on va mettre de côté les descriptions exactes des objets et ne faire ressortir que les traits particuliers que certains d'entre eux ont en commun. En effet, une classification permet de synthétiser les informations contenues dans des groupes plus généraux. Dans un problème de classification on dispose d'un ensemble des données qui reprend une collection d'individus (objets) non étiquetés. Les classes sont encore inexistantes. L'objectif est alors d'obtenir des classes d'objets homogènes, en favorisant l'hétérogénéité entre ces différentes classes. On peut donc dire que toutes les méthodes de classification suivent un même principe général qui consiste à minimiser la dissemblance entre deux individus d'une même classe et à maximiser la cette dissemblance entre les individus de classes distinctes. La classification automatique est populaire pour ses applications et est largement utilisée dans de nombreux domaines. Le succès qu'a rencontré ce type de méthodes et le développement des moyens de stockage des données ont largement contribué aux nombreux travaux liés à la classification comme on peut le constater dans la section suivante.

Dans le domaine de la classification, la plupart des méthodes ont plus ou moins implicitement recours à la mesure de proximité. Dans le cas où le contexte est tel qu'une telle proximité n'a rien de naturel les outils habituels ne sont guère satisfaisants. Supposons par exemple, qu'on dispose d'une population Ω sur laquelle on a observé divers caractères et que, pour chacun de ces caractères, l'observation se traduise par une variable X définie sur Ω et prenant ses valeurs dans l'ensemble E des modalités du caractère. Si le caractère est quantitatif alors l'ensemble E peut naturellement être muni d'une mesure de proximité qui pourra être utilisée pour apprécier la ressemblance ou non de deux objets de Ω. Mais si le caractère est purement qualitatif alors ceci n'est guère possible, sauf à prendre le risque de malmener sérieusement la réalité. Il faut alors construire une procédure spécifique au contexte qualitatif, en vue de proposer une formalisation adaptée du concept de proximité.

Après un bref aperçu de la littérature des méthodes de classification automatique, nous abordons dans ce chapitre trois types de travaux liés à la classification. Le premier travail est la proposition d'une méthode de classification automatique inspirée de l'agrégation des opinions. Après un exposé du problème, une méthode de résolution est proposée, suivie d'un exemple d'application sur des données connues dans la littérature. Le deuxième travail est une proposition d'amélioration de la méthode de Classification Ascendante Hiérarchique en cherchant l'ultramétrique la plus proche de la dissimilarité utilisée dans la méthode. Comme le choix de la mesure de proximité est important lors du processus de classification, nous avons proposé un troisième travail autour de l'équivalence au sens topologique des mesures de

proximité afin de déterminer si elles sont proches ou éloignées. Nous présentons cette approche et montrons les premiers liens que nous avons identifiés entre elle et celle basée sur la préordonnance.

1.2 Problématique et état de l'art

1.2.1 Introduction

Lorsqu'on est confronté à la classification des données, on est amené à faire un **choix d'une méthode** en prenant en compte quelques considérations, tel que le type d'attributs qu'un algorithme peut manipuler, la capacité de travailler sur des données de très grande dimension ou sur un grand volume de données, la capacité de trouver des groupes dans un ensemble des données irrégulières ou encore la complexité de la méthode. Les premières approches proposées en classification étaient algorithmiques, heuristiques ou géométriques et reposaient essentiellement sur la dissimilarité entre les objets à classer. L'approche statistique, plus récente, se base sur des modèles probabilistes qui formalisent l'idée de classe. Cette approche permet en outre d'interpréter de façon statistique la classification obtenue.

A la différence de la classification supervisée, qui, étant donné un ensemble de classes déjà identifiées, consiste à trouver la meilleure classe à laquelle un individu appartient, la classification non supervisée consiste à structurer des classes non encore identifiées qui groupent ces individus. Une définition formelle de la classification qui puisse servir de base à un processus automatisé, amène à se poser un certain nombre de questions : Comment les objets à classer sont-ils définis ? Comment définir la notion de ressemblance entre objets ? Qu'est-ce qu'une classe ? Comment sont structurées les classes ? et comment juger une classification par rapport à une autre ?

1.2.2 Etat de l'art : Bref aperçu des méthodes de classification

Il n'est pas aisé de donner des familles de méthodes de classification car plusieurs descriptions permettent de les caractériser. On peut décrire une méthode par l'homogénéité des classes par rapport aux attributs caractéristiques (variables). On parle de :

- méthode monothétique si, lorsque chaque élément de la partition $\wp$ obtenue possède la même modalité pour au moins une des p variables caractérisant les objets. Ce type des méthodes est appelé : méthodes de segmentation. Elles concernent plutôt les données de type nominal, difficiles à quantifier.
- Méthode polythétique sinon. Dans chaque classe, les éléments se ressemblent au sens du critère de dissimilarité choisi dans la méthode. Les méthodes polythétiques sont nombreuses et couramment utilisées.

Sin on veut obtenir une famille de partitions de l'ensemble à classifier, la comparaison et l'ordre d'obtention des partitions permettent de caractériser la méthode en :

- hiérarchique (les partitions sont totalement ordonnées par la relation d'inclusion) ou non hiérarchique (pas de hiérarchie imposée entre les partitions).
- ascendante (le nombre de classes dans la partition baisse suite à des regroupements des objets ou des classes) ou descendante (par divisions des classes).

Nous nous limitons ici à présenter quarte familles. La première est composée d'algorithmes de classification hiérarchique qui procèdent à la construction des classes par agglomérations successives ou par dichotomies successives. La deuxième est constituée d'algorithmes conduisant directement à une ou plusieurs partitions en optimisant des critères pertinents. La troisième concerne les méthodes de classification par densité qui considèrent des classes comme des régions denses séparées par des régions de basse densité d'objets. La dernière concerne les méthodes basées sur la quantification par grille. D'autres méthodes particulières sont également présentées. La figure 1 donne un aperçu non exhaustif des méthodes de classification couramment rencontrées dans la littérature. La description de ces méthodes n'est pas détaillée ici. L'objectif est de donner au lecteur un bref descriptif de chacune des méthodes. Pour plus de détails, on peut se référer à [8].

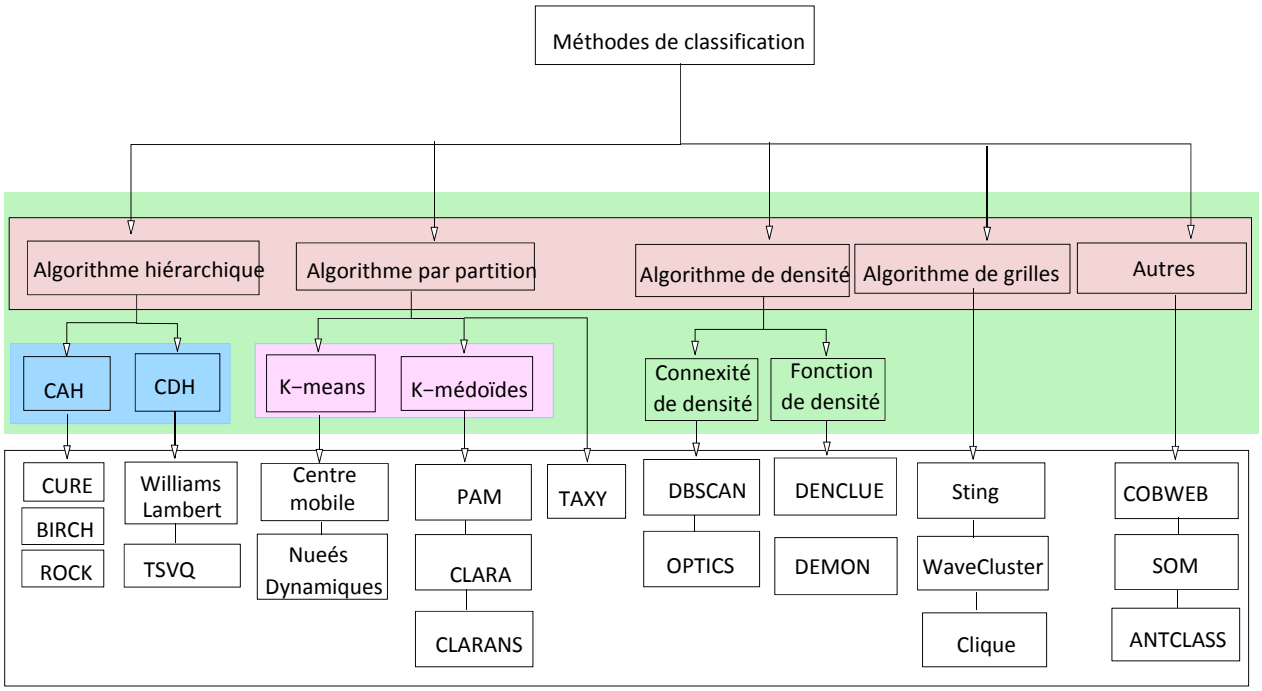


Figure 1 : méthodes de classification automatique

La classification nécessite de disposer d'un outil capable de mesurer les ressemblances ou les dissemblances qui existent entre les individus d'une population. Afin de regrouper les individus homogènes et séparer les individus éloignés, il faut disposer d'un outil permettant de mesurer le degré de similarité ou de dissimilarité entre les objets.

Dans un processus de classification automatique, les classes ne sont pas connues au début de la recherche : le nombre de classes, la définition de leurs caractéristiques et la constitution de leurs objets doivent être déterminés. La classification consiste donc à chercher une partition comportant des classes d'objets semblables, déterminées par les mesures de similarités et dissimilarité. Une condition additionnelle pour une partition est la constitution de classes bien séparées, c.-à-d. un objet doit non seulement être semblable à un autre objet dans la même classe, mais aussi marqué différent de celui dans une autre classe.

Il existe deux familles de représentation de similarité : similarité numérique et similarité symbolique [6], [44], [25], [2] [32]. La première est une quantité mesurable employée pour indiquer le degré ressemblance de deux individus i et j. La deuxième permet de caractériser les ressemblances dans un langage proche du langage naturel. Pour plus de détails sur les mesures de proximités, on peut consulter [8] et la section 1.3.4.2 de ce chapitre.

1.2.2.1 Les méthodes hiérarchiques

Les méthodes hiérarchiques permettent d'obtenir une suite de partitions totalement ordonnées par la relation d'inclusion. On distingue deux principales méthodes

La Classification Ascendante Hiérarchique (CAH)

Soit un ensemble E de n objets à classifier, muni d'une dissimilarité ρ. Partant d'une partition discrète de E, l'algorithme de CAH consiste à construire de façon itérative une suite de partitions de E de telle sorte que la partition obtenue à l'étape (k), soit issue de la partition construite à l'étape (k-1) par réunion de ses deux classes les plus proches au sens d'une dissimilarité Δ définie sur l'ensemble P(E) des parties disjointes de E. cette dissimilarité Δ doit être compatible avec ρ. Cette opération est répétée jusqu'à l'obtention d'une partition grossière. L'ensemble des partitions fournies par l'algorithme constitue une hiérarchie valuée. Pour pouvoir effectuer une CAH, il est donc nécessaire de définir une dissmilarité ρ sur l'ensemble E et une dissmiliarité Δ sur l'ensemble P(E). La dssimilarité Δ est appelée critère de groupement ou stratégie d'agrégation. Plusieurs stratégies d'agrégation ont été proposées dans la littérature [8]. Plusieurs algorithmes de classification basés sur la CAH ont été proposés. On peut citer :

- **CURE (Clustering Using Representatives)** a été proposé par Guha et al dans [21]. Cet algorithme utilise un échantillon représentatif de l'échantillon total pour réduire la complexité temporelle des calculs. Cet échantillon est divisé en sous-ensembles représentant des sous-classes. Les sous-classes seront agrégées hiérarchiquement. La distance entre deux sous-classes C1 et C2 est donnée par la plus petite distance entre un représentant de C1 et un représentant de C2. L'algorithme s'arrête lorsqu'on obtient les k classes demandées.
- **BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)** proposé par Zhang et al en 1996 [58], [59] est un algorithme efficace sur les gros jeux de données. L'idée principale est d'effectuer une classification sur un résumé compact de données originales. Cela permet de traiter un grand volume de données en utilisant une mémoire limitée. Cet algorithme incrémental nécessite un seul balayage des données. Il minimise le coût d'entrée/sortie en organisant les données traitées en une structure d'arbre équilibrée avec une taille limitée. BIRCH est considéré comme l'un des meilleurs algorithmes pour la classification des gros volumes de données.

- **ROCK (RObust Clustering using linKs)** proposé par Guha et al. en 1999 [22] est un algorithme de CAH utilisant le concept de liens (links) pour mesurer la similarité entre des individus au lieu des métriques pour les données numériques. Deux individus sont voisins si leur similarité dépasse un seuil θ. Le nombre de liens entre une paire d'individus est le nombre de leurs voisins communs. En général, les individus se trouvant dans une classe ont de nombreux voisins, par conséquent, de nombreux liens. En se basant sur le concept de « links » entre 2 individus, on définit le lien entre 2 classes par le nombre de liens croisés (cross links). Ce nombre est calculé par la somme de liens entre tous les points situés dans deux classes. ROCK est également adapté au traitement des gros volumes des données car il utilise un échantillon de ces données. L'avantage principal de cet algorithme est la simplicité de l'implémentation. Cependant, il est très coûteux en termes de calculs ($O(n^2)$).

La Classification Descendante Hiérarchique (CDH)

A partir d'un ensemble d'objets E, la Classification Descendante Hiérarchique construit de manière itérative une partition de l'ensemble d'individus. A l'inverse de la CAH, à chaque étape de l'algorithme on cherche la classe à scinder, ensuite on choisit le mode d'affectation des objets aux sous-classes.

La CDH construit sa hiérarchie en commençant par une classe unique contenant les n objets. A chaque étape, elle divise une classe en deux jusqu'à ce que toutes les classes ne contiennent qu'un seul individu. Ainsi, la hiérarchie est construite en (n – 1) étapes. Dans la première étape, les données sont divisées en deux classes au moyen des dissimilarités. Dans chacune des étapes suivantes, la classe avec le diamètre le plus grand se divise de la même façon. Après (n – 1) divisions, tous les individus sont séparés.

Le calcul des dissimilarités se fait de la manière suivante : soit un objet x appartenant à une classe C contenant m objets. La dissimilarité moyenne entre l'individu x et tous les autres individus de la classe C est définie par :

$$d_x = \frac{1}{m} \sum_{y \epsilon C, y \neq x} d(x, y)$$

Parmi les méthodes descendantes, nous décrivons ici l'algorithme de Williams et Lambert (1959) [57] qui est applicable sur des variables qualitatives. Il sélectionne l'une des variables pour servir de critère d'affectation : tous les individus présentant, pour cette variable, la même modalité sont rangés dans la même classe. Si les variables possèdent plus de deux modalités, le nœud correspondant aura plus de deux branches). La variable retenue est celle qui, dans la classe C à scinder, est la plus corrélée à toutes les autres. Comme il s'agit de variables qualitatives, la corrélation est mesurée par le χ^2 d'indépendance de deux caractères. On calcule donc les distances du χ^2 de contingence de toutes les variables prises deux à deux, et l'on retient celle pour laquelle la somme de ses χ^2 est maximum.

Une autre méthode descendante, appelée TSVQ (Tree Structured Vector Quantization) a été proposée par Gersho et Gray (1992) [19]. Elle utilise la méthode des k-means à deux classes pour le partitionnement. On utilise la somme des distances par rapport au centroïde au lieu de la distance moyenne.

1.2.2.2 Méthodes de partitionnement

Au lieu de fournir une suite de partitions hiérarchiques, les méthodes de partitionnement construisent directement une partition de l'ensemble des objets en un nombre de classes choisi a priori. Considérons un nombre k de classes souhaité. L'idée consiste à fournir une partition initiale, puis chercher à l'améliorer en réattribuant les individus d'une classe à l'autre. Etant donné qu'on ne peut pas énumérer toutes les partitions possibles, ces algorithmes recherchent des maximums locaux en optimisant une fonction objectif qui traduit que les objets doivent être semblables au sein d'une même classe, et dissemblables d'une classe à une autre.

Méthode des k-means

Parmi ce type de méthodes, celle des **k-means** est la plus simple et la plus populaire. Plusieurs variantes de cette méthode ont été proposées dans la littérature pour améliorer l'algorithme de base.

Partant d'une partition initiale formée au hasard, on calcule les centroïdes (centres de gravité) de chaque classe. Les individus sont ensuite balayés et affectés à la classe ayant le centre de gravité le plus proche au sens de la distance euclidienne. L'algorithme s'achève lorsqu'aucun individu ne change de classe. Il faut noter que dans sa version la plus classique, l'algorithme consiste à sélectionner aléatoirement k individus qui représentent les centroïdes initiaux.

Le principal avantage de cet algorithme est la simplicité de mise en œuvre. Sa complexité algorithmique est de O (n k t) où t représente le nombre d'itérations et n représente le nombre d'objets à classifier. Ses inconvénients sont nombreux : la solution fournie par l'algorithme est localement optimale, le calcul des centres de gravité est sensible aux données aberrantes, le résultat est influencé par le choix de la partition initiale,….

Méthode des nuées dynamiques

Cette méthode, proposée par Diday [13], est itérative et basée sur l'optimisation d'un critère. Elle consiste à rechercher une partition en k classes d'un ensemble de n individus. Chaque classe est représentée par son centre, également appelé noyau, constitué du petit sous-ensemble de la classe qui minimise le critère de dissemblance.

Soit d'une part, une application f de l'ensemble des classes d'une partition dans un ensemble de représentation dont les éléments sont appelés « noyaux » et d'autre part, une application g qui, à une famille d'individus de l'ensemble de représentation, associe une partition.

Soit une fonction f qui associe à un ensemble des classes $C = \{C_1 , C_2 , \dots . , C_k\}$, un ensemble $N = \{N_1 , N_2 , \dots . , N_k\}$ par f(C) = N où N_j est le noyau de C_j .

De même, à l'ensemble $\{N_1 , N_2 , \dots . , N_k\}$, l'application g associe une partition C par C = g(N)

Le critère H qui suit permet de mesurer l'adéquation d'une famille de noyaux à une partition :

$$H(C,N) = \sum_{j=1}^{k} h(C_j, N_j)$$

Cet algorithme est une succession d'appels aux deux fonctions f et g. on notera que le résultat change selon le choix des conditions initiales. Il faut donc exécuter plusieurs fois l'algorithme et comparer les résultats de manière à extraire les classes stables, c'est à dire à dégager des formes fortes.

L'exécution de l'algorithme se termine après un nombre fini d'itérations. Ce nombre est fixé par l'utilisateur au lancement de l'algorithme.

Méthode des k-médoïdes

Le médoïde d'une classe est l'objet ayant la dissimilarité moyenne la plus faible avec les autres objets de la classe. Dans ce type de méthodes, chaque classe est représentée par un de ses individus (médoïde) au lieu d'être représentée par une moyenne. Le procédé commence avec un ensemble de médoïdes puis itérativement remplace un médoïde par un autre si cela permet de réduire la distance globale.

Il existe plusieurs variantes de la méthode des k-médoïdes [10], [28]. PAM (Partitioning Around Medoids) publiée par Kaufman et Rousseeuw en 1986 joue un rôle important dans la génération d'autres méthodes de partitionnement tel que Clara (Clustering Large Application) utilisé pour les grands volumes de données et CLARANS [43]. Pammedsil et Pamsil sont d'autres variantes basées sur PAM [53].

Cette méthode est simple et couvre n'importe quel type de variables. Les algorithmes issus de cette méthode ont été appliqués à la classification des données hétérogènes [41] et à l'analyse d'une base de graphes issus d'un simulateur en intelligence en essaim [52].

1.2.2.3 Classification par densité

Dans les méthodes de la classification basées sur la densité, les classes sont considérées comme des régions de haute densité, séparées par des régions en faible densité. Ces régions peuvent avoir une forme arbitraire et les points à l'intérieur d'une région peuvent donc être arbitrairement distribués. Les données sont supposées représentées dans un espace métrique.

Le principe repose sur l'idée que les classes sont détectées car la densité des points qui s'y trouvent est supérieure à celle à l'extérieur des classes. En formalisant cette notion, la composition de classe vient de l'idée que pour chaque point dans la classe, son voisinage doit contenir au moins un nombre minimum de points, c'est-à-dire la densité dans le voisinage doit excéder un certain seuil ε.

Parmi les méthodes de classification par densité, nous citons brièvement quelques algorithmes :

L'algorithme DBSCAN (Density Based Spatial Clustering of Applications with Noise) [14], et ses dérivés tels que OPTICS [2] ou DBCLASD [58] sont basés sur l'idée de définir la notion de voisinage de rayon ε

d'un point : tous les points situés à une distance de ce point, inferieure à ε, appartiennent au voisinage. L'algorithme [14] commence par un point arbitraire et cherche tous les objets densité-accessibles. Si c'est un point noyau, alors cette phase forme une classe. S'il c'est un objet de frontière et qu'il n'a aucun point densité-accessible, alors c'est un bruit, l'algorithme passe à un autre objet.

L'algorithme OPTICS (Ordering Points To Identify Clustering Structure) a été proposé par (Ankerst et al) en 1999 [2]. C'est une extension de l'algorithme DBSCAN. L'idée générale consiste à identifier les régions potentielles de début de classe et de fin de classe et ensuite, de combiner ces régions pour former une hiérarchie. L'algorithme ordonne les points en valeurs de voisinage croissantes, et permet une exploration interactive de la hiérarchie ainsi produite.

L'algorithme DENCLUE (Density-based clustering) proposé par Hinneburg et al en 1998 [26], généralise l'approche de DBSCAN puisque celui-ci en est un cas particulier. DENCLUE modélise l'influence de chaque point sur son voisinage par une fonction d'influence, dépendante de la distance entre les objets et d'un paramètre σ réglant l'échelle du voisinage influence.

L'algorithme DEMON est basé sur une structure prétopologique [39] et utilise la fonction d'adhérence. Il fonctionne en deux étapes, une étape descendante et une étape montante :

- Etape Descendante : Le point de départ est la partition grossière $P_k = \{E\}$ (E étant les éléments à classifier). A partir d'une partition donnée $P = \{P_1, P_2 \ldots, P_k, E_k\}$, où $P_1, P_2 \ldots, P_k$ sont des classes obtenues au cours de l'algorithme aux étapes précédentes, et E_k la partie de E non encore explorée par l'algorithme, on sélectionne un point x dans E_k sur des critères "d'intégration" à E_k et on construit l'adhérence (en prétopologie) qui fournit une nouvelle classe après d'éventuelles procédures de réaffectation. On obtient une nouvelle partition plus fine que la précédente. L'algorithme s'arrête lorsque $E_k = \emptyset$, et fournit une partition P_{DE}.
- Etape Montante : Partant de la partition P_{DE}, on construit le plus petit fermé (en prétopologie) contenant la dernière classe obtenue dans P_{DE}., on obtient ainsi une nouvelle partition, et on continue le processus consistant à rechercher les fermetures des classes obtenues dans P_{DE}. dans l'ordre inverse de leur apparition.

Pour plus de détails sur cette méthode, on peut se référer à la thèse de Nicoloyannis [39].

Les algorithmes de classification par densité comme DBSCAN [15], GDBSCAN [46], DCBRD [17], DENCLUE [26],... sont souvent utilisés en analyse des bases de données spatiales [23], afin de détecter des classes ayant une forme arbitraire.

1.2.2.4 Classification basée sur la quantification par grille

L'idée de ces méthodes est qu'on divise l'espace de données en un nombre fini de cellules formant une grille. Ce type d'algorithmes est conçu essentiellement pour des données spatiales. Une cellule peut être un cube, une région ou un hyper rectangle. En fait, avec une telle représentation des données, au lieu de faire la classification dans l'espace de données, on la fait dans l'espace en utilisant des informations

statistiques des points dans la cellule. Les méthodes de ce type sont hiérarchiques ou de partitionnement. Les algorithmes les plus connus sont STING, CLIQUE et WaveCluster.

L'algorithme STING (STatistical INformation Grid) [56] utilise une grille multirésolution. Les données dans chaque cellule sont résumées par le nombre d'objets dans la grille et par leur moyenne. Chaque cellule est récursivement découpée en 4 sous-cellules. On commence en considérant l'espace entier comme une cellule ancêtre de la hiérarchie et on s'arrête si l'on atteint un critère de terminaison : (le nombre des points dans une cellule est inférieur à un seuil). Pour l'affectation d'un point, on mesure des paramètres statistiques pour chaque cellule : moyenne, variance, minimum, maximum des points, distribution de la cellule (normale, uniforme, exponentielle, …). Quand on remonte dans la hiérarchie, ces informations statistiques sont calculées à l'aide de celles du niveau inférieur. La classification est effectuée dans ces cellules de feuille au lieu de toutes les données, le coût de calcul est donc réduit (O(k) au lieu de O(N), k étant le nombre de cellules et N est le nombre de points de données).

L'algorithme WaveCluster proposé par Sheikholeslami et al en 2000 [47] considère les données spatiales comme les signaux multidimensionnels sur lesquelles on applique une transformation wavelet pour transformer l'espace de données dans un espace de fréquences. L'idée est basée sur le fait que les parties de haute fréquence du signal correspondent aux régions de l'espace « spatial » de données où il y a un changement brusque dans la distribution des objets. Ce sont les frontières des classes. Les parties de basse fréquence du signal correspondent aux régions de l'espace de données où les objets sont concentrés, c'est-à-dire, les classes.

L'algorithme CLIQUE (CLustering In QUEst) a été proposé par Agrawal et al en 1998 [1]. Au lieu de construire les classes dans l'espace original, CLIQUE le fait dans des sous-espaces de dimension la plus grande possible. L'idée est basée sur le fait que si un ensemble de points S est une classe dans un espace de k dimensions, alors S est aussi une partie d'une classe dans n'importe quelle projection en (k – 1) dimensions de cet espace. L'algorithme commence par déterminer toutes les unités denses à une dimension en balayant une fois les données. Après avoir déterminé les unités denses en (k - 1) dimensions, il détermine les unités denses candidates en k dimensions en utilisant une procédure qui génère les unités candidates.

1.2.2.5 Autres méthodes

Plusieurs autres méthodes existent, même si elles sont peu utilisées.

L'algorithme COBWEB a été proposé par Fisher en 1987 [18]. Il construit dynamiquement un dendrogramme en passant en revue les individus un à un. Au cours de la construction de dendrogramme, chaque nouvel individu parcourt l'arbre qui est au fur et à mesure mis à jour. COBWEB est rapide avec une complexité de calculs en O(t.n) où t est une constante dépendant des caractéristiques de l'arbre.

L'algorithme SOM (Self Organizing Maps) [29], [30], [31] s'appuie sur la théorie des réseaux à compétition (réseaux de neurones), c'est à dire sur l'établissement d'une liaison de compétition entre les

neurones autres que ceux d'entrée. En pratique, cela signifie que des liens inhibiteurs relient les neurones. Lors de la phase d'apprentissage, le réseau spécialise ses neurones en reconnaissance de catégories d'entrées, ce qui est en effet, une façon d'apprendre des classes. Une classe est définie comme l'ensemble des exemples reconnus par un neurone de sortie d'un réseau à compétition.

L'algorithme SUPER-PARAMAGNETIC-CLUSTERING [7], fonctionne sur un principe de recuit simulé appliqué à une fonction objectif favorisant la formation de clusters de forme quelconque. Cet algorithme fait varier un paramètre de température, et pour chacune de ses valeurs, calcule les zones denses dans les données selon cette fonction objectif. Le résultat est une séquence de partitions imbriquées, équivalente à une hiérarchie.

1.2.2.6 Discussion

Comme on peut le constater, les méthodes de classification sont diverses et se distinguent essentiellement par les mesures de proximités qu'elles utilisent, par la natures des données qu'elles traitent et par l'objectif final de la classification.

Les méthodes hiérarchiques présentent une flexibilité concernant le niveau de granularité, et une facilité de manipulation de toute forme de similarité ou de distance. Le principal inconvénient de ces méthodes est l'absence de révision des classes une fois construites. La classification Ascendante Hiérarchique est facile à implémenter et est souvent utilisée pour classifier des données de petite taille car elle est coûteuse en termes de complexité ($O(n^2)$). Alors que la classification Descendante Hiérarchique présente l'avantage de ne pas recourir à l'utilisation d'un seuil arbitraire pour la formation des classes. Cependant, les résultats sont généralement grossiers et les niveaux des nœuds de la hiérarchie ne sont définis que par l'ordre dans lequel ils apparaissent.

Le principal atout des méthodes de recherche de partitions est leur grande simplicité. Elles sont très intéressantes en termes de complexité. De plus, les classes sont facilement interprétables et représentées naturellement par les Centroïdes. Elles sont aisément applicables à des données de grande taille. Ces méthodes présentent cependant, certains inconvénients : elles fournissent la plus part de temps une partition localement optimale, le résultat dépend de la manière dont les classes sont initialisées, elles ne détectent pas les données bruitées, même si les algorithmes PAM et les k-medoïdes sont plus robustes que les k-means en présence de bruit.

Dans les méthodes basées sur la densité, les algorithmes DBSCAN et DENCLUE présentent l'intérêt de trouver eux-mêmes une évaluation du nombre de classes qui peuvent avoir des formes arbitraires. Ils permettent également de bien gérer les données aberrantes.

Dans les méthodes basées sur la quantification par grille, WaveCluster est capable de traiter efficacement les gros volumes de données spatiales. Il peut découvrir des classes de forme arbitraire de différents niveaux du détail et traite bien le bruit.

1.3 Contributions

Nous avons proposé de nouvelles approches de classification automatique des données et nous les avons expérimentées. Nous avons tout d'abord défini les objectifs, puis nous avons cherché le moyen le plus efficace pour les réaliser. Plusieurs propositions ont été implémentées et testées.

1.3.1 Quelques limites des travaux existants

Comme indiqué dans la section précédente, on recense plusieurs travaux permettant d'améliorer les méthodes de Classification Ascendante Hiérarchique en proposant entre autres, de nouvelles approches pour mesurer la similarité entre individus, de nouvelles stratégies d'agrégation, ou travailler sur des échantillons compacts des données. La CAH ne vise généralement pas à optimiser un critère global portant sur la hiérarchie obtenue sur l'ultramétrique. D'autre part, la majorité des méthodes de classification ne sont pas adaptées à l'analyse des données mixtes ou des données incomplètes.

Avec le développement des moyens de stockage et de collecte des données, le type des données à analyser s'est complexifié et comporte souvent des données mixtes. La plupart des méthodes classiques traitent uniquement des données quantitatives. Lorsqu'on est amené à faire une classification, le résultat peut être influencé par la mesure de similarité (ou dissimilarité). L'équivalence entre les mesures de proximité ou non, n'est pas prise en considération.

Parmi nos travaux sur la classification automatique des données, nous présentons ici trois contributions répondant à certaines à ces limites.

1.3.2 Classification par agrégation des opinions

Nous proposons ici un travail de classification automatique basé sur la notion d'agrégation d'opinions. La méthode proposée consiste à associer à chaque variable une fonction de classement qui va jouer le rôle d'un juge. Ce dernier va classer les individus selon ses propres critères. En rassemblant les classements de toutes les variables, pour tous les couples d'individus, on cherche à construire un classement collectif sur les individus (meilleure agrégation possible). La recherche de la partition résultat est basée sur la maximisation de la fonction de concordance globale.

Considérons l'ensemble des variables, $V = \{V_1, V_2, \ldots, V_p\}$

et l'ensemble des individus, $\mathfrak{I} = \{\omega_1, \omega_2, \ldots, \omega_n\}$

Définition

Etant donné une variable V_k et un couple d'individus (ω_i, ω_j), on définit une application

$$A_k(.,.) = \Im^2 \rightarrow \{0,1\}$$

telle que $\forall\,(\omega_i, \omega_j) \in \Im^2, \omega_i \neq \omega_j$

$$A_k(\omega_i, \omega_j) = \begin{cases} 1 & si\ \omega_i\ et \omega_j\ sont\ dans\ la\ même\ classe \\ 0 & \omega_i\ et \omega_j\ sont\ dans\ deux\ classes\ différentes \end{cases}$$

C'est une fonction de classement qui va jouer le rôle de juge pour la variable V_k. La figure 1 donne une représentation des différents tableaux de classements de toutes les variables.

1.3.2.1 Présentation de la méthode

Etant donné les classements de p variables $A_1, A_2, \ldots, A_p$ nous cherchons à construire sur l'ensemble $\Im$ un classement collectif qui soit la meilleure agrégation possible, qui à la fois va engendrer sur l'ensemble $\Im$ une relation d'équivalence maximisant le nombre de concordances entre le classement collectif et l'ensemble des classements des variables. Une partition sur l'ensemble $\Im$ peut être assimilée à un vote défini sur $\Im$ qui peut prendre autant de modalités qu'il y a d'éléments dans la partition. Nous identifierons cette partition à une application

$$X(.,.) = \Im^2 \rightarrow \{0,1\}$$

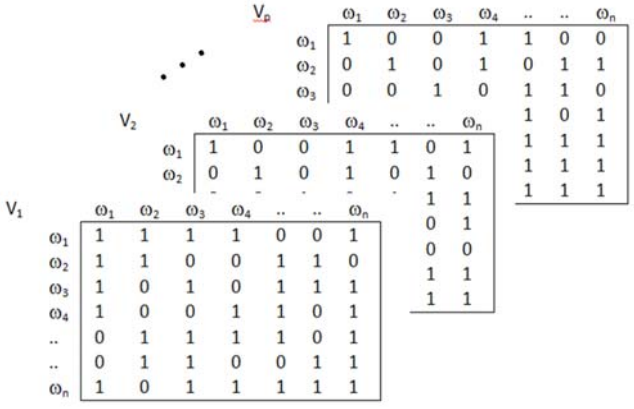


Figure 1 : tableaux des classements des différentes variables

Définition

Etant donné deux applications X(. , .) et Y(. , .) de $\Im^2$ $dans$ {0,1}, on appelle concordance de X(. , .) et Y(. , .) notée C(X,Y), le nombre positif défini par :

$$C(X,Y) = \frac{1}{n^2}\ card\{(\omega_i,\omega_j)\ ;\ \ X(\omega_i,\omega_j) = Y(\omega_i,\omega_j)\}$$

Cette concordance peut bien entendu être décomposée en deux concordances, l'une positive et l'autre négative. Cette définition peut s'appliquer aisément à la concordance $C(X,A_k)$ entre une partition caractérisée par X(. , .) et le classement A_k d'une variable V_k.

Les différents classements A_k des variables, (k=1, . . . , p) étant donnés, nous cherchons X qui maximise l'expression :

$$C(X,\boldsymbol{A}) = \sum_{k=1}^{p} C\ (X,A_k)$$

Il s'agit d'un problème d'optimisation qui peut être résolu par une méthode d'optimisation linéaire en nombres entiers. Si on pose

$$x_{ij} = X(\omega_i,\omega_j)\ ;\quad (\omega_i,\omega_j) \in\ \Im^2$$

$$a_{ij}^k = A_k(\omega_i,\omega_j)\ ;\quad k = 1,\dots,p$$

On obtient

$$C(X,A_k) = \frac{1}{n^2}\ card\{(\omega_i,\omega_j);\ \ X(\omega_i,\omega_j) = A_k(\omega_i,\omega_j)\} = \frac{1}{n^2}\ card\{(\omega_i,\omega_j);\ \ x_{ij} = a_{ij}^k\}$$

Le couple (ω_i,ω_j) contribue donc à la cohérence dans les deux cas suivants :

- $x_{ij} = a_{ij}^k = 1\ \Leftrightarrow\ x_{ij}\ \ a_{ij}^k = 1$
- $x_{ij} = a_{ij}^k = 0\ \Leftrightarrow\ (1-x_{ij})(1-a_{ij}^k)\ = 1$

D'où,

$$C(X,\boldsymbol{A}) = \frac{1}{n^2}\sum_{k=1}^{p}\sum_{(i,j)} x_{ij}\ \ a_{ij}^k + (1-x_{ij})(1-a_{ij}^k)$$

1.3.2.2 Nature des données

Nous allons voir ici que le problème peut traiter des données complètes mais aussi des données à valeurs manquantes. Dans le cas des données complètes, le critère d'optimisation peut être présenté comme suit. Partant des notations suivantes :

$$x_{ij} = \begin{cases} 1 & si\ X(\omega_i, \omega_j) = 1 \\ 0 & sinon \end{cases} \qquad \bar{x}_{ij} = \begin{cases} 1 & si\ X(\omega_i, \omega_j) = 0 \\ 0 & sinon \end{cases}$$

$$a_{ij}^k = \begin{cases} 1 & si\ A_k(\omega_i, \omega_j) = 1 \\ 0 & sinon \end{cases} \qquad \bar{a}_{ij}^k = \begin{cases} 1 & si\ A_k(\omega_i, \omega_j) = 0 \\ 0 & sinon \end{cases}$$

(Pour i = 1,. . . , n ; j=1,. . . , p et k = 1, . . . , p)

On peut dire qu'il y a concordance entre la partition donnée par X et la partition donnée par A_k si $x_{ij} = 1\ et\ a_{ij}^k = 1$ ou si $\bar{x}_{ij} = 1\ et\ \bar{a}_{ij}^k = 1$

Le nombre de concordances entre la partition X et les partitions données par A_k , (k=1,. . p) peut s'écrire :

$$C(X, \boldsymbol{A}) = \sum_{k=1}^{p} C(X, A_k) = \sum_{k=1}^{p} \sum_{(i,j)} \bar{a}_{ij}^k + \sum_{k=1}^{p} \sum_{(i,j)} \left[a_{ij}^k - \bar{a}_{ij}^k\right] x_{ij}$$

En notant respectivement $r_{ij} = \sum_{k=1}^{p} a_{ij}^k$ (et $\bar{r}_{ij} = \sum_{k=1}^{p} \bar{a}_{ij}^k$) le nombre de variables qui classent ω_i et ω_j dans la même classe (et le nombre de variables qui les affectent à deux classes différentes) on peut écrire :

$$C(X, \boldsymbol{A}) = \sum_{(i,j)} \bar{r}_{ij} + \sum_{(i,j)} s_{ij} x_{ij}$$

Avec $s_{ij} = r_{ij} - \bar{r}_{ij}$

Dans les cas des données manquantes, la fonction de classement A_k peut juger que deux individus ne sont pas classifiables en donnant une réponse M (pour manquant) au lieu de donner une valeur 0 ou 1. Par conséquent, toutes les fonctions de classement A_k, (k=1, . . . , p) deviennent des applications :

$$A_k : \Im^2 \to \{0,1,M\}$$

Où $A_k(\omega_i, \omega_j) = M$ signifie que la variable V_k juge que le couple (ω_i, ω_j) est non classifiable.

En associant à la variable V_k le vecteur :

$$v_{ij}^k = \begin{cases} 1 & si\ A_k(\omega_i, \omega_j) = M \\ 0 & sinon \end{cases}$$

La propriété $a_{ij}^k = 1$ ou $\bar{a}_{ij}^k = 1$ n'est vraie que si la variable V_k a donné un avis sur le couple (ω_i, ω_j). D'autre part, nous avons $r_{ij} + \bar{r}_{ij} \leq p$.

En prenant en compte ces remarques, la relation suivante reste toujours valable :

$$C(X, \boldsymbol{A}) = \sum_{(i,j)} \bar{r}_{ij} + \sum_{(i,j)} s_{ij} x_{ij}$$

1.3.2.3 Résolution du problème

La solution de ce problème revient à maximiser l'expression $C(X, \boldsymbol{A})$. Ce problème est connu sous le nom de « Clique partitioning » et un algorithme basé sur la programmation linéaire a été proposé par Grotschel et Wakbayashi [20]. Le nombre de variables est de l'ordre de n^2, le nombre de contraintes est de l'ordre de n^3 et ce problème est NP-complexe.

Nous avons proposé un algorithme de résolution de ce problème en se basant sur une technique métaheuristique : le recuit simulé. Cet algorithme permet d'avoir une solution facile d'application, avec un temps de calcul raisonnable. Remarquons que la donnée du vecteur x_{ij} est équivalente à celle d'une partition $\mathcal{P}$ sur $\Im$, selon le schéma suivant :

$$x_{ij} = 1 \Leftrightarrow \exists\, G \in \mathcal{P} \ tel\ que\ i \in G\ et\ j \epsilon G$$

Si on note $Val(\mathcal{P}) = \sum_{(i,j)} s_{ij} x_{ij}$ et $\mathcal{P} = \{G_1, G_2, \ldots, G_m\}$, on peut écrire :

$$Val(\mathcal{P}) = \sum_{h=1}^{m} \sum_{(i,j) \epsilon Gh} s_{ij}$$

Notre problème revient donc à maximiser l'expression :

$$max\{Val(\mathcal{P}), \mathcal{P} \epsilon \mathbb{P}\}$$

Où $\mathbb{P}$ désigne l'ensemble des partitions de $\Im$.

Définition

Une partition $\mathcal{P}'$ appartient à $\mathcal{V}(\mathcal{P})$ si elle dérive de $\mathcal{P}$ en déplaçant un seul élément i_0 comme suit :

Supposons que $i_0 \in G_h$. Alors,

- Soit i_0 est transféré dans un groupe G_l $(l \neq h), de\ \mathcal{V}(\mathcal{P})$
- Soit i_0 constitue à lui tout seul un nouveau groupe G_l

Il s'en suit : $\Delta = Val(\mathcal{P}') - Val(\mathcal{P})$

Pour simplifier cette écriture, on pose :

$$s_{ij} = \begin{cases} s_{ij} & si\ i < j \\ s_{ij} & si\ i > j \\ 0 & si\ i = j \end{cases}$$

D'où l'expression

$$Val(\mathcal{P}) = \sum_{j \epsilon Gl} s_{i0j} - \sum_{j \epsilon Gh - \{i0\}} s_{i0j}$$

Algorithme

Données :
T_0 : Température initiale
K_{max} Nombre d'itérations dans la boucle interne
r_0 : Coefficient de refroidissement
Résultat : $\mathcal{P}*$
Début
Initialisation :
$T=T_0$
Choisir au hasard une partition $\mathcal{P}$ de $\mathfrak{I}$; $\mathcal{P}*= \mathcal{P}$
Répéter
Change= FAUX
Pour k= 1 à K_{max} faire
Choisir au hasard une partition $\mathcal{P}$' dans $\mathcal{V}(\mathcal{P})$
Calculer $\Delta = Val(\mathcal{P}') - Val(\mathcal{P})$
Tirer au hasard une valeur R dans [0,1] (loi uniforme)
Si $\Delta > 0$ ou $R < exp(\frac{\Delta}{T})$ alors $\mathcal{P}= \mathcal{P}'$; Change=VRAI ;
Si $Val(\mathcal{P}) > Val(\mathcal{P}^*)$ alors $\mathcal{P}^* = \mathcal{P}$;
Fin si
Fin si
Fin pour
$T=T r_0$*
Jusqu'à change=FAUX ;
Fin

Remarques

- La définition de $\mathcal{V}(\mathcal{P})$ permet d'atteindre toutes les partitions possibles à partir d'une partition initiale quelconque en utilisant un procédé itératif.
- Nous avons choisi d'initialiser les paramètres selon les recommandation de (119) : $K_{max} = 10 * Card(\mathfrak{I})$ et $r_0 = 0{,}98$.
- Soient $\mathcal{P}$ une partition choisie au hasard et $\mathcal{P}'$ une partition telle que $\mathcal{P}' \in \mathcal{V}(\mathcal{P})$ et

$$\Delta = Val(\mathcal{P}') - Val(\mathcal{P}) < 0$$

Le choix de la valeur de T_0 doit garantir qu'en moyenne $e^{\Delta/T_0} \approx 1$.

1.3.2.4 Application de la méthode

Nous avons testé cette méthode sur le jeu de données des cétacés utilisées par Grotschel et Wakbayashi [20]. Ces derniers ont fourni une solution optimale. La méthode a donc été appliquée sur le tableau de données des 36 cétacés décrits par 15 variables [8]. Ce tableau comporte quelques données manquantes représentées par le symbole (*). Sur 9096 avis exprimés, le pourcentage de concordances est de 0,705. Une partition en 7 classes donne les résultats suivants :

Classe 1	*{Delphinapterus, Monodon}*
Classe 2	*{Balaenoptea, Balaenoptea Mus, Eschrichtius, Megaptera}*
Classe 3	*{Balaena, Eubalaena, Neophalaena}*
Classe 4	*{Inia, Lipotes, Platanista, Stenodelphis}*
Classe 5	*{Kogia, Physeter}*
Classe 6	*{Berardius, Hyperoodon, Mesoplodon, Tasmacetus, Ziphus}*
Classe 7	*{Cephalorhynehus, Delphinus, Globicephala, Grampus, lagenorhynchus, Lissodelphis, Neophocaena, Orcaella, Orcinus, Phocaena, Pseudorea, Sotalia, Sousa, Stenella, Steno, Tussio}*

Le résultat de la partition de l'ensemble des cétacés en 7 classes coïncide parfaitement avec la partition optimale fournie par Grotschel et Wakbayashi [20]. Le pourcentage de concordances $\tau = \frac{NA(X)}{N} =$

0,705 permet de situer la qualité de la partition obtenue. Le pourcentage de concordance par juge varie de 0,54 à 0,928. Il éclaire sur la contribution des juges à la création de la partition obtenue.

1.3.2.1.5 Généralisation de la méthode

Soient $\Im$ un ensemble muni d'une dissimilarité ρ et une partition $\mathcal{P} = \{G_1, G_2, \dots, G_m\}\ de\ \Im$. On définit une mesure qui est fonction décroissante de la qualité de $\mathcal{P}$ de la manière suivante :

$$Qval(\mathcal{P}) = \sum_{h=1}^{m} \sum_{(i,j)\epsilon Gh} \rho_{ij}$$

Pour un nombre donné de classes, noté m, on cherche à résoudre le problème :

$$\min(Qval(\mathcal{P})\ ;\ \mathcal{P} \in \mathbb{P}_m)$$

$\mathbb{P}_m$ désigne l'ensemble des partitions de $\Im$ en m classes.

Un avantage considérable de cette formulation réside dans le fait que le choix du nombre de classes peut être réglé par le problème d'optimisation cité précédemment.

Lorsqu'on dispose d'une mesure de dissimilarité ρ, on peut ramener un problème de classification à notre formulation précédente de la façon suivante :

$$\bar{\rho} = \frac{1}{card(\Im)} \sum_{(i,j)\epsilon Gh} \rho_{ij} \quad \text{et} \quad s_{ij} = \bar{\rho} - \rho_{ij}$$

Le réel s_{ij} est compris entre $\bar{\rho} - \rho_{max}$ et $= \bar{\rho}$

où $\rho_{max} = max(\rho_{ij},\ (i,j) \in \Im^2)$.

Ainsi, pour une partition $\mathcal{P} = \{G_1, G_2, \dots, G_m\}\ de\ \Im$, on définit

$$Val(\mathcal{P}) = \sum_{h=1}^{m} \sum_{(i,j)\epsilon Gh} s_{ij}$$

Le problème de maximisation de $Val(\mathcal{P})$ est similaire à celui traité précédemment et sa résolution par la méthode du recuit simulé reste valable.

1.3.2.6 Conclusion

Nous avons proposé ici une méthode de classification basée sur l'agrégation d'opinions. Elle consiste à associer à chaque variable une fonction de classement qui va jouer le rôle de juge. Ce dernier va classer

les individus selon ses propres critères. L'ensemble de classement de toutes les variables permet de construire, sur l'ensemble des individus, un classement collectif qui soit le meilleur compromis possible. Notons d'abord que l'algorithme du recuit simulé comporte un aspect aléatoire, ce qui explique que deux exécutions successives sur le même exemple ne donnent pas nécessairement le même résultat. Pour cela nous appellerons "Résultat" de l'algorithme, le meilleur des résultats obtenus lors de plusieurs exécutions successives.

Le principal avantage de cette méthode de classification est que les données ne sont pas nécessairement quantitatives et sa particularité par rapport d'autres méthodes est que nous n'avons pas fait recours aux calculs de distance ou de dissimilarité. Cependant, nous avons vu dans l'extension de la méthode que cette approche est applicable dans le cas où nous voulons utiliser une mesure de proximité. Le deuxième avantage est la possibilité d'appliquer la méthode en présence de données manquantes. Enfin, nous pouvons dire que cette méthode s'appuie sur l'opinion de l'expert du domaine, qui définit les fonctions de classement.

1.3.3 Recherche de l'ultramétrique la plus proche d'une dissimilarité en CAH

1.3.3.1 Introduction

La méthode de Classification Ascendante Hiérarchique (CAH), est une méthode itérative largement utilisée dans divers domaines [12]. L'algorithme de CAH fournit une hiérarchie valuée qui peut être indicée ou non. Cette dernière permet de construire une ultramétrique sur un ensemble d'individus. La CAH ne vise généralement pas à optimiser un critère global portant sur la hiérarchie indicée obtenue sur l'ultramétrique. Nous proposons ici une méthode de recherche de l'ultramétrique la plus proche de la dissimilarité définie sur l'ensemble des individus. Pour minimiser l'écart entre la dissimilarité adoptée et l'ultramétrique, nous avons utilisé l'algorithme du recuit simulé. Après quelques définitions des hiérarchies valuées et indicées, nous présentons le problème à résoudre et nous justifions la décomposition de ce problème en sous-problèmes intermédiaires à résoudre. Lors de la résolution du problème, nous proposons d'utiliser des méthodes spécifiques afin de réduire la complexité des calculs.

1 .3.3.2 Définitions

Considérons un ensemble fini *I* d'individus, de cardinal *n* et notons P l'ensemble des paires d'éléments de *I*

$$P = \{\{i,j\}\ i \in I, j \in I\}$$

On suppose que I est muni d'une dissimilarité ρ. Nous rappelons qu'un ensemble de parties $\mathcal{A}$ non vides de *I*, ordonné par la relation d'inclusion est appelé hiérarchie totale binaire sur *I*, s'il satisfait aux trois propositions suivantes:

(i) $\forall A, B \in \mathcal{A}\ soit\ A \cap B = \phi\ soit\ (A \subset B\ ou\ B \subset A)$

(ii) $\forall I \in \mathcal{A}\ et\ \forall\ i \in I, \{i\} \in \mathcal{A}$

(iii) $\forall A \in \mathcal{A}, Card(A) > 1 \Rightarrow \exists\ A_1, A_2 \in \mathcal{A}\ \ tq\ A = A_1 \cup A_2$

En pratique, la gestion d'un arbre hiérarchique $\mathcal{A}$ est assurée au moyen de deux fonctions *a* et *b,* tels que $\forall A \in \mathcal{A}$ vérifiant

$$A = A_1 \cup A_2 \text{ avec } A_1 \in A \text{ et } A_2 \in A,\ alors\ a(A) = A_1 \text{ et } b(A) = A_2.$$

On appelle $a(A)$ le fils aîné de *A* et $b(A)$ le benjamin de *A*. Une permutation entre $a(A)$ et $b(A)$ ne modifie pas la hiérarchie. Dans la suite, par souci de simplification, nous parlerons de hiérarchie pour désigner une hiérarchie totale binaire.

Soit Δ une dissimilarité sur l'ensemble des parties de I. On dit que la hiérarchie $\mathcal{A}$ est valuée par la fonction f : $\mathcal{A} \rightarrow R^+$ si

- $f(\{i\}) = 0\ \forall\ i \in I$.
- Si Card(A) $> 1, \exists A_1, A_2 \in \mathcal{A}, A = A_1 \cup A_2$, $alors\ f(A) = \Delta(A_1, A_2)$

Nous noterons $\mathcal{H}_0$ l'ensemble des hiérarchies valuées sur *I* et $H \in (\mathcal{A}, f)$ un élément de $\mathcal{H}_0$

Une hiérarchie valuée *(𝒜,f)* est dite indicée si $\forall A, B \in \mathcal{A}, A \subset B \Rightarrow f(A) \leq f(B)$

Nous désignerons par $\mathcal{H}$ l'ensemble des hiérarchies indicées sur *I*. On sait qu'une hiérarchie indicée permet de construire une distance ultramétrique δ sur *I* de la façon suivante :

Etant donné $\boldsymbol{\mathcal{H}} = (A, f)$, $\forall (i,j) \in P$ on considère le plus petit élément $M(i,j)$, au sens de l'inclusion, de A qui contient à la fois *i* et *j*. $\delta(i,j)$ est alors défini par :

$$\delta(i,j) = f(M(a,b)) \qquad (1)$$

1.3.3.3 Formulation du problème

Réciproquement, la donnée d'une distance ultramétrique permet de construire une hiérarchie indicée. Notons que si *(𝒜,f)* est une hiérarchie valuée, l'expression (1) permet seulement de construire une dissimilarité δ sur *I*. Les algorithmes ascendants de construction de hiérarchies, le plus souvent, ne visent pas à optimiser un critère global portant sur la hiérarchie indicée construite sur l'ultramétrique. L'algorithme de *Lance* et *Williams* basé sur le critère du saut minimal est l'exception à cette affirmation. Plusieurs auteurs ont posé le problème de trouver l'ultramétrique δ* qui minimise la quantité :

$$\Phi(\delta) = \sum_{\{x,y\} \in P} \left[\rho(i,j) - \delta(i,j)\right]^2$$

Devant la complexité de ce problème, des algorithmes de recherche d'un optimal local ont été proposés *Chandon et al.* Dans [12] on a proposé un algorithme (*Branch and bound*) fournissant un optimal global, mais inutilisable pour des ensembles comportant plus d'une dizaine d'éléments.

Nous désignons par (P) le problème suivant : trouver une hiérarchie indicée $\boldsymbol{H}^* = (A^*, f^*)$ tell que :

$$\Phi(\delta^*) = min[\Phi(\delta) \mid H \in \mathbf{H}] \qquad (P)$$

où (δ et δ^* désignent respectivement les ultramétriques associées à **H** et **H** *)

C'est le problème (P) que nous proposons de résoudre par la méthode du recuit simulé. Avant de résoudre le problème (P), nous allons d'abord proposer une méthode permettant de résoudre le problème (P_0) suivant:

Trouver la hiérarchie valuée $H_0^* = (A_0^*, f_0^*)$ telle que

$$\Phi(\delta_0^*) = min(\Phi(\delta) \mid \in \mathcal{H}_0) \qquad (P_0)$$

où δ et δ_0 désignent respectivement les dissimilarités associées à *H* et *H**

1.3.3.4 Résolution du problème (P_0)

Principe de l'algorithme.

Le principe consiste à construire une hiérarchie valuée H' par modification élémentaire de H, calculer $\Delta_0 = \Phi(\delta') - \Phi(\delta^*)$ et affecter H' à H si ($\Delta_0 < 0$ ou $Random < e^{-\Delta/T}$). La propriété suivante caractérisant la valuation de l'optimal permet de résoudre le problème (P_0)

Propriété 1:

Soit $H_0^* = (A_0^*, f_0^*)$ une solution du problème (P_0). Pour tout sommet *s* de H_0^*, on a nécessairement:

$$f_0^*(s) = \sum_{\substack{i \in a(s) \\ j \in b(s)}} \rho(i, j) \frac{1}{|a(s)||b(s)|} \qquad (2)$$

Remarque:

On peut noter que la solution H^* du problème (P) vérifie la même propriété.

Nous imposerons donc à la valuation *f* de la hiérarchie valuée H de l'algorithme de toujours vérifier la propriété (2), plus précisément:

- la hiérarchie H_0 sortante de l'algorithme sera construite de façon à vérifier (2)
- Après chaque modification élémentaire, la fonction *f* sera également modifiée, de façon que (2) soit toujours vérifiée.

1.1 Etude d'une modification élémentaire de la hiérarchie valuée $H = (A, f)$

La modification élémentaire envisagée procède en deux temps:

- modification de l'arbre A
- mise à jour de la valuation de façon que (2) soit vérifiée.

Modification de A:

Pour définir la modification de l'arbre A, nous faisons choix de deux éléments *P* et *Q* de *A* vérifiant:

$$Q = b(P) \ \ et \ \ Card(Q) > 1$$

La modification envisagée dépend de *P*, et sera donc notée *τ(P)*. L'arbre obtenu après transformation de A par *τ(P)* sera noté *A'*, nous notons a' et b' les fonctions *"aîné"* et *"benjamin"* attachées à l'arbre A'. Nous notons $A = a(P)$, $B = a(Q)$ *et* $C = b(Q)$. $\tau(P)$ est la transformation qui, à l'arbre A associe l'arbre A' défini par $A' = A - Q + Q'$ où $Q' = A \cup B$. On vérifiera facilement que si A est une hiérarchie totale binaire sur I, il en est de même de A' . Les fonctions a' et b' sont alors définies comme suit:

$$\forall X \in A \text{ tel que } Card(X) > 1 \text{ et } X \neq P, X \neq Q$$

$$a'(X) = a(x) \qquad b'(X) = b(X)$$

$$a'(P) = C \qquad b'(P) = Q'$$

$$a'(Q') = B \qquad b'(Q') = A$$

Mise à jour de la valuation :

Nous notons $H' = (A', f')$ la hiérarchie valuée transformée de H. Il est claire que, compte tenu de (2):

$$f'(P) = \sum_{\substack{i \in A \\ j \in B \cup C}} \rho(i,j) \frac{1}{|A||B \cup C|} \qquad (3)$$

$$f'(Q') = \sum_{\substack{i \in A \\ j \in B}} \rho(i,j) \frac{1}{|A||B|} \qquad (4)$$

Sur le plan algorithmique, il suffira de calculer la quantité $f'(a')$ par la formule (4) ci-dessous la quantité $f'(P)$ pourra s'en déduire en opérant le calcul suivant:

$$f'(P) = \frac{|A|}{|A|+|B|}\beta + \frac{|A|}{|A|+|B|} f'(Q) \text{ avec } \beta = \frac{|B|+|C|}{|C|}\left(f(p) - \frac{|B|}{|B|+|C|} f'(Q')\right) \qquad (5)$$

Détermination de la variation de critère

Soit Δ_0 la quantité $\Phi(\delta') - \Phi(\delta)$, compte tenu de la définition de δ et δ', et des remarques précédentes, on a:

$$\Delta_0 = \sum_{\substack{i \in A, j \in B \\ i \in A, j \in C \\ i \in B, j \in C}} [(\rho(i,j) - \delta'(i,j))^2 - (\rho(i,j) - \delta(i,j))^2]$$

Les relations suivantes permettent un calcul simple de Δ_0 *:*

$$\sum_{i \in A, j \in B} [(\rho(i,j) - \delta'(i,j))^2 - (\rho(i,j) - \delta(i,j))^2] = -(f(P) - f'(Q'))^2 |A||B|$$

$$\sum_{i \in B, j \in C} [(\rho(i,j) - \delta'(i,j))^2 - (\rho(i,j) - \delta(i,j))^2] = (f(Q) - f'(P))^2 |B||C|$$

$$\sum_{i \in A, j \in C} [(\rho(i,j) - \delta'(i,j))^2 - (\rho(i,j) - \delta(i,j))^2] = (f(P) - f'(P))(2\beta - f(P) - f'(P))|A||C|$$

Avec β défini dans (5). Seul le calcul de $f'(Q')$ a une complexité au pire en $O(n^2)$, les autres calculs ayant une complexité constante.

1.3.3.5 Résolution du problème (P)

L'expérimentation montre que la dissimilarité issue de l'exécution de l'algorithme n'est pas, en général, une ultramétrique. Autrement dit, la hiérarchie valuée obtenue présente des inversions. La solution que nous proposons pour résoudre le problème (P) consiste à continuer à travailler sur l'ensemble H_0 des hiérarchies valuées, en utilisant les transformations définies précédemment, mais en pénalisant les hiérarchies présentant des inversions.

Pour éviter des calculs trop complexes, nous allons procéder comme suit:

- Lors d'une transformation $\tau(P)$, nous contrôlons seulement la présence d'une inversion entre P et Q, avant et après la transformation, et suivant le cas:

a) Si $f(Q) \leq f(P)\ et\ f'(Q') > f'(P)$ (Création d'une inversion). Δ_0 est remplacé par $\Delta = \Delta_0 + k\ ,\ (k > 0)$, où k est un réel positif

b) Si $f(Q) > f(P)\ et\ f'(Q') \leq f'(P)$ (Suppression d'une inversion). Δ_0 est remplacé par $\Delta = \Delta_0 - k\ , (k > 0)$

c) Dans tous les autres cas, on prend $\Delta = \Delta_0$, de cette façon les hiérarchies qui ne présentent pas d'inversion sont favorisées.

Le choix de k peut être délicat. En effet s'il est trop petit, les inversions ne seront pas assez pénalisées et il subsistera des inversions dans la hiérarchie obtenue. Au contraire, s'il est trop grand, les inversions seront systématiquement rejetées et l'algorithme pourra ne pas trouver une séquence de transformation conduisant de la hiérarchie initiale à une hiérarchie (proche de) l'optimal.

Une solution consiste à ajuster le choix de k en prenant $k = |\Delta_0| + \varepsilon$, (ε réel positif suffisamment petit devant les valeurs de $|\Delta_0|$ observées en moyenne), ce qui revient à faire le choix de Δ suivant le cas:

a) Création d'une inversion:
Si $\Delta_0 < 0$ *alors* $\Delta = \varepsilon > 0$, *Si* $\Delta_0 > 0$ *alors* $\Delta = 2\Delta_0 < 0 + \varepsilon > 0$
Ainsi l'introduction d'une inversion sera toujours rejetée par une valeur suffisamment basse de la température (elle sera vue comme une dégradation du critère Φ)

b) Suppression d'une inversion:
Si $\Delta 0 < 0$ *alors* $\Delta = +2\Delta 0 - \varepsilon < 0$, *Si* $\Delta 0 > 0$ *alors* $\Delta = -\varepsilon < 0$
Ainsi la suppression d'une inversion sera vue comme une amélioration du critère et sera toujours accepté.

L'introduction d'une inversion sera toujours vue comme une dégradation du critère et sera acceptée pour des valeurs grandes de T et rejetée pour des valeurs faibles de T. Il est clair que, lors de la transformation $\tau(P)$, une autre inversion (ou suppression) peut être introduite entre P et son père, ou entre Q et l'un de ses fils, mais ces autres inversions seront supprimées avec une probabilité égale à 1 ultérieurement. Sur aucun des essais effectués, la hiérarchie finale n'a présenté d'inversion.

1.3.3.6 Conclusion

La méthode proposée permet de déterminer l'ultramétrique la plus proche de la dissimilarité définie sur un ensemble d'individus à classifier. Nous savons que la recherche d'une ultramétrique optimale est un problème NP-complet. Sachant que le recuit simulé est adapté à la résolution de tels problèmes, nous avons montré comment cet algorithme peut être utilisé dans ce contexte particulier. L'inconvénient principal de cette méthode est la complexité des calculs du fait de l'utilisation de l'algorithme du recuit simulé.

1.3.4 Equivalence topologiques des mesures de proximité et classification

Le choix d'une mesure de proximité est un élément important dans plusieurs situations. Le cas particulier de la recherche d'information dans une base de données ou sur internet en est un bon exemple. En soumettant une requête à un moteur de recherche, de manière rapide, celui-ci nous retourne une liste de réponses classées selon leur degré de pertinence par rapport à la requête. Ce degré de pertinence peut alors être perçu comme une mesure de dissimilarité/similarité entre la requête et les objets disponibles dans la base. Est-ce que la façon dont on mesure la similarité ou la dissimilarité entre objets affecte le résultat d'une requête ? Si oui, comment décider de quelle mesure de similarité ou de dissimilarité il faut se servir. Il en est de même quand dans de nombreux domaines on souhaite réaliser un regroupement des individus en classes. La manière de mesurer la distance impacte directement la composition des groupes obtenus.

1.3.4.1 Introduction

Lorsqu'on est amené à faire des opérations de classification, de comparaison ou de structuration des données, on est amené à choisir une mesure de proximité entre les individus ou les éléments étudiés. Les résultats obtenus sont souvent impactés par le choix de la mesure de proximité adoptée. Ces mesures de proximité sont caractérisées par des propriétés mathématiques précises. Sont-elles, pour autant, toutes équivalentes ? Peuvent-elles être utilisées dans la pratique de manière indifférenciée ? Autrement dit, est-ce que, par exemple, la mesure de proximité entre individus plongés dans un espace multidimensionnel comme $\mathbb{R}^P$, influence-t-elle le résultat des opérations comme la classification en groupes ou la recherche des k-plus-proches voisins ?

Le terme proximité recouvre des significations telles que la similarité, la ressemblance, la dissimilarité, la dissemblance, etc. On trouve dans la littérature des dizaines de mesures différentes, notamment si on prend en compte la diversité des types de données (binaires, quantitatifs, qualitatifs, flou...). Dès lors, le choix de la mesure de proximité reste posé. Certes, le contexte d'application, les connaissances a priori, le

type de données etc., peuvent aider à identifier les mesures idoines. Par exemple, si les objets à comparer sont décrits par des vecteurs booléens, on peut se limiter à une catégorie de mesures spécifiquement dédiées. Néanmoins, comment faire quand le nombre de mesures candidates reste grand ? Si toutes les mesures étaient équivalentes, il suffirait d'en prendre une au hasard.

Pour faire face à ce problème de comparaison et de choix entre mesures de proximités, trois approches sont utilisées :

1. Par agrégation de mesures : il s'agit d'éviter de choisir une mesure particulière. Par exemple, Richter [44] utilise, sur un même jeu de données, plusieurs mesures de proximité et agrège ensuite, arithmétiquement, les résultats partiels de chacune en une valeur unique. Le résultat final, peut être perçu comme une synthèse des différents points de vue exprimés par chaque mesure de proximité. Cette approche, évite ainsi de traiter la question de la comparaison qui reste cependant un problème en soi.
2. Par évaluation empirique : de nombreux travaux exposent des méthodologies pour comparer les performances des différentes mesures de proximité. Pour cela il est fait appel soit à des benchmarks comme dans Liu et al. , Strehl et al. [50] dont les résultats attendus sont connus préalablement, soit à des critères jugés pertinents pour l'utilisateur et qui permettent, in fine, d'identifier la mesure de proximité la plus appropriée. On peut citer quelques travaux dans cette catégorie comme Malerba et al. [37], Spertus et al. [49]
3. Par comparaison : l'objectif des travaux qui se situent dans cette catégorie vise à comparer les mesures de proximité entre elles. Par exemple, on vérifie si elles ont des propriétés communes Clarke et al. [11], Lerman [33] ou si l'une peut s'exprimer en fonction de l'autre Zhang et Srihari [61], Batagelj et Bren [5] ou simplement si elles fournissent le même résultat sur une opération de classification Fagin et al. [16], etc. Dans ce cas précis, les mesures de proximité peuvent alors être catégorisées selon leur degré de ressemblance. L'utilisateur peut ainsi identifier les mesures qui sont équivalentes de celles qui le sont le moins Lesot et al. [34], Bouchon-Meunier et al. [9].

Le travail que nous présentons ici se situe dans la troisième catégorie qui vise à comparer les mesures de proximité entre elles afin de détecter celles qui sont identiques de celles qui le sont moins. Il s'agit en fait de les regrouper en classes selon leurs similitudes. Pour comparer deux mesures de proximité, l'approche consiste, jusque-là, à comparer les valeurs des matrices de proximité induites Batagelj et Bren [5], Bouchon-Meunier et al. [9], et, le cas échéant, à établir un lien fonctionnel explicite quand les mesures sont équivalentes. Pour comparer deux mesures de proximité, Lerman [33] s'intéresse aux préordres induits par les deux mesures de proximité et évalue leur degré de ressemblance par la concordance entre les préordres induits sur l'ensemble des couples d'objets. D'autres auteurs, Schneider et Borlund [48] évaluent l'équivalence entre deux mesures par un test statistique entre les matrices de proximité. Les indicateurs numériques issus de ces comparaisons croisées servent alors à catégoriser les mesures. L'idée commune à ces travaux de comparaison s'appuie sur un postulat qui dit que deux mesures de proximité sont d'autant plus proches que les préordres induits sur les couples d'objets ne changent pas. Pour cela, on va s'intéresser à la structure de voisinage des objets que l'on appellera la structure topologique induite par la mesure de proximité. Si la structure de voisinage entre objets, induite par une mesure de proximité u_i ne change pas par rapport à celle d'une autre mesure de proximité u_j, cela signifie que les ressemblances locales entre individus n'ont pas changées. Dans ce cas, on dira que les mesures de proximité u_i et u_j sont en équivalence topologique. On pourra ainsi calculer une mesure d'équivalence topologique entre les

couples de mesures de proximité et effectuer ensuite une classification sur les mesures de proximité. Nous allons définir cette nouvelle approche et montrer les premiers liens que nous avons identifiés entre elle et celle basée sur la préordonnance. Après une description du cadre théorique dans lequel nous nous plaçons, nous introduisons notre approche d'équivalence topologique.

1.3.4.2 Les mesures de proximité

Nous allons nous restreindre ici aux mesures de proximité construite sur R^p même si la généralisation à n'importe quel type de mesure de proximité, qu'elle soit binaire, floue ou symbolique est possible (Batagelj et Bren [5], Lerman [33], Warrens [55], Lesot et al. [34], Zwick et al. [62], Bouchon-Meunier et al. [9], Malerba et al. [37]

Définitions

On considère un ensemble de n individus définis dans un espace à p dimensions. Les individus sont décrits par des variables continues : soient $x = (x_1, x_2 \ldots, x_p)$ et $y = (y_1, y_2 \ldots, y_p)$. Une mesure de proximité u entre deux points x et y de R^p est définie comme suit :

$$u: R^p \times R^p \to R$$

$$(x, y) \to u(x, y)$$

Avec les propriétés suivantes : $\forall (x, y) \in R^p \times R^p$

$$u(x, y) = u(y, x) \qquad (P_1)$$

$$u(x, x) \geq u(x, y) \qquad (P_2)$$

$$u(x, x) \leq u(x, y) \qquad (P'_2)$$

$$\exists\, \alpha \in R, u(x, x) \geq \alpha \qquad (P_3)$$

Une mesure de proximité u vérifiant les propriétés (P1) et (P2) est une mesure de ressemblance. Si elle vérifie les propriétés (P1) et (P2'), c'est une mesure de dissemblance. Il est facile de montrer que toute mesure de ressemblance r peut être transformée en une mesure de dissemblance d comme suit : $r(x, y) = -d(x, y)$.

On peut également définir une mesure de proximité δ : $\delta(x, y) = u(x, y) - \alpha$ qui vérifie les propriétés suivantes : $\forall (x, y) \in R^p \times R^p$

$$\delta(x, y) \geq 0 \qquad (T_1)$$

$$\delta(x, x) = 0 \qquad (T_2)$$

$$\delta(x, x) \leq \delta(x, y) \qquad (T_3)$$

Une mesure de proximité qui vérifie les propriétés T1, T2 et T3 est une mesure de dissimilarité.

On peut également citer d'autres propriétés comme :

$$\delta(x,y) = 0 \quad \Longrightarrow \forall z \in R^p \ \delta(x,z) = \delta(y,z) \qquad (T_4)$$

$$\delta(x,y) = 0 \quad \Longrightarrow x = y \qquad (T_5)$$

$$\delta(x,y) \leq \ \delta(x,z) + \delta(z,y) \qquad (T_6)$$

$$\delta(x,y) \leq \ max\big(\delta(x,z)\,,\delta(z,y)\big) \qquad (T_7)$$

$$\delta(x,y) + \delta(z,t) \leq \ max(\delta(x,z) + \delta(y,t)\,,\delta(x,t) + \delta(y,z)\,) \qquad (T_8)$$

On trouve dans Batagelj et Bren (1992) quelques relations entre ces inégalités :

T7 (Inégalité Ultramétrique) ⇒ T6 (Inégalité Triangulaire) ⇐ T8 (Inégalité de Buneman)

Une mesure de dissimilarité qui vérifie les propriétés T5 et T6 est une distance.

1.3.4.3 Comparaison de deux mesures de proximité

Deux mesures de proximité u_i et u_j fournissent généralement deux matrices de proximité différentes sur un même jeu de données. Peut-on dire que ces deux mesures de proximité sont différentes ? De nombreux articles ont été consacrés à cette question. On peut trouver dans Lerman (1967) une proposition qui consiste à dire que deux mesures de proximité u_i et u_j sont équivalentes dès lors que le préordre induit par chacune des mesures sur tous les couples d'objets sont identiques. D'où la définition suivante :

Equivalence en préordonnance : Soient deux mesures de proximité u_i et u_j sur des objets de R^p. Si pour tout quadruplé d'objets (x, y, z, t), on a :

$u_i(\mathrm{x},\mathrm{y}) \ \leq \ u_i(\mathrm{z},\mathrm{t}) \ \Rightarrow \ u_j(\mathrm{x},\mathrm{y}) \ \leq \ u_j(\mathrm{z},\mathrm{t})$,

alors les deux mesures u_i et u_j sont considérées comme équivalentes.

Cette définition a été ensuite reprise dans de nombreux papiers Batagelj et Bren [5], Bouchon-Meunier et al. [9], Lesot et al. [34] et Schneider et Borlund [48]. Cette définition débouche sur un théorème intéressant dont on peut trouver la démonstration dans Batagelj et Bren [5].

Théorème 1 Soient deux mesures de proximité u_i et u_j, s'il existe une fonction f strictement monotone telle que pour tout couple d'objets (x, y) on a $u_i(\mathrm{x},\mathrm{y}) = \mathrm{f}\,(u_j(\mathrm{x},\mathrm{y}))$ alors u_i et u_j induisent des préordres identiques et par conséquent, elles sont équivalentes : $u_i \equiv u_j$. La réciproque étant également vraie, i.e. deux mesures de proximité dont l'une est fonction de l'autre induisent le même préordre et sont, par conséquent, équivalentes.

On peut alors proposer d'utiliser un indice de discordance entre préordres induits comme mesure de proximité entre deux mesures u_i et u_j. A cet effet, on peut, à l'instar de Rifqi et al. [45] utiliser le « tau »

de Kendall généralisé qui repose sur la mesure de concordance des rangs. Les rangs des n(n − 1) paires de valeurs de proximité entre x et y selon u_i sont comparés à ceux selon u_j. On note $R_i(x,y)$ et $R_j(x,y)$ les rangs respectifs de $u_i(x,y)$ et $u_j(x,y)$.

$$K_{u_i,u_j} = \frac{2}{n(n-1)} \sum_x \sum_{y \neq x} \delta_{ij}(x,y) \qquad \text{avec} \quad \delta_{ij} = \begin{cases} 0 & si\ R_i(x,y) = R_j(x,y) \\ 1 & sinon \end{cases}$$

Cette définition montre ainsi que l'équivalence ne repose pas sur les valeurs numériques des deux matrices mais sur les préordres induits sur les couples de points. La comparaison entre indices de proximité a été étudiée par Schneider et Borlund [48] sous un angle statistique. Les auteurs proposent une approche empirique qui vise à comparer les matrices de proximité obtenues par chaque mesure de proximité sur les couples d'objets. Ils proposent ensuite de tester si les matrices sont statistiquement différentes ou pas en utilisant le test de Mantel, [38]. Le critère utilisé par ces auteurs est le coefficient des rangs de Spearman :

$$\rho_s = 1 - \frac{6 \sum_x \sum_{y \neq x} \left(R_i(x,y) - R_j(x,y) \right)^2}{n(n^2-1)}$$

Les mêmes auteurs proposent de traiter la comparaison des préordres induits par les mesures de proximité dans le cadre de l'analyse de Procuste. Ces techniques visant à comparer directement des matrices de proximité ont été développées pour des domaines appliqués comme l'écologie, les sciences sociales, la géographie, la psychologie et l'anthropologie. Nous ne discutons pas ici du choix de la mesure de comparaison des matrices de proximité. Nous nous contentons d'utiliser l'expression présentée plus haut. Nous reprécisons que notre objectif n'est pas de comparer des matrices de proximité ni les préordres induits mais de proposer une autre notion qui est l'équivalence topologique que nous comparons à l'équivalence préordinale en essayant d'identifier les liens entre les deux approches.

1.3.4.4 Equivalence topologique entre mesures de proximité

L'équivalence topologique est basée sur la notion de graphe topologique que l'on désigne également sous le nom de graphe de voisinage. L'idée de base consiste à dire que deux mesures de proximité sont équivalentes si les graphes topologiques induits sur l'ensemble des objets restent identiques. Mesurer la ressemblance entre mesures de proximité revient à comparer les graphes de voisinage et à mesurer leur ressemblance. Nous allons tout d'abord définir de manière plus précise ce qu'est un graphe topologique et comment le construire. Nous proposons ensuite une mesure de proximité entre graphes topologiques qui servira à comparer les mesures de proximité dans la section suivante.

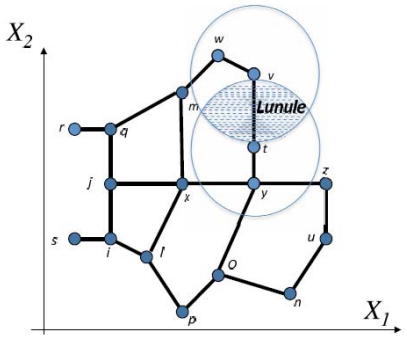
1.3.4.4 Proximité entre graphes topologiques

On considère un ensemble E de n individus définis dans un espace à p dimensions. Il existe de nombreuses possibilités pour construire une relation binaire de voisinage. Dans le cas de de l'arbre de longueur minimale sur $(E \times E)$, deux objets x et y de E vérifient la propriété de voisinage selon l'Arbre de

Longueur Minimale (ALM) s'ils sont reliés par une arête directe. Dans ce cas, $V_u(x,y) = 1$ sinon, $V_u(x,y) = 0$. Où V_u est la matrice d'adjacence associée au graphe ALM, formée de 0 et de 1. Le Graphe de Voisins Relatifs (GVR) est une façon parmi d'autres de construire une relation binaire de voisinage, Toussaint [51]; Preparata et Shamos [39], où les couples de points voisins vérifient la propriété suivante :

$$u_E(x,y) \leq max(u_E(x,z), u_E(y,z)); \qquad \forall z \in E - \{x,y\}$$

La figure 1 est un exemple dans R^2 où la mesure de proximité est la distance euclidienne.



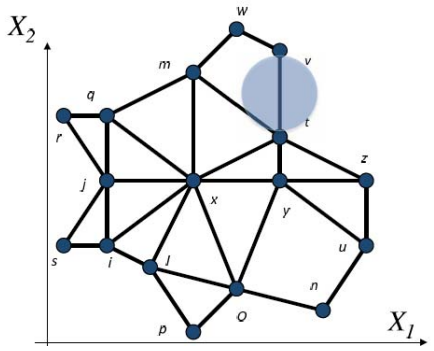
	...	x	y	z	t	u	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
x	...	1	1	0	0	0	...
y	...	1	1	1	1	0	...
z	...	0	1	1	0	1	...
t	...	0	1	0	1	0	...
u	...	0	0	1	0	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...

Figure 1 – Graphe des voisins relatifs et sa matrice d'adjacence

Dans le cas du graphe de Gabriele (GG), les couples de points doivent vérifier :

$$u_E(x,y) \leq \min(\sqrt{u^2{}_E(x,z) + u^2{}_E(y,z)}); \qquad \forall z \in E - \{x,y\}$$

La figure 2 est une illustration du graphe de voisinage dans R^2 et sa matrice d'adjacence.



	...	x	y	z	t	u	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
x	...	1	1	0	1	0	...
y	...	1	1	1	1	0	...
z	...	0	1	1	1	1	...
t	...	1	1	1	1	0	...
u	...	0	0	1	0	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...

Figure 2 – Graphe de Gabriele et sa matrice d'adjacence

Pour illustrer la notion de proximité entre graphes topologiques, considérons deux mesures de proximité u_i et u_j. Soient $D_{u_i}(ExE)$ et $D_{u_j}(ExE)$ les tableaux de distances associés.

Notons V_{u_i} et V_{u_j} les deux matrices d'adjacence associées aux deux structures topologiques engendrées par ces deux distances. Une manière de mesurer le degré de ressemblance entre les deux graphes est de compter le nombre de discordances entre les deux matrices d'adjacence.

$$D\left(V_{u_i}, V_{u_j}\right) = \frac{2\sum_{k=1}^{n}\sum_{l=k+1}^{n}\delta_{kl}}{n(n-1)} \quad \text{avec} \quad \delta_{kl} = \begin{cases} 0 & si\ V_{u_i}(k,l) = V_{u_j}(k,l) \\ 1 & sinon \end{cases}$$

D est une mesure de dissimilarité comprise entre 0 et 1. La valeur 0 signifie que les deux matrices d'adjacence sont identiques et la structure de topologique induite par les deux mesures de proximité est la même. Dans ce cas, on parle d'équivalence topologique entre les deux mesures de proximité. La valeur 1 signifie que la topologie a totalement changé, autrement dit, aucun couple de voisins dans la structure topologique induite par la première mesure de proximité, n'est resté voisin dans la structure topologique induite par la second »e mesure et vice versa. D s'interprète également comme le pourcentage de désaccord entre des tableaux d'adjacence. Grâce à cette mesure de proximité, nous allons enfin pouvoir comparer les mesures de proximité et les classifier selon leur degré de ressemblance. Nous verrons que les résultats obtenus sur ces classifications sont différents. En effet, une équivalence topologique n'implique pas une équivalence en préordonance. En revanche, une équivalence en préordonance entraîne une équivalence topologique.

1.3.4.5 Classification des mesures de proximité :

Nous nous limitons ici à la classification des mesures de proximité dans $\mathbb{R}^p$. Ce travail peut être étendu à toutes les autres mesures dès lors qu'on est capable de construire une structure topologique sur les objets. Nous avons effectué cette classification sur le jeu de données des Iris de Fisher. Pour construire la structure topologique, nous utilisons la propriété du graphe des voisins relatifs Toussaint [51].

Measure	Formula
u_1 : Euclidean	$u_E(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$
u_2 : Mahalanobis	$u_{Mah}(x,y) = \sqrt{(x-y)^t \sum^{-1}(x-y)}$
u_3 : Manhattan (City-block)	$u_{Man}(x,y) = \sum_{i=1}^{p}\lvert x_i - y_i\rvert$
u_4 : Minkowski	$u_{Min_\gamma}(x,y) = (\sum_{i=1}^{p}\lvert x_i - y_i\rvert^{\gamma})^{\frac{1}{\gamma}}$
u_5 : Tchebytchev	$u_{Tch}(x,y) = \max_{1\le i\le p}\lvert x_i - y_i\rvert$
u_6 : Cosine Dissimilarity	$u_{Cos}(x,y) = 1 - \frac{<x,y>}{\|x\|\|y\|}$
u_7 : Canberra	$u_{Can}(x,y) = \sum_{i=1}^{p}\frac{\lvert x_i - y_i\rvert}{\lvert x_i\rvert + \lvert y_i\rvert}$
u_8 : Squared Chord	$u_{SC}(x,y) = \sum_{i=1}^{p}(\sqrt{x_i} - \sqrt{y_i})^2$
u_9 : Weighted Euclidean	$u_{E_w}(x,y) = \sqrt{\sum_{i=1}^{p}\alpha_i(x_i - y_i)^2}$
u_{10} : Chi-square	$u_{\chi^2}(x,y) = \sum_{i=1}^{p}\frac{(x_i - m_i)^2}{m_i}$
u_{11} : Jeffrey Divergence	$u_{JD}(x,y) = \sum_{i=1}^{p}(x_i \log\frac{x_i}{m_i} + y_i \log\frac{y_i}{m_i})$
u_{12} : Histogram Intersection Measure	$u_{HIM}(x,y) = 1 - \frac{\sum_{i=1}^{p}(\min(x_i, y_i))}{\sum_{j=1}^{p} y_j}$
u_{13} : Pearson's Correlation Coefficient	$u_\rho(x,y) = 1 - \lvert\rho(x,y)\rvert$

Tableau 1 : Mesures de proximité utilisées

Le tableau 2 représente les dissimilarités entre les 13 mesures de proximité retenues dans ce test. Les éléments situés au-dessus de la diagonale principale représentent les dissimilarités en préordonnance. Les autres représentent les dissimilarités en topologie. L'application d'un algorithme de construction d'une hiérarchie de partitions basée sur le critère de Ward (Ward Jr [54]), selon le critère d'équivalence topologique fournit le dendrogramme de la figure 3. Si nous comparons les mêmes mesures selon le critère de préordonnance, nous obtenons la matrice de dissimilarité donnée en annexe tableau 2 et le dendrogramme de la figure 4.

$S=1-D$	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}	u_{12}	u_{13}
$u_1 : u_E$	**1**	.776	.973	.988	.967	.869	.890	.942	**1**	.947	.945	.926	.863
$u_2 : u_{Mah}$	.876	**1**	.773	.774	.752	.701	.707	.737	.776	.739	.738	.742	.703
$u_3 : u_{Man}$	.964	.840	**1**	.964	.940	.855	.882	.930	.973	.933	.932	.924	.848
$u_4 : u_{Min_\gamma}$	.964	.876	.947	**1**	.967	.871	.892	.946	.988	.950	.949	.925	.866
$u_5 : u_{Tch}$	.947	.858	.929	.964	**1**	.865	.887	.940	.957	.942	.942	.914	.860
$u_6 : u_{Cos}$	.858	.858	.840	.840	.858	**1**	.893	.898	.869	.899	.899	.830	.957
$u_7 : u_{Can}$	.911	.840	.929	.893	.911	.822	**1**	.943	.890	.940	.942	.874	.868
$u_8 : u_{SC}$	.947	.840	.947	.929	.947	.858	.947	**1**	.942	.957	**1**	.913	.884
$u_9 : u_{E_w}$	**1**	.876	.964	.964	.947	.858	.911	.947	**1**	.947	.945	.926	.863
$u_{10} : u_{\chi 2}$	.947	.840	.947	.929	.947	.858	.947	**1**	.947	**1**	**1**	.912	.885
$u_{11} : u_{JD}$	.947	.840	.947	.929	.947	.858	.947	**1**	.947	**1**	**1**	.914	.884
$u_{12} : u_{HIM}$	.884	.813	.884	.867	.902	.884	.884	.920	.884	.920	.920	**1**	.825
$u_{13} : u_p$	.867	.849	.831	.867	.867	.973	.796	.849	.867	.849	.849	.876	**1**

Tableau 2 : Matrice contenant deux types de similarités $S\left(V_{u_i}, V_{u_j}\right) = 1 - D\left(V_{u_i}, V_{u_j}\right)$ *: en préordonnance au dessus de la diagonale et en topologie en dessous.*

Figure 3 : Classification obtenue sur les mesures de proximité en topologie

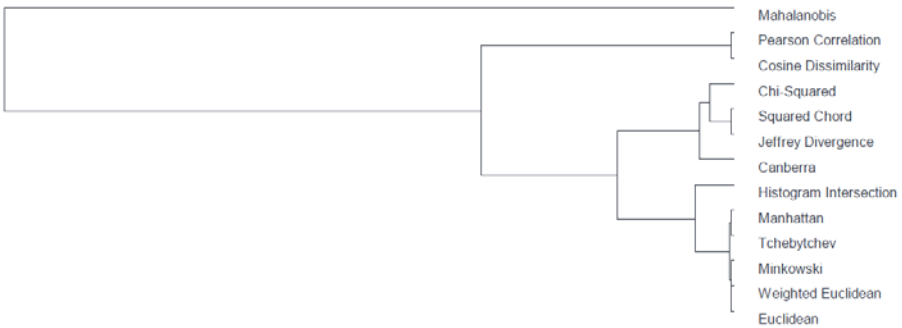


Figure 4 : Classification obtenue sur les mesures de proximité en pré-ordonnance

L'examen de ces deux hiérarchies montre que le résultat de classification est différent selon que l'on compare les mesures de proximité au moyen de l'équivalence de préordonnance ou de l'équivalence topologique.

Nous allons maintenant montrer quelques résultats plus généraux. Du théorème 1 d'équivalence en préordonnance, on en déduit la propriété suivante :

Propriété

Soient f une fonction strictement monotone de R+ dans R+, u_i et u_j deux mesures de proximité telles que $u_i(x,y) \rightarrow f(u_i(x,y)) = u_j(x,y)$ alors,

$$u_i(x,y) \leq \max(u_i(x,z), u_i(y,z)) \iff u_j(x,y) \leq \max(u_j(x,z), u_j(y,z))$$

Démonstration :

Supposons que : $\max(u_i(x,z), u_i(y,z)) = u_i(x,y)$

D'après le théorème 1 d'équivalence en préordonnance,

$$u_i(x,y) \leq u_i(x,z) \implies f(u_i(x,y)) \leq f(u_i(x,z))$$

de plus, $u_i(y,z) \leq u_i(x,z) \implies f(u_i(y,z)) \leq f(u_i(x,z))$

$$\implies f(u_i(x,z)) \leq \max\left(f(u_i(x,z)), f(u_i(y,z))\right)$$

D'où le résultat, $u_j(x,y) \leq \max(u_j(x,z), u_j(y,z))$

L'implication réciproque est vraie, vu que f est continue et strictement monotone alors, son application réciproque f^{-1} est continue et de même sens de variation que f.

On peut ainsi dire, dans le cas où f est strictement monotone, que si le préordre est conservé alors la topologie est conservée et inversement. Cette propriété nous amène à énoncer le théorème suivant :

Théorème 2 (Equivalence en topologie)

Soient deux mesures de proximité u_i et u_j , s'il existe une fonction f strictement monotone telle que pour tout couple d'objets (x,y) on a $u_i(x,y) = f(u_j(x,y))$ alors u_i et u_j induisent des graphes topologiques identiques et par conséquent, elles sont équivalentes : $u_i \equiv u_j$. La réciproque étant également vraie, i.e. deux mesures de proximité dont l'une est fonction de l'autre induisent la même topologie et sont, par conséquent, équivalentes.

La proposition ci-dessous montre que l'équivalence en préordonnance de deux mesures de proximité u_i et $u_j = f(u_i)$ implique nécessairement l'équivalence en topologie, quel que soit la fonction f.

Proposition

Dans le cadre des structures topologiques induites par le graphe des voisins relatifs, si deux mesures de proximité u_i et u_j sont équivalentes en préordonnance, alors elles sont en équivalence topologique.

Démonstration

Si $u_i \equiv u_j$ (équivalence en préordonnance) alors,

$$u_i(\mathrm{x},\mathrm{y}) \leq u_i(\mathrm{z},\mathrm{t}) \implies u_j(\mathrm{x},\mathrm{y}) \leq u_j(\mathrm{z},\mathrm{t}) \qquad \forall x, y, z, t \in R^p$$

On a, en particulier pour t=x=y et $z \neq t$,

$$\begin{cases} u_i(\mathrm{x},\mathrm{y}) \leq u_i(\mathrm{x},\mathrm{z}) \implies u_j(\mathrm{x},\mathrm{y}) \leq u_j(\mathrm{x},\mathrm{z}) \\ u_i(\mathrm{x},\mathrm{y}) \leq u_i(\mathrm{y},\mathrm{z}) \implies u_j(\mathrm{x},\mathrm{y}) \leq u_j(\mathrm{y},\mathrm{z}) \end{cases}$$

On en déduit,

$$u_i(\mathrm{x},\mathrm{y}) \leq \max(u_i(\mathrm{x},\mathrm{z}), u_i(\mathrm{y},\mathrm{z})) \iff u_j(\mathrm{x},\mathrm{y}) \leq \max\big(u_j(\mathrm{x},\mathrm{z}), u_j(\mathrm{y},\mathrm{z})\big)$$

En utilisant la propriété P1 de symétrie,

$$u_i(\mathrm{x},\mathrm{y}) \leq \max(u_i(\mathrm{x},\mathrm{z}), u_i(\mathrm{y},\mathrm{z})) \iff u_j(\mathrm{x},\mathrm{y}) \leq \max\big(u_j(\mathrm{x},\mathrm{z}), u_j(\mathrm{y},\mathrm{z})\big)$$

d'où, $u_i \equiv u_j$ (équivalence topologique).

1.3.4.6 Conclusion

Le choix d'une mesure de proximité est très subjectif, il est souvent fondé sur des habitudes ou sur des critères tels que l'interprétation a posteriori des résultats. Ce travail propose une nouvelle approche d'équivalence entre mesures de proximité. Cette approche que nous appelons topologique est basée sur la notion de graphe de voisinage induit par la mesure de de proximité. D'un point de vue pratique, les mesures que nous avons comparées sont toutes construites sur des données quantitatives. Mais ce travail peut parfaitement s'étendre aux autres en choisissant la structure topologique adaptée.

Nous envisageons d'étendre ce travail à d'autres structures topologiques et d'utiliser un critère de comparaison, autre que les techniques de classification, afin de valider le degré d'équivalence entre deux mesures de proximité. Par exemple, un critère basé sur un test non paramétrique (la corrélation de rang de Spearman, le « tau » de concordance de rang de Kendall, ou encore le test de Mantel, etc.). L'application d'un test par permutations, sur les matrices d'adjacence associées à ces mesures, vont permettre de donner une signification statistique entre les deux matrices de ressemblance et de valider ou pas l'équivalence topologique c'est-à-dire, si vraiment elles induisent ou pas la même structure de voisinage sur les objets.

1.4 Conclusion du chapitre

Nous avons proposé et implémenté des approches de classification automatique. La première approche basée sur l'utilisation de l'agrégation des opinions pour former une partition qui est le résultat de la classification. Ce résultat est obtenu par la construction, sur l'ensemble des individus, d'un classement collectif qui est une agrégation de l'ensemble de classement de toutes les variables. Cette approche permet de classifier des données qualitatives, quantitatives ou mixtes et de traiter le problème des données manquantes dans une classification. La seconde, un peu moins originale, est une amélioration de la méthode CAH par optimisation de l'ultramétrique. Même si cette dernière méthode améliore le résultat, elle présente l'inconvénient de la complexité des calculs. Nous avons étudié ensuite l'équivalence des mesures de proximité et fourni une façon d'apprécier cette équivalence par la classification automatique. Dans cette étude nous avons utilisé une approche topologique et nous nous sommes intéressés uniquement aux données quantitatives, mais elle peut être généralisée à d'autres types de données ou de structures topologiques.

Nous avons proposé d'autres approches de classification non citées dans ce chapitre. On peut citer notamment l'approche prétopologique montrant de quelle manière la prétopologie pouvait être utilisée pour élaborer une méthodologie de classification faisant appel aux concepts de pré-voisinages et applicables dans un contexte plus général où les éléments à classifier sont décrits par des caractéristiques qualitatives. Ceci conduit à ne plus recourir à des métriques pour apprécier la proximité entre les objets. Une des méthodes proposées consiste à structurer l'espace des données en utilisant la fonction d'adhérence qui mène à définir les fermés et les fermés minimaux. Ceci nous permet d'extraire des germes qui serviront par la suite dans les méthodes de classification par réallocation. L'intérêt principal de ce travail est de déterminer une structure forte et de fournir un nombre de classes de manière « naturelle ». Nous avons ensuite utilisé les fermés minimaux pour obtenir la classification prétopologique encadrant la structure prétopologique. Nous avons proposé une autre approche consistant à analyser le problème de classification à la manière des modèles d'agrégation des préférences. Il s'agit d'interpréter le problème de fond de la classification par les concepts de l'agrégation des préférences, de produire des résultats permettant d'apprécier la qualité d'une classification et d'introduire dans cette procédure une part d'intervention humaine permettant d'en contrôler le déroulement. On peut consulter [8] pour prendre connaissance de ces approches.

1.5 Références

[1] Agarwala R., Bana V., Farach M., Narayanan B., Paterson M., and Thorup M. On the approximability of numerical taxonomy. Technical Report 95-46, DIMACS, Rutgers University, Piscataway, NJ 08855, USA, 1995.

[2] Ankerst M., Breunig M. M., Kriegel H.P., and Sander J. Optics : Ordering points to identify the clustering structure. In SIGMOD Conference, pages 49-60, 1999.

[3] Armengol E, et Plaza E. Using symbolic descriptions to explain similarity on cbr. In Beatriz López, Joaquim Meléndez, Petia Radeva, et Jordi Vitrià, éditeurs, Artificial Intelligence Research and Development, volume 131. IOS Press, 2005.

[4] Batagelj, V. et M. Bren (1992). Comparing resemblance measures. Technical report, Proc. International Meeting on Distance Analysis (DISTANCIA'92).

[5] Batagelj, V. et M. Bren (1995). Comparing resemblance measures. Journal of classification 12, 73–90.

[6] Bisson G. La similarité : une notion symbolique/numérique. In E.Diday, P.Brito, Y.Kodratoff, et M.Moulet, éditeurs, Apprentissage symbolique numérique. Editions CEPADUES, 2000.

[7] Blatt M., Wiseman S., and Domany E. Data clustering using a model granular magnet. Neural Computation, 9(8) : 1805-1842, 1997.

[8] Boubou M. Contribution aux méthodes de classification non supervisée via des approches prétopologiques et d'agrégation d'opinions. Thèse de Doctorat, Université Lyon 1, novembre 2007.

[9] Bouchon-Meunier, B., M. Rifqi, et S. Bothorel (1996). Towards general measures of comparison of objects. Fuzzy sets and systems 84(2), 143–153.

[10] Chu S.C, Roddick J.F., et Pan J.S. Efficient K-medoids algorithms using multi-centroids with multi-runs sampling scheme.

[11] Clarke, K., P. Somerfield, et M. Chapman (2006). On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted bray-curtis coefficient for denuded assemblages. Journal of Experimental Marine Biology & Ecology 330(1), 55–80.

[12] CHANDON J.L., LEMAIRE, J., POUGET J., "Construction de l'ultramétrique la plus proche d'une dissimilarité au sens des moindres carrés", R.A.I.R.O. Recherche opérationnelle, vol. 14, n° 2, mai 1980, p. 157-170.

[13] Diday E., Govaert G., Lechevallier Y., and Sidi J. Clustering in pattern recognition. In NATO Advanced study Institute on Digital Image Processing and Analysis, Bonas, 1980.

[14] Ester M., Kriegel H.P., Sander J., and Xu X.. A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD, pages 226-231, 1996.

[15] Ester M, Kriegel HY.P., Sander J, et Xu X. A density based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, et Usama Fayyad, éditeurs, Second International Conference on Knowledge Discovery and Data Mining, pages 226–231, Portland, Oregon, 1996. AAAI Press.

[16] Fagin, R., R. Kumar, et D. Sivakumar (2003). Comparing top k lists. In Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms, pp. 36. Society for Industrial and Applied Mathematics.

[17] Fahim A.M., Salem A.M., Torkey F.A., et Ramadan M.A.. Density clustering algorithm based on radius of data (dcbrd). Georgian Electronic Scientific Journal : Computer Science and Telecommunications, 11(4), 2006.

[18] Fisher D. H.. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 1987.

[19] Gersho A. and Gray R.M. Vector Quantization and Signal Compression. Kluwer Academic Publishers, 1992.

[20] Grotschel M. and Wakbayashi Y. A cutting plane algorithm for a clustering problem. Mathematical Programming Series, B 45 :59-96, 1989.

[21] Guha S., Rastogi R., and Shim K. Cure : an efficient clustering algorithm for large databases. In SIGMOD'98 : Proceedings of the 1998 ACM SIGMOD international conference on Management of data, pages 73-84, New York, NY, USA, 1998. ACM Press.

[22] Guha S., Rastogi R., and Shim K. ROCK : A robust clustering algorithm for categorical attributes. Information Systems, 25(5) : 345-366, 2000.

[23] Güting R.H. An introduction to spatial database systems. The VLDB Journal - The International Journal on Very Large Data Bases, 3(4) :357 – 399, October 1994.

[24] Hartigan J. A. "Clustering algorithms", Wiley 1975.

[25] Heit E., Feature of similarity and category-based induction. In Proc. Of the Interdisciplinary Wprkshop on Categorization and Similarity, University of Edinburgh, 1997.

[26] Hinneburg A et Keim D.A. An efficient approach to clustering in large multimedia databases with noise. In Proc. 4rd Int. Conf. on Knowledge Discovery and Data Mining, pages 58–65, 1998.

[27] JUAN J. "Le programme HIVOR de classification hiérarchique selon les voisins réciproques", Cahiers de l'Analyse des Données, vol ; 7 n° 2, p. 173-184.

[28] Kaufman L. et Rousseeuw P.J. Finding groups in data : An introduction to cluster analysis. WILEY-Interscience, 1990.

[29] Kohonen T. Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43 : 59-69, 1982.

[30] Kohonen T. Self-organized formation of topologically correct feature maps. Pages 509-521, 1988.

[31] Kohonen T. Self-organizing Maps. Springer Series in Information Sciences, 3 edition, Dec 2000.

[32] LE Than Van. Classification prétopologique des données : Application à l'analyse des trajectoires patients. Thèse de Doctorat, Université Lyon 1, Décembre 2007.

[33] Lerman, I. (1967). Indice de similarité et préordonnance associée, Ordres. Travaux du séminaire sur les ordres totaux finis, Aix-en-Provence.

[34] Lesot, M.-J., M. Rifqi, et H. Benhadda (2009). Similarity measures for binary and numerical data: a survey. IJKESDP 1(1), 63–84.

[35] Lin, D. (1998). An information-theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine Learning, Volume 296304. Citeseer.

[36] Liu, H., D. Song, S. Ruger, R. Hu, et V. Uren. Comparing dissimilarity measures for contentbased image retrieval. Information Retrieval Technology, 44–50.

[37] Malerba, D., F. Esposito, V. Gioviale, et V. Tamma (2001). Comparing dissimilarity measures for symbolic data analysis. Proceedings of Exchange of Technology and Know-how and New Techniques and Technologies for Statistics 1, 473–481.

[38] Mantel, N. (1967). A technique of disease clustering and a generalized regression approach. Cancer Research 27, 209–220.

[39] N. Nicoloyannis. Structures prétopologiques et classification automatique : le logiciel DEMON. PhD these de Doctorat, Université. Claud Bernard Lyon I, 1988.

[40] Park, J., H. Shin, et B. Choi (2006). Elliptic gabriel graph for finding neighbors in a point set and its application to normal vector estimation. Computer-Aided Design 38(6), 619–626.

[41] Picard, P. Classification sur des données hétérogènes. Mémoire de D.E.A., Université de la réunion, 2001.

[42] Preparata, F. et M. Shamos (1985). Computational geometry: an introduction. Springer.

[43] Raymond T.Ng et Han J., Efficient and effective clustering methods for spatial data mining. In Proceedings of 20th International Conference on Very Large Databases, pages 144–155, Santiago, Chile, 1994.

[44] Richter M.M. Classification and learning of similarity measures. In Proc. of the Sixteenth Annual Conference of the German Society for Classification, Springer Verlag, 1992.

[45] Rifqi, M., M. Detyniecki, et B. Bouchon-Meunier (2003). Discrimination power of measures of resemblance. IFSA'03.

[46] Sander J, Ester M, Kriegel H.P., et Xu. X. Density-based clustering in spatial databases : The algorithm gdbscan and its applications. Data Mining and Knowledge Discovery, 2(2) :169–194, juin 1998.

[47] Sheikholeslami G, Chatterjee S. and Zhang A. Wavecluster : a wavelet-based clustering approach for spatial data in very large databases. The VLDB Journal, 8(3-4) :289-304, 2000.

[48] Schneider, J. et P. Borlund (2007a). Matrix comparison, part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. Journal of the American Society for Information Science and Technology 58(11), 1586– 1595.

[49] Spertus, E., M. Sahami, et O. Buyukkokten (2005). Evaluating similarity measures: a largescale study in the orkut social network. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 684. ACM.

[50] Strehl, A., J. Ghosh, et R. Mooney (2000). Impact of similarity measures on web-page clustering. In Workshop on Artificial Intelligence for Web Search (AAAI 2000), pp. 58–64.

[51] Toussaint, G. (1980). The relative neighbourhood graph of a finite planar set. Pattern recognition 12(4), 261–268.

[52] Tran N.M.T. Analyse d'une base de graphes issus d'un simulateur en intelligence en essaim. Mémoire de D.E.A., Université de Nantes, 2006.

[53] Van der Laan M.J., Pollard C.S.et Bryan J., A new partitioning around medoids algorithm. Journal of Statistical Computation and Simulation, 73(8) : 575–584, 2002.

[54] Ward Jr, J. (1963). Hierarchical grouping to optimize an objective function. Journal of the American statistical association 58(301), 236–244.

[55] Warrens, M. (2008). Bounds of resemblance measures for binary (presence/absence) variables. Journal of Classification 25(2), 195–208.

[56] Wang W., Yang J., and Muntz R.R. STING : A statistical information grid approach to spatial data mining. In Matthias Jarke, Michael J. Carey, Klaus R. Dittrich, Frederick H. Lochovsky, Pericles Loucopoulos, and Manfred A. Jeusfeld, editors, Twenty-Third International Conference on Very Large Data Bases, pages 186-195, Athens, Greece, 1997. Morgan Kaufmann.

[57] Williams W.T., Lambert J.M. Multivariate methods in plant ecology. Journal of Ecology, 47(1) :83-101, 1959.

[58] Xu X., Ester M., Kriegel H. P., and Sander J. A distribution-based clustering algorithm for mining in large spatial databases. In ICDE'98 : Proceedings of the Fourteenth International Conference on Data Engineering, pages 324-331,Washington, DC, USA, 1998. IEEE Computer Society.

[59] Zhang T., Ramakrishnan R. and Livny. Birch M. an efficient data clustering method for very large databases. In SIGMOD'96 : Proceedings of the 1996 ACM SIGMOD international conference on Management of data, pages 103-114, New York, NY, USA, 1996. ACM Press.

[60] Zhang T., Ramakrishnan R. and Livny. Birch M. A new data clustering algorithm and its applications, 1997.

[61] Zhang, B. et S. Srihari (2003). Properties of binary vector dissimilarity measures. In Proc. JCIS Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing. Citeseer.

[62] Zwick, R., E. Carlstein, et D. Budescu (1987). Measures of similarity among fuzzy concepts: A comparative analysis. INT. J. APPROX. REASON. 1(2), 221–242.

Chapitre 2

Analyse et fouille des données spatiales

Sommaire

2.1 Introduction

Le volume des données ne cesse de croître avec le développement des nouvelles technologies de communication, mais aussi avec les moyens de stockage et de la collecte numérique des données. Ces données sont de plus en plus utilisées dans des applications décisionnelles. Pour une analyse avancée de ces données, on a recours à des méthodes de fouille des données, permettant d'extraire des connaissances noyées dans ces grands volumes de données. Parmi ces données, les données spatiales sont de plus en plus fréquentes et volumineuses. La fouille de données spatiale (FDS) est aujourd'hui identifiée comme un domaine de la fouille de données à part entière. Elle résulte de la combinaison de la fouille de données et des bases de données spatiales.

Dans le domaine spatial, se contenter d'analyser les observations en chaque zone de l'espace ne suffit pas pour en tirer toute l'information souhaitable. En effet, il est très important de s'intéresser aussi aux

influences que les observations exercent les unes sur les autres. Nous allons aborder cette question dans la section relative à l'interaction spatiale. Les méthodes classiques d'analyse et de fouille de données sont insuffisantes pour traiter les données spatiales. Ceci a motivé les recherches d'analyse spatiale, où la prise en compte des interactions spatiales ente observations est centrale.

Les données spatiales et leurs types

Les données spatiales appelées aussi données géo-référencées sont des données pour lesquelles une information géographique est associée à chaque unité statistique. Cette information géographique peut être la position de l'unité sur une carte ou dans un référentiel spatio-temporel, et peut prendre aussi la forme de latitude et longitude ou de coordonnées UTM (Universal Transvers Mercator). La spatialisation d'une donnée revêt plusieurs formes. En effet, les informations recueillies peuvent porter sur des points particuliers répartis dans l'espace ou sur des agrégats, des moyennes ou des taux relatifs à des zones. On peut les répartir en quatre types :

- Les données ponctuelles ou de type géostatistique (hôpitaux, entreprises, villes,…)
- Les données surfaciques ou de type économétrie spatiale (communes, régions,…)
- Les données de type semis de points : la disposition de certaines espèces végétales dans une forêt, les adresses de patients affectés d'une certaine maladie dans une région.
- Les données image : pixels ou motifs

Précautions à prendre pour manipuler les données spatiales

Une analyse des données spatiales qui ignorerait l'aspect de leurs positions spatiales ou qui l'intégrerait de façon inadéquate résulterait en une perte d'information, des erreurs de spécifications et de prédiction. Comme on peut le constater avec le type des données, l'analyse d'une information spatialisée pose d'abord un problème d'hétérogénéité. En effet, toute analyse d'une population statistique suppose que les éléments de cette population ont des points communs, sur lesquels on peut fonder des comparaisons et asseoir des régularités. Or, qu'il s'agisse d'entités ponctuelles ou de zones, les unités spatiales sont généralement fortement hétérogènes. Cette hétérogénéité peut être de taille (villes de tailles différentes en nombre d'habitants), de forme (formes différentes des zones géographiques) ou de structure (comparaison de revenu moyen dans une grande ville et dans une zone rurale). Ceci nous amène à prendre en compte des facteurs de structure. Ceci justifie l'utilisation, en analyse spatiale, de l'analyse structurelle-géographique appelée « shift-share» [41].

Domaines d'application

La fouille des données spatiales concerne divers domaines scientifiques et industriels, notamment en géologie, séismologie, météorologie, économie, géographie, épidémiologie, industrie pétrolière, géomarketing, sciences sociales et en santé. Les applications sont diverses et variées : à titre d'exemples, en prospection pétrolière : prédire la quantité de pétrole potentielle en un lieu donné en fonction de prélèvements effectués en certains points repartis sur une zone pour optimiser l'emplacement des forages. En environnement : produire les cartes de prédictions de niveaux de pollution atmosphérique. En épidémiologie : produire les cartes prédiction de la propagation de la grippe. En géomarketing, prédire les flux de clients d'une zone géographique donnée vers un magasin donné.

2.2 Bref aperçu des méthodes d'analyse spatiale

L'analyse et fouille des données spatiales est définie comme l'extraction de connaissances plongées dans les bases des données, présentant des interactions spatiales. Elle permet également de révéler des relations spatiales et d'autres propriétés implicitement présentes dans ces données. Les méthodes utilisées sont issues des domaines des statistiques, de la fouille des données et des bases de données. Le point commun de ces méthodes est l'exploitation des relations spatiales de voisinage entre les observations. Ces méthodes permettent principalement de faire de l'estimation, de la prédiction et de la structuration des données. L'analyse des données spatiales existe bien avant les années 1950 [47], et s'est développée avec l'avènement des SIG (Systèmes d'Information Géographique) et plus récemment des méthodes de data mining. Les travaux relatifs aux données spatiales sont nombreux et peuvent être répartis en méthodes d'analyse de localisations ou d'individus munis de localisations.

Analyse des localisations avec ou sans attributs

Les localisations sont étudiées de diverses manières : sans attributs, munies de mesures numériques ou munies de mesures catégorielles [66]. Les méthodes basées uniquement sur les localisations explorent généralement un ensemble de localisations pour révéler des tendances ou des concentrations [27], [49], [50]. L'approche de classification spatiale proposée dans [24] peut être classée dans cette catégorie. Parfois, la visualisation cartographique des objets ne reflète pas l'intensité locale ou la tendance générale de leur localisation. C'est particulièrement le cas lorsque le nombre d'objets analysés est élevé et lorsque de nombreuses localisations se superposent. Ces méthodes visent à quantifier les localisations ponctuelles par des mesures de densités dans leur voisinage et d'attribuer une mesure d'intensité du phénomène étudié dans chaque localisation. Les densités sont calculées par un balayage de l'espace par des fenêtres mobiles circulaires [29], de rayon plus ou moins important selon si on s'intéresse à une tendance locale ou globale. Les méthodes de classification automatique ont également été adaptées au domaine spatial. Elles s'appuient sur une mesure de dissimilarité d'objets localisés. L'intérêt de ces méthodes est réduit ici à la détection des concentrations plus ou moins fortes du phénomène étudié.

Généralement les localisations sont munies d'une ou plusieurs mesures, qualitatives ou quantitatives. L'analyse consiste ici à étudier la variation des attributs dans l'espace. Parmi ces méthodes, l'autocorrélation spatiale (locale ou globale) [21] [6], [52], la régression spatiale [25] [41], la co-localisation [58] [40] et la caractérisation sont largement étudiées.

Analyse d'individus munis de localisations

Dans ce type de méthodes, les observations analysées sont munies de variables statistiques. La localisation permet de construire une contiguïté entre les individus. Les variables sont ainsi munies de pondération prenant en compte le voisinage entre observations. Parmi les méthodes les plus courantes, l'auto-régression spatiale [60], intégrant le processus d'autocorrélation dans le modèle de régression, et la classification spatiale (non supervisée) basée sur les localisations des observations [54] [5] [33].

D'autres méthodes comme les règles d'associations ont également été appliquées à des données spatiales Koperski et al. [42], [46]. Ester et al. [25] ont proposé une méthode de classification spatiale (supervisée) utilisant les arbres de décision. Cette méthode combine l'algorithme ID3 et le concept de graphe de voisinage.

2.3 Concepts de l'interaction spatiale

Dans le domaine spatial, se contenter d'analyser les observations en chaque zone (ou point) de l'espace ne suffit pas pour en tirer toute l'information souhaitable. En effet, il est aussi important de s'intéresser également aux influences que les observations exercent les unes sur les autres. Les observations réparties dans l'espace sont souvent interdépendantes [41]. L'observation obtenue dans une localisation donnée dépend des observations obtenues dans d'autres localisations. En effet, ce ne sont pas seulement les dimensions et les structures des observations qui comptent dans l'analyse, mais aussi leurs positions relatives : plus deux observations sont éloignées, plus leurs interactions sont susceptibles d'être faibles (Encore faut-il pouvoir décider de ce qui est proche et de ce qui est éloigné de même que de pouvoir quantifier la force d'une interaction). D'où le besoin d'un outil permettant de représenter cette interaction entre observations et sa décroissance en fonction de leur éloignement. Cette interaction entre observations peut être représentée sous forme de matrice d'interaction, de matrice de contiguïté ou plus généralement de graphe topologique.

Considérons le cas où on dispose de R observations. La mesure de l'interaction entre les deux observations i et j, notée α_{ij} permet de construire une matrice d'interaction A de taille RxR. Dans cette matrice, chaque élément diagonal α_{ii} $(i = 1, \dots., R)$ est nul.

$$A = [\alpha_{ij}], \ \ i = 1, \dots R \ \ et \ j = 1, \dots, R$$

Si on dispose d'une variable Y ayant des valeurs y_i pour l'observation i, on peut représenter ces mesures sous forme d'un vecteur colonne

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \\ y_R \end{bmatrix}$$

Le produit matriciel AY fournit un vecteur où chaque élément $(AY)_i = \sum_{j=1}^{R} \alpha_{ij} y_j$ mesure l'intensité de l'effet global de la variable Y sur l'observation i.

Cependant, on peut constater aisément que lorsqu'on dispose de R observations, il est nécessaire d'estimer les R(R-1)/2 coefficients de la matrice d'interactions. C'est à cette étape de spécification qu'on peut faire intervenir la notion d'espace : on s'attend à priori à une diminution de l'intensité de l'interaction entre deux observations quand leur éloignement augmente. Parmi les indicateurs d'éloignement, on peut citer le coefficient de contiguïté.

Concept de contiguïté

Le concept de contiguïté permet de décider de ce qui est proche ou non, la quantification des interactions passant par une modélisation de type économétrique. Nous définissons donc un coefficient de contiguïté, qui peut se généraliser par le biais de la contiguïté à l'ordre K permettant d'envisager des interactions indirectes, à un ordre plus élevé d'éloignement. D'autres définitions sont proposées dans cette section. La figure 1 illustre un exemple de 5 zones spatiales entre lesquelles le calcul d'une distance permet d'envisager la définition de la contiguïté.

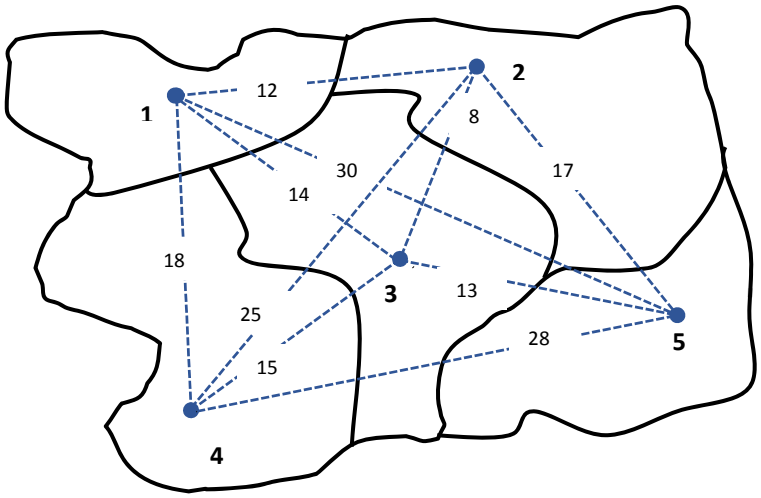


Figure 1 : exemple de 5 zones géographiques et distances entre les points représentant chaque zone

L'interaction entre deux observations géographiques r et s, est d'autant plus importante qu'elles sont proches. Cette interaction peut être exprimée en fonction décroissante $f(.)$ de la distance $d(r,s)$ entre ces deux points. La fonction décroissante $f(.)$ traduit ainsi la baisse d'interaction quand l'éloignement augmente. On dit que deux zones géographiques r et s sont contiguës si elles partagent une frontière commune [41]. Plus généralement, on définit la distance de contiguïté entre deux zones comme le nombre minimal de frontières qu'il faut franchir pour aller de l'intérieur de l'une à l'intérieur de l'autre. A partir d'un tableau de distances entre les observations et pour toute observation r, on construit l'ensemble des zones qualifiées de contiguës ou voisines. Nous avons formulé cette notion de contiguïté pour des observations isolées dans l'espace [12] [13] en fonction décroissante de la distance. Nous avons particulièrement travaillé, en analyse d'images, sur les pixels qui sont des observations isolées et régulières dans l'espace. Plusieurs définitions de distances usuellement utilisées en analyse d'images ont été utilisées pour bâtir la contiguïté [13]. Deux idées peuvent émerger pour définir la contiguïté intuitivement :

- deux zones sont déclarées contiguës si elles se jouxtent,

- deux zones sont déclarées contiguës si elles sont assez proches.

Nous en déduisons donc les définitions de contiguïté suivantes :

Définitions

Considérons E un ensemble fini de zones spatiales. De manière générale, nous définissons la contiguïté de manière binaire par un coefficient, dit de contiguïté, comme suit :

$$\forall\, r \in E, \qquad \forall\, s \in E, \qquad c_{rs} = \begin{cases} 1 & si\ r\ et\ s\ sont\ contiguës \\ 0 & sinon \end{cases}$$

Le problème est la détermination pratique de ce coefficient c_{rs}. Nous avons fait l'hypothèse que l'interaction entre deux points géographiques r et s est considérée d'autant plus importante qu'ils sont proches. Cette interaction peut donc être généralement exprimée en fonction décroissante $f(.)$ de la distance $d(r,s)$ entre ces deux points. La fonction décroissante $f(.)$ traduit ainsi la baisse d'interaction quand l'éloignement augmente.

Si nous suivons les deux idées intuitives exposées ci-dessus, nous pouvons, par exemple, définir la contiguïté de deux manières différentes :

Définition 1 : On dit que deux zones géographiques sont contiguës si elles possèdent une frontière commune. En considérant que la distance entre deux zones géographiques r et s est le nombre de frontières à traverser pour aller d'une zone à l'autre, on peut définir le coefficient de contiguïté de la manière suivante :

$$c_{rs} = \begin{cases} 1 \; si \; d(r,s) = 1 \\ 0 \; sinon \end{cases}$$

Définition 2 : On dit que les observations (ou zones géographiques) r et s sont contiguës si la distance qui les sépare est comprise entre deux seuils donnés t_1 et t_2.

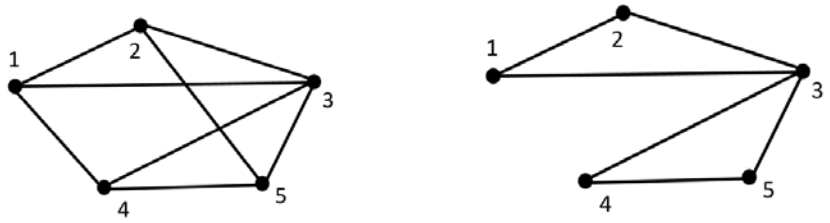
$$c_{rs} = \begin{cases} 1 \; si \; t_1 \leq d(r,s) \leq t_2 \\ 0 \; sinon \end{cases}$$

où $d(r,s)$ est la distance entre r et s, t est le seuil de distance fixé.

Le calcul du coefficient de contiguïté c_{rs} entre deux zones quelconques r et s de E induit un graphe non orienté, appelé graphe de contiguïté et défini par :

$$\forall \, r \in E, \forall \, s \in E, \qquad r \Gamma s \; \Leftrightarrow c_{rs} = 1$$

Dans le cas de l'exemple ci-dessus, selon si on considère les frontières entre les observations ou si on se base sur la distance, on obtient les deux graphes suivants :



(a) (b)

Figure 2. *Graphes relatifs aux deux définitions de contiguïté sur les données de la figure 1. Dans le cas (b), le seuil t est fixé à 15)*

Les boucles sur les sommets ont été omises dans le diagramme, pour une meilleure lisibilité.

La matrice de contiguïté d'un ensemble de R observations est une matrice carrée C où le terme c_{rs} représente le coefficient de contiguïté entre les observations r et s. Les coefficients c_{rr} sont nuls.

Cependant, deux observations non contiguës peuvent être proches l'une de l'autre ou beaucoup plus éloignées. D'où la nécessité de généraliser la notion de contiguïté à un ordre k pour prendre en compte les interactions entre zones non voisines. Le graphe précédent permet de généraliser la contiguïté à l'ordre K de la manière suivante : deux zones r et s de E sont dites contiguës à l'ordre K si et seulement si le plus court chemin entre r et s est de longueur K. Pour K=1, on retrouve ainsi le concept simple de contiguïté, K=0 correspondant à la non contiguïté. Dans notre exemple, compte tenu des deux manières dont la contiguïté a été définie, nous obtenons :

- Deux observations sont dites contiguës à l'ordre K s'il faut traverser K frontières pour aller d'une zone à l'autre.

- Deux observations sont dites contiguës à l'ordre k si on peut les joindre par un plus petit chemin de longueur k d'arcs comprise entre deux seuils donnés t_{1k} et t_{2k}.

*Si nous nous basons sur la notion de distance, d*eux observations sont dites contiguës à l'ordre k si la distance qui les sépare est comprise entre deux seuils donnés t_{1k} et t_{2k}.

$$c_{rs}^{(K)} = \begin{cases} 1 \; si \;\; t_{1k} \leq d(r,s) \leq t_{2k} \\ 0 \; sinon \end{cases}$$

Remarque

Dans le début de cette section, nous avons constaté la difficulté de devoir estimer R(R-1) coefficients lorsqu'on dispose de R observations. La matrice d'interactions A peut être exprimée en fonction de la matrice de contiguïté C par la relation suivante :

$$A(\alpha_1, \alpha_2, \ldots., \alpha_K) = \sum_{k=1}^{K} \alpha_k C^{(k)}$$

Chaque élément $\alpha_k$représente le degré d'interaction pour la contiguïté d'ordre k. $C^{(K)}$ représente la matrice de contiguïté d'ordre K. L'interaction est ainsi réduite aux zones contiguës. Le nombre de paramètres à estimer va ainsi baisser considérablement à K paramètres au lieu de R(R-1)/2.

Autocorrélation et auto-régression spatiale

Considérons un ensemble de R observations décrites par une variable X. On peut se poser la question si pour l'ensemble des observations, les valeurs de X sont proches les unes des autres ou si elles sont dispersées. Les valeurs les plus élevées sont-elles concentrées sur quelques localisations ? L'autocorrélation spatiale traduit l'idée que les valeurs prises par une variable aléatoire X dans un ensemble de localisations ne sont pas disposées au hasard, mais souvent proches les unes des autres. Il y a autocorrélation spatiale si les localisations voisines ont des valeurs proches. Cliff et Ord [21] ont formulé l'hypothèse d'absence d'autocorrélation spatiale sur des exemples simples. Ils ont montré que le rejet ou l'acceptation du test d'absence d'autocorrélation dépend de la définition du graphe de contiguïté. Deux méthodes connues dans la littérature sont souvent utilisées pour tester l'autocorrélation spatiale. Le coefficient de Moran [48] et le coefficient de Geary [30]. Ce qu'il faut retenir de ces études est qu'on ne teste pas l'absence d'autocorrélation spatiale d'une manière générale, mais toujours pour une définition donnée du graphe de voisinage.

L'auto-régression spatiale (AR) est une modélisation des données réelles, définie sur un ensemble spatial discret S. L'ensemble S peut être régulier ($S \subset \mathbb{Z}^2$) ou non. Cette modélisation **explicative** est bien adaptée aux données agrégées par zones spatiales (par exemple, en épidémiologie, X_s est le nombre de personnes malades dans la localisation s).

Parmi les modèles les plus souvent utilisés en analyse spatiale, on peut citer les modèles autorégressifs simultanés (SAR) et les modèles autorégressifs conditionnels (CAR) [36]. Les modèles autorégressifs spatiaux avec ou sans autocorrélation sont largement étudiés dans la littérature [13] [41], notamment en économétrie et en épidémiologie.

2.4 Contributions

2.4.1 Formalisation de la contiguïté

Notons que la définition de la contiguïté définie précédemment peut être formulée différemment. Plus généralement, on s'intéresse à un ensemble d'observations disposées dans le domaine spatial D. Ces observations sont représentées par des points ou par des zones géographiques. Définir une relation binaire $V \subseteq SxS$ revient à associer à chaque observation $s \in S$, l'ensemble de ses voisins noté Vs :

$s \in V_r$ si et seulement si $(r,s) \in V$

Ainsi la donnée de la relation binaire V est équivalente à celle d'un graphe [17], dit de voisinage G=(S , E).

On définit la matrice de contiguïté (de connectivité) de ce graphe, la matrice carrée notée C de terme général :

$$c_{rs} = \begin{cases} 1 \; si \; (r,s) \in V \\ 0 \; si \; (r,s) \notin V \; ou \; si \; r = s \end{cases}$$

Contiguïté directionnelle

Nous avons proposé d'autres définitions de la contiguïté, notamment en analyse d'images. La contiguïté en lignes et en colonnes d'une image permet de prendre en compte séparément les interactions horizontales et verticales :

Définition 4 : contiguïté d'ordre (k_i, k_j)

Deux observations s_1 et s_2 de coordonnées(x_1, y_1) et (x_2, y_2) sont dites contiguës à l'ordre (k_i, k_j) si les différences absolues $|x_1 - x_2|$ et $|y_1 - y_2|$ sont respectivement égales à k_i et k_j. Nous disposons donc de deux décalages (Un décalage vertical k_i et un décalage horizontal k_j).

Nous avons également proposé des définitions de contiguïté directionnelle permettant de mettre en évidence les interactions spatiales dans des directions précises de l'image. En plus de la définition classique de la contiguïté, les contiguïtés d'ordre (k_i, k_j) et les contiguïtés directionnelles ont été utilisées pour tester la présence d'autocorrélation spatiale dans les textures des images [13].

Contiguïté généralisée

Dans les définitions précédentes de la contiguïté, le coefficient de contiguïté prend les valeurs 0 et 1. Cette définition est parfois inadaptée à des situations où on veut exprimer une contiguïté plus ou moins nuancée. Nous avons donc donné une définition où le coefficient peut prendre des valeurs réelles entre 0 et 1.

Définition 5 : contiguïté généralisée

La contiguïté généralisée entre deux observations r et s est définie par :

$$c^*(r,s) = f\big(c(r,s)\big)$$

Ce coefficient prend ses valeurs entre 0 et 1.

L'exemple suivant illustre cette notion de contiguïté généralisée. Soit D l'ensemble de observations s tels que c(r,s)=1. Nous pouvons construire une mesure de dissemblance normée d'une observation r à une observation s ∈ D de la manière suivante :

$$p(r,s) = \frac{x_{max} - |x_r - x_s|}{\sum_{s \epsilon D} x_{max} - |x_r - x_s|}$$

Où x_{max}. est la valeur maximale des observations de D, x_r et x_s sont respectivement les valeurs des observations r et s. Cette mesure est d'autant plus importante que les valeurs des observations r et s sont proches, et devient nulle si la différence des valeurs est égale à x_{max}. Le coefficient de contiguïté généralisée entre r et s est défini par :

$$c^*_{rs} = \begin{cases} p(r,s) \; si \; c(r,s) = 1 \\ 0 \; sinon \end{cases}$$

A titre d'illustration, l'exemple suivant fournit la matrice de contiguïté de 9 observations disposées régulièrement dans l'espace (figure 3-a). La valeur de chaque coefficient dépend de la distribution des valeurs des observations contiguës (figure 3-b).

1	2	3
4	5	6
7	8	9

(a)

1	7	0
3	5	1
2	2	4

(b)

	1	2	3	4	5	6	7	8	9
1	0	0.11	0	0.55	0.33	0	0	0	0
2	0.1	0	0	0.3	0.5	0.1	0	0	0
3	0	0	0	0	0.25	0.75	0	0	0
4	0.2	0.12	0	0	0.2	0	0.24	0.24	0
5	0.09	0.15	0.06	0.15	0	0.09	0.12	0.12	0.18
6	0	0.05	0.31	0	0.15	0	0	0.26	0.21
7	0	0	0	0.23	0.29	0	0	0.47	0
8	0	0	0	0.21	0.14	0.21	0.25	0	0.17
9	0	0	0	0	0.4	0.26	0	0.33	0

(c)

Figure 3. *(a) 9 observations, (b) valeurs de x*

(c) Matrice de contiguïté correspondante selon la distance d_8

Application au lissage spatial

A partir de la définition de la contiguïté généralisée, nous avons proposé une méthode de lissage spatial des données [16]. Le principe consiste à calculer la valeur d'une observation donnée en fonction des valeurs des observations contiguës pondérées par les coefficients de contiguïté généralisée définie ci-dessus. Cette valeur est donc calculée de la manière suivante :

$$\hat{x}_r = \sum_{s \in D_r} c_{rs}^* x_s$$

Dans l'exemple cité précédemment, l'utilisation de cette méthode de lissage fournit le résultat suivant :

1	7	0
3	5	1
2	2	4

(a)

4.07	3.6	2.00
3.00	2.58	2.04
3.08	2.72	2.92

(b)

Figure 4. *(a) valeurs initiales de x, (b) valeurs calculées*

Cette méthode de lissage peut être utilisée dans divers domaines de l'analyse spatiale. Nous l'avons appliqué particulièrement à l'amélioration des images numériques. Les résultats obtenus sont très satisfaisants. Elle permet notamment de réduire le bruit présent dans les images.

La contiguïté généralisée permet de tenir compte non seulement la proximité entre observations disposées dans l'espace, mais aussi des valeurs prises par ces observations. Sur cette définition, nous avons basé une méthode de lissage spatial qui peut s'avérer très utiles dans certains domaines utilisant l'analyse spatiale. Cette méthode de lissage peut être utilisée pour l'estimation des données manquantes dans l'espace. Les modèles autorégressifs spatiaux peuvent être basés sur la contiguïté généralisée. Elle peut également être utilisée pour le lissage des images numériques. La figure 5 montre le résultat de lissage spatial obtenu à partir d'une image originale de taille (214x128). Il faut noter que l'ordre de contiguïté dans cet exemple est limité à 1. Si en revanche, on augmente l'ordre de contiguïté généralisée, la gamme de tons sera de plus en plus différente.

(a) (b)

Figure 5. Résultat du lissage spatial sur une image numérique. a) Image originale b) image résultat

2.4.2 Modèle spatial d'auto-régression logistique

Dans les modèles de régression logistique classique, on s'intéresse à la prédiction des valeurs prises par une variable catégorielle à partir d'une série de variables explicatives (binaires ou continues). Ces modèles ne prennent pas en compte la proximité entre les observations disposées dans un domaine spatial. Afin de prendre en compte les interactions spatiales entre ces observations, nous proposons dans un premier temps un modèle d'autorégression logistique basé sur la notion de contiguïté définie ci-dessus.

Etant donné une variable binaire Y à expliquer au point r, si on note Y_1, Y_2,…Y_R les observations voisines de r, la probabilité de réalisation de l'évènement (Y_r=1) dépend des valeurs prises par Y dans un domaine se situant autour de r.

$$p = P\left(Y_r = 1/\left(Y_1 = y_1, Y_2 = y_2, ...Y_R = y_R\right)\right) = \frac{1}{1+\exp\left(-\left(\sum_{s\in Vr}\alpha_{rs}y_s\right)\right)} \qquad (1)$$

Où les α_{rs} sont les paramètres à estimer. V_r est l'ensemble des points voisins de r. Le nombre d'observations est noté R. Le modèle logit s'écrit alors

$$\text{Logit}(p(x_1,...x_R) = \beta_0 + \sum\beta_i X_i$$

Remarque

Pour un nombre d'observations égal à R, il y a au départ R(R-1) coefficients α_{rs} à estimer. Il y a donc plus de coefficients à estimer que d'observations. C'est à cette étape de spécification qu'intervient la notion d'espace : on s'attend à priori à une **diminution de l'intensité de l'interaction entre deux observations quand leur éloignement augmente**. Un bon indicateur d'éloignement est le coefficient de contiguïté. Le plus simple et le plus fréquent est celui du premier ordre (interaction directe). Les coefficients α_{rs} sont alors exprimés en fonction de la distance d et le vecteur de paramètres θ à estimer.

$$\alpha_{rs} = \alpha\left(d(r,s),\theta\right)$$

Dans le cas général où on prend en compte les interactions à l'ordre k, la probabilité de réalisation de l'évènement (Y_r=1) est donnée par

$$p = \frac{1}{1+\exp\left(-\sum_{k=1}^{K}\theta_k\left(\sum_{s=1}^{R}c_{rs}^{(k)}y_s\right)\right)}$$

$$\text{logit}(p) = \sum_{k=1}^{K}\theta_k\left(\sum_{s=1}^{R}c_{rs}^{(k)}y_s\right)$$

θ_k mesure le niveau global d'autorégression à l'ordre k.

Comme la montre la figure 6, la prise en compte de la contiguïté dans interactions spatiales permet de réduire le nombre de paramètres à estimer. En effet, un seul paramètre est requis pour un ordre donné de contiguïté.

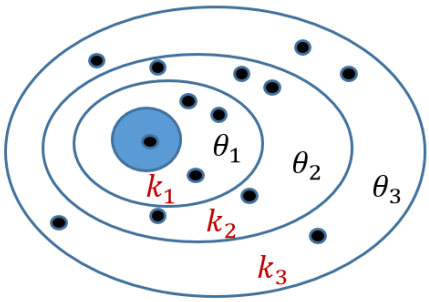


Figure 6. Ordre de contiguïté

Extension du modèle

Le modèle proposé précédemment peut être étendu pour expliquer la variable y à partir d'un ensemble de variables explicatives. Dans ce cas, la variable binaire Y à expliquer au point r dépend d'une part, des valeurs prises par Y au voisinage de r et d'autre part, des valeurs prises par les variables explicatives (x_i, i=1, …m).

$$p = P\left(Y_r = 1/\left(Y_1, Y_2, ...Y_R, X_1, X_2, ..X_R\right)\right) = \frac{1}{1+\exp\left(-\left(\sum_{s=1}^{R}\alpha_{rs}y_s + \sum_j \beta_j x_{rj}\right)\right)} \qquad (2)$$

où β_i sont des paramètres à estimer.

Le modèle logit s'écrit dans ce cas,

$$\text{Logit}(p) = \sum_{s=1}^{R}\alpha_{rs}y_s + \sum_j \beta_j x_{rj}$$

Comme dans la section précédente, en prenant en compte les interactions spatiales, la probabilité p et le modèle logit prennent les formes suivantes :

$$p = \frac{1}{1+\exp\left(-\sum_{k=1}^{K}\theta_k\left(\sum_{s=1}^{R}c_{rs}^{(k)}y_s\right) + \sum_j \beta_j x_{rj}\right)}$$

$$\text{logit}(p) = \sum_{k=1}^{K}\theta_k\left(\sum_{s=1}^{R}c_{rs}^{(k)}y_s\right) + \sum_j \beta_j x_{rj}$$

Estimation des paramètres

Les paramètres du modèle (1) peuvent être estimés par la méthode du maximum de vraisemblance. L'expression de la fonction de vraisemblance pour ce modèle est donnée par

$$L = \prod_{i=1}^{R}\left(P\left(Y_r = 1/\left(Y_{1i}, ...Y_{Ri}\right)\right)^{Y_i}\right)\left(P\left(Y_r = 0/\left(Y_{1i}, ...Y_{Ri}\right)\right)^{1-Y_i}\right)$$

Soit,

$$L = \prod_{i=1}^{R} \left(\frac{\exp\left(\sum_{k=1}^{K} \theta_k \left(\sum_{s=1}^{R} c_{rs}^{(k)} y_{si} \right) \right)}{1 + \exp\left(\sum_{k=1}^{K} \theta_k \left(\sum_{s=1}^{R} c_{rs}^{(k)} y_{si} \right) \right)} \right)^{Y_i} \left(\frac{1}{1 + \exp\left(\sum_{k=1}^{K} \theta_k \left(\sum_{s=1}^{R} c_{rs}^{(k)} y_{si} \right) \right)} \right)^{1-Y_i}$$

Il n'y a pas de méthode donnant une expression explicite des estimations $\hat{\theta}_1, \hat{\theta}_2 .... \hat{\theta}_k$. On sera conduit à maximiser $L(\theta_1, \theta_2 .... \theta_k)$ en utilisant un algorithme usuel de maximisation par balayage (l'algorithme du recuit simulé ou un algorithme génétique).

L'estimation du paramètres du modèle par maximisation de la fonction de vraisemblance étant impossible analytiquement, nous avons proposé une méthode sous-optimale de maximisation qui consiste à discrétiser l'espace des paramètres $\theta_1, \theta_2 .... \theta_k$ et de calculer la valeur de la vraisemblance pour chaque t-uple de valeurs de paramètres compris entre -1 et 1 (voir figure 7). Il faut noter que plus le pas de discrétisation est faible, plus le risque de tomber sur un maximum local est faible.

Le t-uple de paramètres $\hat{\theta}_1, \hat{\theta}_2 .... \hat{\theta}_k$ ainsi obtenu permet d'opérer une nouvelle recherche par discrétisation du sous-espace des paramètres autour de $\hat{\theta}_1, \hat{\theta}_2 .... \hat{\theta}_k$

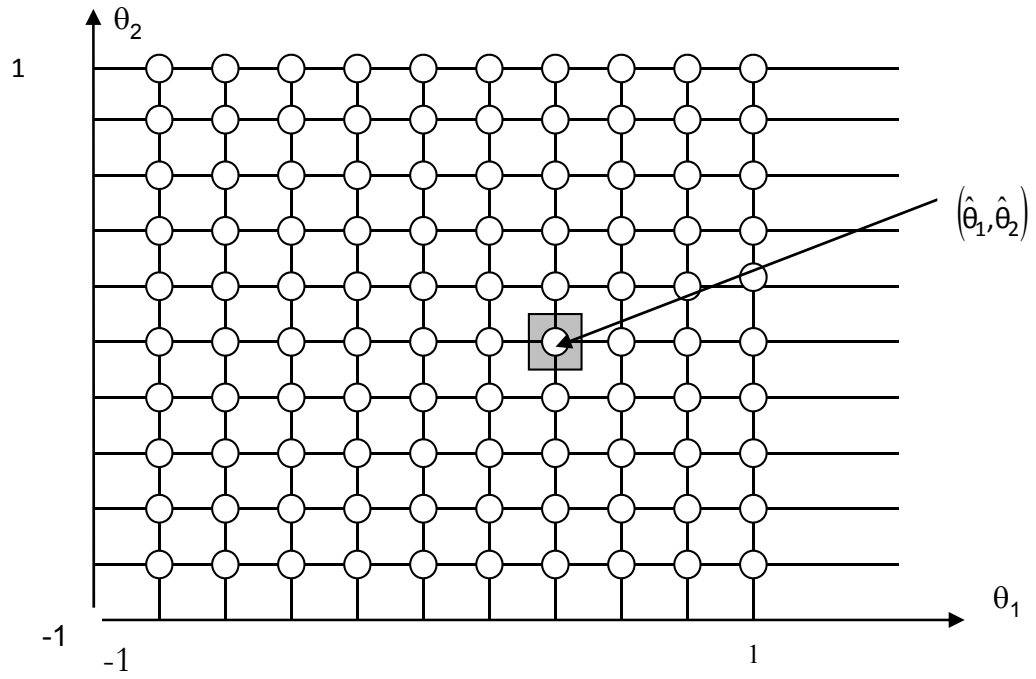


Figure 7. Exemple d'espace à deux paramètres

Dans le cas du modèle (2), l'expression de la fonction de vraisemblance est donnée par :

$$L = \prod_{i=1}^{R} \left(\frac{\exp\left(\sum_{k=1}^{K} \theta_k \left(\sum_{s=1}^{R} c_{rs}^{(k)} y_{si} \right) + \sum_{j} \beta_j x_{rji} \right)}{1 + \exp\left(\sum_{k=1}^{K} \theta_k \left(\sum_{s=1}^{R} c_{rs}^{(k)} y_{si} \right) + \sum_{j} \beta_j x_{rji} \right)} \right)^{Y_i} \left(\frac{1}{1 + \exp\left(\sum_{k=1}^{K} \theta_k \left(\sum_{s=1}^{R} c_{rs}^{(k)} y_{si} \right) + \sum_{j} \beta_j x_{rji} \right)} \right)^{1-Y_i}$$

L'expression de la log-vraisemblance donne une forme simplifiée à maximiser :

$$\ln(L) = \sum Y_i \left(\sum_{k=1}^{K} \theta_k \left(\sum_{s=1}^{R} c_{rs}^{(k)} y_{si} \right) + \sum_{j} \beta_i x_{rji} \right) - \sum (1 - Y_i) \left(\sum_{k=1}^{K} \theta_k \left(\sum_{s=1}^{R} c_{rs}^{(k)} y_{si} \right) + \sum_{j} \beta_j x_{rji} \right)$$

L'estimation des paramètres peut bien se faire par des méthodes d'optimisation combinatoire tel que le recuit simulé ou les algorithmes génétiques. Nous avons proposé tout de même une méthode sous-optimale mais plus simple pour déterminer les paramètres. Selon le degré de précision souhaité, nous pouvons modifier le pas de discrétisation et réitérer l'algorithme pour affiner les valeurs des paramètres.

2.4.3 Modèle autorégressif spatial avec autocorrélation

Dans le modèle proposé, on considère que la valeur y prise par une variable aléatoire endogène Y pour l'observation spatiale r dépend d'une part, des valeurs prises par la variable Y dans un domaine tout autour , selon la définition de la notion de contiguïté mentionnée précédemment, et d'autre part, des valeurs x_{r1}, x_{r2}, …,x_{rI} des I variables explicatives. Ces variables explicatives portent sur des phénomènes qui se déroulent à l'extérieur de l'observation spatiale r et sont donc relatives à d'autres espaces que r. La notation x_{ri} signifie la valeur prise par X_i en un point donné de l'espace influençant le point r. La valeur prise par la variable Y_r en fonction des valeurs prises par les variables Y_s et les variables explicatives est donnée sous la forme

$$y_r = \sum_{s \neq r} \alpha_{rs} y_s + \sum_i x_{ri} \beta_i + \varepsilon_r \quad s = 1,....,R \qquad (1)$$

Les hypothèses habituelles sur les résidus aléatoires ε_r portent sur les moments du premier et du deuxième ordre. On pose

$$E(\varepsilon_r) = 0 \qquad \text{et} \qquad E(\varepsilon_r \, \varepsilon_s) \neq 0$$

On considère que les résidus aléatoires ε_r et ε_s ne sont pas indépendants. Il y a donc autocorrélation spatiale.

$$\varepsilon_r = \sum_{s \neq r} \theta_{rs} \varepsilon_s + \eta_r$$

η_r est un résidu aléatoire de moyenne E_η nulle et de variance $V_\eta = \sigma^2$. (σ^2 est la variance du résidu dans le cas d'homoscédasticité). Il faut noter que dans ce modèle, les interactions spatiales passent à la fois par la partie déterministe et par la partie aléatoire du modèle.

Pour un nombre d'observations égal à R, il y a au départ R(R-1) coefficients α_{rs} à estimer. Il y a donc plus de coefficients à estimer que d'observations. C'est à cette étape de spécification qu'intervient l'espace : on s'attend à priori à une diminution de l'intensité de l'interaction entre deux observations quand leur éloignement augmente. Un bon indicateur d'éloignement est le coefficient de contiguïté. Le coefficient α_{rs} peut être exprimé sous cette forme :

$$\alpha_{rs} = \theta \, c_{rs}$$

θ mesure le niveau global d'autorégression. Quand il est nul, il n'y a pas d'autorégression et quand il est positif, il y a autorégression positive. L'équation (1) peut s'écrire alors sous cette forme :

$$y_r = \theta \left(\sum_{s \neq r} c_{rs} y_s \right) + \sum_i x_{ri} \beta_i + \varepsilon_r \qquad (2)$$

En ajoute une condition de normalisation : la somme des coefficients relatifs à un lieu donné est égale à l'unité.

$$\sum_s c_{rs} = 1$$

L'écriture du modèle spatial peut être facilement généralisée à des indicateurs de contiguïté d'ordre supérieur à 1. Le modèle spatial d'ordre K s'écrit alors sous cette forme :

$$y_r = \sum_{k=1}^{K} \theta_k \sum_{s \neq r} c_{rs}^k y_s + \sum_i x_{ri} \beta_i + \varepsilon_r \qquad (3)$$

Pour simplifier l'écriture du modèle, on va utiliser les notations matricielles suivantes :

Y	Vecteur colonnes centré des valeurs de la variable aléatoire Y.
A	Matrice carrée de taille RxR représentant les effets d'autorégression.
G	Matrice carrée de taille RxR représentant les effets d'autocorrélation.
ε,η	Vecteurs colonnes (Rx1) des résidus aléatoires.
β	Vecteur colonnes des I coefficients des variables explicatives.
μ	Vecteur des paramètres d'hétéroscédasticité.
α	Vecteur des paramètres d'autorégression.
θ	Vecteur des paramètres d'autocorrélation.
$C^{(k)}$	Matrice de contiguïté d'ordre k.

Sous forme matricielle, le modèle linéaire spatial avec autorégression et autocorrélation que nous proposons prend la forme :

$$\begin{cases} Y = A(\alpha).Y + X\beta + \varepsilon \\ \varepsilon = G(\theta).\varepsilon + \eta \end{cases} \qquad (4)$$

avec $A(\alpha)=\alpha_1 C^{(1)} + \alpha_2 C^{(2)} +\ldots..+ \alpha_K C^{(K)}$

et $G(\theta)=\theta_1 C^{(1)} + \theta_2 C^{(2)} +\ldots..+ \theta_K C^{(K)}$

En général A et G ne doivent pas être identiques pour les effets d'autorégression et d'autocorrélation. On peut cependant être amené à considérer que A et G suivent la même forme fonctionnelle et ne diffèrent l'une de l'autre que par la valeur des paramètres à estimer. Les vecteurs de paramètres α et θ sont à estimer. Ce modèle fait intervenir à la fois un processus autorégressif spatial que l'on peut choisir à sa guise en spécifiant précisément la matrice A et à la fois un processus d'autocorrélation spatiale que l'on spécifie par l'intermédiaire de la matrice G. Pour mieux expliquer le taux de naissances dans un point donné, nous avons fait intervenir des variables exogènes (Xi). Ce type de modèle semble plus intéressant que d'autres, plus simples, justement par la prise en compte des phénomènes autorégressifs avec autocorrélation spatiale.

Estimation des paramètres

Les paramètres du modèle spatial sont estimés par maximisation de la fonction de vraisemblance donnée par l'expression :

$$L(\alpha, \theta, \mu, \nu, \beta) = -\left(\frac{R}{2}\right) ln\,(2\pi\,\nu) - \left(\frac{1}{2}\,\nu\right) SRG(\alpha, \theta, \beta) + ln\,(|I - A(\alpha)|) + ln\,(|I - G(\theta)|) - (1/2)\,ln\,(|V(\mu)|)$$

Où :

$$SRG(\alpha, \theta, \mu, \beta) = [(I - A)y - X\,\beta]^t\,(I - G^t)\,V^{-1}(\mu)\,(I - G)\,[(I - A)\,y - X\,\beta]$$
$$SRG(\alpha, \theta, \mu, \beta) = \left(\tilde{y} - \tilde{X}\beta\right)^t\left(\tilde{y} - \tilde{X}\beta\right)$$

Et $\tilde{y}=(I-G).(I-A)\sqrt{V(\mu)}.y$ $\tilde{X}=(I-G)\sqrt{V(\mu)}.X$

Les estimateurs de β et ν sont donnés par :

$$\hat{\beta}(\alpha,\theta,\mu) = \frac{\tilde{X}^t\tilde{y}}{\tilde{X}^t\tilde{X}}$$

$$\hat{\nu}(\alpha,\theta,\mu) = \frac{1}{R}SRG(\alpha,\theta,\mu,\tilde{\beta}(\alpha,\theta,\mu))$$

La valeur de la fonction de vraisemblance à maximiser est donnée par :

$$l_1(\alpha,\theta,\mu) = (R/2)\left[1+\ln(2\pi/R)\right] - (R/2)\ln\left(SRG\left(\alpha,\theta,\mu,\hat{\beta}(\alpha,\theta,\mu)\right)\right) + \ln\left(\left|I-A(\alpha)\right|\right) + \ln\left(\left|I-G(\theta)\right|\right)$$

Comme dans la section précédente, il n'y a pas de méthode analytique permettant d'obtenir une expression explicite des paramètres. Pour maximiser la fonction de vraisemblance l_1, on utilise une méthode d'optimisation combinatoire tel que le recuit simulé.

L'hypothèse d'absence d'autorégression spatiale revient à tester la significativité de l'estimateur $\hat{\alpha}$ du vecteur α (test de l'hypothèse $\alpha = 0$ contre l'hypothèse $\alpha \neq 0$) en utilisant le rapport de vraisemblance. On procède de la même manière pour le test d'absence d'autocorrélation spatiale.

2.4.4 Applications

Estimation du taux de pollution

Le modèle précédent a été utilisé dans le cadre d'un projet pilote du CNRS portant le nom de MOUSSON qui a pour but de mettre en place un système d'alerte à la pollution à Ouagadougou. Il a pour ambition d'améliorer les moyens de lutte contre la pollution et les maladies associées telles que la méningite. Pour cela, le projet a rassemblé toutes les compétences nécessaires afin de renforcer les systèmes d'alerte. L'idée est d'arriver à détecter efficacement les conséquences néfastes de la pollution afin d'informer et sensibiliser les populations. Ce système d'alerte est opéré par la météo nationale et les télécoms et renseigné par des mesures de pollution par aérosols, notamment carbonés et soufrés.

Afin d'obtenir une information sur la qualité de l'air, une campagne de mesures spécifiques a été effectuée. La première étape consiste à dresser une liste, la plus exhaustive possible, des principaux éléments pollueurs aériens (feux domestiques, feux d'ordure, cyclomoteurs, voitures,…), à les situer

pour ceux qui sont fixes et déterminer les trajectoires habituelles de ceux qui sont mobiles. Ces mesures doivent tenir compte des conditions climatiques, météorologiques et socio-économiques.

Dans certaines zones, on ne dispose pas de l'information concernant le taux de pollution. Nous avons proposé dans ce cadre un modèle spatial permettant d'estimer ce taux de pollution en fonction des taux voisins et en tenant compte des conditions météorologiques (direction du vent, température, points de pollution,…).

Estimation du taux de naissances dans les maternités

Nous avons étudié le taux de naissances dans les maternités des communes du département du Rhône. En l'absence de toute information, on peut se demander dans quelle mesure le taux de naissance dans la maternité d'une commune r est expliqué par le taux dans les communes voisines.
Le modèle présenté ci-dessus, basé sur la notion de contiguïté a été utilisé pour l'estimation du taux de naissances dans certaines maternités du département du Rhône.
Sur un ensemble de 16 communes du département nous avons estimé le taux de naissance dans chaque commune. Les estimations obtenues sont très variables : dans certaines communes les données estimées sont proches des données réelles. C'est notamment le cas des communes proches du centre. Ceci peut s'expliquer par le fait que ces communes sont entourées de plusieurs autres communes qui contribuent à l'estimation. Au contraire les communes isolées (se trouvant à la frontière du département) ont moins de valeurs autours permettant de mieux prédire le nombre de naissances. D'autre part, les données représentent un échantillon des communes du département. Une analyse des données collectées au niveau des maternités de tout le département et des données de départements voisins permettra d'améliorer la précision des estimations.

2.5 Conclusion du chapitre

Comme le montre ce chapitre, dans l'analyse spatiale, les données contiennent deux classes d'information. La première classe inclut des attributs des caractéristiques spatiales mesurés sur des variables quantitatives ou nominales. La deuxième classe concerne la position d'une caractéristique spatiale décrite par un référencement sur une carte mesurée dans une ou plusieurs coordonnées géographiques. La prise en compte des positions relatives des objets est déterminante en analyse spatiale. Plusieurs méthodes classiques ont été proposées dans la littérature en les adaptant au contexte spatial. Nous avons développé de nouvelles approches et avons implémenté quelques algorithmes. Ces travaux se traduisent par des contributions dans la façon de prendre en compte la notion de voisinage en proposant d'autres manières de définir la contiguïté et en proposant de la construire dans des espaces prétopologiques. Des modèles de régression, notamment la régression logistique spatiale, ont été proposés et testés sur divers jeux de des données.
De manière générale, en observant les données, on constate dans la pratique qu'une grande partie des données disponibles est rassemblée à des buts autres que l'analyse spatiale. D'autre part, les avancées dans les SIG contribuent à l'accroissement de l'utilisation des méthodes d'analyse spatiale. Un de nos objectifs futurs est de construire (autour SIG combinant des données scio-économiques, des données de santé et autres) des méthodes et des techniques permettant de mieux modéliser certains phénomènes complexes, nécessitant la prise en compte des interactions spatiales, tel que la modélisation des épidémies, la prédiction de la pollution ou la météorologie.

2.6 Références

[1] Anselin, L. Spatial econometrics : Methods and models Kluwer academic publishers, Dordrecht. 1988

[2] Anselin L., « Local indicators of spatial association : LISA », Geographical Analysis, Vol. 27, n° 2, pp. 93-115 1995.

[3] Anselin L., « Interactive Techniques and Exploratory Spatial Data Analysis », Geographical Information Systems : Principles, Techniques, Management and Applications, Cambridge, Geo-information International 1996.

[4] Anselin L., « Exploratory spatial data analysis in a geocomputational environment », Geographical Information Systems : a primer, Cambridge, Geoinformation International, 1998 :.

[5] Ambroise, C., Dang, M. V., Govaert, G., « Clustering of spatial data by the EM algorithm », In Soares, A., Gômez-Hernandez, J., and Froidevaux, F. (Eds), geoENV1-Geostatistics for Environmental Applications, volume 9, 1997, Quantitative Geology and Geostatistics Publisher, Kluwer Academic, pp. 493-504.

[6] Anselin, L., Local indicators of spatial association – LISA, Geographical Analysis, Vol. 27(2), 1995, pp. 93-115.

[7] Arnaud M. et Emery X., 2000 : Estimation et interpolation spatiale, Hermes, Paris, 221 p.

[8] Bailley T. and Gatrell A., Interactive spatial data analysis, Longman, Scientific & Technical, Harlow, 1995, 413 p.

[9] Berglund S. and Karlstrom A.,: Identifying local spatial association in flow data », Journal of Geographical Systems, Vol. 1, pp. 219-236., 1999

[10] Besag J. and Newell J., « The detection of clusters in rare diseases », Journal of the Royal Statistical Society, Série A, Vol. 154, pp. 143-155, 1991

[11] Bolot, « Calibrating models based on anticipation by genetic algorithms », Actes de Colloque, Geocomputation, 6th International Conference on Geocomputation, Brisbane, 5 p. 2001

[12] Bounekkar A., Lamure M., NICOLOYANNIS N., TEXTURE CLASSIFICATION BASED UPON SPATIAL AUTOCORRELATION , SPIE The International Society for Optical Engineering. Visual Communications and Image Processing '96 17 - 20 mars 1996 Orlando, Florida. USA.

[13] Bounekkar A., Analyse statistique de texture : Autocorrélation spatiale et notions de contiguïté. Thèse de doctorat, Université Lyon I. 1997

[14] Bounekkar A., SPATIAL AUTOREGRESSIVE PROCESSES: A STEPWISE ALGORITHM FOR SPECIFICATION, The 7th International Conference on System Science In Health Care. 29 May-2 June, 2000. Budapest, Hungary

[15] Bounekkar A. , Lamure M., Spatial linear model with autoregression and autocorrelation. Application to study of birth rate in Rhône-Alpes commune maternities, Health and System Science, Volume 4, 2000.

[16] Bounekkar A., Contiguïté généralisée et lissage spatial, XVèmes jourrnées de statistiques 2-6 juin 2003 Lyon.

[17] Bounekkar A. Lamure M., Modèle spatial pour la mesure de pollution, International Conference on Systems Science in Health Care, Lyon (France). 2008.

[18] Bounekkar A., Spatial logistic regression based upon contiguity concept. ERCIM'08 - ERCIM Working Group on Computing & Statistics, Neuchatel, Switzerland, 2008.

[19] Brunsdon C., « Estimating probability surfaces in GIS : an adaptive technique », EGIS'91 Proceedings, Second European Conference in GIS, Brussels, Vol. 1, pp. 155-164, 1991.

[20] Brunsdon C., « Exploratory spatial data analysis and local indicators of spatial association with Xlisp-Stat », The Statistician, Vol. 47, n° 3, pp. 471-484, 1998.

[21] Cliff A.D., Ord J.K., "Spatial autocorrelation", Pion, London, 1973.

[22] Cliff, A. D. & J. K. Ord. Spatial processes: Models and applications. Pion Ltd., London. 1981

[23] CRESSIE N., Statistics for spatial data, John Wiley & Sons, New York, 900 p. 1993

[24] Ester M., Kriegel H.P., Sander J., Xu X., "A density-Based algorithm for discovering clusters in lager spatial databases with noise", In proceeding of second international conference on knowledge discovery and data mining, Portland, 1996, pp 226-231.

[25] Ester M., Kriegel H.P., Sander J., "Spatial Data Mining: A Database Approach", in proceedings of 5th Symposium on Spatial Databases, Berlin, Germany, 1997.

[26] Efron B. and Tibshiran R., « Statistical data analysis in the computer age », Science, Vol. 253, pp. 390-395, 1991.

[27] Fotheringham S., Zhan B., "A comparison of three exploratory methods for cluster detection in spatial point patterns", Geographical Analysis, Vol. 28, n° 3, 1996, pp. 200-218

[28] Fotheringham S., « Geocomputational analysis and modern spatial data », Geocomputation, pp. 33-48, 2000.

[29] Gatrell A., Bailey T., Diggle P., Rowlingson B., "Spatial point pattern analysis and its application in geographical epidemiology", Transactions of the Institute of British Geographers, n° 21, 1996, pp. 256-274.

[30] Geary, R.C. (1954) The contiguity ratio and statistical mapping. *The incorporated Statistician* : 5, 3, 115-145.

[31] Getis A. and Ord K., « The analysis of spatial association by use of distance statistics », Geographical Analysis, Vol. 24, n° 3, pp. 189-206, 1992.

[32] Ginns, P. Integrating information: A meta-analysis of the spatial contiguity and temporal contiguity effects. Elsevier, Learning and Instruction 16, 511-525. 2006.

[33] Govaert G., "Classification automatique et modèle de mélange: application aux données spatiales", Revue Internationale de Géomatique, Editions Hermès, Vol. 9, n° 4, 1999, Mai 2000, pp. 457-470.

[34] Griffith D.A. "Towards a theory of spatial statistics". Geographical Analysis Vol. 2 No 4. October 1980

[35] Griffith D., « What is spatial autocorrelation ? Reflections on the past 25 years of spatial statistics », L'Espace Géographique, Vol. 3, pp. 265-280, 1993.

[36] Guyon X., Statistique spatiale, Conférence S.A.D.A.' 07 Cotonou.

[37] Haining R., Spatial data analysis in the social and environmental sciences, Cambridge University Press, Cambridge, 409 p, 1990.

[38] Haining Robert P. “Spatial Data Analysis: theory and practice”. Cambridge University Press. 2007

[39] Haslett J., Bradley R., Craig P. and Unwin A., « Dynamic graphics for exploring spatial data with application to locating global and local anomalies », The American Statistician, Vol. 45, n° 3, pp. 234- 242, 1991

[40] Huang Y., Shekhar Sh., and Xiong H., Discovering Co-location Patterns from Spatial Datasets: A General Approach, IEEE Transactions on Knowledge and Data Engineering (TKDE), 16(12), pp. 1472-1485, December 2004

[41] Jayet H. Analyse spatiale quantitative: Une introduction, Economica, 1993.

[42] Koperski K. and Han J., "Discovery of Spatial Association Rules in Geographic Information Databases", In Advances in Spatial Databases (SSD’95), p. 47-66, Portland, ME, August 1995.

[43] Koperski K., J. Adhikary, J. Han, Knowledge Discovery in Spatial Databases: Progress and Challenges, Proc. SIGMOD Workshop on Research Issues in Data Mining and Knowledge

[44] Discovery (DMKD). Technical Report 96-08, University of British Columbia, Vancouver, Canada, 1996.

[45] Koperski K., Han J., Stefanovic N., "An Efficient Two-Step Method for Classification of Spatial Data", In proceedings of International Symposium on Spatial Data Handling (SDH’98), p. 45-54, Vancouver, Canada, July 1998.

[46] Koperski K., "A progressive refinement approach to spatial data mining", PhD Thesis, Simon Fraser University. April 1999.

[47] Moran, P.A.P. (1948) The interpretation of statistical maps. Journal of the Royal Statistical Society, B : 10, 243-251.

[48] Moran P.A. P. "A test of serial independance of residuals". Biometrica vol. 37, pp 178–181, 1950

[49] Openshaw S., Charlton M., Wymer C., Craft A., 1987 : "A mark 1 geographical analysis machine for the automated analysis of point data sets", International Journal of Geographical Information Systems, Vol. 1, n° 4, pp. 335-358

[50] Openshaw S., 1995, "Developing automated and smart spatial pattern exploration tools for geographical information systems applications", The Statistician, Vol. 44, n° 1, pp. 3-16

[51] Openshaw S., « Geocomputation », Geocomputation, pp. 1-31, 2000 .

[52] Ord J.K., Getis A., 1995, "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application, Geographical Analysis", Ohio State University Press, Vol. 27, n° 4, pp. 287-306

[53] Pumain D. et Saint-Julien T., L’analyse spatiale. Tome 1 : localisations dans l’espace, Armand Colin, Paris, 167 p, 1997.

[54] Sander J., Ester M., Kriegel H.-P., Xu X., Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications, in Data Mining and Knowledge Discovery, An International Journal, Kluwer Academic Publishers, Vol. 2, No. 2, 1998.

[55] Shaw G., D. Wheeler, Statistical Techniques in Geographical Analysis, Edition David Fulton, London, 1994.

[56] Saint-Julien T., Diffusion spatiale, Encyclopédie de Géographie, pp. 559-581, 1995.

[57] Sander J., Ester M., Kriegel H.-P., Xu X., Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications, in Data Mining and Knowledge Discovery, An International Journal, Kluwer Academic Publishers, Vol. 2, No. 2, 1998.

[58] Shaw G., D. Wheeler, Statistical Techniques in Geographical Analysis, Edition David Fulton, London, 1994.

[59] Shekhar Sh. and Huang Y., Discovering Spatial Co-location Patterns : A Summary of Results, In Proc. of 7th Int. Symposium on Spatial and Temporal Databases (SSTD), Springer-Verlag, Lecture Notes in Computer Science, July 2001

[60] Shekhar Sh.., Chawla Sanjay (2003), Spatial Databases: A Tour, Prentice Hall. [189] Yoo J. S.and Shekhar Sh., A Join-less Approach for Mining Spatial Co-location Patterns, to appear in IEEE Transactions on Knowledge and Data Engineering (TKDE), 2006.

[61] Shekhar Sh. and Huang Y., Discovering Spatial Co-location Patterns : A Summary of Results, In Proc. of 7th Int. Symposium on Spatial and Temporal Databases (SSTD), Springer-Verlag, Lecture Notes in Computer Science, July 2001

[62] Shekhar Sh.., Chawla Sanjay (2003), Spatial Databases: A Tour, Prentice Hall.

[63] Sokal R., Oden N. and Thomson B., « Local spatial autocorrelation in a biological model », Geographical Analysis, Vol. 30, n° 4, pp. 331-353, 1998.

[64] Unwin A., Unwin D. and Fisher P., « Exploratory spatial data analysis with local statistics », The Statistician, Vol. 47, n° 3, pp. 415-421, 1998.

[65] Wartenberg, D.E. (1985) Spatial autocorrelation as a criterion for retaining factors in ordinations of geographic data. *Mathematical Geology* : 17, 665-682.

[66] Zeitouni K. Analyse et extraction de connaissances des bases de données spatiotemporelles, rapport d'Habilitation à Diriger des Recherches, Université de Versailles Saint-Quentin-en-Yvelines, 2006

Chapitre 3

Métaheuristiques hybrides évolutionnaires pour l'optimisation multi-objectifs

Sommaire

Abréviations

ACO	Ant Colony Optimization *optimisation par colonies de fourmis*
DE	Differential Evolution *Evolution différentielle*
EC	Evolutionary Computation *Calcul évolutionnaire*
EA	Evolutionary Algorithms *Algorithmes évolutionnaires*
ES	Evolutionary Strategies *Stratégie évolutionnaire*
GA	Genetic Algorithms *Algorithmes génétiques*
GLS	Guided Local Search *Recherche locale guidée*
GRASP	Greedy Randomized Adaptive Search Procedure
DM-GRASP	GRASP with Data Mining
HEMH	Hybrid Evolutionary Metaheuristics *Métaheuristiques hybrids évolutionnaires*
HEMH2	Improved Hybrid Evolutionary Metaheuristics *Métaheuristiques hybrids évolutionnaires améliorées*
HESSA	Hybrid Evolutionary Approach with Search Strategy Adaptation
LS	Local Search *Recherche locale*
ILS	Iterated Local search *Recherche locale itérée*
MA	Memetic Algorithms *Algorithmes mémétiques*
MODM	Multiobjective Decision Making *Aide à la décision multiobjectifs*
MOEA	Multiobjective Evolutionary Algorithm *Algorithme évolutionnaire multiobjectifs*
MOEA/D	Multiobjective Evolutionary Algorithm based on Decomposition *Algorithme évolutionnaire multiobjectifs basé sur la décomposition*
MOGA	Multiobjective Genetic Algorithm *Algorithme génétique multiobjectifs*
MOKP	Multiobjective Knapsack Problems *Problème de sac à dos multiobjectifs*
MOP	Multiobjective Optimization Problem *Problème d'optimisation multiobjectifs*
NPGA	Niched Pareto Genetic Algorithm
NSGA	Nondominated Sorting Genetic Algorithm
PAES	Pareto Archived Evolution Strategy
PESA	Pareto Envelope-based Selection Algorithm
PF	Pareto Front

	Front de Pareto
PR	Path Relinking
	Recomposition de chemin
PSO	Particle Swarm Optimization
	Optimisation par essaims particulaires
SA	Simulated Annealing
	Recuit simulé
SI	Swarm Intelligence
	Intelligence distribuée
SPEA	Strength Pareto Evolutionary Algorithm
SS	Scatter Search
	Recherche dispersée
TS	Tabu Search
	Recherche Tabou

3.1 Introduction

Les problèmes d'optimisation combinatoire issus de problématiques réelles sont nombreux et variés selon les domaines d'applications. Dans un problème d'optimisation combinatoire, on définit un ensemble fini de solutions discrètes D et une fonction objectif f associant à chaque solution une valeur souvent réelle. Le problème consiste en l'optimisation (minimisation ou maximisation) d'un certain critère sous différentes contraintes permettant de délimiter l'ensemble des solutions réalisables. La résolution d'un problème d'optimisation nécessite la prise en compte des points suivants :

- Définir l'ensemble des solutions réalisables en prenant en compte l'ensemble des contraintes du problème.
- Choisir l'objectif à optimiser est déterminant pour le choix de la méthode à utiliser. Cela nécessite une bonne connaissance du problème.
- Choisir la méthode d'optimisation selon la nature et la complexité du problème. Si le problème est de petite taille, un algorithme exact peut être suffisant. Dans le cas de problèmes NP difficiles, on fait appel à des heuristiques fournissant des solutions approchées.

Dans les situations réelles, les décideurs font face à des situations de décision où plusieurs objectifs sont présents simultanément. Les premières études sont basées sur la transformation du problème multi-objectifs en une suite de problèmes mono-objectifs. Une autre approche consistait à faire une agrégation linéaire des objectifs en affectant un poids à chaque objectif. Cependant, il n'est pas toujours possible de déterminer un degré d'importance des objectifs. On cherche alors une solution qui soit le meilleur compromis entre ces objectifs. Dans ce cas, la solution recherchée n'est pas unique mais un ensemble de solutions représentant les compromis possibles. Plusieurs situations réelles sont représentées comme des problèmes d'optimisation multi-objectifs (MOP). Ces problèmes sont souvent caractérisés par leur grande taille et la présence de multiples objectifs contradictoires. En général, la tâche principale dans le processus d'optimisation multi-objectifs consiste à identifier l'ensemble des solutions Pareto optimales, ou obtenir une bonne approximation vis-à-vis du front de Pareto (PF). Plusieurs métaheuristiques ont été introduites durant les trente dernières années [17], comme les algorithmes évolutionnaires (EA), les stratégies d´évolution (ES), le recuit simulé (SA), la recherche Tabou (TS), la Scatter Search (SS), l'optimisation par essaim particulaires (PSO) et l´évolution différentielle (DE) [6].

Les algorithmes évolutionnistes multi-objectifs (MOEAs) représentent des domaines de recherche très prometteurs aujourd'hui. En fait, ces algorithmes permettent de fournir plusieurs avantages pour résoudre les problèmes d'optimisation difficiles. Plusieurs travaux sur la résolution des problèmes MOPs et leurs applications utilisant les algorithmes évolutionnistes sont proposés dans la littérature [23, 33, 61, 120, 118]. Les approches de NSGAII [23] et de SEPA2 [120] sont les approches Pareto les plus populaires à base de MOEAs. Basée sur plusieurs méthodes traditionnelles de programmation mathématiques [81], l'approximation du PF peut être décomposée en plusieurs sous-problèmes mono-objectifs. D'ailleurs, plusieurs approches de MOEAs adoptent ce principe, comme dans les approches de MOGLS [60] et de MOEA/D [115]. Plusieurs algorithmes de recherche visent à obtenir le meilleur de l'ensemble des différentes métaheuristiques qui s'exécutent ensemble, qui sont complémentaires et qui augmentent leurs capacités d'exploration. Ces méthodes sont généralement appelées des méthaheuristiques hybrides. La diversification et l'intensification [6] sont deux éléments majeurs lors de la conception d'une méthode de recherche globale. La diversification signifie la capacité de visiter plusieurs régions de l'espace de recherche, tandis que l'intensification signifie la capacité d'obtenir de bonnes solutions pour ces régions. Un algorithme de recherche permet de satisfaire ces deux éléments, pour faire face aux conflits qui peuvent exister. Les métaheuristiques hybrides permettent de contrôler cet équilibre [77].

Nous proposons ici de nouvelles métaheuristiques hybrides pour faire face aux problèmes combinatoires multi-objectifs et aux problèmes d'optimisation continue multi-objectifs. Parmi les approches de métaheuristiques hybrides proposées, nous citons HEMH, MOEAD/D, HEMH2 pour le domaine de recherche discret et HESSA pour le domaine continu. Ces approches sont abordées dans la section 3.3. Les tests réalisés montrent que ces nouvelles approches permettent d'obtenir de meilleurs résultats vis-à-vis des approches existantes.

3.2 Problématique et état de l'art de l'optimisation multi-objectifs

Dans l'optimisation multi-objectifs, il n'y a généralement pas de solution optimale qui satisfasse tous les objectifs à la fois car ces derniers sont contradictoires. Il faut donc trouver des solutions de compromis, non dominées par d'autres solutions, appelées solutions Pareto optimales [12]. Sans perte de généralité, nous ne considérerons dans les définitions qui suivent, que les problèmes de minimisation. Un problème d'optimisation multi-objectifs (MOP) est Formulé comme suit :

$$\begin{aligned} Optimiser \quad & F(x) = \left(f_1(x), f_2(x), \ldots . f_n(x)\right) \\ sous \quad & x = (x_1, x_2, \ldots . x_n) \in D \\ & g_j(x) \leq 0, \forall\, j = 1, 2, \ldots . k \end{aligned} \qquad (1)$$

Où F est un vecteur de n objectifs, x est le vecteur représentant les variables de décision, D représente l'ensemble des solutions réalisables, $g_j(x)$ est la jème contrainte.

Contrairement à l'optimisation mono-objectif, la solution d'un problème multi-objectif n'est pas unique. C'est un ensemble de solutions non dominées, connu comme l'ensemble des solutions Pareto Optimales.

Définition 1

On dit qu'une solution $y = (y_1, y_2, \ldots . y_n)$ domine une solution $z = (z_1, z_2, \ldots . z_n)$ si et seulement si $f_i(y) \leq f_i(z) \quad \forall\, i = 1, \ldots n$.

Définition 2

Une solution réalisable $x^* \in D$ est pareto optimale (non dominée, efficace) si et seulement s'il n'existe pas de solution $x \in D$ telle que x domine x^*.

Définition 3

L'ensemble des solutions Pareto optimales (P^*) est l'ensemble des solutions de toutes les solutions efficaces $P^* = \{x \in \Omega: \nexists y \in \Omega \text{ tel que } F(y) \leq F(x)\}$.

Définition 4

Le front de Pareto (PF) est l'image de l'ensemble Pareto Optimal P^* dans l'espace objectif :

$$PF = \{F(x) = (f_1(x), f_2(x), \dots, f_m(x)) : x \in P^*\}$$

Définition 5:

Etant donné un point de référence r^*et un vecteur de poids $\Lambda = [\lambda_1, \dots, \lambda_m]$ tel que $\lambda_i \geq 0, \forall i \in \{1, \dots, m\}, \sum_{i=1}^{m} \lambda_i = 1$,

Les fonctions somme pondérée (F^{ws}) et Tchebycheff pondéré (F^{Tc}) correspondant à (1) sont définies respectivement par :

$$Max \ F^{ws}(x, \Lambda) = \sum_{i=1}^{m} \lambda_i \, f_i(x) \qquad (2)$$

$$F^{Tc}(x, r^*, \Lambda) = \underset{1 \leq i \leq m}{Max} \{\lambda_i (r_i^* - f_i(x))\} \qquad (3)$$

Soit un ensemble de m sacs à dos et un ensemble de n objets, le problème du sac-à-dos multiobjectif (MOKP) peut être formulé comme suit :

$$Max \ f_i(x) = \sum_{j=1}^{n} c_{ij} \, x_j, \forall i \in \{1, \dots, m\} \qquad (4)$$

$$s.t.: \ \sum_{j=1}^{n} w_{ij} \, x_j \leq W_i \, , \forall i \in \{1, \dots, m\} \qquad (5)$$

$$x = (x_1, \dots, x_n)^T \in \{0,1\}^n$$

où, $c_{ij} \geq 0$ est le gain du $j^{ème}$ objet dans le $i^{ème}$sac à dos, $w_{ij} \geq 0$ est le poids du $j^{ème}$ objet dans le $i^{ème}$sac à dos et W_i est la capacité du $i^{ème}$sac à dos. Lorsque $x_j = 1$, cela signifie que le $j^{ème}$ objet est sélectionné et mis dans le sac à dos.

On en déduit que toute solution de l'ensemble de Pareto peut être considérée comme optimale. En effet, aucune amélioration de la solution ne peut être faite sur un objectif sans dégrader la valeur d'un autre objectif. L'ensemble de ces solutions forme le front de Pareto.

Remarque

On peut classer les méthodes d'optimisation multi-objectifs en trois catégories, selon le moment où le décideur intervient dans le processus de décision. Dans les méthodes a priori, le compromis à faire entre les objectifs est défini au préalable. Dans les méthodes progressives, le décideur intervient et oriente le processus de recherche de solutions pendant le déroulement du processus. Enfin, dans les méthodes a posteriori, le décideur sélectionne une solution appropriée parmi dans l'ensemble des bonnes solutions obtenues.

3.2.1 Approches de résolution

Les méthodes exactes consistent généralement à une extension des approches monocritères : programmation dynamique (plus court chemin, sac à dos), ou à des approches de séparation et évaluation (Branch and Bound). Les méthodes classiques pour générer les solutions Pareto optimales

agrègent les fonctions objectif en une seule fonction paramétrée par analogie avec la prise de décision avant la recherche. Plusieurs cycles d'optimisation avec différents paramétrages sont effectués afin de parvenir à un ensemble de solutions qui approche l'ensemble optimal de Pareto. Les principales méthodes pour résoudre le problème de type MODM (MultiObjective Decision Making) sont des techniques de Scalarization [114, 81, 38, 81], des approches interactives [99, 8], Goal Programming [15, 13, 56, 54 75, 62] et Fuzzy programming [100]. Ces méthodes classiques présentent quelques inconvénients :

- Les fonctions objectif et/ou les contraintes du problème doivent satisfaire certaines hypothèses telles que la dérivabilité, la continuité, etc.
- Elles ne donnent qu'une bonne solution à la fois.
- Elles nécessitent un temps de calcul long exponentiellement proportionnel à la taille du problème pour atteindre un ensemble de bonnes solutions.
- Elles ne peuvent pas trouver toutes les solutions Pareto optimales lorsqu'il s'agit de problèmes particuliers avec les fronts de Pareto non convexes.
- la plupart de ces méthodes nécessitent une connaissance préalable, par exemple sur les poids appropriés ou valeurs ε [24].

Pour pallier à ces limitations, les chercheurs ont trouvé que les métaheuristiques représentent un outil prometteur qui peut pallier aux inconvénients des méthodes MODM conventionnels. Ces techniques ont la capacité de trouver un ensemble de solutions optimales à chaque exécution de la simulation. Les métaheuristiques représentent un outil prometteur qui peut pallier aux inconvénients des méthodes classiques. Elles ont la capacité de trouver un ensemble de solutions optimales à chaque exécution de la simulation.

3.2.2 Les méthodes heuristiques

La recherche globale et les techniques d'optimisation peuvent être classées en deux catégories de base : déterministes et probabilistes. Les algorithmes déterministes sont le plus souvent utilisées dans le cas de problème de dimension raisonnable [98, 3, 54, 45, 85]. Si la dimension de l'espace de recherche est très élevée, il devient plus difficile de résoudre un problème de manière déterministe. En outre, de nombreux problèmes d'optimisation multi-objectifs sont de grande dimension, discontinus, multimodaux, et/ou NPcomplets ; ils sont appelés irréguliers [107]. Par conséquent, les algorithmes probabilistes (stochastiques) entrent en jeu comme pour produire des solutions de bonne qualité (quasi-optimales) dans un délai raisonnable [97, 4, 9]. Les techniques de recherche probabilistes comprennent au moins une fonction à base de nombres aléatoires [51, 45 ,79]. Les méthodes stochastiques nécessitent une fonction qui attribue une valeur d'ajustement pour chaque solution possible, et un mécanisme de mise en correspondance entre le problème et les domaines de l'algorithme. Bien que certains d'entre eux puissent trouver l'optimum, la plupart ne peuvent pas garantir cette solution optimale [45]. Dans les algorithmes d'optimisation globale, les heuristiques permettent de décider quel ensemble de solutions possibles doit être examinée par la suite. Les heuristiques sont habituellement utilisées dans les algorithmes déterministes pour définir l'ordre de traitement des solutions candidates, comme cela se fait dans une méthode gloutonne. Alors que les méthodes probabilistes ne peuvent tenir compte de ces éléments de l'espace de recherche qui ont été sélectionnés par l'heuristique dans d'autres calculs.

Une heuristique [80] peut être définie comme une "technique qui vise les bonnes solutions (quasi-optimales) à un coût de calcul raisonnable sans être en mesure de garantir la faisabilité ou l'optimalité, voire dans de nombreux cas à indiquer comment approcher de l'optimalité une solution particulière réalisable" [93]. Au cours des trois dernières décennies, de nouveaux algorithmes

heuristiques avancées communément appelés “Metaheuristiques” ont été largement développés et appliqués à une variété de problèmes d’optimisation [93, 111, 44, 1, 87, 92]. Le terme “Métaheuristique” est d’abord introduit par Glover dans [40]. Il peut être décrit comme une stratégie de recherche itérative qui guide le processus en dehors de l’espace de recherche dans l’espoir de trouver la solution optimale.

Selon Voss et al. [111] une métaheuristique est décrite comme “un processus itératif maître qui guide et modifie les opérations d’heuristiques subordonnées à produire efficacement des solutions de grande qualité. Il peut manipuler une solution unique complète (ou partielle) ou un ensemble de solutions à chaque itération. Les heuristiques subordonnées peuvent être de niveau élevé ou faible, comme elle peuvent se résumer à une recherche locale simple, ou tout simplement une méthode de construction”. Les métaheuristiques représentent une nouvelle classe d’algorithmes approximatifs qui tentent de combiner des méthodes heuristiques de base et des méthodes de haut niveau afin d’explorer efficacement l’espace de recherche. Comme les algorithmes approximatifs, les métaheuristiques sacrifient la garantie de trouver des solutions optimales dans le but d’obtenir de « bonnes solutions » en un temps réduit de façon significative. Nous nous référons à [107, 6] pour plusieurs autres définitions proposées. Le succès des métaheuristiques repose sur la fourniture d’un équilibre dynamique et adaptatif entre l’exploitation (intensification) des expériences accumulées de recherche et l’exploration (diversification) de l’espace de recherche pour identifier de nouvelles régions. L’intensification permet de concentrer la recherche dans certaines régions de l’espace, tandis que la diversification permet l’expansion de la recherche en explorant les régions non visités de l’espace. Les mécanismes d’intensification et de diversification sont des éléments fondamentaux de toute méthode de recherche globale. Les métaheuristiques peuvent être classées selon différents aspects qui sont généralement liés à la façon dont elles fonctionnent dans l’espace de recherche. Elles peuvent être classées comme “solution unique” versus “population” [7], “déterministe” versus “stochastique”, “inspiré de la nature” versus “non-inspiré de la nature” [6], “avec mémoire” versus “sans mémoire” [104], etc. Pour les métaheuristiques n’utilisant qu’une solution, le processus de recherche commence par l’amélioration d’une solution initiale unique qui se déplace de manière itérative comme trajectoire dans l’espace de recherche [20]. Les algorithmes à base de solution unique comprennent une heuristique constructive comme les heuristiques gloutonnes [27] et GRASP [44, 34, 35], la recherche locale simple [1] et ses extensions intelligentes (qui améliorent ses capacités d’échapper à l’optimum local tel que la recherche locale réitérée), la recherche à voisinage variable [80], la recherche locale guidée [113, 112], le recuit simulé [69,10] et la méthode tabou [40, 41]. Pour les métaheuristiques à base de population, un ensemble de solutions est adopté plutôt que de considérer une solution unique. Les algorithmes à base de populations les plus couramment utilisés sont les algorithmes évolutionnaires et les algorithmes basés sur l’intelligence distribuée (SI). Les algorithmes évolutionnaires simulent le processus d’évolution naturelle qui se base sur le concept darwinien de la “survie du plus fort”. Ils abordent les problèmes d’optimisation difficiles grâce à l’amélioration d’une population de solutions initiales en utilisant des opérateurs de sélection, de recombinaison et de mutation tels que des algorithmes génétiques (GA) [49], les stratégies évolutives (ES), l’évolution différentielle (DE) [89], la recherche de nuages [71], le Path Relinking [42, 39], les algorithmes mémétiques (MA) [83, 84], etc. Dans les algorithmes SI, l’idée est de produire l’intelligence artificielle en exploitant l’analogie avec l’interaction sociale, plutôt que les capacités cognitives purement individuelles, tels que l’optimisation par essaims particulaires (PSO) [26, 68] et l’optimisation par colonie de fourmis (ACO) [25].

La résolution de problèmes d'optimisation multi-objectifs à l'aide des métaheuristiques devient un champ de recherche très actif, en particulier les métaheuristiques à base de population, car ils peuvent explorer simultanément l'espace de recherche avec la réalisation de la convergence vers le vrai front de Pareto et la diversité uniforme. L'algorithme métaheuristique multi-objectifs comporte trois volets supplémentaires de recherche de base. Il s’agit de la stratégie de l’affectation de l'ajustement, de la préservation de la diversité et de l'élitisme. L’affectation de l'ajustement assigne une forme scalaire à

un vecteur de fonctions objectif afin de guider l'algorithme de recherche vers les solutions optimales de Pareto. Il existe quatre grandes approches dans les stratégies d'affectation de l'ajustement utilisés dans les métaheuristiques multicritères, les approches scalaires [108, 46], les approches fondées sur des critères [36], les approches à base du Pareto [45, 122] et les approches fondées sur des indicateurs [122]. La stratégie de préservation de la diversité contribue à générer un ensemble de solutions efficaces dans l'espace de l'objectif et/ou de décision. L'élitisme ou stratégie d'archivage est un mécanisme servant à maintenir les solutions de bonne qualité rencontrés pendant le processus de recherche. Ainsi, une convergence rapide peut être atteinte [118, 74, 48].

De nombreux chercheurs ont naturellement développé des algorithmes évolutionnaires pour résoudre des problèmes multi-objectifs (MOEAs). Les MOEAs peuvent être classés en quatre catégories. Les algorithmes fondés sur des critères sont marquées à l'aide de schéma de sélection en fonction de chaque objectif séparément comme dans (VEGA) [102]. Les algorithmes de Pareto utilisent un système de sélection basé sur le concept d'optimum de Pareto. Ils peuvent être divisés en deux générations. Dans la première génération [45], l'utilisation du partage d'ajustement et le partage nichage combiné avec classement Pareto est considérée. Elle contient des approches telles que NSGA [103], NPGA [50] et MOGA [37]. La deuxième génération est née avec l'introduction de la notion d'élitisme. Elle contient des approches telles que la SPEA [118], SPEA2 [120], PAES [70], NSGA-II [22], NPGA2 [30, 50, 86], PESA [18] et MOPSO [16, 50, 88, 85]. Les approches fondées sur des indicateurs dont la recherche est guidée par un indicateur de qualité de la performance [117]. Et les approches à base de décomposition où le MOP est divisé en série de N problèmes mono-objectif qui commencent simultanément par les algorithmes tels que MOGLS [60], MOEA/D [115] et dMOPSO [78]. Deux ou plusieurs algorithmes métaheuristiques peuvent être combinés pour développer une approche hybride mieux adaptée pour un problème donné [44, 91, 19, 106, 5, 28, 90]. La motivation principale du concept d'hybridation des métaheuristiques est d'obtenir de meilleurs systèmes performants qui exploitent et combinent les avantages des méthodes employées séparément. Le choix d'une combinaison adéquate de concepts multiples de métaheuristiques est la clé pour atteindre des performances optimales dans la résolution de nombreux problèmes d'optimisation difficiles.

3.2.3 Analyse de la performance

L'absence d'ordre total et l'existence de plusieurs solutions optimales rendent la mesure de la qualité d'un front difficile. En effet, la comparaison d'un ensemble de solutions est difficile. De nombreux indicateur de performances sont proposés dans la littérature. Des travaux ont été présentés [123], [47,69, 116]. Dans nos travaux deux types d'indicateurs d'évaluation de la performance sont utilisés. Le premier concerne les concerne les indicateurs binaires qui sont utilisés pour comparer chaque couple de techniques telles que la couverture d'ensemble (IC) [118]. Le deuxième type concerne les indicateurs unaires qui sont utilisés pour évaluer chaque technique indépendamment des autres, comme les indicateurs : Hypervolume (IHyp) [118], Generational Distance (IGD), Inverted Generational Distance (IIGD), R3-indicator (IR3) [70], Maximum Spread (IMS) [119] et l'indicateur Unary Additive Epsilon (Iε+) [121]. Il n'existe pas de méthode d'analyse universelle pour mesurer la performance en optimisation multi-objectifs. Pour une meilleure analyse, il est important d'utiliser différentes mesures de convergence et de diversité.

3.2.4 GRASP et DMGRASP

GRASP [34] est une métaheuristique ayant un processus itératif à deux phases. Dans la première phase, il s'agit de construire une solution complète. Dans la seconde phase, la recherche locale est

appliquée à cette solution pour garantir un optimum local. Ce processus est répété jusqu'à ce que le critère d'arrêt soit satisfait. On retient comme résultat la meilleure solution trouvée. Pour plus de détails, on peut consulter [67]. Dans GRASP, les itérations sont effectuées indépendamment les unes des autres. Par conséquent, les connaissances acquises dans les itérations passées ne sont pas exploitées dans les itérations suivantes. Le concept de base de l'intégration du datamining dans GRASP est que les motifs trouvés dans les meilleures solutions (obtenues dans les itérations précédentes) puissent être utilisés pour améliorer le processus de recherche. Ceci conduit à une exploration plus efficace de l'espace de recherche, et par conséquent, un comportement coopératif est atteint au lieu de construire chaque solution de manière indépendante. L'heuristique qui en résulte est le DM-GRASP [94] qui comporte deux phases [99]. La première consiste à générer un ensemble d'élites D par l'exécution de n itérations du GRASP pur et de sélectionner les meilleures solutions trouvées. Ensuite, le datamining est appliquée sur D pour extraire l'ensemble de motifs P. Enfin, la phase hybride est effectuée par l'exécution d'un certain nombre d'itérations. Dans ces itérations, la construction reçoit un motif p ∈ P comme une solution partielle à partir de laquelle une solution complète sera construite.

Le Path-Relinking

Le Path-Relinking (ou recomposition de chemin) a été suggéré pour intégrer les stratégies d'intensification et de diversification dans le cadre de Tabou Search (TS) et Scatter Search (SS) [43]. Il génère de nouvelles solutions en explorant les trajectoires qui relient des solutions de qualité. A partir de la solution de départ (x^s), le Path-Relinking génère un chemin dans l'espace de voisinage qui mène vers la solution de guidage (x^t). Ceci peut être accompli grâce à la sélection des mouvements qu'introduisent les attributs contenus dans (x^t) et de les intégrer dans une solution intermédiaire originaire de (x^s). On a observé que les meilleures solutions sont trouvées lorsque la procédure de recomposition commence à partir de la meilleure des (x^s) et (x^t). L'utilisation du Path-Relinking dans GRASP comme une stratégie d'intensification appliquée à chaque solution localement optimale, a été proposé initialement par [70]. Elle a été suivie par plusieurs extensions et applications [92] [93].

3.2.5 Le MOEA/D

La méthode MOEA/D [113] est un Algorithme Evolutionnaire Multi-Objectifs (MOEA : Multiobjective Evolutionary Algorithm) développée récemment où l'idée de décomposition est appliquée à la place de la relation de dominance. Le MOEA/D peut être expliqué comme un MOEA cellulaire [115] avec une structure de voisinage dans l'espace de poids de dimension m. Une cellule unique avec un seul individu se trouve à la même position que chaque vecteur de pondération dans l'espace des poids à m dimensions. Autrement dit, chaque cellule possède son propre vecteur de pondération, qui est utilisée dans la fonction de scalarisation pour évaluer l'individu dans cette cellule. Les voisins d'une cellule sont définis par la distance euclidienne entre les cellules dans l'espace des poids. Les solutions efficaces obtenues au cours du processus de recherche sont conservées dans une archive externe. Pour générer une descendance d'une cellule, deux parents sont sélectionnés au hasard à partir de ses voisins pour effectuer la reproduction. La descendance est comparée à l'individu dans la cellule courante en utilisant la fonction de scalarisation. Si la descendance est meilleure, il remplace l'individu actuel. La descendance est également comparée avec chaque voisin. La fonction de scalarisation avec le vecteur de poids de chaque voisin est utilisée dans la comparaison. Tous les voisins, qui sont inférieurs à la descendance, sont remplacés par les descendants. Ce cadre sera utilisé par HEMH pour réaliser l'hybridation proposée avec DM-GRASP et la recomposition de chemin randomisé (Randomized Path-Relinking) pour améliorer les performances et améliorer les capacités.

3.3 Contributions

3.3.1 Métaheuristiques évolutionnaires hybrides (HEMH)

Dans cette étude, nous proposons une nouvelle approche hybride appelée HEMH [65]. Dans le cadre MOEA/D, la méthode HEMH intègre à la fois le DM-GRASP [94] et la Recomposition de Chemin (Path Relinking). Cela permet de tirer profit des avantages de ces techniques à des fins de coopération, d'intégration et d'équilibre adéquat entre l'intensification et la diversification, pour améliorer les capacités de recherche. L'utilisation de HEMH est motivée par les arguments suivants :

• L'utilisation Data Mining pour extraire les bons motifs, qui seront réutilisés pour construire de nouvelles solutions, permettra d'accomplir la coopération entre les itérations du GRASP.

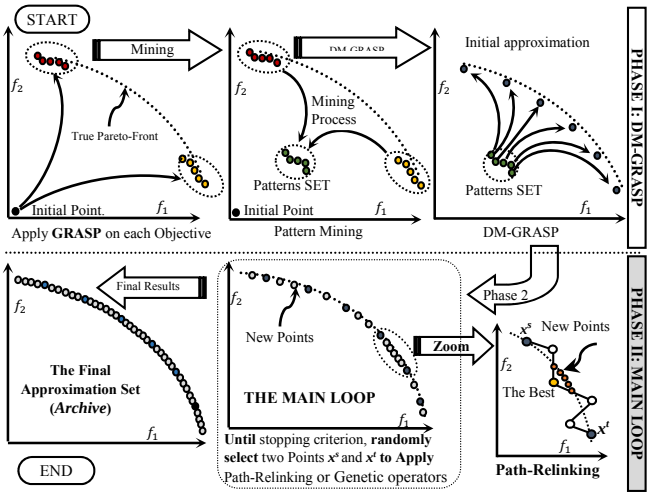
• La reproduction sur des solutions de grande qualité conduit souvent à produire une descendance de meilleure qualité.

• L'intégration de la recomposition de chemin aidera à trouver des solutions au-delà des points d'élite comme une stratégie de post optimisation et donc augmentera l'intensification dans ces régions.

• La recomposition de chemin donne la possibilité d'explorer les régions non-convexes pour découvrir des solutions prometteuses.

Cadre de travail

La méthode HEMH utilise une technique de décomposition (approche basée sur la somme pondérée) pour convertir MOKP formulée dans l'équation (5) par N problèmes mono-objectifs en utilisant un ensemble de N vecteurs poids uniformément répartis $\{\Lambda^1, \dots, \Lambda^N\}$. HEMH vise à optimiser simultanément ces N sous-problèmes. L'ensemble des voisins du i^{ème} sous-problème inclut tous les sous-problèmes avec les T vecteurs de poids $\{\Lambda^{i1}, \dots, \Lambda^{iT}\}$, les plus proches (en terme de distance euclidienne) du vecteur de poids Λ^i en cours.



Mécanisme du HEMH

Dans la méthode HEMH, il y a deux populations : la population principale et l'archive. La population principale se compose de N membres dans lesquels une solution est maintenue dans chaque direction de recherche. Chaque sous-problème possède T voisins. L'Archive est utilisée pour collecter toutes les solutions efficaces explorées au cours de l'ensemble du processus de recherche. Elle est périodiquement mise à jour par les nouvelles solutions explorées en ajoutant des solutions non dominées et en enlevant les dominées. Dans la méthode HEMH, le processus de recherche est divisé en deux phases de base : "l'initialisation" et "La boucle principale.

Algorithme 1 - HEMH (*critère d'arrêt*,$N, W_V, T, t, \delta, \alpha, \beta, \sigma, \varepsilon$)
Inputs:
N: *Taille de la population ou nombre de sous problèmes retenu.*
$W_V = \{\Lambda^1, \dots, \Lambda^N\}$: *Ensemble de N vecteurs de poids.*
T: *Nombre de voisins pour chaque vecteur de poids.*
$t \leq T$: *Nombre de voisins en compétition.*
$\delta \in [0,1]$: *Probabilité de sélection de parents à partir du voisinage*
$\alpha \in [0,1]$: *paramètre utilisé lors du processus de construction.*
$\beta \in [0,1]$: *Paramètre utilisé dans le processus de recherche locale.*
σ: *Ensemble de supports minimum pour le pattern-mining*
ε: *Distance de Hamming minimale pour l'application de la recomposition de chemin*
Sortie : $Archive$: *toutes les solutions efficaces trouvées au cours des générations.*
Début: *// Phase d'initialisation*
01. **pour** $i \in \{1, \dots, N\}$ **faire :** *//Définir un ensemble de T voisins pour chaque* Λ^i
02. $Neighbors^i \leftarrow \{i1, \dots, iT\} : \Lambda^{i1}, \dots, \Lambda^{iT}$ sont les T plus proches de Λ^i
03. **Fin-Pour**
04. Soit $\{\Lambda_{f_1}, \dots, \Lambda_{f_m}\} \subseteq W_V$ l'ensemble des vecteurs poids extrêmes.
05. $Archive \leftarrow \emptyset$;
06. **Pour** $i \in \{1, \dots, m\}$ **faire :** *//Exécuter le GRASP séparément pour chaque objectif*
07. $sol \leftarrow \emptyset$;
08. $sol \leftarrow$**Construction** $(sol, \alpha, \Lambda_{f_i}, Archive)$;
09. $sol \leftarrow$**Local-Search** $(sol, \beta, \Lambda_{f_i}, Archive)$;
10. **Fin-Pour**
11. $\mathcal{P} \leftarrow$**PattarnMining**$(\sigma, Archive)$; *//Construit l'ensemble des motifs*
12. **Pour** $i \in \{1, \dots, N\}$ **faire :** *//Initialiser la population utilisant DM-GRASP*
13. Tirer au hasard p à partir de $\mathcal{P}$ *//Choisir un motif*
14. $x^i \leftarrow$**Construction** $(p, \alpha, \Lambda^i, Archive)$; *//Construire* x^i *en utilisant p.*
15. $x^i \leftarrow$**Local-Search** $(x^i, \beta, \Lambda^i, Archive)$; *//Améliorer* x^i.
16. $FV^i \leftarrow F(x^i)$; *// Evaluation de* x^i
17. **Fin-Pour**
18. **Tant que** *le critère d'arrêt n'est pas satisfait* **faire:** *//Phase de la boucle locale*
19. **Pour** $i \in \{1,2, \dots, N\}$ **faire:**
20. Générer aléatoirement $r \in [0,1]$;
21. **Si** ($r < \delta$) **alors**:
22. $Pop \leftarrow Neighbors^i$;
23. **sinon:** $Pop \leftarrow \{1, \dots, N\}$;
24. **Fin-Si**
25. Tirer au hasard *j* et k à partir *Pop*.
26. **Si** ($\Delta(x^j, x^k) < \varepsilon$) **alors**:
27. $y \leftarrow$**Reproduction** (x^j, x^k); *// Croisement et mutation*
28. $y \leftarrow$ **GreedyRepair** (y, Λ^i)**;**
29. **Sinon:**
30. $y \leftarrow$ **GRPathRelinking**$(x^j, x^k, \Lambda^i, \alpha, \beta, Archive)$**;**
31. **Fin-Si**
32. Comparer y avec t
33. **Update-Solutions** (y, t, Pop);
34. $Archive \leftarrow$ **Update-Archive**(y); *// mise à jour de l'Archive*
35. **Fin->Pour**
36. **Fin-Tant que**
37. **Retourner** $Archive$**;**

Dans la phase d'initialisation, le DM-GRASP génère une première série de solutions de bonne qualité pour obtenir la population principale. Dans un premier temps, le GRASP initial [44] est appliqué sur chaque fonction objectif séparément pour construire un ensemble de solutions « d'élite ». A partir de cet ensemble on extrait un ensemble de bons motifs en utilisant le data mining. Ensuite, pour chaque

sous-problème, l'un des motifs extraits est choisi comme une solution partielle pour construire la solution en cours. Dans la phase de la « boucle principale », on applique la recomposition de chemin aléatoire randomisé ou l'opérateur de reproduction classique sur les solutions précédemment obtenues dans la phase d'initialisation jusqu'à ce que le critère d'arrêt soit atteint. Ceci permet d'intensifier le processus de recherche dans les régions entourant le front de Pareto et à concentrer les efforts de recherche sur les régions prometteuses pour découvrir de nouvelles solutions de bonne qualité. Pour plus de détails sur la méthode HEMH on peut consulter [61].

Expérimentations

Pour vérifier les performances de HEMH, une partie des méthodes MOEAs citées dans l'état de l'art est considérée (NSGAII [6], SPEA2 [118], GRASPM [108] et MOEA/D [113]). Les exemples de tests énumérés le tableau 1 sont couramment utilisés dans la littérature. Les différents paramètres utilisés pour chaque MOEA sont discutés dans [63]. Les métriques d'évaluations utilisées sont la s-métrique, la GD-métrique, la IGD-métrique et MS-métrique. Ces métriques, ainsi que d'autres paramètres d'initialisation pour chaque méthode sont définies dans [61].

Instances			$N(H)$	HMEH $N(H)$	$MaxEvals$
Name	Knaps(m)	Items(n)			
KSP252	2	250	150(149)	75(74)	75000
KSP502	2	500	200(199)	100(99)	100000
KSP752	2	750	250(249)	125(124)	125000
KSP253	3	250	300(23)	153(16)	150000
KSP503	3	500	300(23)	153(16)	150000
KSP753	3	750	300(23)	153(16)	150000
KSP254	4	250	364(11)	165(8)	182000
KSP504	4	500	364(11)	165(8)	182000
KSP754	4	750	364(11)	165(8)	182000

Tableau 1: Jeu d'essai Knapsack

Les résultats de simulation sont présentés brièvement ici. Ils sont présentés en détail dans [67]. La figure 2 montre les résultats de la C-métrique. Elle contient des graphiques pour les différents couples de méthodes MOEAs utilisées (avec échelle 0 en bas et 1 en haut). Chaque graphique contient 9 boxplots (boîtes à moustaches) résumant la distribution des C-valeurs. Chaque boxplot représente un des neuf exemples du tableau 1. Il est clair d'après les résultats de la figure 2 que HEMH et GRASPM sont meilleures que les autres MOEAs. Il est également évident que la méthode HEMH fonctionne mieux ou un peu mieux que le GRASPM dans tous les exemples.

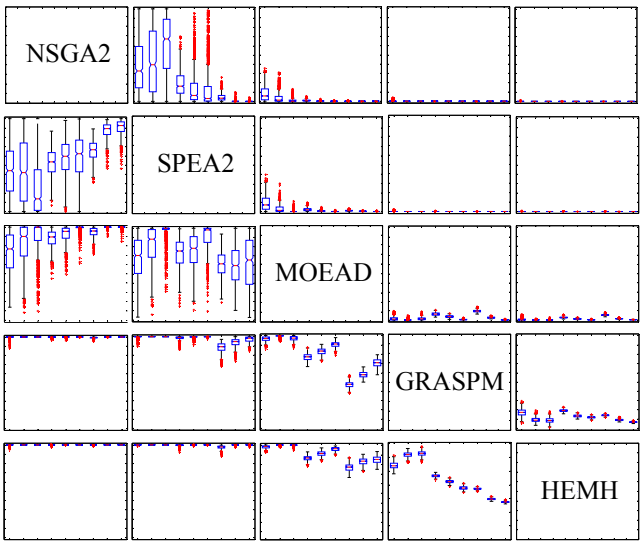


Figure 2: Comparaison des C-Métriques.

Les tableaux de 3 à 6 contiennent les valeurs calculées des différentes métriques. Ces valeurs sont visualisées dans les figures correspondantes (de 3 à 6). Le tableau 3 et la figure 3 représentent les résultats de la s-métrique. Ils contiennent les valeurs moyennes de l'indicateur après 30 essais indépendants. On peut constater que HEMH est meilleure que toutes les autres méthodes. Les s-métriques moyennes les plus élevées sont celles de la méthode HEMH. Cela signifie la capacité qu'elle a pour améliorer à la fois la convergence et la diversité. GRASPM et MOEAD ont respectivement le deuxième et le troisième rang dans tous les cas de tests.

Les valeurs moyennes de l'indicateur de distance générationnelle, sont listées dans le tableau 4 et on peut les visualiser dans la figure 4. Dans ce cas, HEMH surpasse toutes les méthodes MOEA. GRASPM atteint le deuxième rang, suivi de MOEAD qui prend la troisième place à l'égard de tous les cas de test. Cela signifie que HEMH a les capacités de la découverte des solutions les plus proches possible du Front de Pareto.

Instance	NSGAII	SPEA2	MOEAD	GRASPM	HEMH
KSP252	6.680E-01	6.576E-01	7.763E-01	7.948E-01	**7.976E-01**
KSP502	5.889E-01	5.842E-01	7.492E-01	7.710E-01	**7.757E-01**
KSP752	5.516E-01	5.469E-01	7.540E-01	7.702E-01	**7.751E-01**
KSP253	4.129E-01	3.994E-01	5.342E-01	5.538E-01	**5.580E-01**
KSP503	3.175E-01	3.070E-01	4.982E-01	5.247E-01	**5.308E-01**
KSP753	2.665E-01	2.599E-01	4.861E-01	5.211E-01	**5.270E-01**
KSP254	2.122E-01	2.094E-01	3.334E-01	3.502E-01	**3.553E-01**
KSP504	1.325E-01	1.498E-01	2.922E-01	3.235E-01	**3.306E-01**
KSP754	9.766E-02	1.145E-01	2.666E-01	3.124E-01	**3.216E-01**

Tableau 3 : s-métrique

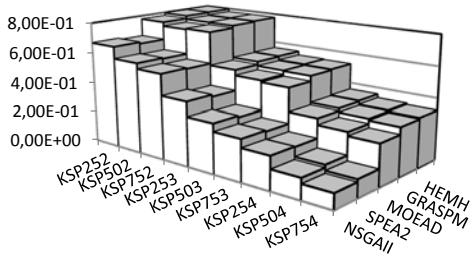


Figure 3 : s-métrique

Instance	NSGAII	SPEA2	MOEAD	GRASPM	HEMH
KSP252	3.240E-03	3.142E-03	1.457E-03	4.020E-04	**2.307E-04**
KSP502	4.424E-03	4.555E-03	1.458E-03	3.500E-04	**1.747E-04**
KSP752	4.171E-03	4.993E-03	1.009E-03	2.889E-04	**1.462E-04**
KSP253	1.622E-03	1.377E-03	4.457E-04	1.771E-04	**1.261E-04**
KSP503	2.369E-03	1.984E-03	4.468E-04	1.312E-04	**9.126E-05**
KSP753	3.345E-03	2.912E-03	4.760E-04	1.041E-04	**7.739E-05**
KSP254	1.538E-03	1.042E-03	2.849E-04	1.516E-04	**1.140E-04**
KSP504	2.571E-03	1.576E-03	3.203E-04	9.534E-05	**8.983E-05**
KSP754	3.173E-03	2.017E-03	4.009E-04	8.047E-05	**6.864E-05**

Tableau 4 : GD-métrique

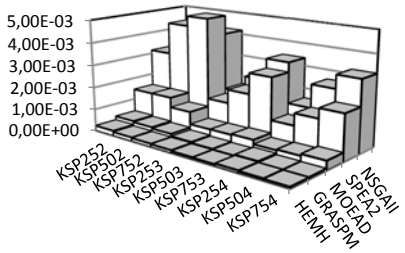


Figure 4 : GD-métrique

Instance	NSGAII	SPEA2	MOEAD	GRASPM	HEMH
KSP252	7.899E-03	8.608E-03	8.094E-04	3.468E-04	**3.161E-04**
KSP502	8.438E-03	8.595E-03	8.236E-04	2.467E-04	**1.717E-04**
KSP752	8.295E-03	8.126E-03	5.864E-04	2.055E-04	**1.378E-04**
KSP253	1.007E-03	1.153E-03	1.921E-04	9.910E-05	**8.606E-05**
KSP503	1.028E-03	1.143E-03	1.791E-04	9.015E-05	**7.263E-05**
KSP753	1.045E-03	1.124E-03	1.673E-04	8.099E-05	**6.300E-05**
KSP254	3.838E-04	4.127E-04	1.075E-04	8.232E-05	**7.264E-05**
KSP504	3.899E-04	3.968E-04	1.037E-04	7.686E-05	**6.203E-05**
KSP754	4.040E-04	4.081E-04	1.084E-04	7.057E-05	**5.611E-05**

Tableau 5 : distance générationnelle inversée IGD-métrique

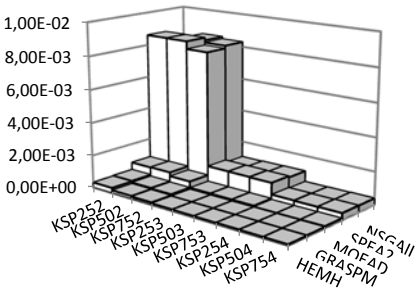


Figure 5 : distance générationnelle inversée IGD-métrique

Instance	NSGAII	SPEA2	MOEAD	GRASPM	HEMH
KSP252	5.168E-01	4.705E-01	1.373E+00	1.360E+00	**1.374E+00**
KSP502	3.788E-01	3.678E-01	1.309E+00	1.371E+00	**1.393E+00**
KSP752	2.598E-01	2.736E-01	1.317E+00	1.354E+00	**1.367E+00**
KSP253	8.916E-01	7.604E-01	1.650E+00	1.677E+00	**1.702E+00**
KSP503	6.653E-01	5.536E-01	1.653E+00	1.703E+00	**1.708E+00**
KSP753	4.758E-01	3.851E-01	1.644E+00	1.713E+00	**1.725E+00**
KSP254	1.234E+00	9.954E-01	1.903E+00	1.944E+00	**1.981E+00**
KSP504	1.066E+00	7.832E-01	1.902E+00	1.975E+00	**1.985E+00**
KSP754	8.273E-01	5.803E-01	1.838E+00	1.958E+00	**1.960E+00**

Tableau 6 : Maximum-spread métrique

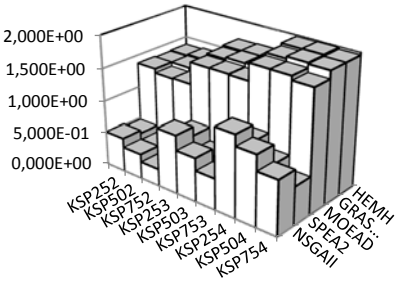


Figure 6 : Maximum-spread métrique

Les valeurs moyennes des résultats des comparaisons de la distance générationnelle inversée sont listées dans le tableau 5 et représentés dans la figure 5. Là aussi, on constate clairement que HEMH surpasse toutes les autres méthodes. Ce qui reflète sa capacité à obtenir des solutions avec une bonne propagation sur la frontière de Pareto. Les résultats indiquent également que le GRASPM atteint le deuxième rang, suivi du MOEAD qui prend le troisième rang.

Le tableau 6 et la figure 6 montrent les valeurs moyennes de l'indicateur de propagation maximum (Maximum Spread). Sur la base de ces résultats, HEMH a la supériorité sur les autres méthodes, suivi par GRASPM. Ceci montre leurs capacités à explorer les régions extrêmes dans l'espace de recherche, en raison de la recherche locale utilisée dans les deux, qui intensifie la recherche aux extrémités.

La figure 7 contient les ensembles d'approximations obtenues par chaque méthode qui sont visualisées pour les exemples bi-objectifs KSP502 et KSP752. Chaque sous-figure contient deux diagrammes de dispersion. Le grand diagramme représente l'ensemble des approximations tandis que le petit est un zoom sur la partie représentée par le petit rectangle. Dans les deux cas, KSP502 et KSP752, il est évident que les solutions obtenues par HEMH sont de meilleure qualité.

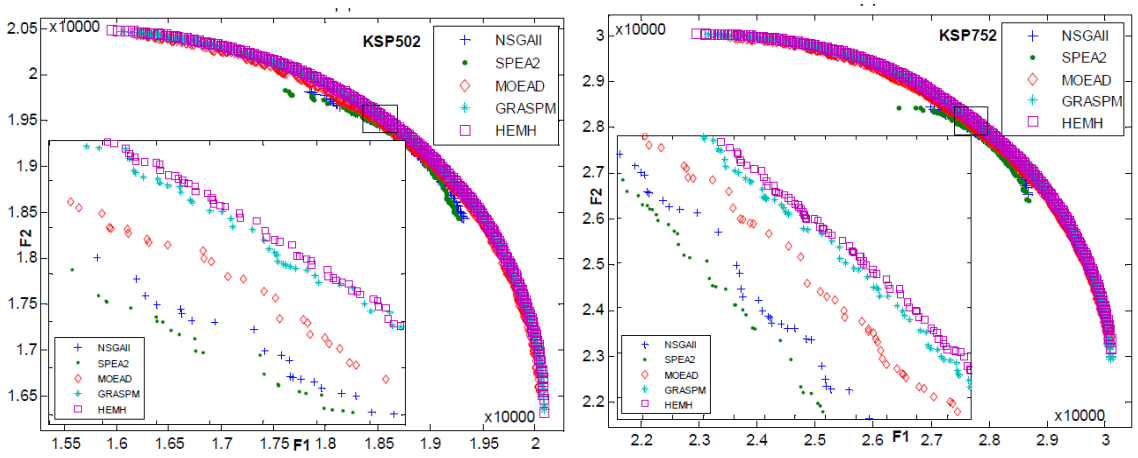


Figure 7 : Les ensembles d'approximation obtenus pour KSP502 et KSP752 pour chaque méthode en 30 exécutions

Conclusion

Dans ce travail, une métaheuristique hybride évolutionnaire (HEMH) basée sur le DM-GRASP et recomposition de chemin randomisé a été présentée et testée sur des jeux de données couramment utilisées dans la littérature. Le HEMH comparé à quatre autres méthodes (les plus populaires) à travers un ensemble d'indicateurs d'évaluation de la qualité. Les résultats montrent la supériorité de des MOEA basés sur la recherche locale et particulièrement le HEMH.

La principale contribution de notre méthode est la combinaison entre les différentes métaheuristiques techniques qui intensifient le processus de recherche dans la découverte des régions les plus prometteuses dans l'espace de recherche et améliorent la capacité d'explorer des solutions de bonne qualité. La seconde contribution est la capacité à trouver un bon ensemble de solutions de haute qualité à l'aide d'un petit ensemble de directions de recherche réparties uniformément en raison de l'utilisation de Path-Relinking et les stratégies de recherche locale.

3.3.2 Métaheuristiques hybride basées sur le MOEA/D

Dans cette partie, nous étudions l'effet de l'utilisation à la fois de l'évolution différentielle adaptative discrète proposée dans [114] et/ou du Path-Relinking glouton [32] comme opérateurs de reproduction à la place de la reproduction standard (croisement et mutation). La motivation principale de ce travail est d'obtenir les combinaisons appropriées d'opérateurs de recherche améliorant la performance des MOEA/D et particulièrement la méthode HEMH.

Nous proposons donc quatre variantes de l'algorithme, la première variante appelée MOEADde, dans laquelle l'évolution différentielle discrète adaptative remplace complètement les opérateurs de croisement et de mutation dans MOEA/D. La seconde variante est appelée MOEADpr dans laquelle l'opérateur de Path-Relinking glouton est appliqué, avec le croisement et mutation standard, après un certain nombre d'évaluations pour garantir l'existence de solutions de bonne qualité. Dans la troisième et la quatrième variante, l'évolution différentielle et la recomposition de chemin remplacent le croisement et la mutation, ils sont appelés respectivement MOEADdp1 et MOEADdp2.

Dans les méthodes MOEADde et MOEADpr présentés respectivement dans Algorithmes ci-dessous, on génère l'ensemble Λ de vecteurs uniformes de poids et on construit la structure de voisinage. La population initiale est également générée de façon aléatoire. Ensuite, la boucle principale est exécutée jusqu'à la réalisation des évaluations maximales. Afin de générer une nouvelle descendance pour chaque sous-problème i, la région de reproduction/mise-à-jour (M) est déterminée pour être à la fois le voisinage du ième sous-problème (localement), ou l'ensemble de la population (globalement) en fonction d'une certaine probabilité (δ). Cela peut donner une meilleure chance de sélection des parents distincts, ce qui encourage l'utilisation du Path-Relinking dans MOEADpr, ou permettre à l'évolution différentielle de fonctionner sur des individus distincts dans MOEADde. La sélection des parents est ensuite effectuée. Dans le cas de MOEADpr, les deux parents x^j et x^ksont choisis au hasard à partir de M. Ensuite, l'opérateur de Path-Relinking n'est utilisée que si la distance de Hamming entre les deux parents sélectionnés est supérieure à une certaine valeur ε, et le nombre d'évaluations Eval dépasse un certain ratio γ. Dans le cas contraire, l'opérateur de reproduction standard est appliqué pour générer la nouvelle descendance. Dans le cas de la variante MOEADde, trois individus parents distincts sont choisis au hasard pour appliquer évolution différentielle adaptative discrète. La nouvelle descendance générée est évaluée et utilisée pour mettre à jour le point de référence z. On met à jour également la population en fonction du paramètre t. Celui-ci permet de limiter le nombre de solutions remplacées. Enfin, l'ensemble des solutions efficaces mis dans la population finale est retourné en sortie. Dans les deux variantes MOEADdp1 et MOEADdp2, quelques modifications sont apportées au MOEADde pour intégrer le Path-Relinking après un certain nombre d'évaluations effectuées pour

assurer l'existence de solutions de bonne qualité. Ces modifications peuvent être opérées de la manière suivante : lorsque le nombre d'évaluations *Eval* dépasse une certaine valeur ($\gamma \times$ MaxEvals) préalablement déterminée pour utiliser le Path-Relinking, nous avons trois parents sélectionnés x^a, x^b et x^c dans l'étape de sélection. Si nous choisissons au hasard deux d'entre eux, qui ont distance de Hamming supérieure à une certaine valeur (ε) pour appliquer la recomposition de chemin au lieu de l'évolution différentielle, nous obtiendrons la variante MOEADdp1. D'autre part, en supposant que les conditions de distance de Hamming soient remplies, si nous appliquons le Path-Relinking sur les trois parents sélectionnés (x^a, x^b et x^c) de la manière suivante : en choisissant aléatoirement deux individus (x^a, x^c) pour appliquer la recomposition de chemin générant un nouvel individu y, ensuite en appliquant le Path-Relinking de nouveau sur Y et x^b nous obtenons la variante MOEADdp2.

Algorithmes

Algorithme 2 - MOEAD*pr* $(N, T, t, \delta, \varepsilon, \gamma)$

N: Nombre de sous-problèmes considérés.
T: Taille du voisinage pour chaque vecteur de poids.
$t \leq T$: Nombre maximal de solutions remplacées.
$\delta \in [0,1]$: Probabilité de sélection des parents à partir du voisnage
ε: Distance de Hamming minimale autorisée.
γ: Minimum d'évaluations autorisées pour appliquer le Path-Relinking

$\Lambda \leftarrow$ InitializeWeightVectors( );
$B \leftarrow$ InitializeNeighborhood( );
$P \leftarrow$ InitializePopulation( );
$z \leftarrow$ InitializeRefrencePoint( ); $NEval \leftarrow 0$;
Tant que $(NEval < MaxEvals)$ **faire :** *// Boucle principale*
  Pour $i \in \{1,2,\dots,N\}$ **faire :**
    $M \leftarrow \begin{cases} B(i) \text{ , } \textbf{Si } (rnd \in [0,1] < \delta) \\ P \qquad\qquad otherwise \end{cases}$
    $x^j, x^k \leftarrow$ Selection(M, i);
    $dist \leftarrow$ HamDist(x^j, x^k);
    Si $(dist \geq \varepsilon \wedge NEval \geq \gamma \times MaxEvals)$ **alors**:
      $y \leftarrow$ GreedyPathRelinking $(x^j, x^k, \Lambda^i,)$
    Sinon : $u \leftarrow$ Reproduction(x^j, x^k);
      $y \leftarrow$ GreedyRepair(u, Λ^i);
    Fin-Si
    FitnessEvaluation(y);
    $z \leftarrow$ UpdateRefrencePoint(y);
    $M \leftarrow$ UpdateSolutions(y, t);
    $NEval \leftarrow NEval + 1$;
  Fin-Pour
Fin Tant-que
Retourner P.

Algorithme 3 - MOEAD*de* $(N, T, t, \delta, F_0, CR_0, a_1, a_2)$

N : Nombre de sous-problèmes considérés.
T : Taille du voisinage pour chaque vecteur de poids.
$t \leq T$: Nombre maximal de solutions remplacées.
$\delta \in [0,1]$: Probabilité de sélection des parents à partir des voisins
$F_0, CR_0 \in [0,1]$: Facteur d'échelle et taux de croisement.

$\Lambda \leftarrow$ InitializeWeightVectors( ) ;
$B \leftarrow$ InitializeNeighborhoods( ) ;
$P \leftarrow$ InitializePopulation( ) ;
$z \leftarrow$ InitializeRefrencePoint( ) ; $NEval \leftarrow 0$;
Tant que $(NEval < MaxEvals)$ **faire :** *// Boucle principale*
  Pour $i \in \{1,2,\dots,N\}$ **faire :**
    $M \leftarrow \begin{cases} B(i) \text{ , } \textbf{Si } (rnd \in [0,1] < \delta) \\ P \qquad\qquad otherwise \end{cases}$
    $x^a, x^b, x^c \leftarrow$ Selection$(M, i) : x^i \neq x^a \neq x^b \neq x^c$
    $u \leftarrow$ Diff_Evolution$(x_i, x^a, x^b, x^c, F_0, CR_0, a_1, a_2)$;
    $y \leftarrow$ GreedyRepair(u, Λ^i) ;
    FitnessEvaluation(y) ;
    $z \leftarrow$ UpdateRefrencePoint(y) ;
    $M \leftarrow$ UpdateSolutions(y, t) ;
    $NEval \leftarrow NEval + 1$;
  Fin Pour
Fin Tant que
Retourner P.

Algorithme 4 - $\text{MOEAD}_{dp1}(\text{N; T; t; } \delta; \epsilon; \gamma ; F_0, CR_0, a_1, a_2)$
Remplacer les lignes (10-11) dans l'algorithme 4 par les lignes suivantes :
1: $x^i, x^k \leftarrow$ RandomSelection(x^a, x^b, x^c);
2: Si ($H_{dist}(x^j, x^k) \geq \varepsilon$ *et* $Eval \geq \gamma\, MaxEvals$) alors
3: $y \leftarrow$ PathRelinking ($x^j, x^k, \Lambda^i,$)
4: Sinon
5: $u \leftarrow$ Diff_Evolution($x_i, x^a, x^b, x^c, F_0, CR_0, a_1, a_2$)
6: $y \leftarrow \text{Repair}(u, \Lambda^i)$
7: Fin si

Algorithme 5 - $\text{MOEAD}_{dp2}(\text{N; T; t; } \delta; \epsilon; \gamma ; F_0, CR_0, a_1, a_2)$
Remplacer les lignes (10-11) dans l'algorithme 4 par les lignes suivantes :
1: $x^a, x^c \leftarrow$ RandomSelection (x^a, x^b, x^c);
2: Si ($H_{dist}(x^a, x^c) \geq \varepsilon$ *et* $Eval \geq \gamma\, MaxEvals$) alors
3: $y \leftarrow$ PathRelinking ($x^j, x^k, \Lambda^i,$)
4 : Si ($H_{dist}(x^b, y) \geq \varepsilon$) alors
5 : $y \leftarrow$ PathRelinking ($x^j, x^k, \Lambda^i,$)
6 : Fin si
7: Sinon
8: $u \leftarrow$ Diff_Evolution($x_i, x^a, x^b, x^c, F_0, CR_0, a_1, a_2$)
9: $y \leftarrow \text{Repair}(u, \Lambda^i)$
10: Fin si

Résultats

Les exemples de tests utilisés sont ceux du tableau 1. Les différents paramètres utilisés pour chaque algorithme sont fixés dans [64]. On peut également y trouver les descriptifs des d'indicateurs d'évaluation de la qualité (Hypervolume, Genrational distance, Inverted genrational distance, et R3-Indicator). A partir des tableaux et des figures ci-dessous, on constate que les résultats expérimentaux montrent la supériorité de toutes les variantes hybrides proposées sur le MOEA/D initial et SPEA2. Dans le cas des tests bi-objectifs, nous avons constaté que le MOEADpr est meilleur, tandis que le MOEADde a une moins bonne performance. D'autre part, dans le cas des exemples avec trois ou quatre objectifs, la performance de l'évolution différentielle est améliorée. Par conséquent, toutes les variantes proposées permettent d'atteindre une meilleure performance. Ils ont un rendement moyen très compétitif par rapport à la MOEA/D et SPEA2 (en se basant sur les indicateurs d'évaluation utilisés dans cette étude). La conclusion que nous avons retenue est pour les exemples de tests MOKP bi-objectifs, l'opérateur de Path-Relinking est le mieux classé, suivi du croisement standard, de la mutation et de l'évolution différentielle. Cependant, dans les cas des tests MOKP avec trois ou quatre objectifs, l'évolution différentielle et la recomposition de chemin ont un meilleur rendement que le croisement standard et la mutation.

Inst.	Algorithm					
	SPEA2	MOEAD	MOEAD$_{de}$	MOEAD$_{pr}$	MOEAD$_{dp1}$	MOEAD$_{dp2}$
KS252	2.79E-03	1.39E-03	2.25E-03	**1.12E-03**	2.22E-03	2.33E-03
KS502	3.98E-03	1.53E-03	2.34E-03	**1.10E-03**	2.18E-03	2.36E-03
KS752	4.62E-03	1.28E-03	1.38E-03	**1.23E-03**	1.27E-03	1.24E-03
KS253	3.71E-03	1.88E-03	7.39E-04	9.77E-04	6.55E-04	**6.30E-04**
KS503	4.06E-03	2.13E-03	7.12E-04	1.24E-03	**5.93E-04**	6.08E-04
KS753	4.18E-03	2.10E-03	8.06E-04	1.05E-03	6.55E-04	**6.37E-04**
KS254	5.17E-03	2.37E-03	9.84E-04	1.33E-03	8.76E-04	**8.61E-04**
KS504	6.34E-03	3.18E-03	9.58E-04	1.02E-03	**5.98E-04**	6.17E-04
KS754	6.75E-03	3.69E-03	9.73E-04	1.39E-03	6.08E-04	**5.51E-04**

Tableau 8. distance générationnelle (***GD***-metric)

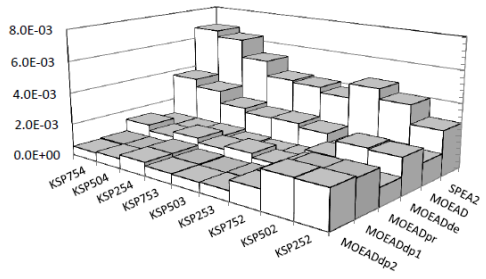


Fig. 9. distance générationnelle (***GD***-metric)

Inst.	Algorithm					
	SPEA2	MOEAD	MOEAD$_{de}$	MOEAD$_{pr}$	MOEAD$_{dp1}$	MOEAD$_{dp2}$
KS252	1.09E-02	1.00E-03	2.31E-03	**8.37E-04**	2.40E-03	2.25E-03
KS502	1.16E-02	9.72E-04	1.89E-03	**7.05E-04**	1.88E-03	1.89E-03
KS752	1.27E-02	7.96E-04	1.29E-03	**7.04E-04**	1.20E-03	1.16E-03
KS253	2.24E-03	6.11E-04	4.48E-04	4.99E-04	4.38E-04	**4.35E-04**
KS503	2.83E-03	6.12E-04	4.11E-04	4.82E-04	**3.99E-04**	4.01E-04
KS753	3.15E-03	5.93E-04	3.89E-04	4.27E-04	3.79E-04	**3.70E-04**
KS254	1.45E-03	6.80E-04	5.54E-04	6.07E-04	5.53E-04	**5.49E-04**
KS504	1.69E-03	6.51E-04	4.56E-04	4.82E-04	4.41E-04	**4.37E-04**
KS754	1.91E-03	6.88E-04	4.40E-04	4.81E-04	4.14E-04	**4.07E-04**

Tableau 9. distance générationnelle inverse (***IGD***-metric)

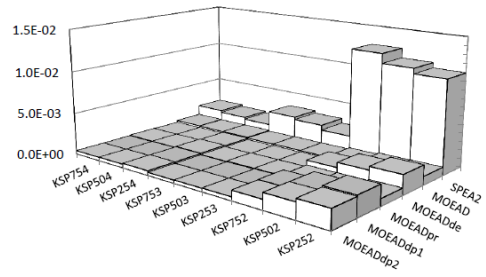


Fig. 10. distance générationnelle inverse (***IGD***-metric)

Inst.	Algorithm					
	SPEA2	MOEAD	MOEAD$_{de}$	MOEAD$_{pr}$	MOEAD$_{dp1}$	MOEAD$_{dp2}$
KS252	4.28E-02	5.27E-03	1.08E-02	**3.85E-03**	1.09E-02	1.12E-02
KS502	7.98E-02	6.66E-03	1.40E-02	**4.64E-03**	1.33E-02	1.40E-02
KS752	9.61E-02	6.46E-03	8.11E-03	**5.64E-03**	7.23E-03	7.16E-03
KS253	6.07E-02	1.09E-02	6.29E-03	6.65E-03	5.82E-03	**5.79E-03**
KS503	9.74E-02	1.32E-02	7.08E-03	7.26E-03	**5.91E-03**	6.27E-03
KS753	1.21E-01	1.45E-02	8.34E-03	**6.52E-03**	7.25E-03	7.06E-03
KS254	7.24E-02	1.55E-02	1.03E-02	1.11E-02	9.66E-03	**9.63E-03**
KS504	1.06E-01	2.05E-02	1.08E-02	9.09E-03	**8.93E-03**	9.00E-03
KS754	1.38E-01	2.54E-02	1.21E-02	1.02E-02	9.17E-03	**8.81E-03**

Tableau 10. Indicateur R_3 (I_{R3})

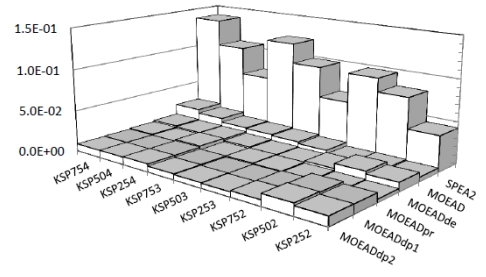


Fig. 11. Indicateur R_3 (I_{R3})

Conclusion

Dans les quatre variantes d'hybridation proposées, MOEADde utilise l'évolution différentielle discrète adaptative comme opérateur de recombinaison dans le cadre MOEA/D. le MOEADpr, utilise l'opérateur Path-Relinking avec les opérateurs de reproduction standard. Les deux dernières variantes MOEADdp1 et MOEADdp2 utilisent à la foi l'évolution différentielle et le Path-Relinking.

Ces quatre propositions ont été comparées avec MOEA/D et SPEA2 en utilisant un ensemble d'exemples couramment utilisés dans la littérature. Un ensemble d'indicateurs d'évaluation de la qualité a également été utilisé pour évaluer la performance de ces méthodes. Les résultats expérimentaux montrent la supériorité de toutes les variantes hybrides proposées sur MOEA/D et SPEA2 dans la plupart des tests. Dans le cas bi-objectif, nous avons constaté que MOEADpr est meilleur, alors que MOEADde donne de mauvaises performances. Cependant, dans le cas d'exemples avec trois ou quatre objectifs, la performance de l'évolution différentielle est améliorée. Par conséquent, toutes les variantes proposées permettent d'atteindre de meilleures performances. Pour les tests de MOKP bi-objectif, L'opérateur Path-Relinking occupe le premier rang, suivi par le croisement standard et la mutation et enfin l'évolution différentielle en dernier rang.

3.3.3 Métaheuristiques évolutionnaires hybrides améliorées (HEMH2)

Motivés par les résultats obtenus précédemment, ce travail est une extension de HEMH [61] basé sur l'élaboration d'une nouvelle version appelée HEMH2 avec deux autres variantes : HEMHde et HEMHpr. Les principales motivations de ce travail ont pour but de surmonter les limites des performances de HEMH. La méthode HEMH proposée précédemment présente quelques limitations :

- Malgré l'utilisation du DMGRASP qui réalise des solutions initiales de haute qualité, il consomme plus de temps et opère plus d'évaluations en particulier avec les populations de grande taille. Ainsi, dans la deuxième phase, les chances d'améliorer le processus de recherche seront réduites.
- La collecte de toutes les solutions efficaces représente un gaspillage de temps et d''espace de stockage. Contrôler le processus d'archivage devient alors indispensable.

- Le Path-Relinking adopte un « bit-fip » par coup et utilise la recherche locale pour améliorer la solution générée. Cela provoque un temps de calcul plus important.

Les principales différences entre HEMH2 et HEMH sont brièvement présentées ici :

• La population initiale est créée en utilisant le glouton inverse simple dans chaque direction de recherche plutôt que le DMGRASP. La qualité des solutions initiales obtenues sera affectée, mais cela nous donnera une meilleure chance lors de la seconde phase pour améliorer et renforcer le processus de recherche.

• Au lieu de recueillir toutes les solutions efficaces, l'epsilon-dominance du Pareto-adaptatif (paε-dominance) [48] est adoptée pour contrôler la quantité des solutions efficaces recueillies dans les archives.

• La taille du voisinage dynamique, qui permet de diminuer ou augmenter le voisinage pour chaque sous-problème, est prise en compte.

• L'évolution différentielle binaire adaptative est utilisée comme opérateur de reproduction au lieu du croisement/mutation avec le Path-Relinking.

• Le Path-Relinking est appliqué seulement après un certain nombre d'évaluations comme stratégie de post-optimisation. Cette action garantit l'existence de solutions de meilleure qualité, mais le Path-Relinking retourne deux bits à chaque étape.

• En HEMH2, la recherche locale est évitée soit après le Path-Relinking soit après la construction du glouton inverse tel que proposé dans HEMH.

Voici quelques arguments qui nous ont amenés à améliorer la méthode HEMH :

1. Il ne fait aucun doute que la génération de la population initiale à l'aide DMGRASP peut atteindre de meilleures solutions de qualité, mais elle nous contraint à utiliser des petites populations. Dans certains cas, la recherche locale nécessite beaucoup plus de temps et d'évaluations pour explorer une petite région spécifiée dans l'espace de recherche. Par conséquent, la phase de la boucle principale a une petite chance d'améliorer le processus de recherche. Pour contourner cette limitation, la construction du glouton inverse a été proposée. A partir des résultats empiriques, le glouton inverse donne des solutions aussi proches que possible des régions limitrophes qu'un glouton simple.

2. L'utilisation de paε-dominance va contrôler la taille de l'archive. Par conséquent, on économise davantage de ressources en temps et en espace de stockage tout en conservant la qualité des solutions recueillies.
3. Les faibles performances de la mutation différentielle binaire se produit lors du traitement des individus différentiels avec une grande distance de Hamming. Sélectionner les parents à partir de toute la population peut encourager ce scénario. En HEMH2, les parents de chaque sous-problème sont toujours choisis à partir de son voisinage ayant une taille dynamique. Ceci garantit l'obtention des individus avec des distances de Hamming appropriées.
4. L'Evolution Différentielle (DE) binaire adaptative a empiriquement la possibilité d'explorer l'espace de recherche mieux que le croisement et la mutation classique. Ainsi, la performance de HEMH sera améliorée par l'adoption du DE binaire adaptative pour la reproduction plutôt que le croisement.

Algorithme HEMH2

La méthode HEMH2 est détaillée dans l'algorithme 7. On crée d'abord un ensemble de N vecteurs de poids uniformément répartis. Ensuite, la structure de voisinage est construite pour chaque sous-problème i. Ces sous-problèmes sont triés par ordre croissant de la distance euclidienne entre les vecteurs de poids et le vecteur de poids courant Λ_i. Ensuite, la population initiale P est créée en appliquant le glouton inverse dans chaque direction de recherche. La boucle principale est exécutée jusqu'à la réalisation des évaluations maximales $Mevls$. Pour chaque sous-problème i, la routine de sélection est utilisée pour déterminer la taille actuelle de voisinage B_i tels que $|B_i| = T + r$, où T et R représentent respectivement le nombre de solutions différentes et répétés dans B_i. Cela signifie que la routine de sélection agrandit la taille de B_i pour garantir l'existence d'au moins T solutions différentes et choisit au hasard trois d'entre eux : x^a, x^b et x^c pour la reproduction. Deux des trois parents sélectionnés X_j et X_k sont choisis au hasard. Ensuite, le Path-Relinking n'est utilisé que si la distance de Hamming Δ (X_j , X_k) est supérieure à une certaine valeur ε et que le nombre d'évaluations $Eval$ dépasse un certain ratio γ des évaluations maximales $Mevls$ permettant de garantir l'application du Path-Relinking sur des solutions de bonne qualité. Dans le cas contraire, l'Evolution Différentielle binaire adaptative est appliquée pour générer une nouvelle descendance y. Cette nouvelle descendance générée est évaluée et utilisée pour mettre à jour le voisinage (B_i) en fonction du paramètre t, ce qui détermine le nombre des solutions remplacées. L'archive est également mise à jour par y selon paε-domination [48]. Enfin, les solutions extrêmes sont mises à jour dans l'archive et retournées en sortie.

Pour étudier les effets de l'Evolution Différentielle binaire adaptative et les opérateurs de Path-Relinking distinctement, deux variantes supplémentaires de l'algorithme appelées HEMHde et HEMHpr sont considérées. Toutes les deux ont la même procédure que HEMH2 à la seule différence que le HEMHde utilise l'Evolution Différentielle binaire adaptative pour la reproduction. Tandis que le HEMHpr remplace le l'Evolution Différentielle binaire adaptative dans la procédure HEMH2 par le croisement et la mutation.

Algorithme 7 - HEMH2(N; T; t ; ε ; γ ; CR_0; a)

Entrées:
N : Taille de la population ou nombre de sous-problèmes
T : Taille min du voisinage
t : Max des solutions remplacées
ε : Distance de Hamming minimale
γ : Contrôle l'exécution du Path-Relinking
$CR_0 \in [0; 1]$, a *//a: Taux de croisement, constant*

1: Début
2: $W_u \leftarrow \{\Lambda^1, \Lambda^2, \dots \Lambda^N\}$ *//Ensemble de N vecteurs poids*
3: Pour i = 1 à N faire *//Construire le voisinage*
4: $i \leftarrow \{i_1, i_2, \dots i_N\}$ *// Où $\Lambda^{i1}, \Lambda^{i2}, \dots \Lambda^{iN}$ sont triés par ED par ordre croissant*
5: Fin Pour
6: $Arch \leftarrow \emptyset$ *// Archive vide*
7: $Evl \leftarrow 0$
8: Pour i = 1 à N faire *//Phase d'initialisation*
9: $x^i =$ InverseGreedy(x^i ; Λ^i)
10: $P =$ AddSubProblem(x^i ; Λ^i)
11: Extremes = Update(x^i); Update(Evl);
12: Fin Pour
13: Tant que (Evl < Mevls) faire *// Boucle principale*
14: Pour i = 1 à N faire *// Pour chaque sous-problème i*
15: x^a; x^b; x^c = Selection(B_i, i)
16: // avec $x^i \neq x^a \neq x^b \neq x^c$
17: $x^i, x^k = RANDSELECTION(x^a, x^b, x^c)$
18: $D = \Delta(x^j, x^k)$ *// Distance de Hamming*
19: $E = \gamma\, Mevls$) *//min. eval pour PR*
20: Si ($D \geq \varepsilon \wedge Eval \geq E$) alors
21: $y = PathRelinking(x^j, x^k, \Lambda^i)$
22: Sinon
23: $u = ABDEvol(x^i, x^a, x^b, x^c, CR_0)$
24: $y = Reapir(u, \Lambda^i)$
25: Fin Si
26: $P = UpdateSolution(y, t, B_i)$
27: $Arch = UpdateArchive^{PA\varepsilon}(y)$;
28: $Extremes = Update(y); Update(Evl)$;
29: Fin Pour
30: Fin Tant que
31: Arch = AddExtremes(Extremes);
32 : Retourner Arch
32: Fin

Résultats

Comme dans la méthode HEMH, les métriques d'évaluations (I_C [118], *IRhyp* [118], *IGD*, *IIGD* et *IR*3 [70]), les méthodes MOEA et les différents paramètres utilisés sont détaillés dans [65]. Les valeurs moyennes des indicateurs IRH, IgD, IIGD et IR3 sont énumérées respectivement dans les tableaux 3, 4, 5 et 6 et représentées dans les figures 2, 3, 4 et 5. Chacune de ces tables contient les valeurs moyennes obtenues sur 30 essais indépendants pour chaque exemple de test et pour chaque algorithme. Sur la base de ces résultats, il est évident que les propositions HEMH2, $HEMH_{de}$ et $HEMH_{pr}$ sont meilleures que les MOEA/D et HEMH. D'autre part, HEMH2 et $HEMH_{de}$ sont meilleurs dans tous les cas de tests, suivis par $HEMH_{pr}$. Selon les résultats, nous pouvons déduire que l'Evolution Différentielle binaire adaptative, incluse dans HEMH2 et $HEMH_{de}$ a de meilleures capacités d'exploration qui pallient aux capacités de recherche locales contenues dans le HEMH originel. Dans certains cas, $HEMH_{de}$ peut atteindre des résultats très compétitifs par rapport à HEMH2 (en se basant sur l'Evolution Différentielle binaire adaptative).

Inst.	Algorithms				
	MOEA/D	HEMH	$HEMH_{de}$	$HEMH_{pr}$	HEMH2
KP252	4.66E-02	1.02E-02	**6.07E-03**	9.33E-03	6.08E-03
KP502	5.67E-02	1.55E-02	5.57E-03	1.16E-02	**5.56E-03**
KP752	4.73E-02	1.64E-02	**3.84E-03**	7.34E-03	3.94E-03
KP253	2.24E-01	1.21E-01	8.85E-02	1.09E-01	**8.77E-02**
KP503	2.76E-01	1.16E-01	**7.39E-02**	9.01E-02	**7.39E-02**
KP753	2.89E-01	8.85E-02	6.48E-02	7.63E-02	**6.39E-02**
KP254	8.56E-01	5.55E-01	**4.26E-01**	4.90E-01	4.27E-01
KP504	1.07E+00	4.53E-01	3.96E-01	4.10E-01	**3.84E-01**
KP754	1.23E+00	3.93E-01	3.54E-01	3.64E-01	**3.41E-01**

Tableau 11. Average Referenced Hypervolume (IRH)

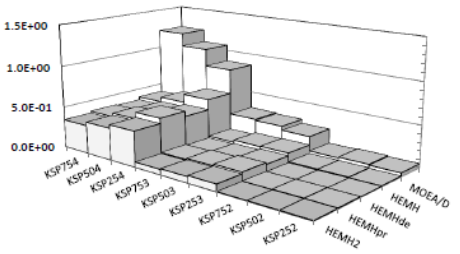


Fig. 12. IRH

Inst.	Algorithms				
	MOEA/D	HEMH	$HEMH_{de}$	$HEMH_{pr}$	HEMH2
KP252	1.50E-03	5.17E-04	3.40E-04	5.74E-04	**3.37E-04**
KP502	1.53E-03	4.41E-04	**1.44E-04**	3.12E-04	1.54E-04
KP752	1.33E-03	4.77E-04	**7.52E-05**	1.91E-04	7.81E-05
KP253	1.98E-03	6.83E-04	3.92E-04	6.82E-04	**3.88E-04**
KP503	2.25E-03	4.95E-04	2.76E-04	4.25E-04	**2.70E-04**
KP753	2.16E-03	3.63E-04	2.17E-04	2.77E-04	**2.07E-04**
KP254	2.52E-03	1.14E-03	**6.07E-04**	9.13E-04	6.14E-04
KP504	3.45E-03	6.63E-04	4.08E-04	5.14E-04	**3.62E-04**
KP754	3.79E-03	4.91E-04	2.88E-04	3.56E-04	**2.53E-04**

Tableau 12. Average Generational Distance (IGD)

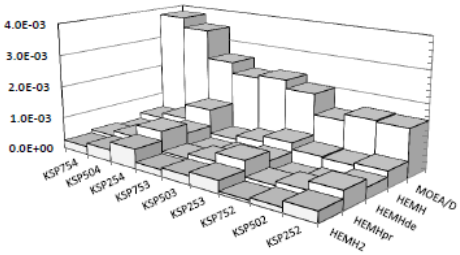


Fig. 13. IGD

Inst.	Algorithms				
	MOEA/D	HEMH	$HEMH_{de}$	$HEMH_{pr}$	HEMH2
KP252	9.32E-04	4.83E-04	**3.49E-04**	4.51E-04	3.50E-04
KP502	7.31E-04	2.72E-04	**1.31E-04**	2.10E-04	1.32E-04
KP752	5.41E-04	2.43E-04	**7.89E-05**	1.14E-04	7.95E-05
KP253	6.31E-04	4.48E-04	**3.83E-04**	4.41E-04	3.84E-04
KP503	5.28E-04	3.33E-04	**2.58E-04**	2.99E-04	2.60E-04
KP753	4.55E-04	2.46E-04	**1.93E-04**	2.24E-04	1.95E-04
KP254	6.75E-04	5.69E-04	**4.88E-04**	5.33E-04	4.91E-04
KP504	5.95E-04	4.04E-04	**3.46E-04**	3.67E-04	3.46E-04
KP754	5.46E-04	3.28E-04	2.71E-04	2.85E-04	**2.70E-04**

Tableau 13. Average Inv. Generational Dist. (IIGD)

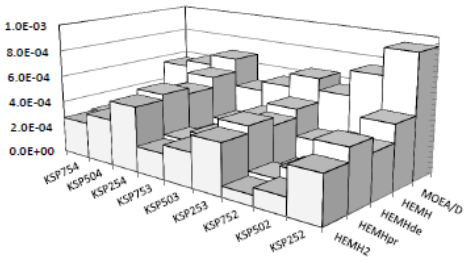


Fig. 14. IIGD

Inst.	Algorithms				
	MOEA/D	HEMH	$HEMH_{de}$	$HEMH_{pr}$	HEMH2
KP252	6.05E-03	1.88E-03	**1.29E-03**	2.32E-03	1.30E-03
KP502	7.30E-03	2.06E-03	**6.89E-04**	1.62E-03	7.12E-04
KP752	6.88E-03	2.47E-03	**5.46E-04**	1.21E-03	5.54E-04
KP253	1.11E-02	5.62E-03	4.49E-03	5.24E-03	**4.48E-03**
KP503	1.34E-02	5.16E-03	3.83E-03	4.60E-03	**3.78E-03**
KP753	1.42E-02	4.25E-03	3.38E-03	3.92E-03	**3.32E-03**
KP254	1.61E-02	9.70E-03	7.86E-03	8.84E-03	**7.84E-03**
KP504	2.02E-02	7.91E-03	7.18E-03	7.40E-03	**7.06E-03**
KP754	2.39E-02	7.02E-03	6.02E-03	6.26E-03	**5.82E-03**

Tableau 14. Average R3 indicator (IR3)

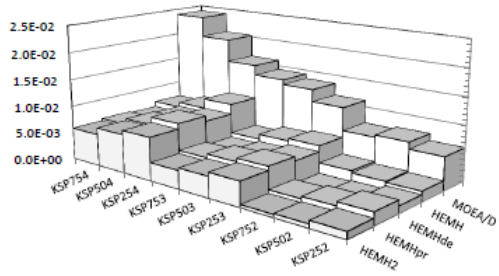


Fig. 15. IR3

Conclusion

Afin d'améliorer les performances HEMH présentée précédemment, nous avons proposé des métaheuristiques hybrides évolutionnaires HEMH2 et deux variantes ($HEMH_{de}$ et $HEMH_{pr}$). Le HEMH2 adopte la procédure du glouton inverse dans sa phase d'initialisation. L'Evolution Différentielle binaire adaptative et le Path-Relinking sont tous les deux utilisés. L'algorithme $HEMH_{de}$ utilise seulement l'Evolution Différentielle binaire adaptative, tandis que le $HEMH_{pr}$ utilise le procédé de croisement/mutation en plus du Path-Relinking. La comparaison de ces méthodes avec les méthodes de la littérature montrent des résultats très encourageants. En effet, l'Evolution Différentielle binaire adaptative incluse à la fois dans HEMH2 et dans HEMHde possède de meilleures capacités d'exploration qui surclassent les capacités de recherche locale utilisées dans la HEMH.

3.3.4 Approche évolutionnaire hybride avec adaptation de la stratégie de recherche (HESSA)

Introduction

Dans les travaux précédents, détaillés dans [63, 65, 64], l'influence de l'intégration des déférentes métaheuristiques coopératives dans le MOEA/D a été examinée pour les espaces de recherche discrets. Les résultats obtenus nous ont incités à étendre l'idée au cas continu, en élaborant une approche évolutionnaire hybride (HESSA) qui intègre un ensemble de stratégies de recherche adaptatives dans le MOEA/D. L'objectif principal est de tirer parti des avantages de ces stratégies grâce à la coopération et à l'intégration. L'objectif est également de rendre l'approche capable de sélectionner la stratégie de recherche adaptée en fonction du problème étudié.

Adaptation de la stratégie de recherche

Dans HESSA, au lieu d'utiliser une seule stratégie, un groupement de plusieurs stratégies de recherche est adopté pour générer les nouvelles solutions (descendances). Pour générer une nouvelle descendance, l'ensemble des candidats est accessible pour sélectionner une stratégie de recherche pour chaque individu cible dans la population actuelle. Au cours de l'évolution, chaque élément du bassin est considéré lors de la période d'apprentissage (LP). La meilleure stratégie obtenue durant la période d'apprentissage précédente est utilisée pour générer les solutions prometteuses. La plus probable sera choisie dans la période d'apprentissage actuelle et sera utilisée pour générer les nouvelles solutions des descendants. A chaque phase d'apprentissage, la somme des probabilités de sélection de chaque stratégie dans le bassin des candidats est égale à 1. Ces probabilités sont adaptées progressivement au cours du processus de l'évolution. Dans la période initiale d'apprentissage, toutes les stratégies ont la même chance d'être choisies, à savoir, chaque stratégie k a une probabilité $p_k = \frac{1}{K}$, où K est le nombre total de stratégies dans le bassin de candidats. Au cours de chaque période d'apprentissage, chaque stratégie k peut être choisie afin de générer la nouvelle solution en fonction de sa probabilité p_k en utilisant la sélection stochastique universelle [2]. Le nombre de sélections de chaque stratégie k est représenté par $calls_k$. Chaque stratégie est considérée pour obtenir un succès si elle a la capacité de générer une descendance capable de mettre à jour la population actuelle. Le nombre d'appels réussis pour chaque stratégie k est inscrit dans Suc_k Le nombre total dans le bassin des candidats est exprimé par :

$$calls_{tot} = \sum_{l=1}^{L} \sum_{k=1}^{K} calls_{k,l}$$

où L est le nombre total de périodes d'apprentissage dans l'ensemble du processus d'évolution. Cependant, après chaque période d'apprentissage l ($Calls_{tot} \% LP = 0$), la probabilité de sélection de chaque stratégie k pour la prochaine période d'apprentissage $p_{k,l+1}$ sera adoptée selon les formules suivantes :

$$p_{k,l+1} = \frac{SucR_{k,l}}{\sum_{k=1}^{K} SucR_{k,l}}$$

$$SucR_{k,l} = \begin{cases} \dfrac{Suc_{k,l}}{calls_{k,l}} + \varepsilon & si\ \ calls_{k,l} > 0 \ \ \forall\, k, l \\ \varepsilon & sinon \end{cases}$$

où $SucR_{k,l}$ est le taux de réussite de la i[ème] stratégie dans la période d'apprentissage l. La valeur $\varepsilon = 0.01$ est utilisé pour éviter les taux de réussite nuls. Par conséquent, les stratégies ayant un taux de réussite nul ont une chance d'être choisies pour générer une descendance. Les deux quantités $Suc_{k,l}$ et $calls_{k,l}$représentent le nombre de succès et le nombre total de succès de la k[ème] stratégie dans la période d'apprentissage l.

Principe de la méthode HESSA

La méthode HESSA utilise le Tchebycheff pondéré comme technique de décomposition pour convertir un problème multi-objectifs en un ensemble de sous-problèmes mono-objectif. Si nous avons un ensemble de N vecteurs poids uniformément répartis $\{\Lambda^1, \dots, \Lambda^N\}$ suite à la décomposition, nous avons N sous-problèmes mono-objectif. HESSA tente d'optimiser simultanément ces sous-problèmes. Chaque sous-problème i a son propre ensemble de voisins appelé B_i, qui inclut tous les sous-problèmes avec les T vecteurs poids les plus proches $\{\Lambda^{i1}, \dots, \Lambda^{iT}\}$ de Λ^i en termes de distance euclidienne. La structure de la méthode proposée est résumée comme suit :

- une population P de N individus, = $P = \{x^1, \dots, x^N\}$ où x^i représente la solution en cours du ième sous-problème. Chaque individu x^i a sa propre vitesse v^i, sa meilleure position personnelle x^i_{pb} et son âge a_i.
- un ensemble de N vecteurs de poids uniformément répartis $\{\Lambda^1, \dots, \Lambda^N\}$, correspondrait aux N sous-problèmes. Chaque $\Lambda = \{\lambda_1, \dots, \lambda_m\}$ possède m composantes correspondant aux m objectifs, tels que,

$$\sum_{i=1}^{m} \lambda_i = 1, \quad \forall\, \lambda_i \in \left\{\frac{0}{H}, \frac{1}{H}, \dots . \frac{H}{H}\right\} \quad et \quad N = C_{m-1}^{H+m-1} \qquad \forall\, H \in Z^+$$

- un voisinage B_i pour chaque sous-problème $i \in \{1, \dots, N\}$, qui comprend tous les sous-problèmes avec les T vecteurs poids les plus proches $\Lambda^{i1}, \dots, \Lambda^{iT}\}$ de Λ^i.
- un ensemble de stratégies de reproduction adaptative contenues dans un bassin pour générer de nouvelles solutions. Chaque stratégie est choisie en fonction de sa probabilité, comme mentionnée ci-dessus. Le tableau 15 résume l'ensemble des stratégies adoptées.
- une archive externe pour collecter les bonnes solutions explorées lors du processus de recherche. L'archive joue également le rôle de référentiel des "leaders" globaux.

Stratégie	Description
SBXPM	Le croisement SBX [22] est appliqué sur les deux parents suivi d'une mutation polynomiale [22].
DEXPM	L'évolution différentielle [11, 89] est appliquée sur trois parents sélectionnés, suivie d'une mutation polynomiale [22].
MPCPM	Le croisement multiple de parents [68] est appliqué sur trois parents sélectionnés, suivi de la mutation polynomiale.
GM	La mutation guidée [52] est utilisée pour produire une descendance depuis son parent et la solution globale guidée.
PSO	Les essaims particulaires [68] calculent une nouvelle position à partir du parent actuel, de son record personnel et du guide global.

Tableau 15. Ensemble des stratégies de reproduction utilisées

Algorithme

Après la construction du cadre proposé, l'algorithme HESSA comporte deux phases principales. La première est la phase d'initialisation, dans laquelle une population initiale est générée de manière aléatoire. La seconde phase représente la boucle principale dans laquelle les recherches sont effectuées afin d'améliorer la population initiale. L'algorithme 8 explique la procédure HESSA. Tout d'abord, un ensemble de N vecteurs de poids uniformément répartis est initialisé. Ensuite, la structure de voisinage B_i est construite pour chaque sous-ensemble i en attribuant tous les sous-problèmes correspondants aux T vecteurs de poids les plus proches de Λ^i. Le bassin de candidats est également construit en utilisant les stratégies de reproduction adoptées. Les archives et le compteur d'évaluation sont initialisés. Dans un second temps, la population initiale est construite. Pour chaque sous-population i, la solution actuelle x^i est initialisée aléatoirement. Ensuite, x^i est évaluée et utilisée pour mettre à jour le point de référence r^* [115], le record personnel x_{pb}^i et l'archive. La vitesse v^i et l'âge a_i sont également initialisés à 0. Le i^{ème} sous-problème est affecté à la population P. Par la suite, la boucle principale est exécutée jusqu'à la réalisation des évaluations maximales MaxEvals. Pour chaque sous-ensemble i, l'intervalle de reproduction/mise à jour M_i est choisi pour être soit le voisinage B_i soit la population entière. Puis, trois solutions parentes différentes sont choisies au hasard à partir de M_i pour la reproduction. Le « leader » global x_{gb}^i est sélectionné aléatoirement à partir des archives. Une stratégie globale de reproduction S_k est également sélectionnée dans le bassin pour générer la nouvelle descendance y.

Selon la stratégie S_k choisie, la descendance y est générée. En cas de l'utilisation de la mutation guidée ou des essaims particulaires, le paramètre âge a_i contrôle le processus de génération. Dans ce cas, si a_i dépasse le l'âge maximum autorisé T_a, une valeur gaussienne :

$$N\left(\frac{1}{2}\left[x_{gb}^i, x_{pb}^i\right], \quad \left|x_{gb}^i, x_{pb}^i\right|\right)$$

est affectée à y. Ensuite, le descendant y est évalué et utilisé pour mettre à jour le point de référence r^*. La population actuelle P est mise à jour en invoquant le module UpdateSolutions. L'archive est également mise à jour par y en fonction de la distance de surpeuplement. Le compteur d'évaluation est mis à jour et vérifié. A la fin de chaque période d'apprentissage, le bassin est adapté en calculant la probabilité p_k pour chaque stratégie k. A la fin de l'évolution, l'archive est retournée.

Algorithme 8 - HESSA (N; T; t; δ; ϵ; γ ; $\eta_c, \eta_m, CR, F, T_a$)
N : Taille de la population ou nombre de sous-problèmes
T : Taille min du voisinage
t : Max des solutions remplacées
$\delta \in [0,1]$ *Probabilité de choisir les parents à partir du voisinage*
Début
1: $W_V = \{\Lambda^1, \dots, \Lambda^N\}$: *Ensemble de N vecteurs de poids.*
2: $B_i \leftarrow \{i_1, i_2, \dots i_T\} \ \forall\, i = 1, \dots N$
3:Pool ← ConstructPool (sbxpm, dexpm, mpcpm, gm, pso) *// 5 stratégies*
4: $Eval \leftarrow 0$; $Arch \leftarrow \emptyset$
5: Pour i = 1 à N faire *//Phase d'initialisation*
6: $x_j^i \leftarrow \mathrm{U}(a_j, b_j) \ \forall\, i = 1, \dots N$ / *tirer une valeur aléatoire uniforme dans* $[a_j, b_j]$
7: $r^* \leftarrow EvaluateUpdate\ (x^i)$
8 : $x_{pb}^i \leftarrow x^i$; $v^i \leftarrow 0$; $a_i \leftarrow 0$
9 : $P \leftarrow$ AddSubProblem$(x^i, \Lambda^i, v^i, x_{pb}^i, a_i)$
10: $Arch \leftarrow UpdateArchive\ (x^i)$; Eval++ ;
11 : Fin Pour
12 : Tant que $(Eval < Mevals)\ faire$
13 : Pour i = 1 à N faire

$$M_i = \begin{cases} B_i & Si\ (rnd \in [0,1] \\ 1, \dots . N & Sinon \end{cases}$$

14: $x^a, x^b, x^c \leftarrow$ Selection(M_i, i)
15: $x_{gb}^i \leftarrow SelectGlobalBest(Arch)$
16: $S_k \leftarrow SelectStrategy(Pool)$
17: $Calls_k \leftarrow Calls_k + 1$
18 : Si $(S_k = SXBPM)$ alors
19: $y \leftarrow Crossover(x^a, x^b)$
20: $y \leftarrow PolyMutation$(y)
21 : Sinon, si $(S_k =$ "$DEXPM$"$)$ alors
22 : $y \leftarrow$ DiffEvolution$(x^i, x^a, x^b, x^c, CR, F)$
23 : $y \leftarrow PolyMutation$(y)
24 : Sinon, si $(S_k =$ "$MPCPM$"$)$ alors
25 : $y \leftarrow$ MPCrossover$(x^a, x^b, x^c, x_{gb}^i)$
26 : $y \leftarrow PolyMutation$(y)
27 : Sinon, si $(S_k =$ "GM"$)$ alors
28 :

$$y = \begin{cases} GuidedMutation(x^i, x_{gb}^i) & Si\ (a_i < T_\alpha) \\ N\left(\frac{1}{2}\left[x_{gb}^i, x_{pb}^i\right],\ \left|x_{gb}^i, x_{pb}^i\right|\right) & sinon \end{cases}$$

29 : Sinon, si $(S_k =$ "PSO"$)$ alors
30 :

$$y = \begin{cases} PSO(x^i, x_{pb}^i, x_{gb}^i,\ v^i, a_i) & Si\ (a_i < T_\alpha) \\ N\left(\frac{1}{2}\left[x_{gb}^i, x_{pb}^i\right],\ \left|x_{gb}^i, x_{pb}^i\right|\right) & sinon \end{cases}$$

31 : Fin si
32: $r^* \leftarrow EvaluateUpdate\ (y)$
33: $P \leftarrow UpdateSolutions\ (y, t, M_i, P, S_k, r^*)$
34: $Arch \leftarrow UpdateArchive\ (y, S_k)$
35: $Calls_{tot}++$; $Eval++$;
36 : $Si\ (Calls_{tot}\%LP = 0)$ alors
37: $Pool \leftarrow AdaptPool\ (Pool)$
38 : Fin Si
39 : Fin Pour
40 : Fin Tant que
41: Retourner Arch
42 : Fin

Algorithme 9 - UpdateSolutions (y ; T; M_i; P; S_k; r^*)
Entrées
P : Population
M_i : intervalle du sous-problème i
y : Nouvelle solution
t : Max des solutions remplacées
S_k : Stratégie sélectionnée
r^*: point de référence
Début
1: $c \leftarrow 0$
2: Tant que ($c < t$ et $M_i \neq \emptyset$) $faire$
3:j ← SelectRandomIndex (M_i)
4: Si $F^{TC}(y, \Lambda^j, r^*) = F^{TC}(x^j, \Lambda^j, r^*)$ alors
5: $x^j \leftarrow y$; $c++$; $a^j \leftarrow 0$;
6: $Suc_k \leftarrow Suc_k + 1$
7: Si $F^{TC}(x^j, \Lambda^j, r^*) = F^{TC}(x^i_{pb}, \Lambda^j, r^*)$ alors
8: $x^j_{pb} \leftarrow x^j$
9: Fin Si
10: M_i ← RemoveIndex (M_i, j)
11: Sinon $a_i \leftarrow a_i + 1$
12: Fin Si
13: Fin Tant que
14: Retourner P
15: Fin

Résultats

Les résultats détaillés et le descriptif des indicateurs de qualité (Set Coverage, Hypervolume, Genrational distance, Inverted genrational distance, unary additive epsilon), Sont présentés dans [66]. Notre méthode HESSA a été comparée à trois méthodes de l'état de l'art : MOEA/D1 [115], MOEA/D2 [76] et dMOPSO [78]. Les résultats expérimentaux montrent la supériorité de HESSA à la fois sur MOEA/D et dMOPSO dans la plupart des tests effectués. Ils indiquent également que HESSA a une performance moyenne très compétitive par rapport aux MOEAs basés sur les indicateurs d'évaluation utilisés dans cette étude. La contribution de HESSA réside dans la combinaison des différents opérateurs de recherche coopératifs qui intensifient le processus de recherche pour découvrir les régions prometteuses dans l'espace de recherche et d'améliorer la capacité d'explorer des solutions de bonne qualité. La seconde contribution est la capacité d'adapter le processus de recherche en utilisant l'opérateur de recherche approprié au problème étudié.

MOPs	$MOEAD_1$	$MOEAD_2$	dMOPSO	HESSA
Fonse	$5.03e-3_{6.5e-4}$	$3.64e-3_{4.3e-5}$	$1.19e-2_{1.4e-3}$	$4.08e-3_{1.1e-4}$
Kursa	$2.94e-1_{1.8e-2}$	$3.83e-1_{3.7e-2}$	$1.49e+0_{2.1e-1}$	$2.77e-1_{1.3e-2}$
ZDT1	$2.92e-2_{1.5e-2}$	$4.50e-1_{8.9e-2}$	$2.36e-2_{2.5e-3}$	$5.50e-3_{2.0e-4}$
ZDT2	$1.87e-1_{8.1e-2}$	$3.33e-1_{0.0e+0}$	$1.23e-1_{1.4e-1}$	$5.48e-3_{4.7e-4}$
ZDT3	$2.28e-2_{2.3e-2}$	$6.22e-1_{6.3e-2}$	$3.09e-2_{6.2e-3}$	$5.65e-3_{2.9e-4}$
ZDT4	$1.05e-1_{7.5e-2}$	$6.66e-1_{1.1e-8}$	$1.01e-1_{1.2e-1}$	$5.53e-3_{4.3e-4}$
ZDT6	$1.54e-2_{2.6e-3}$	$1.32e-1_{8.8e-2}$	$8.76e-3_{3.6e-3}$	$2.16e-4_{1.2e-5}$
DTLZ2	$4.61e-2_{1.3e-3}$	$4.84e-2_{1.1e-3}$	$1.21e-1_{6.7e-3}$	$4.64e-2_{1.1e-3}$
DTLZ4	$1.40e-1_{1.3e-1}$	$2.05e-2_{3.0e-2}$	$5.81e-2_{1.2e-2}$	$4.01e-4_{1.1e-3}$
DTLZ6	$1.11e-2_{6.9e-3}$	$2.67e-4_{7.6e-6}$	$7.71e-4_{1.6e-4}$	$3.11e-4_{3.4e-6}$
DTLZ7	$1.71e-1_{2.1e-2}$	$5.36e-1_{1.2e-1}$	$1.67e-1_{2.0e-2}$	$1.69e-1_{5.2e-3}$

Tableau 16. Résultats pour l'indicateur I_{RH} (moyenne, écart-type)

MOPs	$MOEAD_1$	$MOEAD_2$	dMOPSO	HESSA
Fonse	$1.56e-3_{2.4e-4}$	$9.96e-4_{3.3e-5}$	$4.49e-3_{5.1e-4}$	$1.14e-3_{5.8e-5}$
Kursa	$8.56e-3_{1.2e-3}$	$1.09e-2_{1.6e-3}$	$4.89e-2_{8.7e-3}$	$7.46e-3_{6.6e-4}$
ZDT1	$5.15e-3_{1.4e-3}$	$3.67e-1_{9.7e-2}$	$1.29e-2_{1.8e-3}$	$8.30e-4_{1.2e-4}$
ZDT2	$1.06e-3_{1.4e-3}$	$4.58e-1_{1.6e-1}$	$1.35e-2_{1.1e-2}$	$9.12e-4_{2.8e-4}$
ZDT3	$5.83e-3_{6.5e-3}$	$5.06e-1_{1.0e-1}$	$8.87e-3_{1.6e-3}$	$2.70e-3_{1.6e-4}$
ZDT4	$7.15e-2_{8.1e-2}$	$1.75e+1_{6.4e+0}$	$1.87e-3_{1.3e-3}$	$8.38e-4_{2.4e-4}$
ZDT6	$1.35e-2_{2.1e-3}$	$2.15e-1_{1.9e-1}$	$5.45e-3_{9.0e-3}$	$2.62e-3_{3.7e-5}$
DTLZ2	$5.85e-3_{1.1e-4}$	$7.85e-3_{2.4e-4}$	$6.85e-2_{5.0e-3}$	$6.33e-3_{1.7e-4}$
DTLZ4	$2.51e-2_{1.1e-2}$	$3.60e-2_{3.0e-3}$	$4.58e-2_{5.2e-3}$	$3.51e-2_{1.4e-3}$
DTLZ6	$4.19e-2_{2.8e-2}$	$3.72e-3_{8.4e-5}$	$4.42e-3_{2.3e-4}$	$3.85e-3_{6.8e-5}$
DTLZ7	$2.25e-2_{1.7e-3}$	$1.69e-1_{7.4e-2}$	$5.20e-2_{6.2e-3}$	$2.19e-2_{4.0e-4}$

Tableau 17. Résultats pour l'indicateur I_{GD} (moyenne, écart-type)

MOPs	$MOEAD_1$	$MOEAD_2$	dMOPSO	HESSA
Fonse	$4.21e-3_{4.5e-4}$	$3.57e-3_{2.2e-5}$	$1.63e-2_{5.3e-3}$	$3.70e-3_{5.9e-5}$
Kursa	$4.24e-2_{1.2e-3}$	$4.42e-2_{1.2e-3}$	$1.06e-1_{2.1e-2}$	$4.20e-2_{5.2e-4}$
ZDT1	$3.87e-2_{3.2e-2}$	$3.90e-1_{9.6e-2}$	$1.48e-2_{1.5e-3}$	$4.05e-3_{7.1e-5}$
ZDT2	$2.46e-1_{1.2e-1}$	$8.98e-1_{1.5e-1}$	$1.95e-1_{2.7e-1}$	$4.00e-3_{1.2e-4}$
ZDT3	$2.67e-2_{2.7e-2}$	$4.66e-1_{7.6e-2}$	$1.75e-2_{2.4e-3}$	$1.06e-2_{1.1e-4}$
ZDT4	$1.05e-1_{6.8e-2}$	$6.48e+0_{2.8e+0}$	$1.60e-1_{1.8e-1}$	$4.11e-3_{1.3e-4}$
ZDT6	$1.93e-3_{3.4e-3}$	$1.63e-1_{2.2e-1}$	$3.63e-3_{1.1e-3}$	$1.89e-3_{1.6e-5}$
DTLZ2	$3.72e-2_{2.0e-4}$	$3.81e-2_{3.4e-4}$	$7.33e-2_{4.6e-3}$	$3.73e-2_{2.3e-4}$
DTLZ4	$1.69e-1_{1.4e-1}$	$2.99e-2_{5.2e-3}$	$4.41e-2_{6.5e-3}$	$2.98e-2_{2.0e-3}$
DTLZ6	$3.82e-2_{2.6e-2}$	$4.39e-3_{3.2e-5}$	$7.72e-3_{7.2e-4}$	$4.52e-3_{1.5e-5}$
DTLZ7	$2.24e-1_{1.5e-1}$	$3.76e-1_{1.8e-1}$	$8.94e-2_{5.1e-2}$	$1.15e-1_{3.5e-3}$

Tableau 18. Résultats pour l'indicateur I_{IGD} (moyenne, écart-type)

Conclusion

Dans ce travail, une nouvelle approche évolutionnaire hybride fondée sur l'adaptation de la stratégie de recherche (HESSA) a été présentée. Dans HESSA, le processus de recherche est effectué par l'adoption d'un ensemble de stratégies de recherche différentes, où chacune a un taux de succès spécifié. Une nouvelle descendance est générée en utilisant une stratégie choisie au hasard. Ensuite, en fonction du succès de la descendance générée pour la mise à jour la population ou de l'archive, le taux de succès de la stratégie sélectionnée est adapté. Cela donne la possibilité à la méthode HESSA d'adopter la stratégie de recherche appropriée en fonction du problème à traiter. La méthode HESSA proposée est comparée à quelques méthodes de l'état de l'art MOEA en utilisant un ensemble de jeu de données couramment utilisés dans la littérature. Les résultats obtenus sont très satisfaisants.

3.4 Conclusion du chapitre

Dans les travaux mentionnés dans ce chapitre nous avons proposé plusieurs approches de métaheuristiques hybrides (HEMH, MOEADde, MOEADpr, MOEADdp1, MOEADdp2, HEMH2, HEMHde et HEMHpr pour le domaine de recherche discret et HESSA pour le domaine continu). Ces approches ont été testées et validées sur un ensemble de problèmes d'optimisation multi-objectifs utilisés dans la littérature. Sur un ensemble d'indicateurs d'évaluation de la qualité, nous avons testé ces méthodes. Toutes les approches proposées donnent des résultats très satisfaisants et compétitifs par

rapport aux autres méthodes équivalentes de la littérature. Les approches proposées sont en mesure d'intensifier le processus de recherche de régions prometteuses de l'espace de recherche et d'améliorer la capacité d'explorer des solutions de bonne qualité.

Dans les travaux futurs, nous envisageonsb d'étudier les paramètres de réglage de HEMH2 et ses variantes, ainsi que l'analyse de leur convergence. En outre, d'autres métaheuristiques seront étudiées pour améliorer la performance de HEMH2 et gérer d'autres types de problèmes d'optimisation combinatoire. Nous allons nous pencher également sur la manière d'inclure le décideur dans le processus de recherche.

3.5 Références

[1] [Aarts & Lenstra 1997] Emile Aarts and Jan K. Lenstra, editeurs. Local Search in Combinatorial Optimization. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1997.

[2] [Baker 1987] James E. Baker. Reducing bias and inefficiency in the selection algorithm. In Proceedings of the Second International Conference on Genetic Algorithms on Genetic algorithms and their application, pages 14-21, Hillsdale, NJ, USA, 1987. L. Erlbaum Associates Inc.

[3] [Bednorz 2008] Witold Bednorz. Advances in greedy algorithms. Wienna: I-Tech Education and Publishing KG, 2008.

[4] [Bledsoe & Browning 1959] W. W. Bledsoe and I. Browning. Pattern recognition and reading by machine. In Papers presented at the December 1-3, 1959, eastern joint IRE-AIEE-ACM computer conference, IRE-AIEE-ACM '59 (Eastern), pages 225-232, New York, NY, USA, 1959. ACM.

[5] [BLUM et al. 2005] CHRISTIAN BLUM, Andrea Roli and Enrique Alba. An Introduction to Metaheuristic Techniques. Parallel Metaheuristics: A New Class of Algorithms, vol. 47, page 1, 2005.

[6] [Blum & Roli 2003] Christian Blum and Andrea Roli. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. ACM Computing Surveys, vol. 35, no. 3, pages 268-308, 2003.

[7] [Boussaid et al. 2013] Ilhem Boussaïd, Julien Lepagnot and Patrick Siarry. A survey on optimization metaheuristics. Information Sciences, 2013.

[8] [Branke et al. 2008] Jurgen Branke, Kalyanmoy Deb, Kaisa Miettinen and Roman Slowinski. Multiobjective optimization: Interactive and evolutionary approaches, volume 5252. Springer, 2008.

[9] [Bremermann 1962] H. J. Bremermann. Optimization through evolution and recombination. In M. C. Yovits, G. T. Jacobi and G. D. Golstine, editeurs, Proceedings of the Conference on Self-Organizing Systems - 1962, pages 93-106, Washington, DC, 1962. Spartan Books.

[10] [Cerny 1985] VLADIMIR Cerny. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. Journal of optimization theory and applications, vol. 45, no. 1, pages 41-51, 1985.

[11] [Chakraborty 2010] U.K. Chakraborty. Advances in Differential Evolution. Studies in Computational Intelligence. Springer, 2010.

[12] [Chankong & Haimes 1983] V. Chankong and Y.Y. Haimes. Multiobjective Decision Making Theory and Methodology. Elsevier Science, New York,1983.

[13] [Charnes & Cooper 1961] A. Charnes and W.W. Cooper. Management models and industrial applications of linear programming. Numeero v. 1 de Management Models and Industrial Applications of Linear Programming. Wiley, 1961.

[14] [Charnes & Cooper 1977] A. Charnes and W.W. Cooper. Goal programming and multiple objective optimizations: Part 1. European Journal of Operational Research, vol. 1, no. 1, pages 39 - 54, 1977.

[15] [Charnes et al. 1955] Abraham Charnes, William W Cooper and Robert O Ferguson. Optimal estimation of executive compensation by linear programming. Management science, vol. 1, no. 2, pages 138-151, 1955.

[16] [Coello Coello & Lechuga 2002] Carlos A. Coello Coello and M.S. Lechuga. MOPSO: a proposal for multiple objective particle swarm optimization. In Evolutionary Computation, 2002. CEC '02. Proceedings of the 2002 Congress on, volume 2, pages 1051-1056, 2002.

[17] Coello et al. 2006] Carlos A. Coello Coello, Gary B. Lamont and David A. Van Veldhuizen. Evolutionary Algorithms for Solving Multi-Objective Problems (Genetic and Evolutionary Computation). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[18] [Corne et al. 2000] David Corne, Joshua D. Knowles and Martin J. Oates. The Pareto Envelope-Based Selection Algorithm for Multi-objective Optimisation. In Proceedings of the 6th International Conference on Parallel Problem Solving from Nature, PPSN VI, pages 839-848, London, UK, UK, 2000. Springer-Verlag.

[19] [Cotta-Porras 1998] Carlos Cotta-Porras. A study of hybridisation techniques and their application to the design of evolutionary algorithms. AI Communications, vol. 11, no. 3, pages 223-224, 1998.

[20] [Crainic & Toulouse 2003] TeodorGabriel Crainic and Michel Toulouse. Parallel Strategies for MetaHeuristics. In Fred Glover and GaryA. Kochenberger, editeurs, Handbook of Metaheuristics, volume 57 of International Series in Operations Research & Management Science, pages 475-513. Springer US, 2003.

[21] [Das & Dennis 1998] Indraneel Das and J. E. Dennis. Normal-Boundary Intersection: A New Method for Generating the Pareto Surface in Nonlinear Multicriteria Optimization Problems. SIAM Journal on Optimization, vol. 8, no. 3, pages 631+, 1998.

[22] [Deb & Agrawal 1995] Kalyanmoy Deb and Ram Bhushan Agrawal. Simulated Binary Crossover for Continuous Search Space. Complex Systems, vol. 9, pages 115-148, 1995.

[23] [Deb et al. 2000] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal and T. Meyarivan. A Fast Elitist Multi-Objective Genetic Algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation, vol. 6, pages 182-197, 2000.

[24] [Deb 2001] K. Deb. Multi-Objective Optimization using Evolutionary Algorithms. Wiley Interscience Series in Systems and Optimization. Wiley, 2001.

[25] [Dorigo et al. 1996] Marco Dorigo, Vittorio Maniezzo and Alberto Colorni. The Ant System: Optimization by a colony of cooperating agents. IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART B, vol. 26, no. 1, pages 29-41, 1996.

[26] [Eberhart et al. 2001] Russell C. Eberhart, Yuhui Shi and James Kennedy. Swarm Intelligence (The Morgan Kaufmann Series in Evolutionary Computation). Morgan Kaufmann, 1 edition, April 2001.

[27] [Edmonds 1971] J. Edmonds. Matroids and the greedy algorithm. Mathematical Programming, vol. 1, pages 127-136, 1971.

[28] [El-Abd & Kamel 2005] Mohammed El-Abd and Mohamed Kamel. A taxonomy of cooperative search algorithms. In Hybrid Metaheuristics, pages 32-41. Springer, 2005.

[29] [Elsayed et al. 2011] Saber M. Elsayed, Ruhul A. Sarker and Daryl Essam. GA with a new multi-parent crossover for solving IEEE-CEC2011 competition problems. In IEEE Congress on Evolutionary Computation, pages 1034-1040. IEEE, 2011.

[30] [Erickson et al. 2001] Mark Erickson, Alex Mayer and Jeffrey Horn. The Niched Pareto Genetic Algorithm 2 Applied to the Design of Groundwater Remediation Systems. In Eckart Zitzler, Lothar Thiele, Kalyanmoy Deb, CarlosArtemio Coello Coello and David Corne, editeurs, Evolutionary Multi-Criterion Optimization, volume 1993 of Lecture Notes in Computer Science, pages 681-695. Springer Berlin Heidelberg, 2001.

[31] [Falkenauer 1996] Emanuel Falkenauer. A hybrid grouping genetic algorithm for bin packing. Journal of heuristics, vol. 2, no. 1, pages 5-30, 1996.

[32] Faria, H., Jr. Binato, S. Resende, M.G.C. and Falcao, D.M. Power transmission network design by greedy randomized adaptive path relinking approach. IEEE Transactions on Power Systems, 20, 1 (2005) 43-49.

[33] [Farina et al. 2004] Marco Farina, Kalyanmoy Deb and Paolo Amato. Dynamic multiobjective optimization problems: test cases, approximations, and applications. IEEE Trans. Evolutionary Computation, pages 425-442, 2004.

[34] [Feo & Resende 1989] Thomas A. Feo and Mauricio G. C. Resende. A probabilistic heuristic for a computationally difficult set covering problem. Operations Research Letters, vol. 8, no. 2, pages 67 - 71, 1989.

[35] [Feo & Resende 1995] Thomas A Feo and Mauricio GC Resende. Greedy randomized adaptive search procedures. Journal of global optimization, vol. 6, no. 2, pages 109-133, 1995.

[36] [Fishburn 1974] Peter C Fishburn. Lexicographic orders, utilities and decision rules: A survey. Management science, pages 1442-1471, 1974.

[37] [Fonseca et al. 1993] Carlos M Fonseca, Peter J Fleminget al. Genetic algorithms for multiobjective optimization: Formulation, discussion and generalization. In Proceedings of the fifth international conference on genetic algorithms, volume 423, pages 416-423. San Mateo, California, 1993.

[38] [Gass & Saaty 1955] S. Gass and T.L. Saaty. The computational algorithm for the parametric objective function. Naval Research Logistics Quarterly, vol. 2, page 39, 1955.

[39] [Glover et al. 2000] Fred Glover, Manuel Laguna and Rafael Marti. Fundamentals of scatter search and path relinking. CONTROL AND CYBERNETICS, vol. 39, pages 653-684, 2000.

[40] [Glover 1986] Fred Glover. Future paths for integer programming and links to artificial intelligence. Comput. Oper. Res., vol. 13, no. 5, pages 533-549, May 1986.

[41] [Glover 1989] Fred Glover. Tabu search-part I. ORSA Journal on computing, vol. 1, no. 3, pages 190-206, 1989.

[42] [Glover 1996] Fred Glover. Tabu search and adaptive memory programing Advances, applications and challenges. In Interfaces in Computer Science and Operations Research, pages 1-75. Kluwer, 1996.

[43] Glover, F. and Laguna, M. Fundamentals of scatter search and path relinking. Control and Cybernetics 29, 3 (1999) 653-684.

[44] [Glover & Kochenberger 2003] F.E. Glover and G.A. Kochenberger. Handbook in Metaheuristics. International Series in Operations Research & Management Science, 57. Kluwer Academic Publishers, 2003.

[45] [Goldberg 1989] D.E. Goldberg. Genetic Algorithms in Search, Optimization, and Machine Learning. Artificial Intelligence. Addison-Wesley, 1989.

[46] [Hansen 1997] Michael Pilegaard Hansen. Tabu search for multiobjective optimization: MOTS. In Proceedings of the 13th International Conference on Multiple Criteria Decision Making, pages 574-586. Citeseer, 1997.

[47] M.P. Hansen and A. Jaszkiewicz. Evaluating the quality of approximations of non dominated set. Tech. Rep., Institute of Mathematical modeling, Tech. Univ of Denmark, 1998. IMM Tech. Rep. IMM-REP-1998-7.

[48] [Hernandez-Diaz et al. 2007] Alfredo G. Hernandez-Diaz, Luis V. Santana-Quintero, Carlos A. Coello Coello and Julian Molina Luque. Pareto-adaptive epsilon-dominance. Evolutionary Computation, vol. 15, no. 4, pages 493-517, 2007.

[49] [Holland 1975] J.H. Holland. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. University of Michigan Press, 1975.

[50] [Horn et al. 1994] J. Horn, N. Nafpliotis and D.E. Goldberg. A niched Pareto genetic algorithm for multiobjective optimization. In Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on, pages 82-87 vol.1, 1994.

[51] Hromkovic 2005] J. Hromkovic. Design and Analysis of Randomized Algorithms: Introduction to Design Paradigms. Texts in Theoretical Computer Science. An EATCS Series. Springer, 2005.

[52] [Hsieh et al. 2007] Chang-Tai Hsieh, Chih-Ming Chen and Ying-ping Chen. Particle swarm guided evolution strategy. In Proceedings of the 9th annual conference on Genetic and evolutionary computation, GECCO '07, pages 650-657, New York, NY, USA, 2007. ACM.

[53] [Hu & Eberhart 2002] Xiaohui Hu and R. Eberhart. Multiobjective optimization using dynamic neighborhood particle swarm optimization. In Evolutionary Computation, 2002. CEC '02. Proceedings of the 2002 Congress on, volume 2, pages 1677-1681, 2002.

[54] [Hwang & Masud 1979] C.L. Hwang and A.S.M. Masud. Multiple objective decision making, methods and applications: a state-of-the-art survey. Lecture notes in economics and mathematical systems. Springer-Verlag, 1979.

[55] [Hwang et al. 1980] Ching-Lai Hwang, SR Paidy, K Yoon and ASM Masud. Mathematical programming with multiple objectives: A tutorial. Computers & Operations Research, vol. 7, no. 1, pages 5-31, 1980.

[56] [Ijiri 1965] Y. Ijiri. Management goals and accounting for control. Studies in mathematical and managerial economics. North Holland Pub. Co., 1965.

[57] Ishibuchi, H., Sakane, Y., Tsukamoto, N. and Nojima, Y. Effects of using two neighborhood structures on the performance of cellular evolutionary algorithms for many-objective optimization, In Proc. of IEEE Congress on Evolutionary Computation, 2508-2515, (2009)

[58] [Jaszkiewicz 2003] Andrzej Jaszkiewicz. Do multiple-objective metaheuristics deliver on their promises? A computational experiment on the set-covering problem. IEEE Trans. Evolutionary Computation, vol. 7, no. 2, pages 133-143, 2003.

[59] [Jaszkiewicz 2002a] Andrzej Jaszkiewicz. Genetic local search for multi-objective combinatorial optimization. European journal of operational research, vol. 137, no. 1, pages 50-71, 2002

[60] [Jaszkiewicz 2002b] Andrzej Jaszkiewicz. On the performance of multiple objective genetic local search on the 0/1 knapsack problem - a comparative experiment. IEEE Trans. Evolutionary Computation, vol. 6, no. 4, pages 402-412, 2002.

[61] [Jaszkiewicz 2003] Andrzej Jaszkiewicz. Do multiple-objective metaheuristics deliver on their promises? A computational experiment on the set-covering problem. IEEE Trans. Evolutionary Computation, vol. 7, no. 2, pages133-143, 2003.

[62] [Jones & Tamiz 2010] Dylan Jones and Mehrdad Tamiz. Practical goal programming, volume 141. Springer, 2010.

[63] A.Kafafy, A. Bounekkar, S.Bonnevay. « A Hybrid Evolutionary Metaheuristics (HEMH) Applied On 0/1 Multi-objective Knapsack Problems ». Genetic and Evolutionary Computation Conference, Dublin (Irlande), July, pages 497-504, 2011.

[64] A.Kafafy, A. Bounekkar, S.Bonnevay. « Hybrid Metaheuristics based on MOEA/D for 0/1 Multiobjective Knapsack Problems : A comparative Study ». IEEE Congress on Evolutionary Computation, Brisbane (Australia), June, pages 3616-3623, 2012.

[65] A.Kafafy, A. Bounekkar, S.Bonnevay. « HEMH2: An Improved Hybrid Evolutionary Metaheuristics for 0/1 Multiobjective Knapsack Problems ». the 9th International Conference on Simulated Evolution And Learning, Hanoï (Vietnam), December, 2012.

[66] A. Kafafy, S. Bonnevay, A. Bounekkar. « A Hybrid Evolutionary Approach with Search Strategy Adaptation for Mutiobjective Optimization », Genetic and Evolutionary Computation Conference (GECCO), Amsterdam, October, 2013.

[67] [Kafafy 2013] Kafafy A. A Hybrid Evolutionary Metaheuristics based on DM-GRASP, Path-Relinking and genetic operators applied on 0/1 Multi-objective Knapsack Problems. Thèse de doctorat Université Lyon 1 Nov. 2013.

[68] [Kennedy & Eberhart 1995] J. Kennedy and R. C. Eberhart. Particle swarm optimization. In IEEE international conference on neural networks IV, pages 1942-1948, 1995.

[69] [Kirkpatrick et al. 1983] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi. Optimization by Simulated Annealing. Science, Number 4598, 13 May 1983, vol. 220, 4598, pages 671-680, 1983.

[70] Knowles & Corne 2000] Joshua D. Knowles and David W. Corne. Approximating the Nondominated Front Using the Pareto Archived Evolution Strategy. Evol. Comput., vol. 8, no. 2, pages 149-172, June 2000.

[71] J.D. Knowles and D.W. Corne. On metrics for comparing non-dominated sets. In Congress on Evolutionnary Computation (CEC'02), IEEE Press, pages 711–716, 2002.

[72] Laguna, M. and Marti R. GRASP and path relinking for 2-layer straight line crossing minimization. INFORMS Journal on Computing, 11, 1 (1999) 44-52.

[73] [Laguna et al. 2003] Manuel Laguna, Rafael Marti and Rafael Cunquero Marti. Scatter search: methodology and implementation in C, volume 24. Springer, 2003.

[74] [Laumanns et al. 2002] Marco Laumanns, Lothar Thiele, Kalyanmoy Deb and Eckart Zitzler. Combining convergence and diversity in evolutionary multiobjective optimization. Evolutionary computation, vol. 10, no. 3, pages 263-282, 2002.

[75] [Lee & Olson 1999] Sang M Lee and David L Olson. Goal programming. In Multicriteria Decision Making, pages 203-235. Springer, 1999.

[76] [Li & Zhang 2009] Hui Li and Qingfu Zhang. Multiobjective optimization problems with complicated Pareto sets, MOEA/D and NSGA-II. Trans. Evol. Comp, vol. 13, no. 2, pages 284-302, April 2009.

[77] [Lozano & Garcia-Martinez 2010] Manuel Lozano and Carlos Garcia-Martinez. Hybrid metaheuristics with evolutionary algorithms specializing in intensication and diversication: Overview and progress report. Computers & OR, vol. 37, no. 3, pages 481-497, 2010.

[78] Martinez & Coello 2011] Saul Zapotecas Martinez and Carlos A. Coello Coello. A multi-objective particle swarm optimizer based on decomposition. In Krasnogor & Lanzi [Krasnogor & Lanzi 2011], pages 69-76.

[79] [Michalewicz 1996] Z. Michalewicz. Genetic Algorithms + Data Structures = Evolution Programs. Articial intelligence. Springer, 1996.

[80] [Michalewicz & Fogel 2004] Z. Michalewicz and D.B. Fogel. How to Solve It: Modern Heuristics. Springer, 2004.

[81] [Miettinen 1999] K. Miettinen. Nonlinear multiobjective optimization. Kluwer Academic Publishers, Boston, 1999.

[82] [Mladenovic & Hansen 1997] Nenad Mladenovic and Pierre Hansen. Variable neighborhood search. Computers & Operations Research, vol. 24, no. 11, pages 1097-1100, 1997.

[83] [Moscato 1989] Pablo Moscato. On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms. Caltech concurrent computation program, C3P Report, vol. 826, page 1989, 1989.

[84] [Moscato 1999] Pablo Moscato. Memetic algorithms: a short introduction. In David Corne, Marco Dorigo, Fred Glover, Dipankar Dasgupta, Pablo Moscato, Riccardo Poli and Kenneth V. Price, editeurs, New ideas in optimization, pages 219-234. McGraw-Hill Ltd., UK, Maidenhead, UK, England, 1999.

[85] [Mostaghim & Teich 2003] S. Mostaghim and J. Teich. Strategies for finding good local guides in multi-objective particle swarm optimization (MOPSO). In Swarm Intelligence Symposium, 2003. SIS '03. Proceedings of the 2003 IEEE, pages 26-33, 2003.

[86] [Oei et al. 1991] C. K. Oei, David E. Goldberg and S. Chang. Tournament selection, niching, and the preservation of diversity. Rapport technique IlliGAL Report 91011, University of Illinois, 1991.

[87] [Osman & Laporte 1996] Ibrahim Osman and Gilbert Laporte. Metaheuristics: A bibliography. Annals of Operations Research, vol. 63, no. 5, pages 511-623, October 1996.

[88] [Parsopoulos & Vrahatis 2002] K.E. Parsopoulos and M.N. Vrahatis. Recent approaches to global optimization problems through Particle Swarm Optimization. Natural Computing, vol. 1, no. 2-3, pages 235-306, 2002.

[89] Price et al. 2005] K. Price, R.M. Storn and J. Lampinen. Differential Evolution: A Practical Approach to Global Optimization. Natural Computing Series. Springer, 2005.

[90] [Puchinger & Raidl 2005] Jakob Puchinger and G unther R Raidl. Combining metaheuristics and exact algorithms in combinatorial optimization: A survey and classification. In Artificial intelligence and knowledge engineering applications: a bioinspired approach, pages 41-53. Springer, 2005.

[91] [Raidl 2006] Gunther R. Raidl. A Unied View on Hybrid Metaheuristics. In Francisco Almeida and Maria J. Blesa Aguilera andChristian Blum and J. Marcos Moreno-Vega and Melquades Perez Perez and Andrea Roli and Michael Sampels, editeur, Hybrid Metaheuristics, volume 4030 of Lecture Notes in Computer Science, pages 1-12. Springer, 2006.

[92] [Rayward-Smith 1996] V.J. Rayward-Smith. Modern heuristic search methods. Wiley, 1996.

[93] [Reeves 1993] C.R. Reeves. Modern heuristic techniques for combinatorial problems. Advanced topics in computer science series. Halsted Press, 1993.

[94] Resende, M. G. C. and Werneck, R. F. A hybrid multistart heuristic for the uncapacitated facility location problem. European Journal of Operational Research, 174, 1(Oct 2006) 54-68.

[95] Resende, M. G. C., Marti, R., Gallego, M. and Duarte, A. GRASP and path relinking for the max-min diversity problem. Technical report, AT&T Labs Research, Florham Park, NJ, 2008.

[96] Ribeiro, M. H., Plastino, A. and Martins, S. L. Hybridization of GRASP metaheuristics with data mining techniques. Mathematical Modeling and Algorithms, 5 (2006) 23-41.

[97] [Robbins & Monro 1951] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. The Annals of Mathematical Statistics, vol. 22, no. 3, pages 400-407, 1951.

[98] [Russell & Norvig 2010] S.J. Russell and P. Norvig. Articial intelligence: a modern approach. Prentice Hall series in articial intelligence. Pearson Education/Prentice Hall, 2010.

[99] [Sakawa 1982] Masatoshi Sakawa. Interactive multiobjective decision making by the sequential proxy optimization technique: SPOT. European Journal of Operational Research, vol. 9, no. 4, pages 386-396, 1982.

[100] [Sakawa 1993] M. Sakawa. Fuzzy sets and interactive multiobjective optimization. N° v. 1 de Applied information technology. Plenum, 1993.

[101] Santos, L. F., Martins, S. L. & Plastino, A. Applications of the DM-GRASP heuristic: A survey. International Trans. On Operational Research, 15, 4(2008) 387-416.

[102] [Schaffer 1985] J. David Schaffer. Multiple Objective Optimization with Vector Evaluated Genetic Algorithms. In Proceedings of the 1st International Conference on Genetic Algorithms, pages 93-100, Hillsdale, NJ, USA, 1985. L. Erlbaum Associates Inc.

[103] [Srinivas & Deb 1994] N. Srinivas and Kalyanmoy Deb. Muiltiobjective optimization using nondominated sorting in genetic algorithms. Evol. Comput., vol. 2, no. 3, pages 221-248, September 1994.

[104] [Taillard et al. 2001] Eric D Taillard, Luca M Gambardella, Michel Gendreau and Jean-Yves Potvin. Adaptive memory programming: A unified view of metaheuristics. European Journal of Operational Research, vol. 135, no. 1, pages 1-16, 2001.

[105] [Taillard 1991] E Taillard. Robust taboo search for the quadratic assignment problem. Parallel computing, vol. 17, no. 4, pages 443-455, 1991.

[106] [Talbi 2002] E.-G. Talbi. A Taxonomy of Hybrid Metaheuristics. Journal of Heuristics, vol. 8, no. 5, pages 541-564, 2002.

[107] [Talbi 2009] El-Ghazali Talbi. Metaheuristics - From Design to Implementation. Wiley, 2009.

[108 Ulungu et al. 1999] EL Ulungu, JFPH Teghem, PH Fortemps and D Tuyttens. MOSA method: a tool for solving multiobjective combinatorial optimization problems. Journal of Multi-Criteria Decision Analysis, vol. 8, no. 4, pages 221-236, 1999.

[109] [Veldhuizen 1999] David A. Van Veldhuizen. Multiobjective Evolutionary Algorithms: Classications, Analyses, and New Innovations. PhD thesis, Department of Electrical and Computer Engineering. Graduate School of Engineering. Air Force Institute of Technology, Wright-Patterson AFB, Ohio, May 1999.

[110] Vianna, D. S. and Claudio Arroyo, J. E. A GRASP Algorithm for the Multiobjective Knapsack Problem, In Proceedings of XXIV International Conference of the Chilean Computer Science Society (SCCC'04), (2004) 69-75.

[111] [Voss et al. 1999] Stefan Voss, Ibrahim H. Osman and Catherine Roucairol, editeurs. Meta-Heuristics: Advances and Trends in Local Search Paradigms for Optimization. Kluwer Academic Publishers, Norwell, MA, USA, 1999.

[112] [Voudouris & Tsang 1999] C. Voudouris and E. Tsang. Guided local search. European Journal of Operational Research, vol. 113, pages 469-499, 1999.

[113] [Voudouris 1997] C. Voudouris. Guided Local Search for Combinatorial Optimization Problems. PhD thesis, Department of Computer Science, University of Essex, 1997.

[114] [Zeleny 1982] M. Zeleny. Multiple Criteria Decision Making. McGraw-Hill, New York, 1982.

[115] [Zhang & Li 2007] Qingfu Zhang and Hui Li. MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition. IEEE Trans. Evolutionary Computation, vol. 11, no. 6, pages 712-731, 2007.

[116] Zhang, M., Zhao, S. and Wang, X., Multi-objective evolutionary algorithm based on adaptive discrete Differential Evolution. IEEE Congress on Evolutionary Computation (CEC 2009), pp. 614-621, (2009)

[117] [Zitzler & Kunzli 2004] Eckart Zitzler and Simon Kunzli. Indicator-based selection in multiobjective search. In Parallel Problem Solving from Nature-PPSN VIII, pages 832-842. Springer, 2004.

[118] [Zitzler & Thiele 1999] Eckart Zitzler and Lothar Thiele. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. IEEE Trans. Evolutionary Computation, vol. 3, no. 4, pages 257-271, 1999

[119] [Zitzler et al. 2000] Eckart Zitzler, Kalyanmoy Deb and Lothar Thiele. Comparison of Multiobjective Evolutionary Algorithms: Empirical Results. Evolutionary Computation, vol. 8, no. 2, pages 173-195, 2000.

[120] [Zitzler et al. 2001] E. Zitzler, M. Laumanns and L. Thiele. SPEA2: Improving the strength pareto evolutionary algorithm for multiobjective optimization. In K. C. Giannakoglou, D. T. Tsahalis, J. Periaux, K. D. Papailiou and T. Fogarty, editeurs, Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problems, pages 95-100, Athens, Greece, 2001. International Center for Numerical Methods in Engineering.

[121] [Zitzler et al. 2003] Eckart Zitzler, Lothar Thiele, Marco Laumanns, Carlos M. Fonseca and Viviane Grunert da Fonseca. Performance assessment of multiobjective optimizers: an analysis and review. IEEE Trans. Evolutionary Computation, vol. 7, no. 2, pages 117-132, 2003

[122] [Zitzler et al. 2004] Eckart Zitzler, Marco Laumanns and Stefan Bleuler. A tutorial on evolutionary multiobjective optimization. In Metaheuristics for Multiobjective Optimisation, pages 3-37. Springer, 2004.

[123] [Zitzler et al. 2008] Eckart Zitzler, Joshua Knowles and Lothar Thiele. Quality Assessment of Pareto Set Approximations. In Jurgen Branke, Kalyanmoy Deb, Kaisa Miettinen and Roman Slowinski, editeurs, Multiobjective Optimization, volume 5252 of Lecture Notes in Computer Science, pages 373-404. Springer Berlin Heidelberg, 2008.

Chapitre 4

Modélisation et analyse des données en santé

Sommaire

4.1 Modélisation de la diffusion des épidémies

4.1.1 Introduction

Les maladies infectieuses sont de plus en plus présentes dans notre quotidien et représentent un problème majeur en santé publique. Les individus sont amenés à rencontrer un certain nombre de personnes au cours d'une journée que ce soit dans le milieu professionnel ou lors de sorties diverses (voyages, cinéma, centres commerciaux, école,…). Ainsi, une épidémie peut très rapidement être propagée tant à l'échelle régionale que mondiale si les mesures nécessaires ne sont pas prises. Cela peut s'illustrer par la pandémie intestinale (gastro-entérite) ou encore celle de la grippe (influenza) que nous observons chaque année. Face à ces diverses pandémies, le pouvoir public doit réagir c'est-à-dire prendre les décisions adéquates de manière à réduire la propagation de la pandémie. Ces mesures prises sont fonction de l'offre de soins fournie par les services médicaux en regard de l'ampleur de l'épidémie. La figure ci-dessous illustre deux situations où le seuil de saturation du système de santé est dépassé, sur des périodes très différentes, engendrant ainsi des situations devant être gérées différemment.

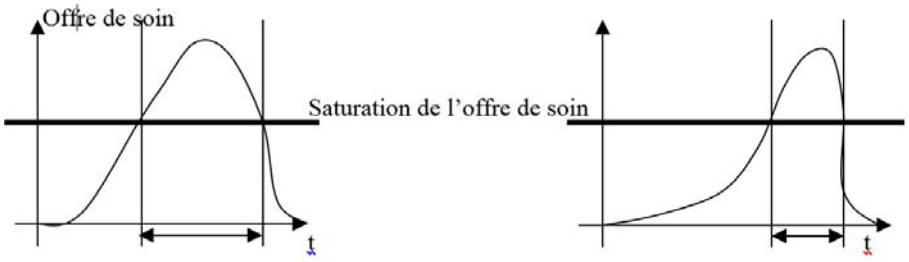


Figure 1. Présentation de deux cas différents d'évolution d'une pandémie.

Le pouvoir public, en tant que décideur, a pour objectifs de gérer toute pandémie en préservant au mieux l'économie de la région concernée. En particulier, le décideur peut être amené à gérer les déplacements des individus qu'ils soient intra ou inter régions tout en conservant le fonctionnement de l'économie.

La distinction entre l'endémie, l'épidémie et la pandémie est importante pour évaluer la proportion de personnes atteintes par l'agent infectieux. On parle d' « endémie » lorsque la maladie infectieuse est présente habituellement dans une zone géographique (fièvre jaune en Thailande). Une endémie se développe en fonction des conditions géographiques et climatiques, des facteurs socio-économiques et de l'alimentation des individus. Le terme « épidémie » est plus approprié lors d'une forte croissance du nombre de cas d'une même maladie contagieuse ou non en un lieu donné et à un moment précis (épidémie de la grippe espagnole, épidémie du SIDA,…). Le terme de « pandémie » quant à lui, désigne une épidémie dont la diffusion géographique est très étendue [45], (grippe H1N1). Nous présentons dans ce chapitre quelques contributions dans la modélisation de diffusion des épidémies. Ces travaux sont liés à des projets de recherche ou à des thèses de doctorat.

4.1.2 Problématique et état de l'art

Devant l'augmentation des résistances bactériennes, l'émergence de nouveaux pathogènes et la propagation rapide de l'épidémie, la prévention de la transmission de la maladie devient particulièrement importante et indispensable. Face à une telle menace, la société doit se préparer à l'avance pour réagir rapidement et efficacement si une telle épidémie est déclarée. Ce contexte épidémiologique souligne la nécessité de coupler les champs disciplinaires de l'épidémiologie et de la modélisation du phénomène afin d'identifier et de mieux comprendre comment ces maladies sont transmises dans leur environnement et d'étudier quelles sont les stratégies de contrôle les plus efficaces faces à la progression de cette épidémie.

La modélisation en épidémiologie a été utilisée dans l'évaluation des préventions, des programmes de contrôle et de lutte. Comprendre le phénomène de transmission des virus dans les populations humaines est une question fondamentale en épidémiologie. Les modèles traditionnels épidémiologiques supposent que la transmission d'une infection dans une population hôte homogène augmente bien avec le nombre d'individus et avec un risque d'infection au hasard (Anderson et May, 79) (May 97). Dans notre étude, on admet que la transmission d'un virus se fait par un contact direct entre un susceptible et un infecté. Récemment, la recherche a reconnu l'importance de l'étude du comportement de l'hôte (individu) et des moyens de contact entre les hôtes dans la transmission de parasites. Les deux concepts : comportement de l'individu et degré de contact entre les individus dans une population seront influencés par la structuration sociale de la population.

En épidémiologie et dans les études des maladies infectieuses, nous cherchons à identifier les agents infectieux et à comprendre leur mode de propagation. Un certain nombre de questions se pose : Comment l'épidémie se propage d'un individu vers un autre ? Quels sont les facteurs qui favorisent cette transmission ? Peut-on prédire l'apparition future de l'épidémie ? Répondre à ces questions nous permettra de mieux cerner le problème et de proposer une approche de modélisation adéquate.

La diffusion ou la propagation de l'épidémie est définie comme étant la croissance du nombre d'individus infectés dans le temps et dans l'espace. Il faut noter qu'une épidémie a tendance à se propager d'autant plus rapidement que chaque individu est amené à rencontrer un certain nombre d'individus (lieu du travail, établissements scolaires, voyages, le cinéma, centres commerciaux, transports en commun,. . .). La transmission de la grippe d'un individu à un autre s'effectue surtout par les aérosols émis lors des éternuements ou la toux [17]. Cette transmission sous-entend une proximité des individus mais les particules posées sur une surface restent vivantes durant un à deux jours et constituent une source de diffusion. Les réseaux de contacts sont probablement différents selon la culture, la taille de communauté et l'appartenance à un milieu rural ou urbain [36]. Les transports modernes ont tendance à aggraver les risques de contracter une épidémie.

On distingue trois types de diffusion de l'épidémie :

1. La diffusion par contagion : c'est la transmission directe interhumaine c'est-à-dire de l'agent pathogène d'un individu contaminé à un individu susceptible. L'individu est donc le seul réservoir et le seul transmetteur du virus.

2. La diffusion hiérarchique ascendante (propagation s'effectuant d'une petite ville à une grande ville) ou descendante : la contamination s'effectue de manière ordonnée. Par exemple, d'un centre métropolitain à un village distant. Elle est engendrée par l'utilisation des transports aériens, ferroviaires, maritimes. . .

3. La combinaison de la diffusion par contagion et la diffusion hiérarchique.

Suite à l'apparition d'une grippe en Chine en 1957 qui se propagea à l'échelle mondiale, les géographes Hunter et Young [44] proposent de cartographier la diffusion de l'épidémie. Ils étudient l'évolution de l'épidémie en Angleterre et au Pays de Galles. Ils constatent que le virus est peu virulent et que les personnes à risque restent les personnes âgées, les femmes enceintes et les personnes présentant des complications pulmonaires. Ce type de modèle vise à évaluer la variation de la quantité d'opportunités de relation en fonction de la position. Il est en de même dans [54] où M. Loytonen et S.I. Arbona souhaitent modéliser et prédire la diffusion de l'épidémie du SIDA à Porto Rico. Ils utilisent la méthode de régression linéaire multivariée pour analyser l'épidémie.

Une première étude sur la coqueluche dans un pays en voie de développement a été effectuée à une échelle spatiale [18]. Dans ce travail, les auteurs déterminent l'impact de l'hétérogénéité locale de la diffusion de l'épidémie de la coqueluche et sa persistance dans un environnement spatio-temporel. Ensuite, ils mettent en évidence l'impact de la taille et de la densité de population sur la diffusion ainsi que la persistance de la coqueluche en réalisant une analyse des séries temporelles.

D'autres travaux [3,19, 23, 24, 64, 65] utilisant la notion de voisinage tirée de la théorie des graphes ont été réalisés dans le but d'étudier la propagation d'une épidémie. Dans [3], J. Arino et al. décrivent un modèle où la propagation spatiale d'une épidémie peut être transmise entre divers espèces. Chacun des nœuds du graphe correspond à une espèce. Les arcs du graphe représentent les interactions entre les espèces. En se positionnant sur la notion de voisinage, les auteurs considèrent indirectement un modèle de diffusion par contagion. Ainsi, la notion de proximité est un facteur très important car le

risque qu'un individu soit contaminé est d'autant plus élevé qu'il est en interaction avec un individu infectieux. Dans [19, 65], les auteurs utilisent la notion d'automates cellulaires basée sur les approches de Von Neumann et Moore.

D'autres travaux ont couplé les notions de densité, de masse et de voisinage [15, 30]. Dans [37], les auteurs O.N. Bjornstad et al. traitent de la dynamique de l'épidémie de la rougeole. Ils ont établi un modèle statistique TSIR (Time-series Susceptible Infected Recovered) en se basant sur la taille de la population afin de produire des dynamiques endémiques et épisodiques. Ce modèle fait la transition entre les modèles théoriques et les données empiriques. Il permet de comprendre la dynamique endémique où le processus de transmission est dominant.

Par ailleurs, en épidémiologie, il faut noter que la densité de la population rapprochant les contacts entre humains influe sur la persistance et la propagation de l'épidémie [39, 40, 63]. Plusieurs études ont été réalisées dans le but de démontrer cette influence [6, 39]. La combinaison de l'épidémiologie et de l'écologie a permis de mieux comprendre la dynamique des maladies, leurs persistances et leurs transmissions.

Dans [39], B. Grenfell et J. Harwood s'intéressent à la dynamique de la métapopulation (une population au sein d'une population) en cas de maladies infectieuses et précisément dans le cas de la rougeole. De manière générale, l'analyse de la métapopulation requiert des données spatio-temporelles mais ces dernières étant difficiles à obtenir, la plupart des développements sont théoriques. Les métapopulations et les théories épidémiques sont fortement liées. Si la densité des personnes susceptibles est forte alors le nombre d'infectés croît fortement de manière exponentielle. Les modèles homogènes simples prédisent trop de mortalité, [38] d'où l'introduction d'hétérogénéité réaliste tels que le groupe d'âge, la disposition spatiale des groupes . . . dans les transmissions du virus pour remédier à ce problème.

Divers modèles ont été créés dans le but de modéliser des situations épidémiques. Nous citons ici brièvement les modèles les plus connus :

- Dans un **modèle SI**, il est considéré qu'un individu traverse deux phases durant la période épidémique. Il est susceptible (S) puis il est infectieux (I). En épidémiologie, un individu susceptible est une personne saine apte à contracter la maladie. Un individu est dit infectieux lorsqu'il peut propager autour de lui l'agent pathogène. Dans le cas d'un modèle de type SI, les chercheurs émettent une hypothèse forte : dès qu'un individu susceptible entre en contact avec un individu infectieux, il devient de suite infectieux.
- Le **modèle SIS** est tel que : l'individu est susceptible (S) puis il est infectieux (I) et ensuite il redevient susceptible (S) et il peut, durant la période épidémique, redevenir infectieux.
- Dans un **modèle SIR**, l'individu est susceptible (S) puis il est infectieux (I) et ensuite il dispose d'une immunisation (R) permanente c'est-à-dire l'immunisation est valable durant toute la période épidémique.
- Dans le **modèle SIRS**, l'individu est susceptible (S) puis il est infectieux (I) et dispose ensuite d'une immunisation (R) temporaire puis il redevient susceptible. Dans ce cas, l'immunisation est valable pour une courte durée durant la période épidémique.
- Dans le modèle **SEIR**, l'individu est susceptible (S) puis il est infecté (E) mais pas infectieux. Par la suite, il devient infectieux (I) et dispose d'une immunisation permanente (R).
- Dans le **modèle SEIRS**, l'individu est susceptible (S) puis il est infecté (E) mais pas infectieux. Par la suite, il devient infectieux (I) et dispose d'une immunisation temporaire (R) puis il redevient susceptible. Le SEIRS est un modèle SEIR avec perte d'immunité.

La période latente est considérée comme une variable non négligeable. Rappelons que la période latente ou la période d'incubation est la période en unité de temps durant laquelle un individu susceptible contracte le virus et devient infectieux. L'individu est infecté mais il ne transmet pas

encore l'agent pathogène. Cette période varie en fonction de l'âge de l'individu. Pour un adulte, on observe une moyenne de deux jours environ. L'individu est infectieux 24 heures avant l'apparition des symptômes jusqu'à 5 voire 10 jours après le début de la maladie [50]. S'il s'agit d'un enfant, il est infectieux très tôt et la durée de celle-ci est longue [41]. Les enfants peuvent être contagieux durant plus de 10 jours.

Dans [42], W.H. Hethcote et P. Van Den Driessche proposent un modèle SIS englobant les naissances, les mortalités naturelles, celles liées à la maladie et une structure démographique exponentielle. Dans ce modèle, les auteurs prennent en compte un délai pour la période infectieuse et considèrent que la taille de la population est variable. Ils émettent l'hypothèse que la période infectieuse est constante pour tous les individus. Le choix de faire varier la taille de la population se justifie du fait que les processus épidémiologiques et démographiques interagissent et entraînent de nouveaux comportements qui n'apparaissent pas lorsque la taille de la population est constante.

Dans [67], les auteurs S. Towers et Z. Feng souhaitent prédire la trajectoire de la pandémie de la grippe H1N1 et déterminer l'efficacité de la campagne de vaccination aux Etats-Unis. Ils utilisent un modèle de type SIR basé sur des données réelles provenant de l'Organisation Mondiale de la Santé (OMS) et du Centre de prévention et de contrôle des maladies des Etats-Unis (CDS US). Ils prennent en compte les effets de la campagne de vaccination et une immunité est acquise deux semaines après la prise d'une dose de vaccin ce qui entraine une diminution du nombre d'individus susceptibles. Ils supposent que 100 % de la population vaccinée obtient une immunité suite à la prise du vaccin. Leur modèle compare l'évolution de l'épidémie d'une part sans la prise de vaccins puis avec la prise du vaccin. En tenant compte de la campagne de vaccination, le modèle prédit une réduction de 6 % du nombre total de personnes infectées à la fin de l'année 2009. Une des failles de ce modèle est que la force du virus de la grippe n'a pas été étudiée et de plus, les auteurs émettent une hypothèse forte où 100 % de la population vaccinée obtient une immunité.

Dans le cadre de la modélisation dans l'article [3], les auteurs J. Arino et al. utilisent un modèle de type SEIR qui caractérise l'évolution de l'état de l'infection qu'un membre d'une espèce peut être amené à traverser durant la période épidémique.

Dans [74], Z. Zhao et al. utilisent un modèle de type SEIR ayant un délai qui est la période latente et un taux d'incidence non linéaire. Ils concluent que la période latente et la vaccination massive ont un effet sur l'éradication de l'épidémie. Flahaut et al., dans [34], utilisent un modèle stochastique de type SEIR dans le but d'analyser la diffusion de la pandémie A (H1N1) dans 52 villes.

Plusieurs modèles épidémiologiques de type SEIRS [46, 57, 53, 27, 75] ont été proposés. Dans [46], les auteurs J. Jiao et al. concluent qu'une vaccination massive ou encore une courte période de vaccination effectuée de manière périodique peut permettre d'éradiquer la maladie. Dans [53], M.Y. Li et al. se sont penchés sur les problèmes de stabilités globales pour les modèles épidémiologiques de type SEIRS. Dans [27], K.L. Cooke et P. Van Den Driessche procèdent à l'analyse d'un modèle épidémiologique de type SEIRS où la structure démographique suit une loi exponentielle, la période latente et la période où on conserve l'immunité sont des constantes. Les temps d'attente au sein de chacune des classes S, E, I, R suivent aussi une loi exponentielle. Il est important de noter que le choix d'un modèle dépend essentiellement de la maladie épidémiologique que l'on souhaite étudier. Par exemple, dans le cas de la varicelle, un modèle de type SEIRS ne serait pas adapté puisqu'un même individu ne peut avoir à deux reprises la varicelle.

Par ailleurs, l'arrivée de la mondialisation a conduit certains chercheurs à tenter une approche plus fine dans leurs modèles. Il s'agit de l'intégration d'un facteur non négligeable à prendre en considération lors de la modélisation d'une épidémie : la « mobilité ». Il s'agit de la capacité pour un individu de se déplacer d'une zone géographique à une autre quel que soit le type de transport utilisé. Il existe donc un risque non négligeable pour les voyageurs, de contracter de nouvelles pathologies pendant la durée

du transport (ou du séjour) et de les propager dans des régions à leur arrivée [66]. Cela s'illustre assez bien par l'épidémie voire la pandémie de la grippe. Selon P. Lepine, la rapidité des transports aériens a multiplié dans des proportions considérables les risques de diffusion à distance des sujets contagieux ou en incubation de la maladie contagieuse [51].

Dans [25], V. Colizza et al. étudient le rôle du réseau des transports aériens dans le modèle de diffusion globale des maladies émergentes ainsi que la fiabilité des prévisions et des scénarios où la maladie débute en tenant compte de la transmission de la maladie de façon aléatoire et des flux de mobilité. Ils utilisent un modèle épidémique stochastique incluant les bases de données de l'association des transports aériens internationaux. Ce modèle vise à étudier les connexions du réseau aérien et les caractéristiques stochastiques des dynamiques d'infection. Ils utilisent dans [26] un modèle épidémique stochastique de métapopulation considérant les données de flux de transport aérien sur des zones urbaines. Ce modèle leur permet de fournir l'évolution spatio-temporelle de la pandémie avec une analyse de sensibilité de niveau d'infectiosité du virus différent. L'utilisation thérapeutique à l'échelle mondiale de médicaments d'antiviraux pourrait atténuer l'effet de la pandémie avec un taux reproducteur pouvant atteindre 1,9 la première année. Dans ce cas, ils démontrent que plus la stratégie est coopérative, mieux on parviendra à contenir la maladie infectieuse dans toutes les régions du monde.

J.M. Epstein et al, dans [32], utilisent un modèle stochastique afin d'étudier la propagation de la pandémie de grippe, les effets de restrictions de voyage et la vaccination. Ils prennent en compte les coûts économiques qui peuvent survenir suite à une intervention. De plus, de ce modèle il ressort que réduire uniquement les voyages aériens internationaux engendrerait un petit retard au niveau de la propagation. En ajoutant à la restriction de transport aérien d'autres mesures, le retard pourrait être beaucoup plus long. Si de plus, divers pays se coordonnent pour l'établissement de mesures de contrôle, il peut y avoir une réduction significative du nombre de cas. Cependant, si les restrictions de voyage ne sont pas combinées avec d'autres mesures, cela peut induire une épidémie locale sévère. A la suite de l'apparition du syndrome respiratoire aigu sévère (SRAS) en 2003, D.M. Goedecke et al., dans [37], utilisent un modèle de type SEIR pour mettre en évidence l'impact du transport aérien sur la diffusion de l'épidémie. Ils parviennent aux mêmes conclusions. En effet, d'après leur modèle, la réduction du transport aérien engendre uniquement un léger retard au niveau de la propagation de l'épidémie. Toutefois, ce modèle fait ressortir que les mesures de restriction de voyage seules peuvent mener à une épidémie plus sévère. Ils précisent que le choix des villes a une importance capitale au niveau de la modélisation car il a une grande influence sur les résultats du modèle.

Les travaux de Cliff et Haggett [24], indiquent une pertinence de l'intégration des transports quotidiens (déplacements vers le lieu de travail, les écoles, les loisirs. . .) au sein d'une étude épidémiologique par rapport à une recherche mêlant l'utilisation des transports aériens ou des réseaux régionaux terrestres. Une étude réalisée par Beaujouan et al. [11] suite à une épidémie de bronchiolite en Ile De France fait ressortir que durant les vacances scolaires, l'épidémie régressa. Puis, à la reprise des cours, on observe une augmentation du nombre de cas. En 2007, à la fin des vacances scolaires, une grève a eu pour conséquence d'accentuer la diminution du nombre de cas.

D'après les travaux de J.C. Desenclos [29], la transmission d'un agent pathogène par voie aérienne est l'un des modes de transmission des agents infectieux. Dans le cas de la grippe, la transmission d'un agent infectieux peut être à la fois directe (d'un individu infectieux à un individu susceptible), indirecte (d'un objet infecté à un individu susceptible) et par voie aérienne (généralement par voie respiratoire).

L'environnement joue un rôle important dans l'étude d'une épidémie. Des chercheurs se sont intéressés à prendre en compte cette variable, et notamment le climat, au sein de leur modèle [58, 22, 14, 51, 35]. Dans [14], selon J.P. Besancenot, beaucoup de maladies infectieuses dépendent des conditions météorologiques et du climat. De plus, l'étude réalisée par P.T. Nastos et A. Matzarakis

[59] fait ressortir que plus la température et la masse de vapeur d'eau contenu dans l'air augmentent, plus le nombre de consultations liées aux infections respiratoires aiguës diminue. Il semble que les variations climatiques dues aux changements des saisons jouent un rôle dans l'éclosion des foyers endémo-épidémiques [51]. Effectivement, le froid a tendance à favoriser le développement de la grippe.

La formation d'une épidémie nécessite simultanément une population sensible et des conditions météorologiques telles que l'humidité, le printemps précoce, la température . . . [51]. En effet, les épidémies de grippe débutent dans les collectivités d'enfants [1]. De plus, la formation de l'épidémie requiert une probabilité supérieure à une chance sur deux que le vecteur soit en contact avec un individu susceptible. L'introduction d'un nouvel agent pathogène dans une population n'ayant jamais rencontré cet agent produit des effets dévastateurs. A l'inverse, l'apparition d'un même virus au sein d'une zone géographique engendre une diminution de la force du virus ce qui réduit l'impact au niveau de la population.

La littérature sur les études de diffusion de la grippe est très riche et a abouti à une multitude de mesures dans les différents pays pour faire face à l'épidémie. Toutefois, les effets de ces mesures sur la dynamique de l'épidémie sont moins bien connus. Ainsi, cette ignorance engendre des difficultés tant au niveau des priorités que des stratégies à adopter en cas de crise épidémique. On peut retrouver de nombreux travaux dans la thèse de BASILEU [10].

4.2 Contributions

4.2.1 Modèle SEIR-SW

Afin de modéliser le phénomène de transmission des virus grippaux dans les populations humaines et prédire l'évolution future de la maladie, nous proposons ici un modèle SEIR-SW (Susceptible, Exposed, Infected, Removed-Small World) combinant l'approche SEIR et le modèle réseau social petit monde.

Le modèle réseau social "petit-monde"

Les réseaux complexes sont présents dans de nombreux domaines aussi divers les uns que les autres : biologie, sociologie, psychologie, informatique, etc. Dans cette partie, nous utilisons les réseaux complexes pour tenir compte des interactions permettant la transmission des maladies infectieuses entre les individus. De tels réseaux permettent de prédire l'issue d'une épidémie et donc d'aider les décideurs à améliorer les politiques de santé publique. Un réseau social peut être défini comme : «un ensemble de personnes ou de groupes de personnes possédant des schémas de contacts ou d'interactions entre eux », [28]. C'est à partir de ce type de réseaux que la modélisation du monde réel a été introduite de façon empirique grâce à l'expérience de Milgram [55].

Dans un graphe, l'effet petit monde de Watts et Strogatz [71], signifie que la plupart des nœuds sont connectés par un plus court chemin à travers le réseau, et qu'il y a un effet de regroupement (clustering) signifiant qu'il y a une grande probabilité pour que deux nœuds soient connectés directement à un autre s'ils ont un nœud voisin en commun. Le réseau petit monde occupe une place intermédiaire entre réseaux réguliers et les réseaux aléatoires c.-à-d. il est ni totalement aléatoire ni parfaitement régulier, la probabilité p de recablage (rewiring) joue un rôle important dans le passage d'un type de réseau à un autre.

Dans ce travail, nous avons choisi le modèle petit monde (SW), car ce réseau est le plus proche de la réalité des contacts entre les individus ou les groupes d'individus. Il combine les deux caractéristiques d'un réseau réel : le fort coefficient de clustering et le petit diamètre. Dans le modèle de Watts-Strogatz [71], le degré de nœud est fixe pour tous les nœuds du réseau. Or, dans la réalité le nombre de contacts est différent d'un individu à un autre, cela nous amène à proposer une distribution de degré des nœuds suivant une loi de puissance [49]. Dans ce qui suit, nous formalisons l'approche de Watts-Strogatz [71] sous forme d'algorithme décrivant les étapes de génération d'un SW.

L'algorithme de génération du réseau Small World (SW)
Entrée : *Noeud* le nombre initial de noeuds qui représente le nombre d'individus (N)
Entrée : *Degré* le nombre de degré de noeud (k)
Entrée : *Recablage* le paramètre spécial de recablage (p)
Sortie : *Réseau* Small World (G)
Debut
Copier les Noeuds dans List.Noeud
/**Création d'un réseau en anneau : Réseau.Anneau* */
Si (k < (N - 1) / 2) alors
  Pour i = 0 à Taille.List.Noeud faire
  Pour j =1 à i <= k faire
    d =n (i+j)
    Connecter List.Noeud [i] avec List.Noeud [d]
  Fin pour
  Fin pour
Fin Si
/**Recâbler les arrêtes au hasard avec une probabilité* (*p*)*/
Si Réseau.Anneau= existe alors
  Pour i =1 à N faire
  Pour j = (i + 1) à (i+ k/2) faire
    Si j > N alors j =j − N
    Fin Si
    C = une variable aléatoire uniforme entre 0 et 1
    Si p > C alors
    Choisir Nœud [l] uniformément de l'ensemble des noeuds
    Déconnecter: Noeud [i] et Noeud [j]
    Créer un arrête Noeud [i] et Noeud [l]
    Fin Si
  Fin pour
  Fin pour
Fin Si
FIN

Le modèle SEIR

Le modèle mathématique pour l'épidémiologie SEIR (Susceptible-Exposed-Infected- Removed) présenté brièvement dans la partie état de l'art est une extension du modèle compartimental SIR (Susceptible- Infected - Removed) simple. Ce modèle est fréquemment utilisé pour étudier l'évolution d'une épidémie. Il est également recommandé dans le cas d'étude d'une maladie caractérisée par une période de latence. En effet, dans le cas de la grippe, le modèle SEIR est le plus approprié. Il permet de modéliser les différents statuts de la maladie et divise la population en quatre compartiments : susceptibles, exposés, infectés et retirés au cours de l'infection. Le passage d'un état vers un autre se fait selon des probabilités α, β et γ. La figure ci-dessous (Fig.2) illustre le processus SEIR.

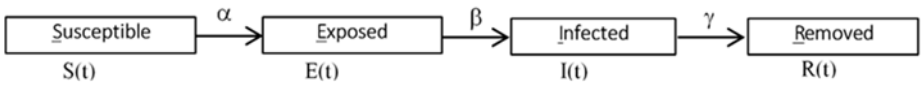


FIG. 2 – Représentation de la transmission de la maladie dans le modèle SEIR

Les paramètres α, β et γ représentent respectivement le taux de transmission de la maladie, le taux de latence et le taux de guérison.

De par leur simplicité conceptuelle, les modèles compartimentaux peuvent être aisément adaptés à plusieurs situations épidémiologiques en faisant varier le nombre de catégories dans lesquelles la population est divisée [1]. Un individu susceptible infecté par le virus de la grippe, passe par une période de latence, qui est nécessaire pour passer de l'état de contamination, à celui de contagion. Cette période précède l'apparition des symptômes. L'individu reste infecté pendant une durée qui s'appelle la période d'infection puis il passe à la phase de retiré.

Dans la présente étude, nous optons pour le modèle SEIR car durant la saison 2009-2010, un seul type de virus grippal a circulé : les virus de type A. Aucune souche de type B n'a été détectée durant la période de surveillance de la grippe saisonnière selon le bilan annuel. En effet, nous supposant que les immunisés ne redeviennent pas susceptibles car ils sont immunisés de cette souche après leurs guérison. Afin de mettre en place le modèle SEIR-SW, nous avons dû poser quelques hypothèses :

H1 : La taille de la population égale à N, supposée fixe ;

H2 : La variable temps t est de type discret, tel que $t \in T$ ou T est la durée totale;

H3 : La période de temps Δt peut représenter des jours ;

H4 : A chaque instant t, la population est partitionnée en quatre classes aléatoires : Ps : ensemble d'individus susceptibles, Pe : ensemble d'individus exposés, Pi : ensemble d'individus infectés et Pr : ensemble d'individus retirés.

H5 : Nous admettons que chaque individu susceptible dans une période Δt soit infecté puis guéri.

H6 : Un individu infecté ne peut plus redevenir susceptible.

Dans le modèle SEIR-SW, on prend en considération la notion de voisinage ainsi que la distribution de degré des nœuds qui suit une loi de puissance. L'algorithme SEIR-SW proposé se déroule comme suit:

L'algorithme SEIR-SW
Entrée : Réseau Small World (G)
Entrée : taux de transmission (α), taux de latente (β), taux de guérison (γ)
Sortie : Liste des infectés (List.infecté)
Début
List.infecté ← ϕ
Infecter quelques Nœuds
Pour I(t) ≠0 faire
Nœud.S devient Nœud.E avec une probablité $1 - (1 - \alpha)^k$
Nœud.E devient Nœud.I après une période $t_1 \sim E(1/\beta)$
Ajouter Nœud.I à List.infecté
Nœud.I devient Nœud.R après une période $t_2 \sim E(1/\gamma)$
G évolue à G'
Fin
Retourner List.infecté
FIN

Dynamique et seuil épidémique (le taux de reproduction de base)

Le taux de reproduction de base (R_0) est défini comme : « *un concept clé en épidémiologie. On le définit « heuristiquement » comme le nombre moyen de nouveaux cas d'infection, engendrés par un individu infecté moyen (au cours de sa période d'infectiosité), dans une population entièrement constituée de susceptibles*» (Sallet G. R0. EPICASA09,INRIA & IRD). Il joue un rôle très important pour la prédiction, car il est relié par les trois paramètres qui peuvent diminuer l'évolution d'épidémie. La transmission, le nombre de contact d'un individu et la période d'infectiosité (contagiosité).

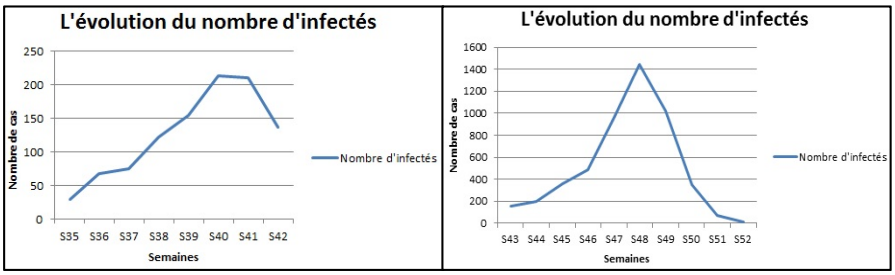
On calcule le seuil épidémique : $R0 = \beta * k * D$ avec D : période de contagiosité

Si la valeur de $R_0 < 1$: l'épidémie décroit, si $R_0 > 1$: l'épidémie s'étend.

Dans notre travail, on n'utilise pas k comme une valeur fixe mais plutôt comme une distribution de degré suivant une loi de puissance pour calculer la valeur de R_0.

Expérimentation

Ce travail rentre dans le cadre d'une étude concernant la maladie de la grippe saisonnière de l'année 2009 de la région d'Oran. Nos expérimentations sont effectuées sur des données médicales et socio-économiques obtenues de la Direction de la Santé et de la Population (DSP) d'Oran. La base de données est composée de 5504 enregistrements (déclarations) pour les 26 communes de la wilaya d'Oran. Pendant cette période la région oranaise a connu deux vagues de grippe saisonnière, la première a eu lieu entre fin Août et fin d'Octobre (S35-S42) et la deuxième a eu lieu de fin Octobre à fin de décembre (S43-S52). La figure 3 représente les deux vagues de grippe dans la population.



Vague 1 *Vague 2*

FIG. 3 – Représentation des deux vagues de grippe saisonnière en 2009

Les chercheurs ont fourni des estimations fiables pour un certain nombre de paramètres tel que la durée moyenne des périodes de latence et d'infectiosité. Cependant, il n'existe pas de mesure directe permettant de disposer de la vitesse de transmission, et très peu d'estimations de la fraction des individus initialement susceptibles dans la population [21]. Nous donnons ici des estimations de quelques paramètres à partir de notre base de données. Le **modèle SEIR-SW** dépend des huit paramètres. Certains paramètres relatifs à la population (N = nombre total de la population supposée susceptible, I_0 = nombre initialement infecté) d'autres sont relatifs au réseau small-world (k : degré d'un nœud, p : probabilité de recablage « rewiring ») et les derniers paramètres sont liés à l'infection (α : La probabilité infection, β : La probabilité de latence, γ : La probabilité de guérison).

Parmi ces paramètres du modèle SEIR-SW, la durée de latence et d'infectiosité peuvent être obtenues à partir des caractéristiques du virus de la grippe. D'autres paramètres sont fournis à partir des méthodes estimations ou à partir de la base de données réelle (BD). Le tableau 1 contient les paramètres estimés pour le modèle SEIR-SW.

Paramètres	Valeurs pour vague 1				Valeurs pour vague 2				source
Age	0-5	6-17	18-59	60+	0-5	6-17	18-59	60+	BD
N	260	170	670	510	750	1850	3200	470	estimé
I_0	2	1	1	1	52	21	75	25	BD
k	6 - 8	4-8	2- 8	2-8	2-6	2-8	2-8	2-8	Estimé
P	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	Estimé
α	+	+	+	+	+	+	+	+	Estimé
Période de latence	2	1	1	2	2	1	1	2	Valleron [2012]
Période d'infectiosité	3	3	1-4	3	1-3	2-3	1-3	1-3	Valleron [2012]

Tableau 1. Paramètres du modèle SEIR-SW

Afin de mieux comprendre les facteurs qui favorisent la transmission et prévoir l'issue finale d'une épidémie, nous avons effectué quelques études sur le taux de transmission, le degré de contact entre individus (k) et la période de contagiosité par rapport aux tranches d'âges d'individus pour les deux vagues. La figure 4, représente la variation du taux de transmission par rapport à l'âge d'individus pour les deux vagues.

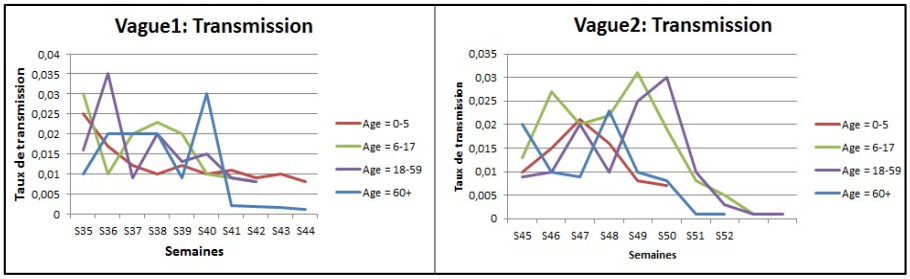


FIG. 4 – La variation du taux de transmission pendant les deux vagues pour les 4 tranches d'âge

À partir des résultats obtenus, nous constatons que le taux de transmission varie d'une tranche d'âge à l'autre et d'une vague à l'autre, pour la vague 1, il existe peu de pics par rapport à la vague 2 et le taux de transmission plus faible de la vague 2. Dans la vague1, toutes les tranches d'âges ont presque les mêmes valeurs autour de 0.01 et 0.025. La transmission s'est arrêtée dans la semaine S42 pour les tranches d'âges 6-17 et 18-59 et s'est poursuivie pour les tranches d'âges 0-5 et 60+, suite à leurs sensibilités devant la maladie (personnes fragiles). La transmission n'a pas beaucoup diminué ce qui indique une apparition future probable de la grippe**.** Dans la vague 2, la tranche d'âges 6-17 a eu un taux de transmission moins important par rapport aux autres. La transmission a diminué dans les quatre tranches d'âges ce qui indique la fin de l'épidémie. Un autre facteur qui peut aussi influencer la transmission de la maladie est le nombre de voisins pour chaque individu. La figure 5 représente la variation du nombre de degré k dans chaque tranche d'âges durant la période de la grippe.

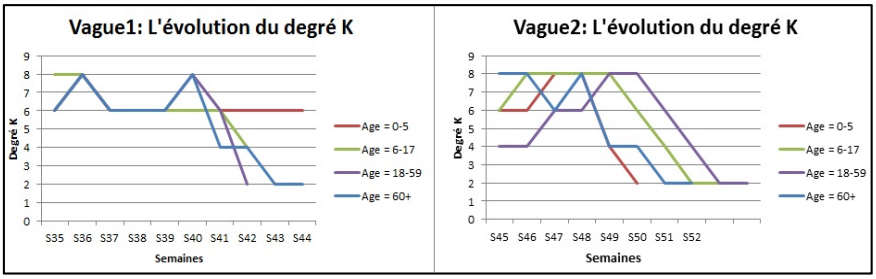


FIG. 5 – La variation du degré K pendant les deux vagues pour les 4 tranches d'âge

La différence dans la variation du nombre de contacts entre les deux graphes est très claire. Dans la 1er vague : le nombre de contacts est fixe pour la tranche d'âge 0-5 ans, la majorité de cette tranche reste chez elle et

n'exerce aucune activité. La tranche d'âge 6-18 ans, le nombre de contacts varie entre 4-8, cette tranche représente les élèves des écoles, du fondamental et des lyciens. Le nombre de contacts varie entre 2-8 pour les deux tranches d'âge 18-59 ans et 60+ ans. La première est la tranche la plus dynamique (employés, étudiants, privés, etc.) et la deuxième représente une catégorie de personnes où une partie d'entre elles exerce toujours son activité et d'autre se rencontrent dans les cafétérias ou dans les lieux publiques. Pour la vague 2 : la variation dans le nombre de contacts reste presque la même que dans la vague 1, sauf à partir de la semaine S50 où une diminution bien claire est observée. Ceci peut-être expliqué par l'arrivée des vacances d'hiver.

La période de contagiosité joue un rôle primordial dans la propagation d'épidémie, plus la durée est grande plus la maladie reste plus long temps dans la population. A partir des résultats obtenus et représentés dans la figure 6 on constate dans les deux vagues, que la durée est presque stable pour les tranches d'âge 0-5 et 60+ car ce sont des personnes sensibles à la maladie et leurs anticorps sont plus faibles par rapport aux autres tranches d'âges. Ces derniers (6-17 et 18-59 ans) avaient une période de contagiosité de 3 jours en moyenne. Ce sont des personnes qui ont une certaine défense contre le virus, nous remarquons une diminution dans cette période dans les deux dernières semaines de l'année 2009, suite aux vacances d'hiver où la majorité des personnes reste chez elle.

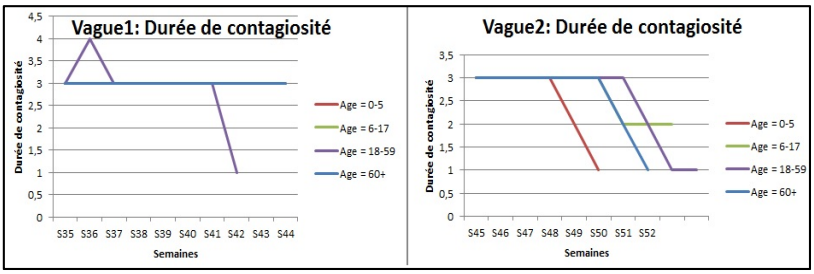


FIG. 6 – La variation de la durée de contagiosité pendant les deux vagues pour les 4 tranches d'âge

Nous rappelons ici, que la période de contagiosité et le taux de transmission ont une relation directe avec le type de virus, sa structure et ses caractéristiques.

Prédiction

L'éradication d'une infection par la vaccination peut être comprise en termes de réduction des sujets susceptibles d'être infectés en dessous d'un seuil. Cet effet est appelé «immunité collective» puisque la population peut être protégée contre les épidémies, même si il y a quelques susceptibles dans la population. Ainsi, l'élimination est théoriquement possible avec un taux de vaccination important. Nous avons estimé l'indicateur du risque épidémique R_0. Ce calcul est effectué pour quelques semaines. Le tableau 2 récapitule les résultats obtenus pour l'ensemble de la population.

Semaines	**Nombre (k)**	**Période d'infection**	**Taux de transmission**	**R_0**
S47	6	3	0.056	0.79
S48	8	3	0.07	1.37
S 51	2	2	0.01	0.06

TABLEAU 2– Représentation du taux de reproduction de base avec ses paramètres par rapport aux semaines

A partir de ce tableau, on constate que le nombre de contacts et le taux de transmission ont augmenté. De même, le taux de reproduction de base R_0 a augmenté. Nous remarquons dans la S47 un risque d'épidémie (R_0=0.79) et dans S48 une épidémie signalée (R_0>1). Dans ce cas, nous pouvons prévenir une évolution future de l'épidémie de la grippe (figure 4). Dans la semaine S51, le taux de reproduction de base a diminué à 0.06. Cela suppose que c'est la fin de l'épidémie : ce qui est confirmé dans la figure (FIG. 3).

Comme mesure de prévention, les responsables sanitaires peuvent imaginer quelques scénarios : si une population est atteinte d'une maladie, que la transmission est élevée et que la période de contagiosité est élevée dans cette population, il est recommandé d'identifier les solutions de contrôle les plus efficaces (traitement, quarantaine, vaccinations, etc.) afin de réduire son évolution. Aussi dans le cas où le nombre de contacts est élevé de cette population, il est recommandé de fermer les écoles et les lieux de rencontre. Ces mesures sont généralement prises avec la prise en compte de facteurs d'activités économiques et sociales.

Conclusion

Les maladies infectieuses transmissibles comme la grippe, tuberculose, etc. sont de complexités multiples. On peut citer d'une part la complexité de la maladie elle-même, la structure et les caractéristiques du virus, les modes de transmission de la maladie et d'autre part les structures sociales d'individus, les facteurs socio-économiques et démographiques, qui facilitent la transmission. En réponse à la problématique, dans ce travail, nous avons essayé de présenter un modèle pour la propagation de la grippe au sein de la population oranaise, ce modèle résulte de la combinaison du modèle mathématique d'épidémie SEIR et du modèle réseaux social petit monde. Les résultats fournis par le modèle SEIR-SW sont satisfaisants et représentatifs des situations réelles.

4.2.2 Modèle de diffusion des épidémies basé sur la prétopologie stochastique

4.2.2.1 Introduction

Cette partie concerne l'étude de la propagation des épidémies en proposant un modèle mathématique de réseaux stochastiques basée sur une extension de la théorie des graphes aléatoires. Il s'agit de la prétopologie stochastique qui est issue du couplage de deux théories mathématiques à savoir la prétopologie et les ensembles aléatoires. Cette approche de la simulation des comportements sociaux prend en compte deux aspects : les diverses relations sociales et leur aspect aléatoire et complexe. Ce travail est suivi de la mise en œuvre du modèle proposé par la conception et le développement d'un outil de simulation basé sur les systèmes multi-agents (SMA) et les systèmes d'information géographiques (SIG) permettant de prendre en compte l'aspect spatial. Les rôles sociaux des agents (individu, personnel médical, décideur ou autre acteur) sont ainsi aisément représentés dans cette modélisation. De plus, l'approche multi-agents permet de prendre en compte de manière simultanée les comportements individuels, les interactions entre les individus et les hypothèses dynamiques formulées dans le modèle. Des simulations ont été effectuées à partir de données épidémiologiques (GROG) et socio-économiques (INSEE).

4.2.2.2 Modèle de réseaux stochastiques

L'objectif principal de notre travail est de proposer un modèle mathématique pour décrire de manière efficace les réseaux sociaux. Cela nous semble en effet indispensable si l'on veut pouvoir simuler de manière pertinente des phénomènes de diffusion dans lesquels les comportements individuels des membres de la société interviennent fortement dans le phénomène étudié. Ceci est le cas dans les situations d'épidémie ou de pandémie.

La seule virulence du virus ne suffisant pas à expliquer le phénomène de diffusion, les contacts entre individus restent un élément primordial de cette diffusion. Ces contacts se font à travers diverses relations entre les individus et selon des modalités difficiles à prévoir. Les graphes aléatoires [33, 16, 62] se sont révélés un outil intéressant pour modéliser les réseaux sociaux complexes. Ils souffrent cependant d'une insuffisance importante

: ils ne prennent en compte, de fait, qu'une relation entre les individus et, de surcroit, la supposent la plupart du temps non orientée. Nous postulons que, dans la réalité, les individus d'une société sont reliés par plusieurs relations et que celles-ci ne sont pas nécessairement symétriques. En revanche, nous pouvons faire l'hypothèse qu'elles sont réflexives.

Nous proposons donc de développer un modèle de la prétopologie stochastique qui généralise les graphes aléatoires en prenant en compte une famille de relations binaires réflexives entre individus et en les plongeant dans un contexte stochastique. Ce modèle s'appuie sur deux théories : la prétopologie d'une part, la théorie des ensembles aléatoires d'autre part. Aussi, dans un premier temps, nous rappellerons les concepts de prétopologie nécessaires à la compréhension du modèle, et dans un second temps les définitions de base sur les ensembles aléatoires. Enfin, nous présenterons le modèle, les premiers résultats obtenus ainsi que les indicateurs utiles qui peuvent en être dérivés pour l'analyse « topologique » du réseau. Ces résultats seront ensuite associés au modèle de simulation que nous proposons.

4.2.2.3 Espaces prétopologiques

Étant donné un ensemble E non vide, on définit deux applications $a(.)$ et $i(.)$ de $P(E) \rightarrow P(E)$ de la manière suivante :

Définition 1

La fonction adhérence $a(.)$ est telle que :

(P1) : $a(\emptyset) = \emptyset$

(P2) : $\forall\, A \subset E,\ A \subset a(A)$

Définition 2

La fonction intérieur $i(.)$ est telle que :

$i(E) = E$

$\forall\, A \subset E,\ i(A) \subset A$

Définition 3

On appelle espace prétopologique le triplet $(E, i(.), a(.))$

Les fonctions $i(.)$ et $a(.)$ peuvent être munies des propriétés supplémentaires permettant de définir des espaces prétopologiques particuliers. Considérons pour l'adhérence les propriétés suivantes :

(P3) $\forall\, A \subset E,\ \forall\, B \subset E,\ A \subset B \implies a(A) \subset a(B)$

(P4) $\forall\, A \subset E,\ \forall\, B \subset E,\ a(A \cup B) = a(A) \cup a(B)$

(P5) $\forall\, A \subset E,\ a(A) = \bigcup_{x \in A} a(\{x\})$

(P6) $\forall\, A \subset E,\ a\big(a(A)\big) = a(A)$

Définition 4 : $(E, i(.), a(.))$ est dit un espace de type V si et seulement si $a(.)$ vérifie les propriétés P1, P2 et P3.

Définition 5 : $(E, i(.), a(.))$ est dit un espace de type V_D si et seulement si $a(.)$ vérifie les propriétés P1, P2 et P4.

Définition 6 : $(E, i(.), a(.))$ est dit un espace de type V_S si et seulement si $a(.)$ vérifie les propriétés P1, P2 et P5.

Définition 7 : $(E, i(.), a(.))$ est un espace topologique si et seulement **si** $a(.)$ vérifie les propriétés P1, P2, P4 et P6.

A l'inverse de ce qui se passe en topologie mathématique, en prétopologie, les applications $a(.)$ et $i(.)$ ne sont pas supposées idempotentes. En effet, l'application $a(.)$ permet de modéliser tout phénomène d'extension ou de diffusion : $a(A)$ peut s'interpréter comme la zone influencée par A ou comme la zone atteinte à partir de A dans un phénomène de diffusion. La non idempotence de $a(.)$ permet donc de suivre pas à pas le phénomène de diffusion qu'elle modélise : $a(A)$ représente la première étape du processus, $a^2(A) = \mathrm{a(a(A))}$ la seconde étape du processus, et ainsi de suite.

Compte tenu, de leur importance dans les applications, nous allons nous intéresser plus particulièrement aux espaces de type V. En effet, ceux-ci sont suffisamment généraux pour pouvoir être utilisés dans une multitude d'applications et la propriété P3 qu'ils possèdent permet de définir de manière utile le concept de voisinage.

Définition 8 :

Pour tout $x \in E$, on dit que V, $V \subset E$ est un voisinage de x si et seulement si $x \in i(V)$. L'ensemble des voisinages de x est alors noté $\boldsymbol{V}(x)$.

Comme nous le constatons dans [10], la prétopologie se révèle un outil tout à fait adapté à l'analyse «topologique» des réseaux. Cependant, nous avons constaté qu'elle ne suffit pas pour prendre pleinement en compte des phénomènes liés à la diffusion d'une épidémie au sein d'un réseau. En effet, de multiples facteurs peuvent influencer la structure même du réseau, facteurs qui ne sont pas nécessairement contrôlables. C'est le cas notamment des comportements des membres du réseau qui sont souvent non prédictibles et peuvent obéir à des logiques diverses et non connues de l'observateur. Ces logiques peuvent amener les membres du réseau à modifier leurs connections avec d'autres membres, donc à modifier aléatoirement, pour l'observateur, la structure du réseau.

Face à cette interrogation, nous avons jugé nécessaire d'introduire un modèle stochastique au sein de la prétopologie, de manière à prendre en compte cette incertitude sur la structure du réseau pour proposer un modèle de réseau stochastique, généralisant le concept de graphes aléatoires.

Pour cela, nous proposons la prétopologie stochastique qui est fondée sur la prétopologie d'une part et sur des résultats connus relatifs aux ensembles aléatoires. Nous rappelons les éléments nécessaires sur les ensembles aléatoires dans le paragraphe suivant avant de proposer le modèle de prétopologie stochastique.

4.2.2.4 Les ensembles aléatoires

La définition des ensembles aléatoires, généralisation de la définition d'une variable aléatoire, utilise le concept de mesurabilité puisqu'un ensemble aléatoire est une correspondance mesurable.

Etant donnés une population finie E de n individus, et un espace probabilisé $(\Omega, \mathrm{A}, \mathrm{p})$, on considère l'opérateur $\mathcal{R}(.)$ tel que :

$$\mathcal{R}(.): (\Omega, \mathrm{A}, \mathrm{p}) \rightarrow \mathcal{R}(E)$$

Où $\mathcal{R}(E)$ est l'ensemble des relations binaires sur E.
Par définition, $\mathcal{R}(E)$ est la famille des sous-ensembles de $E \times E$. On suppose que $\mathcal{R}(.)$ Est un ensemble aléatoire, c-a-d une correspondance mesurable de $(\Omega, \mathrm{A}, \mathrm{p})$ dans $E \times E$.

Définition 9 : $\mathcal{R}(.)$ est appelé opérateur de graphe stochastique.

Définition 10 : un réseau est défini comme une famille $\{R_i\}, i = 1, .., p$ de relations binaires sur E.

Définition 11 : Un réseau stochastique est une famille $\{R_i\}, i = 1,..,p$ d'opérateurs de graphes aléatoires.

Pour plus de détails et exemples, on peut consulter [48], [9]

Modélisation

Le modèle de la prétopologie stochastique proposé est basé sur les concepts de la prétopologie et ceux relatifs aux ensembles aléatoires. Il sera développé dans le cadre de la donnée initiale d'une famille de relations binaires réflexives définie sur un ensemble fini E. Il peut être défini dans d'autres cadres [12, 13, 52], mais nous nous limiterons à la présentation proposée compte tenu du domaine d'application.

Dans toute la suite, nous considérons un espace probabilisé $(\Omega, \mathrm{A}, \mathrm{p})$, un ensemble E non vide de cardinal fini n. Pour tout i = 1, ..., p, pour tout $\omega \epsilon \Omega$, $\mathcal{R}_i(\omega)$ désigne une relation binaire réflexive définie sur E. $\mathcal{R}(\mathrm{E})$ désigne l'ensemble de toutes les relations binaires définies sur E.

Tout élément $\mathcal{R}_i(\omega)$ de $\mathcal{R}(\mathrm{E})$ est caractérisé par la correspondance suivante :

$$\Gamma_i(\omega, x) = \{y \epsilon E \ / \ x \mathcal{R}_i(\omega) \mathrm{y}\}$$

E est muni de la topologie discrète T_d

Par conséquent, $\forall x, \Gamma_i(.,x)$ est un ensemble aléatoire si et seulement si :

$$\forall F, F \epsilon \boldsymbol{F}(E), \{\omega \epsilon \Omega / \Gamma(\omega) \cap \mathrm{F} \neq \emptyset\} \in A$$

Dans notre cas, E étant de cardinal fini, cette condition se réduit à :

$$\forall F, F \epsilon \wp(E), \{\omega \epsilon \Omega / \Gamma(\omega) \cap \mathrm{F} \neq \emptyset\} \in A$$

Nous supposons par la suite que $\Gamma(.,x)$ est un ensemble aléatoire pour tout xϵE

Propriété 1

Étant donné un espace probabilisé $(\Omega, \mathrm{A}, \mathrm{p})$ et une relation binaire réflexive aléatoire $R(.)$ définie sur$(\Omega, \mathrm{A}, \mathrm{p})$, l'application définie par

$$\forall A \epsilon P(E), \forall \omega \epsilon \Omega, \mathrm{a}(\omega, \mathrm{A}) = \{x \epsilon \mathrm{E} / \Gamma(\omega, \mathrm{x}) \cap \mathrm{A} \neq \emptyset\}$$

est une fonction d'adhérence et également un ensemble aléatoire.

Définition 12 :

On appelle réseau stochastique complexe défini sur un ensemble E fini, la donnée d'une famille de relations binaires aléatoires $\{R_i\}, i = 1,..,p$ définies sur $(\Omega, \mathrm{A}, \mathrm{p})$ et à valeurs dans l'ensemble des relations binaires réflexives définies sur E.

Propriété 2

Étant donné un espace probabilisé $(\Omega, \mathrm{A}, \mathrm{p})$, l'application définie par :

$$\forall A \epsilon P(E), \forall \omega \epsilon \Omega, \mathrm{a}(\omega, \mathrm{A}) = \{x \epsilon \mathrm{E} \ / \ \exists \mathrm{i} = 1,..\mathrm{p}, \Gamma(\omega, \mathrm{x}) \cap \mathrm{A} \neq \emptyset\}$$

est un ensemble aléatoire.

En effet, Il suffit d'utiliser le résultat précédent et de remarquer que l'application $\forall A \epsilon E$, $a(.,A)$ est une réunion d'ensembles aléatoires, donc un ensemble aléatoire.

Propriété 3

La fonction d'adhérence ainsi définie vérifie toutes les propriétés d'espace de type V.

Définition 13

Le triplet $(E, i(.), a(.))$ est tel que :

$\forall \omega \epsilon \Omega, (E, i(.), a(.))$ est un espace prétopologique de type V.

$\forall A \subset$ E, $i(., A)$ et $a(., A)$ sont des ensembles aléatoires.

Le couple $(i(.,.), a(.,.))$ est appelé prétopologie stochastique et le triplet $(E, i(.), a(.))$ est un espace prétopologique stochastique.

On peut trouver dans [10] les détails du modèle, les preuves des propriétés énoncées et des exemples d'espace prétopologique stochastique. On peut également y trouver une analyse prétopologique d'un réseau stochastique en considérant les notions d'ouverts et de fermés, telles qu'elles sont définies en prétopologie. Cette analyse permet de caractériser des cas particuliers, en utilisant le rapport d'adhérence et le rapport d'intérieur, et des cas intermédiaires.

Nous avons présenté un nouveau modèle mathématique couplant la théorie des graphes et les ensembles aléatoires appelé prétopologie stochastique. Cette dernière permet de généraliser les graphes aléatoires, notamment en considérant non pas une relation mais des familles de relations entre les individus. D'autre part, l'intégration des ensembles aléatoires nous permet de prendre en compte les facteurs incontrôlables dans le modèle.

4.2.2.5 Modèle de simulation

Le modèle mathématique présenté précédemment est mis en œuvre en s'appuyant sur un système multi-agents. L'objectif est de montrer comment ce modèle permet de construire un outil d'aide à la décision en cas d'épidémie (ou de pandémie). Il s'agit donc davantage de proposer au décideur un outil capable de lui mettre en évidence les conséquences, en matière socioéconomique, de décisions qu'il pourrait prendre plutôt que d'un outil de suivi de l'épidémie (ou de la pandémie). A travers le modèle de simulation présenté brièvement ici et développé dans [10], nous visons en effet à fournir au décideur un moyen de tester les mesures adéquates de manière à préserver les fonctions vitales de la société en cas de crise sanitaire, en plaçant l'individu et la société au cœur de la réflexion. Nous nous sommes focalisés, dans ce modèle de gestion de crise, sur les aspects relationnels entre les individus dans la logique du modèle de la prétopologie stochastique. Après une description du modèle, une première implémentation via l'outil Repast sera proposée et quelques premiers résultats de simulation commentés.

Nous avons constaté que dans le cas de la gestion d'une épidémie, les autorités disposent de plusieurs outils de simulation qui ne reflètent chacun qu'une vision globale de son évolution. Ces modèles sont assez pauvres en ce qui concerne la prise en compte des réalités de la société et des individus qui la composent. En effet la quasi-totalité des modèles existants ne distinguent pas les individus en fonction de leur rôle au sein du système « société ». Or ce dernier est capital à prendre en compte dans le phénomène de propagation de l'épidémie, mais aussi dans le fonctionnement de la société : la contamination d'un individu « ordinaire » et celle d'un médecin n'ont pas le même impact. Selon le cas, c'est un élément de la lutte contre l'épidémie duquel on se prive, aggravant ainsi le phénomène. De la même manière, si un décideur de haut rang est atteint, toute la société peut en être affectée. L'individu inséré dans son réseau social est ainsi un pilier central du modèle. Comme nous l'avons souligné, les individus sont reliés par diverses relations de tous ordres (professionnelles, amicales, loisirs. . .). Ce constat est à la base du modèle mathématique de la prétopologie stochastique qui permet, rendre simultanément en compte ces diverses relations et d'en déduire la structure topologique du réseau social ainsi modélisé.

Le modèle de simulation proposé est basé sur une approche multi-agents dans laquelle on distingue les différents types d'agents en fonction de leur rôle, trois principaux dans notre cas : l'individu (ordinaire), le décideur, le

personnel médical (médecins, infirmiers, etc.). Le modèle intègre également l'aspect spatial en étant fondé aussi sur une approche géographique par intégration d'un SIG. Nous aurons ainsi, dans cette première version, la vision au jour le jour et spatialisée, de l'état de santé des individus des trois différentes catégories. Chaque agent a la possibilité d'utiliser le moyen de transport de son choix : utilisation de transport en commun ou utilisation de transport privé. Nous disposons donc de deux matrices de déplacements. En intégrant au cours de la journée, les moments de transport entre le domicile et le lieu de travail ou autre, il est possible de comptabiliser dans les différentes périodes de la journée les contaminations qui ont lieu et de suivre ainsi, au jour le jour, en fonction des comportements des agents, la progression de l'épidémie et le niveau de charge du système de santé. Notre modèle est fondé sur deux sous-modèles en interaction : le modèle monde et le modèle individu.

Dans le modèle monde, l'environnement socio-économique englobe l'offre de soins et est représenté par des structures telles que les lieux d'enseignement, les crèches, les garderies, les lieux de soins, les lieux de travail,... Chacune de ces structures, considérée comme système multi-agents, est située dans une zone géographique (l'environnement situé de l'agent) composée d'une population qui elle, est formée d'agents.

Dans le modèle individu, un agent change d'état de santé selon le modèle SEIR présenté en début de chapitre. Ce modèle a été retenu car il représente le modèle le plus complet et le plus réaliste dans le cadre du traitement d'une épidémie de grippe. En effet, une même souche de grippe ne peut être contractée deux fois par un même agent au cours d'une même période épidémique. De plus, à la suite d'une rencontre entre deux agents dont l'un est infectieux et l'autre susceptible, il existe une période dite latente où le virus se développe au sein du corps humain. Il s'agit des 24 heures précédant l'apparition des symptômes. A ce stade, l'agent infecté ne propage pas encore autour de lui le virus. Chaque agent appartient à une famille dont nous connaissons la taille et la localisation. Il se déplace d'un lieu géographique à un autre au moyen d'un type de transport (privé ou commun). Les bases de données de l'INSEE nous permettent de connaître la matrice des flux de transport entre ces différents lieux. L'agent est caractérisé par son statut par rapport à la maladie : s'il est vacciné ou non, son état de santé (S, E, I ou R), sujet à risque (asthme, femme enceinte,…) et par d'autres données épidémiologiques. Il est également décrit par un ensemble de caractéristiques sociodémographiques (Sexe, âge, secteur d'activité,…). Les différents modes de déplacement sont pris en compte dans le modèle de manière à intégrer les possibilités de contamination dans les transports en commun.

Notre modèle est basé sur la combinaison d'un environnement physique et socio-économique. L'espace physique peut être schématisé par un graphe où les sommets constituent les zones géographiques considérées et les arcs représentent les flux de transports. Nous distinguons les flux caractérisant les transports collectifs et ceux caractérisant les transports privés. Dans ce modèle, une dizaine d'hypothèses ont été émises, ainsi qu'une trentaine d'informations issues des données du GROG ou de l'INSEE concernant les départements (nombre d'agents par type, nombre d'infectés, immunisés, susceptibles, …) pour plus de détails, voir [10]. L'unité de temps, la journée, est décomposée en 4 états : Départ de la zone i vers la zone j. Arrivée dans la zone j. séjour dans la zone j. retour vers la zone i. Séjour dans la zone i. Avant chacune de ces états, nous comptabilisons le nombre d'agents infectés. Le modèle est simplifié du fait que l'on suppose qu'il n'y a pas de contamination nocturne et que les agents empruntent le même type de transport le soir et le matin. Le processus de comptabilisation des agents est bien détaillé dans [10].

4.2.2.6 Résultats

Le modèle de simulation proposé intègre des données épidémiologiques provenant de l'institut Groupes Régionaux d'Observation de la Grippe (GROG) et des données sociodémographiques réelles issues de l'Institut National de la Statistique et des Études Économiques (INSEE). Notre démarche a été testée en procédant à trois types d'expérience. La première a consisté à simuler le fonctionnement du système sur cinq jours, avec une population donnée, dans le but de s'intéresser au calcul des indicateurs définis à partir du modèle prétopologique. La seconde a consisté à faire tourner notre modèle en intégrant trois scénarios relatifs à de possibles mesures contre la pandémie que pourrait prendre les décideurs. Dans cette expérience, nous nous sommes intéressés aux conséquences que ces mesures ont sur la distribution des personnes infectés, en d'autres termes, sur leur capacité à contenir la pandémie. La troisième et dernière expérience a pour objectif de suivre l'évolution d'une pandémie sur une population donnée, en tenant compte des différentes natures des individus qui composent cette population. Pour cela, nous avons simulé cent jours sans intégrer dans la simulation de

mesures particulières, mais en nous focalisant sur les trois grandes catégories de populations qui nous semblent importantes : les citoyens, les personnels du service de santé et les personnels à « emplois sensibles ». Nous ne présentons ici que le premier type d'expériences. On peut se reporter à la thèse de C. BASILEU [10] pour les détails.

Pour effectuer le calcul des indicateurs prétopologiques, nous avons réalisé donc une simulation de 5 jours. Chacun des agents est réparti dans une zone géographique qui est, dans ce cas, un des quatre départements choisis (Rhône, Loire, Isère et Ain) en utilisant l'outil logiciel ARCGIS. Nous disposons également d'informations sur ces départements [10] issues des données de l'INSEE. Chaque agent est dans un état de santé : S, E, I ou R. L'état de santé de l'agent est caractérisé par une couleur dans le modèle : S : vert, E : orange, I : infectieux, R : bleu. La figure 7 représente 20 agents répartis dans les quatre départements. Sur cet exemple, nous avons deux habitants sains, un habitant infectieux, un personnel médical immunisé et un emploi prioritaire sain par département. La représentation du type de l'agent est caractérisé par une forme géométrique dans le modèle : **H**abitant : rond, **P**ersonnel Médical : triangle, décideurs et **E**mplois prioritaires : carré.

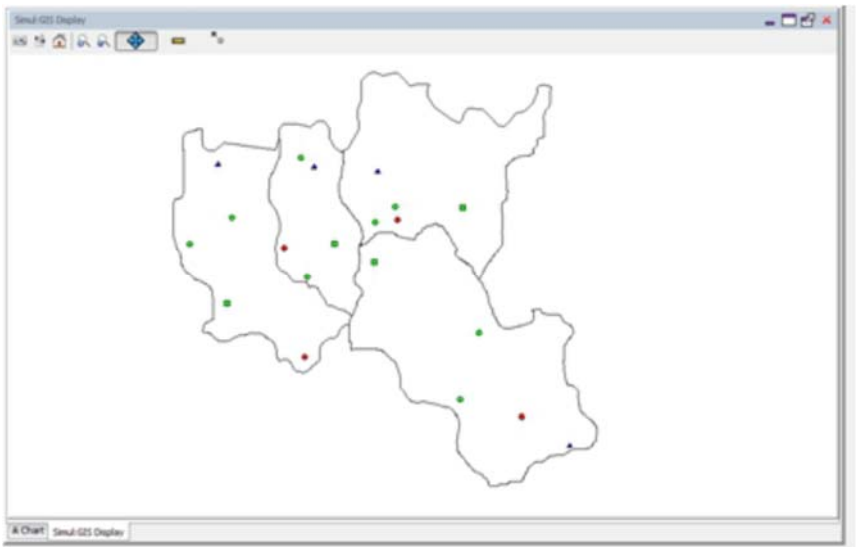


Figure 7 – Répartition des agents dans des zones géographiques

Ce modèle nécessite la comptabilisation journalière des malades. On se base essentiellement sur les flux de mobilité pour établir cette comptabilisation.

A partir de cette répartition des agents, et tenant compte des cinq types de relations (professionnelles, ménage, loisirs. . .) intégrées dans le modèle, nous obtenons, au cinquième jour, la distribution d'états des agents suivante (vert=S, orange=E, rouge=I et bleu=R) :

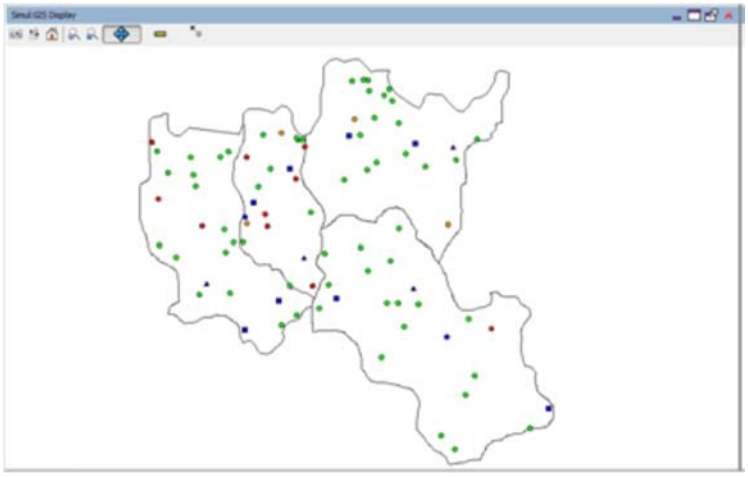


Figure 8 – Etat des agents après le cinquième jour

Les résultats obtenus sont détaillées et commentées dans [10] notamment, l'évolution des infectieux par département, l'illustration des calculs des paramètres prétopologiques, ….D'autres résultats concernant l'analyse de l'impact de mesures prise pour la gestion de crise et l'analyse de l'évolution d'une pandémie par type d'agents sont également exposés et commentés.

L'exemple illustré dans cette expérimentation montre l'avantage qu'il y a à utiliser les concepts prétopologiques pour analyser l'évolution d'un réseau social. Le premier point est que l'on dispose d'une structure formelle, de type topologique, sur l'ensemble des nœuds du réseau, ce qu'on ne peut pas réellement faire avec la théorie des graphes. Le second avantage est qu'il est possible de s'intéresser, en tant que tel, à des ensembles particuliers de nœuds et non pas à des nœuds pris isolément, tout en généralisant des paramètres caractéristiques de la structure du réseau.

Conclusion

Nous avons présenté dans cette section un nouveau modèle mathématique couplant la théorie des graphes et les ensembles aléatoires appelé prétopologie stochastique. Cette dernière permet de généraliser les graphes aléatoires, notamment en considérant d'une part des familles de relations entre les individus et non pas une relation. D'autre part, l'intégration des ensembles aléatoires nous permet de prendre en compte les facteurs incontrôlables (rencontres entre des amants. . .) dans le modèle.

Nous avons par la suite intégré la prétopologie stochastique au sein d'un outil de simulation qui couple le SIG et le SMA. Nous avons effectué diverses analyses qui, rassemblées, fournissent des éléments opportuns aux décideurs politiques tant sur la structure de la population que sur l'impact des mesures prises. Les résultats obtenus montrent clairement la capacité du modèle à tester et évaluer les mesures pouvant être prises en cas de situation de crises sanitaires liées à une épidémie de grippe.

Discussion

Les travaux présentés ici font suite à des recherches que nous avons entamées en 2005 sur la modélisation pour la gestion des crises sanitaires et pour diffusion d'épidémies en collaboration avec la préfecture du Vaucluse. Dans ces recherches, nous avons notamment travaillé sur l'analyse de la

mobilité journalière de la population entre les différentes communes et l'estimation de l'évolution de l'épidémie [7]. Nous avons proposé entre autres un modèle de diffusion basé sur une approche Markovienne [16]. Le travail présenté dans la section précédente trouve son origine dans les problèmes posés à une société lors d'une épidémie ou d'une pandémie virale. Nous avons fourni les premiers éléments de modélisation permettant des simulations de prise de mesures afin d'en évaluer l'impact, dans l'optique de mettre en place un outil d'aide à la décision efficace, prenant en compte les caractéristiques de la société et ses comportements. Dans la première partie de ce travail, nous avons généralisé l'approche proposée par F. Carrat en posant les bases de la prétopologie stochastique dans le cas discret. Cela nous a amené à prendre en compte une famille de graphes aléatoires définis sur un ensemble fini en lieu et place d'un seul graphe. Dans la seconde partie de nous avons proposé une implémentation de la modélisation prétopologique comme modélisation du réseau social sous-jacent à l'ensemble étudié. Nous avons également fourni des fonctionnalités pour simuler la diffusion d'une épidémie et tester l'impact de mesures prises. Dans la deuxième section, nous avons abordé et analysé le problème de la propagation de l'épidémie dans les réseaux sociaux par la construction d'un modèle SEIR basé sur la notion réseau social petit monde. Ceci nous a permis de mieux caractériser l'évolution de l'épidémie par la prise en compte des interactions sociales entre les individus. Ce travail, actuellement en cours, a été suivi de simulations basées sur des données épidémiologiques et socio-démographiques. A court terme nous allons travailler sur l'amélioration de cette étude en intégrant de modèle de simulation SEIR-SW dans un entrepôt de données afin d'améliorer ses fonctionnalités.

4.3 Autres travaux

- Nous avons proposé une méthodologie rigoureuse s'appuyant sur un cadre conceptuel approprié dans le but de tester la validité transculturelle des instruments de Qualité de Vie. Ce travail est motivé par l'internationalisation de la recherche clinique et le besoin grandissant d'instruments subjectifs pouvant être utilisés de manière transculturelle, la recherche transculturelle prend un caractère essentiel dans le domaine de la qualité de vie. Dans ce cadre, si les processus de traduction sont particulièrement bien étudiés, les méthodes statistiques permettant d'évaluer de manière quantitative la validité transculturelle des questionnaires sont encore assez peu connues.

- Dans le cadre de la perspective de réalisation des objectifs du millénaire pour le développement, nous avons mené une étude pour la conception d'un système d'information sanitaire et la mise en place d'une base d'indicateurs statistiques en santé publique pour le mali. Ce système d'information est capable de gérer, d'acquérir d'évaluer et de rendre visible et lisible les nombreux projets et activités qui donnent un aspect d'appartenance au secteur informel comme l'immense majorité des efforts des pays en voie de développement.
- Dans le cadre du projet européen FLURESP nous avons contribué à redéfinir les principaux scénarios d'une pandémie humaine au niveau européen, à décrire des stratégies de réponses possibles et d'évaluer ces stratégies d'intervention dans un cadre d'analyse multicritères et des analyses coût-efficacité.
- Dans le cadre du projet interdisciplinaire MOUSSON du CNRS ayant pour objectif d'élaborer un système d'alerte à la pollution de l'air, nous avons mis en place un modèle spatial de propagation de la pollution en utilisant des méthodes d'analyse spatiale (Régression, autocorrélation) basées sur la contiguïté. Nous avons également étudié les corrélations entre les lieux de pollution et certaines pathologies (maladies de la peau, des yeux, respiratoires,..) dans la région de Ouagadougou.
- Dans le cadre des projets thématiques de recherche de la région Rhône-Alpes, nous avons élaboré de règles prédictives des durées interventions dans les blocs opératoires en fonction

des paramètres ayant une influence significative sur les durées. Nous avons également mis en place une méthodologie d'évaluation des performances dans les établissements hospitaliers. Nous avons également étudié les facteurs influençant le taux d'occupation des salles dans les blocs opératoires et fourni une déclinaison et une analyse des indicateurs stratégiques dans le contexte de regroupement de plateaux médicotechniques.

4.4 Conclusion du chapitre

Nous avons présenté dans ce chapitre un aperçu des travaux mathématiques et informatiques proposant des modèles applicables dans le domaine des systèmes complexes, particulièrement en santé. Nous avons mis l'accent sur les travaux concernant la modélisation et la simulation de propagation d'épidémies, mais d'autres travaux cités brièvement dans la section 4.3, aussi importants ont été menés, même s'ils ne sont pas détaillés ici.

4.5 Références

[1] Allard A. *Modélisation mathématique en épidémiologie par réseaux de contacts*. Mémoire de maîtrise en Physique, 2008, l'Université Laval.

[2] Anderson R. M. et R.M. May (1991). *Infectious diseases of humans: dynamics and control*. 757 pages, Oxford University Press, New York.

[3] Arino J., Davis J., Hartley D., Jordan R., Miller J., Van der Driessche P. A multi-species epidemic model with spatial dynamics. 2005, http ://www.math.mcmaster.ca/arino/papers/ADHJMvdD.pdf

[4] Br Med J., A combined study group. Some aspects if the recent epidemic of influenza in Dundee. vol. 1, p. 908-913, 1958.

[5] Barthélemy M., Barrat A., Pastor-Satorras R., et A. Vespignani (2005). Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. Journal of Theoretical Biology, vol. 235, no. 2, pp. 275–288

[6] Bartlett M.S. The critical community size for measles in United States. J.R. Stat. Soc. A 123, p. 37-44, 1960

[7] Basileu C., Bounekkar A., Kabachi N., Lamure M. Pingeon J.M., Modèle de diffusion spatiale d'une pandémie basé sur les flux migratoires quotidiens. International Conference on Systems Science in Health Care, Lyon (France). 2008.

[8] Basileu C. Bounekkar A, Kabachi N, Lamure M. (2010). Vers un modèle de diffusion spatiale d'une pandémie, 4eme Colloque international de Veille Stratégique Scientifique et Technologique (VSST 2010), Toulouse

[9] Basileu C, Ben Amor S., Bui M., Bounekkar A., Kabachi N., Lamure M. Toumi M. « Stochastic networks ». 10ème journées Francophones "Extraction et Gestion des Connaissances" (EGC 2010). Hammamet, 26-29 janvier 2010. France

[10] Basileu C., Modélisation structurelle des réseaux sociaux : Application à un système d'aide à la décision en cas de crise sanitaire, Thèse de doctorat, Université Lyon 1, 2011

[11] Beaujouan L., Brun-Ney D., Chéron G., Joubert P., Midan S., Camphin P. L'épidémie de bronchiolite en île-de-France en 2005, 2006 et 2007 : impact des vacances scolaires et de la grève des transports sur le recours aux urgences. Journal Europeen des Urgences, vol. 21, numéro S1, p. A74-A75, mars 2008.

[12] Ben Amor S. Percolation, prétopologie et multialéatoires, contributions à la modélisation des systèmes complexes : exemple du contrôle aérien. LaISC - Ecole Pratique des Hautes Etudes, Juin 2008.

[13] Ben Amor S., Bonnevay S., Bui M., Lamure M. Un modèle prétopologique stochastique pour la simulation de la pollution aérienne urbaine. 9th international conference on system science in health care (ICSSHC), September 2008, Lyon.

[14] Besancenot J.P. Maladies infectieuses et climat. Méd. et mal. infect., vol. 37, no 1, p. s37-s39, 2007.

[15] Bjornstad O.N., Finkenstadt B., Grenfell B.T. Dynamics of measles epidemics : Estimating scaling of transmission rates using a time series SIR model. In Ecological Monographs, 72, p. 169-184, 2002. http ://www.zoo.cam.ac.uk/zoostaff/grenfell/research/research.htm (University of Cambridge)

[16] Bounekkar A., Lamure M., PINGEON J.M., ROY D.. « Modélisation pour la gestion d'une crise sanitaire due à une pandémie ». CALASS'07, Marseille FRANCE. 2007.

[17] Brankston G., Gitterman L., Hirji Z., Lemieux C., Gardam M. Transmission of influenza A in human beings. Lancet Infect Dis, vol. 7, p. 257-265, 2007.

[18] Broutin H., Elguera E., Simondon F., Guegan J-F. Spatial dynamics of pertussis in a small region of Senegal. In Proceeding of the Royal Society, Lond B., 271, p. 2901-2998, 2004.

[19] Brownlea A.A. Modelling the geographic epidemiology of infectious hepatitis. In Medical Geography : Techniques and Field Studies, N.D. McGlashan (Ed.), p. 279-300, 1972.

[20] Carrat F., Luong J., Lao H., Sallé A.-V., Lajaunie C. et Wackernagel H. A 'small-worldlike'model for comparing interventions aimed at preventing and controlling influenza pandemics. BMC Medecine, vol. 4, no 26, 2006.

[21] Cauchemez S, Valleron AJ, Boelle PY, Flahault A et NM Ferguson (2008). *Estimating the impact of school closure on influenza transmission from Sentinel data.* Nature 2008; 452:750–754.

[22] Chastel C. Émergence de virus nouveaux en Asie : les changements climatiques sont-ils en cause? Médecine et maladies infectueuses. vol. 34, no 11, p. 499-505, 2004.

[23] Cliff A.D. et Smallman-Raynor M.R.War Epidemics : a Geography of Infectious Diseases in Military Conflict and Civil Strife, 1850-2000. Oxford University Press, p. 800, 2004.

[24] Cliff A.D., Haggett P. et Smallman-Raynor M.R. Island Epidemics Oxford University Press, p. 563, 2000.

[25] Colizza V., Barrat A., Barthelemy M., Vespignani A. The role of the airline transportation network in the prediction and predictability of global epidemics. PNAS, vol. 103, no 7, p. 2015-2020, 14 février 2006.

[26] Colizza V., Barrat A., Barthelemy M., Valleron A.J. et Vespignani A. Modeling the worldwide spread of pandemic influenza : Baseline case and containment interventions. PLoS Medicine, vol. 4, no 1 : e13, 2007.

[27] Cooke K.L. et Van Den Driessche P. Analysis of an SEIRS epidemic model with two delays. J. Math. Biol., vol. 35, no 2, p. 240-260, 1996.

[28] Degenne,A. Forsé,M (1994), *Les réseaux sociaux*. Armand Colin, Paris.

[29] Desenclos J.C. La transmission aérienne des agents infectieux. Méd. et mal. infect., vol. 38, no 8, p. 449-451, août 2008.

[30] Dibble C. The GeoGraph 3D Computational Laboratory : Network and Terrain Landscapes for RePast. In Journal of Artificial Societies and Social Simulation, vol. 7, no 1, 2004. http://jasss.soc.surrey.ac.uk/7/1/7.html

[31] Dorjee S., Poljak Z., Revie C. W., Bridgland J., McNab B., Leger E., Sanchez J.(2013). *A Review of Simulation Modelling Approaches Used for the Spread of Zoonotic Influenza Viruses in Animal and Human Populations*. Zoonoses and Public Health, 60, 383–411

[32] Epstein J.M., Goedecke D.M., Yu F., Morris R.J., Wagener D.K. et Bobashev G.V. Controlling Pandemic Flu : The Value of International Air Travel Restrictions. PLoS ONE, vol. 2, no 5 : e401, 2007.

[33] Erdös P., Rényi A. On random graphs. Publicationes Mathematicae, vol. 6, p. 290-297, 1959.

[34] Flahault A., Vergu E., Boëlle P.Y. Potential for a global dynamic of Influenza A (H1N1). BMC Infectious Diseases, vol. 9, no 129, 2009.

[35] Gilbert M., Slingenbergh J., Xiao X. Climate change and avian influenza. Rev. sci. tech. off. int. epiz., vol. 27, no 2, p. 459-466, 2008.

[36] Glass R., Glass L.M., Beyeler W.E. et Min J.H. Targeted social distancing design for pandemic influenza. Emerg. Infect. Dis., vol. 12, no 11, 2006.

[37] Goedecke D.M., Bobashev G.V., Yu F. A stochastic equation-based model of the value of international air-travel restrictions for controlling pandemic flu. Presented at Winter Simulation Conference, Washington DC, Décembre 2007.

[38] Grenfell B.T. Chance and chaos in measles dynamics. Journal of the Royal Statistical Society, serie B, vol. 54, p383-398, 1992.

[39] Grenfell B.T. et Harwood J. (Meta)population dynamics of infectious diseases. Tree 12, p. 395-399, 1997.

[40] Grenfell B.T., Bjornstad O.N. et Kappey J. Travelling waves and spatial hierarchies in measles epidemics. Nature 414, p. 716-723, 2001.

[41] Hall C.B., Douglas R.G. Nosocomial influenza infection as a cause of intercurrent fevers in infants. Pediatrics, vol. 55, p. 673-677, 1975.

[42] Hethcote Herbert W. et Van Den Driessche P. An SIS epidemic model with variable population size and a delay. In Journal of Mathematical Biology, vol. 34, p. 177-194, 1995.

[43] Hsu C. I. et H. H. Shih (2010). *Transmission and control of an emerging influenza pandemic in a small-world airline network*. Accident Analysis and Prevention, vol. 42, no. 1, pp. 93–100.

[44] Hunter J.M. et Young J.C. Diffusion of influenza in England and Wales. In annals of the Association of American Geographers, 61, p. 637-653, 1971.

[45] info' pandémie grippale, online : http ://www.pandemie-grippale.gouv.fr.

[46] Jiao J., Chen L. et Cai S. An SEIRS epidemic model with two delays and pulse vaccination. Journal of Systems Science and Complexity, vol. 21, no 2, p. 217-225, 2008.

[47] Kleinberg J. (2000). *The Small-World Phenomenon : An Algorithmic Perspective*. In Proceedings of the 32nd ACM Symposium on Theory of Computing (STOC), pages 163–170.

[48] Lamure M. Contribution à la théorie de la multiestimation. Thèse de Doctorat de l'Université Lyon 1, 1978.

[49] Lebhar E. (2005). Algorithmes de routage et modèles aléatoires pour les graphes petits mondes. Thèse de doctorat, Ecole Normale Supérieure de Lyon.

[50] LeeKha S., Zitterkopf N.L., Espy M.J., Smith T.F., Thompson R.L., Sampathkumar P. Duration of influenza A virus shedding in hospitalized patients and implications for infection control. Infect Control Hosp Epidemiol, vol. 28, p. 1071-1076, 2007.

[51] Lepine P. Genèse et périodicité des grandes épidémies. Méd. mal. infect., vol. 1, no 9, p. 357-368, 1971.

[52] Levorato V., Ahat M. Modélisation de la Dynamique des Réseaux Complexes associée à la Prétopologie, Proceedings of ROADEF'08, p. 299-300, 2008.

[53] Li M.Y., Muldowney J.S., Van Den Driessche P. Global stability of SEIRS models in epidemiology. Can. Appl. Math. Q., vol. 7, no 4, p.409-425, 1999.

[54] Loytonen M. et Arbona S.I. Forecasting the AIDS epidemic in Puerto Rico. In Social Sciences and medicine, 42(7), p. 997-1010, 1996.

[55] Milgram S. (1967).*The small-word problem.* Psychology Today, 2, p. 60-67.

[56] Miller J. C., (2011). *A note on a paper by Erik Volz: SIR dynamics in random networks.* Journal of Mathematical Biology, vol. 62, no. 3, pp. 349–358.

[57] Mishra B.K et Saini D.K. SEIRS epidemic model with delay for transmission of malicious objects in computer network. Applied Mathematics and Computation, vol. 188, no 2, p. 1476-1482, 2007.

[58] Momas I. Épidémiologie et environnement. Revue Française des Laboratoires, vol. 2001, no 336, p. 53-58, octobre 2001.

[59] Nastos P.T. et Matzarakis A.Weather impacts on respiratory infections in Athens, Greece. Int J Biometeorol, vol. 50, p. 358-369, 2006.

[60] Pastor-Satorras R. and A. Vespignani, (2001). *Epidemic spreading in scale-free networks*. Physical Review Letters, vol. 86, no. 14, pp. 3200–3203

[61] Rizzo C., Lunelli A., Pugliese A., Bella A., Manfredi P., Scalia-Tomba G., Iannelli M., Ciofi Degli Atti M.L. (2008). *Scenarios of diffusion and control of an influenza pandemic in Italy, Epidemiol Infect*. 136(12): 1650–1657.

[62] Reed B., Molloy M. A critical point for random graphs with a given degree sequence. 1995.

[63] Rohani P., Earn D.J. et Grenfell B.T. Impact of immunisation on pertussis transmission in England and Wales. Lancet 355, p. 285-286, 2000.62

[64] Santos C.B. Dos, Barbin D., Caliri A. Percolation and the epidemic phenomenon : a temporal and spatial approach of the illness spread. In Scienta Agricola, vol. 55, no 3, p. 418-427, 1998.

[65] Simoes J.M. A complex system approach to spatial epidemic. In www.conferences.unimelb.edu.au/smocs05/SMOCS_Papers/simoes.pdf, 2005.

[66] The committee to advise on tropical medicine and travel (CATMAT) and the national advisory committee on immunization (NACI). Travel, influenza and prevention. Can Com DIs Report, vol. 22, no 17, p. 141-145, 1996.

[67] Towers S, Feng Z. Pandemic H1N1 influenza : predicting the course of a pandemic and assessing the efficacy of the planned vaccination programme in the United States. In Eurosurveillance, Vol. 14, Issue 41, 15 October 2009.

[68] Vazquez,A. (2006). *Spreading dynamics on small-world networks with connectivity fluctuations and correlations*, PHYSICAL REVIEW E 74, 056101.

[69] Volz E. (2008). *SIR dynamics in random networks with heterogeneous connectivity*. Journal of Mathematical Biology, vol. 56, no. 3, pp. 293–310.

[70] Warren C. P., Sander L. M., Sokolov I., Simon C., et J. Koopman (2002). *Percolation on disordered networks as a model for epidemics*. Mathematical Biosciences, vol. 180, no. 1-2, pp. 293–305.

[71] Watts DJ et Strogatz SH (1998). *Collective dynamics of 'small-world' networks*. Nature-Jun 4;393(6684):440-2. PubMed PMID: 9623998.

[72] Yoneyama. T et Krishnamoorthy, S (2012). *Simulating the spread of influenza pandemic of 2009 considering international traffic.* Simulation 88(4): P:437-449

[73] Younsi F.Z. Bounekkar A. Hamdadou D., SEIR-SW : A model for monitoring the spread of epidemics, ASD'2014 Hammamet, Tunisie, mai 2014

[74] Zhao Z., Chen L., Song X. Impulsive vaccination of SEIR epidemic model with time delay and nonlinear incidence rate. Mathematics and Computers in Simulation, vol. 79 , no 3, p. 500-510, 2008.

[75] Zongo P. Modélisation mathématique de la dynamique de la transmission du paludisme. Thèse d'université, Ouagadougou, LANIBIO, 2009.

Chapitre 5

Curriculum Vitae

Sommaire

5.1 Renseignements administratifs

5.1.1 Etat civil

Nom : Ahmed BOUNEKKAR
Date de naissance : 8 mars 1964
Situation familiale : Marié, deux enfants
Affectation : Maître de conférences,
Université Lyon 1 – Polytech Lyon
15, Boulevard Latarjet – 69622 Villeurbanne Cédex
Téléphone 04 72 43 27 23
Adresse électronique : ahmed.bounekkar@univ-lyon1.fr

5.1.2 Cursus et diplômes obtenus

Diplôme de Doctorat en informatique, obtenu à l'université Claude Bernard Lyon 1 en juillet 1997. Cette thèse intitulée : « Analyse statistique de texture, autocorrélation spatiale et notion de contiguïté », fut effectuée au Laboratoire LASS (UMR 5823). Ce travail se situe dans le domaine de l'analyse des images numériques. L'objectif de ce travail est de fournir des modèles permettant de caractériser les textures des images numériques. Une première partie de ce travail traite la modélisation des textures par le périodogramme de Shuster. Cette approche de compression des textures est basée sur principe d'ajustement d'une fonction périodique à la ligne image (ou à une fenêtre image) observée. La deuxième partie concerne l'analyse de texture à partir de l'autocorrélation spatiale basée sur la notion de contiguïté. Des modèles spatiaux autorégressifs avec autocorrélation ont été proposés. Les travaux effectués au cours de cette thèse ont été publiés dans plusieurs communications nationales et internationales (voir **5.5.3**).

D.E.A. Ingénierie Informatique, obtenu à l'INSA de Lyon (Laboratoire d'Informatique Appliquée) en 1990. Le stage de DEA est effectué au centre de recherche de la cimenterie LAFARGE (Viviers). L'objectif du stage est de fournir un modèle d'estimation de la qualité du clinker avant sa sortie du four en fonction d'un certain nombre de paramètres (Température du four, vitesse de rotation, …).

Diplôme d'ingénieur en électronique, option contrôle des systèmes, obtenu à l'université de Constantine en 1989.

5.2. Synthèse de ma carrière et situation professionnelle actuelle

Immédiatement après la soutenance de thèse, j'ai intégré la société GROUPESOL, spécialisée en conseil et en formation, en tant qu'ingénieur en informatique. En septembre 1999, j'ai occupé un poste ATER en informatique à l'université Lyon 1. Depuis septembre 2000, j'occupe le poste de Maître de Conférences en informatique à l'université Lyon 1. Ce poste était rattaché au département d'informatique jusqu'en 2008. Je suis rattaché depuis, à Polytech Lyon, composante de l'université Lyon1. Après le laboratoire LASS UMR 5823 et le laboratoire LIRIS UMR 5205, notre équipe de recherche est rattachée depuis 2009 au laboratoire ERIC EA 3083 (Entrepôts, Représentation et Ingénierie des Connaissances). En 2009, j'ai pris la responsabilité pédagogique du Master MIAGE (Méthodes Informatiques Appliquées à la Gestion des Entreprises). Dans le cadre de cette formation j'ai mis en place deux accords entre l'université Lyon 1 et deux établissements d'enseignement supérieur à Rabat (High-Tech) et à Alger (ISTAM) pour le double diplôme MIAGE. Depuis 2012, deux centres associés créés à Rabat et à Alger accueillent des étudiants qui préparent le diplôme de MIAGE de l'université Lyon1, en suivant les contenus pédagogiques de la MIAGE. Ces étudiants sont évalués par l'équipe pédagogique de l'université Lyon 1. J'ai été élu président du Consortium International e-MIAGE (IEM) en juin 2011. Je m'occupe de la coordination de l'e-MIAGE dans 7 établissements français et 12 à l'étranger dispensant la formation MIAGE à distance.

5.3 Principaux contrats de recherche

Dans le cadre de mes recherches, j'ai travaillé sur divers projets souvent liés à des problématiques dans le domaine de la santé.

Projets thématiques de la Région Rhône-Alpes HRP, HRP2, HRP3

HRP (Hospital Resource Planning) : 2000-2003
Ce projet concerne la modélisation des connaissances, l'organisation des moyens et la gestion prévisionnelle des ressources humaines et matérielles en milieu hospitalier.
Contributions : Elaboration de règles prédictives des durées interventions dans les blocs opératoires en fonction des paramètres ayant une influence significative sur les durées. Méthodologie d'évaluation des performances dans les établissements hospitaliers.

HRP2 (Hôpitaux : Regroupement, Partage et Pilotage) : 2003-2006
HRP2 concerne l'aide à la décision, la mutualisation et le pilotage des plateaux médico-techniques dans les établissements hospitaliers de la région Rhône-Alpes.
Contributions : Etudes des facteurs influençant le taux d'occupation des salles dans les blocs opératoires. Déclinaison et analyse des indicateurs stratégiques dans le contexte de regroupement de plateaux médico-techniques.

HRP3 (Hôpitaux en Réseau, Prévoir, Partager, Piloter) : 2006-2009
Ce projet a pour but d'apporter des outils aux filières d'urgence pour faire face aux défis auxquels elles se trouvent confrontées : engorgement des services d'urgence, détournement du fonctionnement des filières dans leur ensemble, organisation parfois difficile à construire, etc. Il concerne l'étude des dimensions de la complexité de la prise en charge soins dans les filières d'urgences. Les approches proposées intègreront les problématiques liées au dimensionnement et à la configuration des réseaux, au dimensionnement des ressources humaines, à l'organisation ainsi qu'au pilotage des filières d'urgence.

Projet européen FLURESP

Le consortium FLURESP est composé des principaux experts européens dans les stratégies d'alerte contre la grippe et d'intervention, dans la santé publique et économie de la santé. La menace constante de nouveaux virus de la grippe humaine capable de provoquer une pandémie humaine impose aux Etats membres de l'UE d'élaborer des réponses efficaces et adaptées à la phase d'alerte (planification de la pandémie). L'objectif du consortium FLURESP http://www.fluresp.eu/ est de redéfinir les principaux scénarios d'une pandémie humaine au niveau européen, de décrire des stratégies de réponses possibles et d'évaluer ces stratégies d'intervention dans un cadre d'analyse multicritères et des analyses coût-efficacité, compte tenu des enseignements de la pandémie de 2009 en Europe. Si les scénarios de pandémie humaine et principales réponses connexes sont bien documentés et étudiés par des projets européens, ils n'ont jamais été évalués et classés en utilisant à la fois le côté multicritères et les approches coût-efficacité. Ensuite, il apparaît urgent d'élaborer une stratégie s'appuyant sur une évaluation des leçons apprises en ce qui concerne l'amélioration de la collaboration intersectorielle et la coordination transfrontalière dans la réponse aux urgences sanitaires. L'approche intégrée de la prise de décision proposée par le consortium FLURESP constituerait une première au niveau européen et mondial, qui viendrait appuyer les états membres à choisir la réponse la plus adéquate et la plus efficace du pouvoir public à divers scénarios de pandémie humaine.
Contributions : Collecte et analyse des données récoltées dans 5 pays européens. Comparaisons des stratégies des réponses des différents états. Analyse multicritère des stratégies des réponses en cas de pandémie. Développement d'un outil en ligne d'aide à la décision multicritères des interventions. Cet outil s'adresse à des décideurs et aux chercheurs intéressés par cette problématique.

Projet européen ECHOUTCOME

L'objectif du projet ECHOUTCOME http://www.echoutcome.eu/ est d'étudier les systèmes de santé européens afin d'évaluer les critères de prise de décision dans le cadre des besoins nationaux et les

attentes entre les États membres concernant les résultats des soins de santé et les analyses coûts-efficacité.
L'utilisation des approches descriptives et expérimentales, permet au consortium ECHOUTCOME (composé de 8 partenaires des états membres) d'étudier la relation entre la qualité des soins en fonction des coûts, de l'efficacité et de l'accessibilité en identifiant et en évaluant les approches existantes, mais avec la capacité de développer de nouvelles approches pour la prise de décision.
Contributions : Conception d'un questionnaire de qualité de vie et participation à l'analyse des données du questionnaire.

Projet CNRS MOUSSON

Ce projet interdisciplinaire concerne l'étude de la dynamique de la pollution aérienne à Ouagadougou. Il fédère des mathématiciens, des informaticiens, des météorologues, des géographes, des médecins, des chimistes et autres pour élaborer un système d'alerte à la pollution de l'air à Ouagadougou. L'objectif est de contrôler la pollution pour l'amélioration de la santé publique et pour la qualité de vie des populations. Il s'agit d'élaborer un système d'alerte sur la pollution par

- La construction d'un Système d'Information Géographique (SIG)
- La modélisation de la pollution de l'air en zone urbaine
- La Géo-simulation par la construction de scénarii de pollution à partir des informations du SIG
- La mise en place d'un système de vigies automatiques (capteurs, indicateurs de brumes sèches, feux de forêts)
- Définition et construction d'indicateurs d'alerte en fonction de la structure sociale et spatiale

Contributions : Mise en place d'un modèle spatial de propagation de la pollution en utilisant des méthodes d'analyse spatiale (Régression, autocorrélation) basées sur la contiguïté. Etude des corrélations entre les lieux de pollution et des pathologies (maladies de la peau, des yeux, respiratoires,..) dans la région de Ouagadougou.

Projet ETADAM

Cette étude porte sur l'analyse d'une base de données construite dans le cadre d'une étude intitulée « Bruit et santé en Ile-de-France ». Le but de l'enquête est de rechercher s'il existe des liens statistiques privilégiés entre un agent « stresseur » et une manifestation pathologique, ce qui permettra de mieux identifier les méfaits spécifiques du bruit, des différentes nuisances et des autres sources de stress. Ainsi il a été mis en évidence que la fréquence à laquelle était survolé le domicile des patients pouvait avoir des effets sur l'anxiété, l'appétit, le sommeil et sur la tension artérielle. L'altitude de survol agit sur la tension artérielle uniquement. Le temps passé dans les transports en commun agit sur le sommeil et sur la fréquence d'hospitalisation. L'exposition au bruit ferroviaire agit sur l'appétit et sur la tension artérielle. Et enfin, l'écoute d'un baladeur pendant plus d'une heure par jour à des effets sur la fréquence des arrêts de travail.

Projet PARAD (Patients en difficulté avec l'Alcool, à Risque, Abuseurs et Dépendants).

Le réseau PARAD est un dispositif de soins expérimental promu par l'assurance maladie et a pour but de prendre en charge des patients ayant des difficultés avec l'alcool. Les patients du réseau de soins PARAD sont de regroupés en trois catégories appelées dépendants, abuseurs et à risque. Dans le cadre de l'évaluation médio-économique du réseau, nous avons et dégagé des trajectoires patients, défini les critères d'efficacité du réseau et la métrologie associée à ces critères et enfin, estimé le coût de prise en charge d'un patient au sein du réseau

5.4 Responsabilités pédagogiques, électives et autres

- **Président du Consortium International e-MIAGE** (Depuis juin 2011). Ce Consortium fédère 7 universités françaises dispensant la formation MIAGE à distance et une douzaine centres associés dans des pays francophones

- **Responsable du Master MIAGE** (Méthodes Informatiques Appliquées à la Gestion des Entreprises) de l'université Claude Bernard Lyon 1 (Depuis septembre 2009)

- **Responsable de la VAE du Master MIAGE** de l'université Claude Bernard Lyon 1 (Depuis septembre 2009)

- **Membre du Conseil d'Administration** de l'université Lyon 1 (2010-2012)

- **Membre du conseil scientifique** de l'université Lyon 1 (2002-2006)

- **Membre du CNU section 27** (Depuis 2012)

- **Membre du Conseil du SUAS** (Service Universitaire d'Action Sociale) de l'université Lyon 1 (depuis Septembre 2014)

- **Membre du comité de direction** de Polytech Lyon (Depuis 2009)

- **Membre du comité consultatif** de l'université Lyon 1, section 27 : informatique (Depuis 2009)

- **Membre de la commission de spécialité** de l'université Lyon 1, section 27 : informatique (2001-2008)

- **Membre du comité d'organisation** "International Conference in System Sciences in Health Care: New Information Technologies and Governance of Health Systems" (2008)

- **Membre du comité scientifique de GISEH** (Gestion et Ingénierie des SystEmes Hospitaliers) (Depuis 2003

- **Membre du comité Scientifique** "International Conference in System Sciences in Health Care: New Information Technologies and Governance of Health Systems" (2008)

5.5 Publications, encadrements

5.5.1 Encadrement de Stagiaires de thèses

❖ **Ahmed KAFAFY**

« A Hybrid Evolutionary Metaheuristics based on DM-GRASP, Path-Relinking and genetic operators applied on 0/1 Multi-objective Knapsack Problems »

Ahmed BOUNEKKAR (50%), Stéphane BONNEVAY (50%),

Thèse soutenue en novembre 2013.

- **Cynthia BASILEU**

« Modélisation structurelle des réseaux sociaux : Application à un système d'aide à la décision en cas de crise sanitaire ».

Ahmed BOUNEKKAR (50%), Michel LAMURE (50%)

Thèse soutenue en décembre 2011

- **Issa Bara BERTHE**

« Analyse des systèmes de santé Malien dans la perspective de réalisation des objectifs du millénaire pour le développement »

Ahmed BOUNEKKAR (50%), Gérard DURU (50%)

Thèse soutenue en juillet 2008

- **Mounzer BOUBOU**

« Contribution aux méthodes de classification non supervisée via des approches prétopologiques et d'agrégation d'opinions »

Ahmed BOUNEKKAR (50%), Michel LAMURE (50%)

Thèse soutenue en novembre 2007

- **Antoine REGNAULT**

« Méthodes quantitatives pour l'évaluation de la validité interculturelle des instruments de mesure subjective évaluée par les patients »

Ahmed BOUNEKKAR (50%), Michel LAMURE (50%)

Thèse soutenue en septembre 2007

Thèse en cours

- **Fatima-Zohra YOUNSI**

« Vers un modèle d'évolution de la propagation d'épidémies »

Ahmed BOUNEKKAR (50%), Omar BOUSSAID (50%). Co-encadrement à l'université Lyon 2

Thèse en cours

5.5.2 Encadrements de stagiaires de Masters

J'ai encadré 13 étudiants de Master recherche dans le domaine de l'informatique, des statistiques et de la santé publique. J'ai également encadré une trentaine d'étudiants de Master Professionnel en informatique de gestion (MIAGE), en formation ingénieur informatique de Polytech Lyon et en Master Ingénierie Mathématique. Enfin, dans le cadre de la formation par apprentissage j'ai suivi 5 étudiants de Master MIAGE et 1 étudiant de la formation Ingénieur Informatique à Polytech Lyon. Le tableau ci-dessous est un extrait de la liste d'encadrements en Master Recherche.

Dorothée DEVOLFE	Analyse comparée des patients de la CMU complémentaire et du Régime Général dans le Service de consultations et de traitements dentaires de Lyon	2004	Master S3
Mona LAILA	Proposition pour la mise en œuvre d'un système d'information pour une application de télé médecine	2004	Master S3
Marie PERCEVAL	Evaluation médico-économique d'un réseau de soin pour la prise en charge des personnes alcoolo-dépendants.	2004	Master S3
Amal GHARIB	Extraction des connaissances pour l'optimisation de l'utilisation des blocs opératoires	2001	Master Informatique
Ba Huy TRAN	High Dimensional Data Analysis	2009	Master ECD (Extraction des Connaissances à partir des Données)
Rasmata TRAORE	Recours aux soins de santé au Burkina Faso: Importance de la médecine traditionnelle dans le circuit du patient	2011	Master Santé et Populations
Cécile BLEIN	Etude de l'impact des évolutions de classification sur les établissements de santé du secteur privé	2005	Master S3

5.5.3 Publications

Ouvrages

A. BOUNEKKAR, G. DURU. « Les nouvelles organisations des systèmes de santé : Nouvelles technologies de l'information, évaluation et financement » Editions Hermès-Lavoisier , janvier 2009, ISBN 978-2-7462-2315-8

A. BOUNEKKAR. Systèmes d'informations en santé. Editions Hermès-Lavoisier, Janvier 2007, 2007, ISBN 978-2-7462-1912-0

A. BOUNEKKAR, A. GUINET. « Méthodologies pour la gestion des ressources hospitalières. » Editions Hermès-Lavoisier, juin 2005, ISBN 978-2-7462-0892-6

Chapitre d'ouvrage

D. ZIGHED, R. ABDESSELAM et A. BOUNEKKAR. « Comparison of proximity measures: a topological approach » in « Advances In Knowledge Discovery and Management - Vol 3 », 2013, Springer.

Revues internationales

F. MAUNOURY, J.L. VANHILLE, N. VÉRON, A. BOUNEKKAR, B. FANTINO, J.P. AURAY, R. LAUNOIS, French gatekeeping cost-effectiveness impact on treated patients with chronic asthma, in a real way. Value in Health, 2009.

F. MAUNOURY, J.L. VANHILLE, A. BOUNEKKAR, B. FANTINO, J. AURAY, O. MOLINIER, R. LAUNOIS, Can the French general practitioner as a gatekeeper be cost-effective for managing chronic patients treated with inhaled corticosteroids? European Journal of Health Economics, 2009.

F. MAUNOURY, J.L. VANHILLE, A. BOUNEKKAR, B. FANTINO, J.P. AURAY, O. MOLINIER, R. LAUNOIS, Can the French reform be efficient for patients treated with inhaled corticosteroids? European Respiratory Journal, 2009.

A. BOUNEKKAR, M. LAMURE, « Spatial linear model with autoregression and autocorrelation. Application to study of birth rate in Rhône-Alpes commune maternities » « Health and System Science » Volume 4, 2000.

Revues nationales ou francophones

V. DESLANDRES, A BOUNEKKAR. « Spécification du système d'information hospitalier dans le cadre de regroupement d'établissements ». Santé et sytémique 10(1-2/2). 2007.

I.B. BERTHE, A. BOUNEKKAR, G. DURU. « Systèmes d'information et oralité » Santé et systémique 10(1-2/2). 2007

A. REGNAULT, C. DE LA LOGE, A. BOUNEKKAR, M. LAMURE « Le fonctionnement différentiel de l'item dans la démarche d'évaluation de la validité transculturelle des questionnaires patients » Santé et Systémique 2006, vol 9 (1-2) : 175-203 (CES 2006 : 4.23)

A. BOUNEKKAR, M. LAMURE, « Indicateurs de performances dans les établissements hospitaliers. Journal d'économie médicale », Vol.22, N.7-8, pages 393-402, Decembre 2004.

G. GAVIN, A. BOUNEKKAR, M. LAMURE « Prédiction des durées opératoires pour la planification hospitalière ». Santé et systémique, Hermès Volume 7, n° 1-2/2003.

A. BOUNEKKAR « Modélisation des pratiques médicales dans un service de chirurgie » journal Santé et systémique, pp.51-66, VOL 6/4 – 2002 Hermès

S. BONNEVAY, A. BOUNEKKAR, D. CLOT, M. EGEA, F. FESCHET, M. LAMURE, « Méthode informatique et biotechnologie », 2001, Santé et Systémique, Hermès Publication Volume 5, Numéro 1-2, pages 180-231,

Conférences internationales avec comité de lecture

A. KAFAFY, S. BONNEVAY, A. BOUNEKKAR. « A Hybrid Evolutionary Approach with Search Strategy Adaptation for Mutiobjective Optimization », Genetic and Evolutionary Computation Conference (GECCO), Amsterdam, October, 2013.

A. KAFAFY, A. BOUNEKKAR, S. BONNEVAY. « HEMH2: An Improved Hybrid Evolutionary Metaheuristics for 0/1 Multiobjective Knapsack Problems ». the 9th International Conference on Simulated Evolution And Learning, Hanoï (Vietnam), December, 2012.

A. KAFAFY, A. BOUNEKKAR, S. BONNEVAY. « Hybrid Metaheuristics based on MOEA/D for 0/1 Multiobjective Knapsack Problems : A comparative Study ». IEEE Congress on Evolutionary Computation, Brisbane (Australia), June, pages 3616-3623, 2012.

A. KAFAFY, A. BOUNEKKAR, S. BONNEVAY. « A Hybrid Evolutionary Metaheuristics (HEMH) Applied On 0/1 Multi-objective Knapsack Problems ». Genetic and Evolutionary Computation Conference, Dublin (Irlande), July, pages 497-504, 2011.

A. BOUNEKKAR. « Spatial logistic regression based upon contiguity concept ». ERCIM'08 - ERCIM Working Group on Computing & Statistics, Neuchatel, Switzerland. 2008.

A. BOUNEKKAR, M. LAMURE, « Modèle spatial pour la mesure de pollution », International Conference on Systems Science in Health Care, Lyon (France). 2008.

F. MAUNOURY, A. BOUNEKKAR, S. BONNEVAY, JP. AURAY. « Le processus de décision markovien : Adaptation des routines ». International Conference on Systems Science in Health Care, Lyon (France). 2008.

(E) C. BASILEU, A. BOUNEKKAR, N. KABACHI, M. LAMURE – Jean-Michel Pingeon. Modèle de diffusion spatiale d'une pandémie basé sur les flux migratoires quotidiens. International Conference on Systems Science in Health Care, Lyon (France). 2008.

(E) I.B. BERTHE, A. BOUNEKKAR, G. DURU « Problématique de réalisation d'un système d'information d'aide à la décision » International Conference on Systems Science in Health Care, Lyon (France). 2008.

J.M. COHEN, A. BOUNEKKAR. « An epidemiological study on noise in Paris area based on GPS practice: methods and preliminary results » 2007 WONCA EUROPE Conference, Paris. 2007.

A. REGNAULT, C. DE LA LOGE, A. BOUNEKKAR, M. LAMURE « Detection of Differential Item Functioning with Ordinal Response Format: Application of Ordinal Log-Linear Models » 5th Conference of the International Test Commission, Brussels(Belgium), July 2006.

M. BOUBOU, A. BOUNEKKAR, M. LAMURE « Clustering method based on the aggregation of preferences » 3rd world conference on Computational Statistics & Data Analysis- Limassol, Cyprus, 28-31 October, 2005

M. PERCEVAL, S. BONNEVAY, A. BOUNEKKAR, N. KABACHI, and M. LAMURE. Un réseau de soins pour la prise en charge de personnes alcoolo-dépendantes. In International Conference on System Science in Health Care, Genève (Suisse), September 2004.

B. BATRANCOURT, S. BONNEVAY, A. BOUNEKKAR, M. LAMURE, « Proposal for a modelling of brain activation in cognitive functions »; Information Processing and Management of uncertainty'2000; Madrid, Juillet 2000, 1016-1020

S. BONNEVAY, A. BOUNEKKAR, M. LAMURE, N. NICOLOYANNIS « Experimentation analysis of color representation spaces. An attempt of synthetization » International computer science conventions ICSC Symposium on neural computation NC'2000. May, 23-26, 2000 Berlin Germany.

A. BOUNEKKAR « SPATIAL AUTOREGRESSIVE PROCESSES: A STEPWISE ALGORITHM FOR SPECIFICATION ». The 7th International Conference on System Science In Health Care. 29 May-2 June, 2000. Budapest, Hungary.

A. BOUNEKKAR, M. LAMURE, N. NICOLOYANNIS « TEXTURE CLASSIFICATION BASED UPON SPATIAL AUTOCORRELATION » SPIE The International Society for Optical Engineering. Visual Communications and Image Processing '96 17 - 20 mars 1996 Orlando, Florida. USA

Conférences francophones ou nationales avec comité de lecture

[73] F.Z. YOUNSI, A.BOUNEKKAR D. HAMDADOU, SEIR-SW : A model for monitoring the spread of epidemics, ASD'2014 Hammamet, Tunisie, mai 2014

D. ZIGHED, R. ABDESSELAM et A. BOUNEKKAR. « Equivalence topologique entre mesures de proximité » 11ème journées Francophones "Extraction et Gestion des Connaissances" (EGC 2011), Brest 25 - 28 janvier 2011

C. BASILEU, A. BOUNEKKAR, N. KABACHI, M. LAMURE « Vers un modèle de diffusion spatiale d'une pandémie » VSST'2010 Colloque international Veille Stratégique Scientifique et Technologique, Toulouse, 25 - 29 octobre 2010

C. BASILEU, S. BEN AMOR, M. BUI, A. BOUNEKKAR, N. KABACHI, M. LAMURE et M. TOUMI. « Stochastic networks ». 10ème journées Francophones "Extraction et Gestion des Connaissances" (EGC 2010). Hammamet, 26-29 janvier 2010.

A. BOUNEKKAR, M. LAMURE, J.M. PINGEON, D. ROY. « Modélisation pour la gestion d'une crise sanitaire due à une pandémie ». CALASS'07, Marseille FRANCE. 2007.

M. BOUBOU, A. BOUNEKKAR, M. LAMURE. « Classification basée sur l'agrégation d'opinions par la méthode de recuit simulé ». XIVe Rencontre de la Société francophone de classification-SFC 2007 , Ecole nationale supérieure des télécommunications, Paris-France. 2007.

IB. BERTHE, A. BOUNEKKAR, G. DURU « Etude de la viabilité des centres de santé communautaire maliens en fonction des zones de pauvreté ». CALASS 2006, 4-6 octobre 2006, Milan

A. BOUNEKKAR, V. DESLANDRES, D. LEMAGNY, L. TRILLING « Etude des facteurs influençant le taux d'occupation des salles dans le contexte de regroupement de plateaux médico-techniques » Conférence GISEH 2006, 13-15 septembre 2006. Luxembourg

M. BOUBOU, A. BOUNEKKAR, D. TOUNISSOUX, M. LAMURE « Utilisation du recuit simulé pour la recherche d'une ultramétrique optimale ». 12-èmes Rencontres de la SFC Pages 71-74-Montréal mai-juin 2005

A. BOUNEKKAR, B. BREMOND, N. KABACHI, M. LAMURE, D. ROBERT. « Analyse des trajectoires patients au sein d'un établissement hospitalier » Giseh 2004, Mons Belgique septembre 2004

A. BOUNEKKAR, M. LAMURE « Evaluation des performances dans les établissements hospitaliers» ISSHC'2004 Genève, Septembre 2004

A. BOUNEKKAR « Contiguïté généralisée et lissage spatial » XVèmes jourrnées de statistiques 2-6 juin 2003 Lyon.

G. GAVIN., A. BOUNEKKAR, M. LAMURE, « Etude de l'influence de la composition de l'équipe médicale sur les durées opératoires », Conférence GISEH 2003, 17-18 Janvier, actes, 564-570. Lyon.

A. BOUNEKKAR, M. EGEA, « Télésanté et nouvelles technologies de la communication » CALASS 2002, Toledo, Espagne, Septembre 2002.

M. EGEA, A. BOUNEKKAR, « Middlewares appliqués aux données hétérogènes en Biotechnologie », XIIIème Congrès Econométrie de la santé, Louxor, Avril 2001

A. BOUNEKKAR, A. GHARIB, M. LAMURE « Acquisition des connaissances pour l'optimisation de l'utilisation des blocs opératoires. CALASS 2001 Lyon ».

Autres communications

M. LAMURE, S. BONNEVAY, A. BOUNEKKAR, N. NICOLOYANNIS « APPROCHE PRÉTOPOLOGIQUE POUR L'AIDE À LA DÉCISION DANS UN CADRE MULTIAGENTS » GDR I3 LIP6 12 mai 1999 Paris XVème

A. BOUNEKKAR M. LAMURE « AMELIORATION D'IMAGES PAR LA CONTIGUÏTE GENERALISEE » 22ème journée ISSF France. 4 Fev. 1999 Ecole des Mines, Paris.

A. BOUNEKKAR « CLASSIFICATION DES TEXTURES BASEE SUR LA CONTIGUÏTE DIRECTIONNELLE » Cinquièmes rencontres de la société francophone de classification 17-19 septembre 1997 Université Lyon II France

A. BOUNEKKAR, M. LAMURE, N. NICOLOYANNIS « ANALYSE DE TEXTURE PAR LE PERIODOGRAMME DE SCHUSTER » 18ème réunion de la section française de la Société Internationale de Stéréologie- 8 fev.1995 Ecole des Mines, Paris

A. BOUNEKKAR, M. LAMURE, N. NICOLOYANNIS « MODELE SPATIAL POUR L'ANALYSE DE TEXTURE » 17ème réunion de la section française de la Société Internationale de Stéréologie-3 Fev. 1994 Ecole des Mines, Paris

5.6 Activités pédagogiques

5.6.1 Responsabilités pédagogiques

Mes responsabilités pédagogiques concernent essentiellement la formation e-MIAGE et le consortium International e-MIAGE :

- Depuis septembre 2009, je suis chargé de la gestion pédagogique du Master MIAGE (Méthodes Informatiques Appliquées à la Gestion des Entreprises). Environ 75 étudiants sont inscrits en 2014 dans ce Master. Ces étudiants sont généralement des salariés et suivent la formation à distance avec peu de présentiel. Ces étudiants disposent d'une plateforme pédagogique où se trouvent notamment les contenus pédagogiques numériques à suivre. Ils bénéficient aussi d'un tutorat personnalisé assuré par les enseignants du Master. Environ la moitié des étudiants suivent

la formation en présentiel et sont rattachés à des centres associés situés à Rabat, Fès et Alger. Ces centres ont signé une convention de double diplôme avec l'université Lyon 1.

- Dans le cadre du Master MIAGE, je m'occupe également de la VAE MIAGE (Validation des Acquis de L'Expérience). Ma mission concerne la sélection des candidats à la procédure de VAE, de l'accompagnement de ces candidats dans leur démarche jusqu'à la soutenance et de l'organisation du jury de VAE (choix d'un jury composé d'enseignants et de professionnels, programmation de la date et présidence du jury

- Depuis juin 2011, je coordonne le Consortium International e-MIAGE qui fédère 7 universités françaises dispensant la formation MIAGE à distance. A ces 7 centres de référence s'ajoute une douzaine de centres associés dans des pays francophones. Ils sont liés par convention à ces centres de référence français. L'objectif du Consortium est de mutualiser les ressources humaines et matérielles (tuteurs, responsables nationaux des modules, organisation commune des examens, contenus pédagogiques numériques, traitement mensuel des candidatures,…). Le consortium a aussi pour mission de promouvoir la formation en France et à l'étranger, de rénover graduellement les contenus des programmes, de créer de nouveaux contenus numériques de modules en fonction des évolutions de programmes et des préconisations de la Commission Pédagogique Nationale des MIAGE de France. La formation e-MIAGE s'adresse essentiellement à des salariés, à des personnes éloignées géographiquement et à des personnes à mobilité réduite. Environ 850 étudiants sont inscrits actuellement en Licence et en Master.

- Dans le cadre de la formation MIAGE à distance, j'ai piloté un projet de mise en en place des classes virtuelles permettra aux étudiants d'être en relation (comme dans des formations présentielles) avec les acteurs de la formation. L'objectif est de d'améliorer le parcours de formation actuel en créant un cursus où s'enchaînent des cours asynchrones, des regroupements présentiels, des classes virtuelles et des tests. Il permet également aux formateurs de suivre précisément les parcours de chaque apprenant. En effet, pour une meilleure efficacité de la formation, la programmation de classes virtuelles avec un agenda précis, permet aux apprenants de sortir de leur « isolement » en ayant un peu plus d'interaction avec les tuteurs et les autres apprenants. Aujourd'hui, l'outil logiciel est opérationnel et accessible par un lien web aux enseignants chercheurs de l'université.

- Afin de veiller au bon fonctionnement des centres associés, je rends visite annuellement aux centres de Rabat et Alger. J'organise également les missions d'enseignements des enseignants de l'université Lyon 1 dans ces centres soit par des interventions sur place en présentiel soit par des outils de classes virtuelles.

- Dans le cadre du Consortium e-MIAGE, nous organisons mensuellement des réunions téléphoniques d'examen des candidatures déposées en lignes et étudiées préalablement par les directeurs des centres de référence. Les candidats retenus sont affectés aux centres e-MIAGE les plus proches.

5.6.2 Responsabilités d'Unités d'Enseignements

Unité d'Enseignement	Public concerné	Dates
Analyse données multidimensionnelles	Ingénieurs informatique Polytech Ingénieurs informatique par apprentissage	Depuis 2008
Analyse et fouille de données	Matser e-MIAGE	2006
Analyse de données, classification	Master Ingénierie Mathématique (SITN)	2001-2013
Introduction aux méthodes probabilistes et statistiques	Ingénieurs informatique Polytech Ingénieurs informatique par apprentissage	Depuis 2008
Statistiques monodimensionnelles	Ingénieurs informatique Polytech Ingénieurs informatique par apprentissage	Depuis 2008
Modélisation statistique et aide à la décision	Master Informatique	Depuis 2005
Statistiques, Traitement et analyse de l'information	Master ISM	Depuis 2010
Informatique décisionnelle	Master e-Miage	Depuis 2007
Data Mining	Master Ingénierie Mathématique (SITN)	Depuis 2013
Modélisation des systèmes complexes en santé	Master Santé et Populations	2005-2012
Séries chronologiques, processus stochastiques, modèles et applications	Master Informatique	Depuis 2009
Mathématiques pour l'informatique	Master e-Miage	Depuis 2006

5.6.3 Enseignements dispensés

Passionné par l'enseignement, j'ai effectué un grand volume d'enseignement durant la préparation de la thèse de doctorat. Les enseignements dispensés avant et après mon recrutement à l'université Lyon 1 concernent essentiellement les statistiques, l'analyse et la fouille des données. J'ai dispensé également des enseignements en algorithmique programmation et en bases des données.

Lors de mes enseignements, et dans la mesure du possible, j'oriente souvent la discussion autour de sujets ouverts qui alimentent la réflexion collective et qui poussent les étudiants à se familiariser avec des questions de recherche. D'autre part, des exemples d'applications concrets et des tests de méthodes sur des outils logiciels sont omniprésents pour préparer les étudiants au milieu professionnel. Convaincu que le système d'éducation doit évoluer avec le temps, je suis partisan du travail sur des projets en petits groupes, de la diffusion du savoir en incitant les étudiants à faire eux-mêmes des présentations et des vidéos pédagogiques. Des supports de cours, des diapositives et d'autres contenus numériques pédagogiques sont mis à disposition des étudiants.

Ces enseignements sont synthétisés dans le tableau suivant :

	Public	Lieu	Type	Années
Probabilités	L3 MIAGE	Informatique	TD, TP	1999-2007
	3ème année ingénieur Informatique	Polytech Lyon	CM, TD, TP	Depuis 2008
	3ème année ingénieur Informatique (Apprentis)	Polytech Lyon	CM, TD, TP	Depuis 2011
Statistiques inférencielles	L3 MIAGE	Informatique	TD, TP	1999-2007
	3ème année ingénieur Informatique	Polytech Lyon	CM, TD, TP	Depuis 2008
	3ème année ingénieur Informatique (Apprentissage)	Polytech Lyon	CM, TD, TP	Depuis 2011
	L3 Psychologie	Lyon 2	TD	1995-1999
Méthodes de sondages	M2 S3 (Sciences et systèmes de Santé)	Lyon 1	CM, TD	2003-2006
Analyse des données	L3	Informatique	CM, TD, TP	1999-2006
	M1 MASS (Mathématiques Appliquées Aux Sciences Sociales)		CM, TD	2000-2006
	4ème année ingénieur Informatique	Polytech Lyon	CM, TD, TP	2008-2012
	4ème année ingénieur Informatique (Apprentissage)	Polytech Lyon	CM, TD, TP	Depuis 2012
	M1 e-MIAGE	Polytech Lyon	CM,TD	Depuis 2007
	M1 DMKM (Data Mining and Knowledge Management) ERASMUS MUNDUS (Cours en anglais)	Lyon 2	CM,TD,TP	Depuis 2010
	Maîtrise de Sciences et Gestion	Lyon 3	TD	1997-1999
Méthodes de classification	M2 SITN (Statistiques, Informatique et Techniques Numériques)	Mathématiques	CM, TD, TP	2002-2012
	L3 MIASHS	Lyon 2	CM, TD, TP	Depuis 2003
Processus stochastiques	M2 MIAGE	Informatique	TD, TP	2003-2005
Datamining	M2 EQUADES	Lyon 2	CM, TD, TP	2003-2005
	M2 e-MIAGE	Polytech Lyon	CM, TD	Depuis 2008
	M2 SITN (Statistiques, Informatique et Techniques Numériques)	Mathématiques	CM, TD, TP	Depuis 2013
Bases des données	M2 S3 (Sciences et systèmes de Santé)	Lyon 1	CM, TD, TP	2003-2006
	M2 SP (Santé et Populations)	Lyon 1	CM, TD, TP	2006-2010
Réseaux de Neurones	M2 IIDEE (Ingénierie Informatique de la Décision et de l'Evaluation Economique)	Lyon 2	CM,TD	Depuis 2008
Programmation mathématique	L3 MASS (Mathématiques Appliquées Aux Sciences Sociales)	Lyon 1	CM,TD	2001-2003
Algorithmique	L3 MIAGE	Informatique	CM, TD	2001-2004
Langage Scheme	DEUG MIAS	Lyon 1	TD, TD	2001-2003
Langage Pascal	DEUG MASS	Lyon 1	TD, TP	1998-2001

Conclusion

Nous dressons ici un bilan et des perspectives des travaux de recherche post-thèse.

Nous avons présenté dans ce mémoire quelques travaux de l'état de l'art abordant des problèmes d'analyse des données dans un contexte spatial ou pour classifier les données. Nous avons aussi présenté des méthodes hybrides d'optimisation multiobjectifs et de modélisation pour l'aide à la décision. Nous avons apporté à chaque problématique nos contributions.

Concernant les problèmes de classifications, nous avons notamment traité cette question sous divers aspects : l'amélioration de méthodes existantes tels que la recherche de l'ultramétrique optimale en CAH, la proposition de nouvelles méthodes tels que la classification par agrégation des opinions, ou encore l'équivalence topologique des mesures de proximités, indispensables dans les algorithmes de classification. Le fait d'utiliser la méthode d'agrégation des opinions pour résoudre des problèmes de classification pallie à certains problèmes tels que la présence de données manquantes ou la présence de variables mixtes (qualitatives et quantitatives). D'autres contributions aux méthodes de classifications ont été apportées mais non abordées dans ce mémoire. En particulier, la classification des données hétérogènes basée sur les concepts de prétopologie et la classification basée sur l'agrégation des préférences.

Dans l'analyse des données spatiales, nous avons mis l'accent sur l'importance de la définition de la contiguïté. Des propositions de définitions ont été présentées et testées sur différents types de données : images, zones géographiques, ou observations isolées dans l'espace. La définition de la contiguïté basée sur la prétopologie est une autre manière de décrire le voisinage entre les observations spatiales. Des modèles autorégressifs avec ou sans corrélation ont également été proposé. La méthode de régression logistique spatiale est une contribution importante, qui nécessite davantage d'approfondissement, notamment sa généralisation à la régression logistique spatiale multinomiale.

Concernant le domaine d'optimisation multiobjectifs, riche en littérature, nous avons proposé plusieurs méthodes de métaheuristiques hybrides évolutionnaires. Globalement, nous avons constaté que chaque méthode proposée et comparée à des méthodes équivalentes de la littérature, présente une supériorité plus ou moins nette par rapport aux méthodes existantes. D'autres métaheuristiques restent à étudier pour améliorer la méthode HEMH2.

La problématique de l'étude et de la modélisation de la propagation des épidémies a été abordée dans ce mémoire du point de vue de la théorie des graphes, en se basant sur les recherches dans le domaine des réseaux sociaux (graphes aléatoires et Small World). D'autres contributions ont été apportés à cette problématique notamment la modélisation et la simulation de la propagation dans un domaine spatial. D'autres travaux de modélisation appliqués à la santé ont été cités.

Les travaux cités dans ce mémoire ne sont pas exhaustifs mais donnent un aperçu global sur les différents sujets abordés dans mes recherches. Par exemple, dans l'optique de l'évaluation de la validité interculturelle des instruments de mesure de la qualité de vie, nous avons proposé une procédure d'évaluation combinant des aspects méthodologiques encadrant cette évaluation. Nous avons également proposé une méthodologie aborder un aspect de la validité interculturelle, celui du Fonctionnement Différentiel de l'Item (FDI). Nous avons utilisé la modélisation log-linéaire ordinale pour traiter ce problème de FDI.