GEOXP-AGMC :
Le Couplage de Méthodes Statistiques et d'Algorithmes Evolutionnistes pour l'Analyse de Dynamiques Spatiales

Coelho Sandrine*, Berro Alain**, Duthen Yves**
** Géosignal, ** IRIT-Université de Toulouse*
Sandrine.coelho@univ-tlse1.fr, berro@univ-tlse1.fr, duthen@univ-tlse1.fr

Abstract

Ce papier présente les résultats de recherches appliquées dans le cadre d'un projet « l'homme et la société » soutenu par la Région Midi-Pyrénées. Trois laboratoires de l'Université Toulouse I Sciences Sociales, en collaboration avec la société Géosignal, ont apporté leurs compétences respectives en informatique, mathématiques, économie et statistique, pour le développement d'une plateforme de simulation et de visualisation des dynamiques spatiales : la plateforme DYNASPAT. L'objectif global est la création d'un outil d'aide à la décision pour les acteurs politiques et sociaux, mais aussi un outil de recherche pour toute discipline ayant une dimension spatiale. La plateforme permet d'étudier les tendances dans une série de données, de détecter les phénomènes d'auto-corrélation spatiale…, puis de lancer des simulations pour optimiser certains critères. Plusieurs applications réelles ont été réalisées en partenariat avec certaines institutions, notamment avec la DRAF (Direction Régionale de l'Agriculture et de la Forêt) qui sera présentée dans cet article. L'exploitation du couplage entre les bibliothèques GEOXP et AGMC permet de comprendre l'influence réciproque entre la structuration du territoire et les comportements des acteurs économiques et sociaux et de répondre à des questions de placements (localisation).

1. Introduction

De nombreux outils actuels sont spécialisés pour la visualisation ou bien le stockage ou le traitement de données. L'utilisateur est souvent obligé de manipuler un système d'informations géographiques, un outil de traitement de données, un logiciel de création de carte. L'incompatibilité des formats et l'adaptation des données en sortie de chaque outil est un problème. La plateforme DYNASPAT a été développée pour réunir plusieurs modules, permettant de stocker et d'analyser des données géoréférencées, et de proposer une visualisation adaptée et cohérente des résultats.

Dans de nombreuses disciplines de recherche, la visualisation adaptée des données, qu'elles soient brutes ou issues de calcul, reste un problème commun et récurrent. Les données ont une signification dans un contexte précis et l'être humain ne peut en analyser qu'une quantité limitée. La découverte des données repose sur des outils d'extraction et de synthèse de l'information stratégiques qui doivent résoudre des problèmes de sélection de données, de mise en conformité pour des données hétérogènes (transformation, nettoyage), d'analyse pour en extraire de nouvelles connaissances cachées, et enfin de visualisation adaptée et explicite.

En particulier, pour des données dites « géoréférencées ou géolocalisées », les Systèmes d'Informations Géographiques, SIG ou GIS en anglais, sont très utilisés comme outil de stockage, de visualisation et parfois de traitement.

Dans les paragraphes qui suivent, une première partie définit les systèmes d'informations géographiques, et leur évolution. La constatation de faiblesses des fonctionnalités pour le traitement de données, nous a conduit à développer et à adopter les solutions exposées ensuite.

2. Les SIG

2.1. Quelles fonctions ?

Le traitement des données géographiques est souvent effectué par des logiciels spécialisés appelés SIG. La première définition, américaine, émane du comité fédéral de coordination inter-agences pour la cartographie numérique (FICCDC, 1988)*:* « Système informatique de matériels, de logiciels, et de processus conçus pour permettre la collecte, la gestion, la manipulation, l'analyse, la modélisation et l'affichage de données à référence spatiale afin de résoudre des problèmes complexes d'aménagement et de gestion ». La définition française est due à l'économiste Michel Didier (1990), dans une étude réalisée à la demande du CNIG : Un système d'information géographique est un « ensemble de données repérées dans l'espace, structuré de façon à pouvoir en extraire commodément des synthèses utiles à la décision ». Il semble difficile d'en établir une, cela dépend avant tout de l'utilisation que l'on en fait. Quoiqu'il en soit, nous pouvons dire qu'il s'agit d'un système d'information qui utilise un ensemble de procédures, matériel, et logiciel permettant d'acquérir, de stocker, de structurer et de communiquer des informations spatialement localisées et des données graphiques. Il représente un environnement spatial à l'aide de primitives graphiques : des points, des vecteurs,

des polygones ; auxquelles sont associées des informations qualitatives ou contextuelles. Il est en mesure de répondre à une question posée, et de produire une réponse sous forme de carte : on parle de sémiologie graphique [1].

Dans le paragraphe qui suit, nous faisons un court historique de l'évolution des SIG.

2.2. Evolution

Les premières applications cartographiques automatiques sont apparues dans les années 1970 avec les développements informatiques [2] [3] [4] et ceux du domaine particulier de la gestion des données urbaines. Des outils de Dessin Assisté pas Ordinateur permettaient de créer des plans ou des cartes auxquels des systèmes de gestion de fichiers étaient associés. Les systèmes ont considérablement évolué grâce à la gestion des données avec l'apparition des bases de données. L'ancêtre du SIG couplait un outil de Dessin Assisté par Ordinateur et un gestionnaire de bases de données. C'est à la fin des années 1980 qu'apparaissent les premiers SIG [5] [6], s'appuyant sur des concepts orientés-objets. Les données géométriques sont associées à des données attributaires constituant des objets.

Il en existe une dizaine en France : ArcView d'ESRI, GeoConcept de la société homonyme, MapInfo importé par Axciom, StarGIS et Apic de la société STAR-APIC, GeoMedia de chez Intergraph, AutoCad Map chez AutoDesk, MicroStation de Bentley.

Les SIG sont bien souvent lourds à utiliser surtout pour des applications de type consultation, pour cette raison des développeurs se sont intéressés à la diffusion de l'information géographique sur INTERNET et INTRANET dès 1994. Les SIG évoluent lentement vers des architectures « n-tiers » intéropérables grâce au respect de normes de web-services édictées par un regroupement mondial des éditeurs, l'Open GIS Consortium (OGC) [7]. Internet joue donc un rôle de plus en plus important dans le développement des SIG, en particulier grâce à des logiciels libres. Souvent très chers, ce domaine connaît une croissance exponentielle des logiciels libres tels que Grass [8][9], Jump, Udig, QuantumGIS... [10] [11]

2.3. Les nouvelles utilisations

Le but de ces outils est non seulement de pouvoir afficher les informations géographiques sur différents référentiels, mais aussi de réaliser des croisements, d'interroger les bases, de créer des cartes synthétiques ou thématiques, ou des graphes.

Aujourd'hui les collectivités territoriales réclament des systèmes évolutifs d'extraction et d'analyse de l'information mettant l'accent sur leurs besoins spécifiques. Elles veulent pouvoir comprendre, informer, anticiper et réagir : « Savoir pour prévoir afin de pouvoir » [12]. La collecte d'informations ne suffit pas. Il faut d'une part isoler l'information utile et pertinente d'une grande masse de données, mais aussi disposer d'outils permettant d'extraire de nouvelles connaissances. Pour cela il est indispensable de réunir le savoir-faire de plusieurs champs disciplinaires : l'informatique, les sciences et technologies de l'information, l'économie, la sociologie, les statistiques et bien d'autres encore.

3. La plateforme DYNASPAT

Nous exposons ici le résultat d'un projet de recherche commun à plusieurs équipes interdisciplinaires de l'Université Toulouse I Sciences Sociales et à la société Géosignal, afin de répondre aux problèmes de traitement, d'analyse, de représentation et de visualisation adaptée de données géoréférencées.

Nous avons développé un outil complet, la plate-forme DYNASPAT (figure 1), permettant d'effectuer de l'exploration de données, de l'extraction de connaissances pertinentes, de la représentation synthétique de données, de la veille, et de l'aide à la décision.

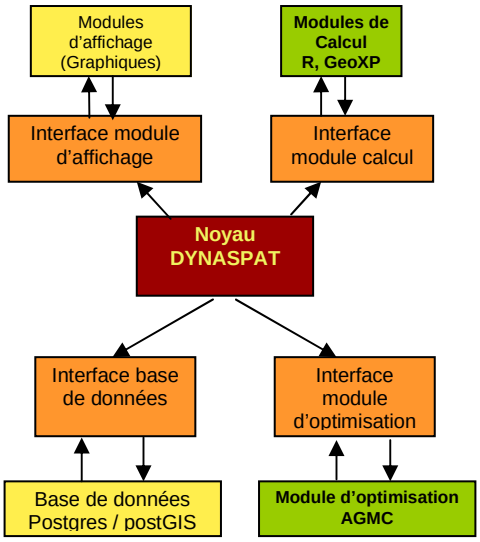

Figure 1 : Architecture de la plate-forme DYNASPAT.

Elle est constituée d'un système de plugins pour permettre l'utilisation de plusieurs modules et son évolution. Elle intègre principalement deux modules issus de la recherche que sont la bibliothèque d'Algorithmes Génétiques Multi-critères : AGMC et la bibliothèque d'analyse spatiale exploratoire de données géoréférencées, GeoXp.

Ces deux modules peuvent communiquer à travers le noyau DYNASPAT permettant d'analyser un ensemble de données puis de lancer plusieurs scénarios d'optimisation prospectifs.

4. GEOXP

Les Systèmes d'Information Géographique sont utilisés pour saisir, conserver et cartographier des données géoréférencées. Ils permettent de faire de la cartographie très évoluée mais n'intègrent pas d'outils statistiques très sophistiqués et en particulier d'outils adaptés aux données spatiales [13]. Afin de palier cette faiblesse, l'équipe statistique du GREMAQ a développé une bibliothèque d'analyse statistique exploratoire de données, GeoXp[1] [14] [15]. Ce module regroupe les fonctions aidant à l'analyse de données spatiales [16] exclusivement. Couplés à un SIG [17], ces deux modules sont complémentaires. GeoXp propose un bon nombre d'outils d'analyse exploratoire provenant de deux champs distincts, à la fois de la géostatistique et de l'économétrie spatiale, mais ne gère en aucun cas le stockage, la hiérarchisation ou le typage des données.

L'originalité de ce module provient de sa conception qui associe à tout instant carte et calcul pour une interaction spatialisée. L'utilisation des fonctionnalités de GeoXp fait apparaître deux interfaces graphiques ; l'une présentant une carte constituée des unités spatiales ayant servi au calcul et l'autre présentant le graphique. L'utilisateur peut sélectionner des unités spatiales sur la carte ce qui provoque la mise en évidence d'objets sur le graphique, et vise versa. Cette interaction en fait sa force, car en associant carte et calcul elle met en évidence des résultats et ainsi facilite la perception et l'analyse de relations complexes.

Il existe quatre catégories de fonctions interactives traitant de domaines statistiques différents :

- analyse descriptive :
  - univariée : histogramme, estimateur non paramétrique de la densité, boîte à moustaches, courbe de concentration, diagramme en barre,
  - bivariée : deux histogrammes, deux estimateurs non paramétriques de la densité, un diagramme en barre et un histogramme, des boîtes à moustaches parallèles, un nuage de points,
- analyse multivariée : analyse en composantes principales, classification ascendante hiérarchique,
- étude économétrique spatiale : diagramme de Moran, diagramme des plus proches voisins,
- géostatistique : courbe des moyennes et médianes, un nuage variographique, un diagramme avec en ordonnées la valeur absolue des différences observées entre deux sites i et j pour une variable quantitative et en abscisses l'angle entre l'axe des abscisses et le vecteur d'origine i et d'extrémité j.

Les buts principaux des méthodes proposées sont d'étudier les tendances (ou variations à large échelle) dans une variable donnée, des phénomènes d'auto-corrélation spatiale (agrégations spatiales de points semblables du point de vue de la variation étudiée, ou au contraire dissemblables), ou d'identifier des points atypiques.

Les interfaces graphiques mettent en évidence des résultats ou des informations pertinentes, en créant une unité entre les graphiques. Elle est caractérisée par l'application d'une charte de couleurs et l'utilisation de symboles, le but étant de fournir à l'analyste un maximum d'informations lors de la première lecture.

Le paragraphe suivant présente une analyse spatiale [18].

5. Problème de l'attractivité territoriale des sites scolaires d'enseignement agricole

Dans le cadre du projet de l'enseignement agricole Midi-Pyrénées, l'adaptation de l'offre de formations aux besoins de la société et des territoires demeure primordiale. Ainsi la DRAF[2] a montré un intérêt particulier à l'utilisation de la plateforme pour apporter une dimension spatiale à l'analyse de données « classique ». L'objectif est de construire des indicateurs d'attractivité des sites de formation scolaire agricole et de les intégrer dans un processus d'optimisation utilisant l'AGMC afin de déterminer le site de formation optimal pour l'ouverture d'une formation.

5.1. Données

L'étude se porte sur 44 sites d'enseignement scolaire en Midi-Pyrénées, offrant des formations allant de la 4ème à la licence professionnelle.

Dans les paragraphes qui suivent, l'analyse est faite sur un seul type de donnée, le taux de réussite. Pour mener cette analyse, un graphique de la bibliothèque GeoXp est utilisé, la boîte à moustaches.

5.2. La boîte à moustaches

Ce graphique résume une variable quantitative à partir de ses valeurs extrêmes, de ses quartiles et de sa médiane. La boîte à moustaches (figure 2) utilise 5 valeurs qui résument des données : le minimum, les 3 quartiles Q_1, Q_2 (médiane), Q_3, et le maximum.

Le quartile inférieur est la valeur du milieu du premier ensemble, dans lequel 25 % des valeurs sont inférieures à Q_1 et 75 % lui sont supérieures. Le premier quartile prend la notation Q_1. Le quartile supérieur est la valeur du milieu du deuxième ensemble, dans lequel 75 % des valeurs sont inférieures à Q_3 et 25 % lui sont inférieures. Le troisième quartile prend donc la notation Q_3. Le deuxième quartile correspond tout simplement à la médiane. L'étendue interquartile est la distance entre le premier et le troisième quartile.

[1] Dans un cadre de valorisation, un transfert a été effectué avec la société Géosignal. Elle commercialise le module GeoXpCommander, qui est une extension de la bibliothèque GeoXp traduite dans le langage R.

[2] DRAF : Direction Régionale de l'Agriculture et de la Forêt.

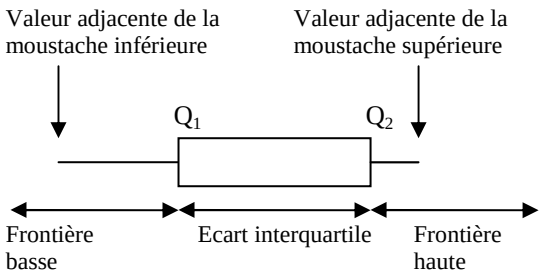

Figure 2 : Représentation d'une boîte à moustaches.

L'extrémité de la moustache inférieure est la valeur minimum dans les données qui est supérieure à la valeur frontière basse : *Q1 -1,5*(Q3-Q1).* L'extrémité de la moustache supérieure est la valeur maximum dans les données qui est inférieure à la valeur frontière haute : *Q3 +1,5*(Q3-Q1).*

L'écart interquartile est utilisé comme indicateur de dispersion. Il correspond à 50% des effectifs situés dans la partie centrale de la distribution :

Ecart Interquartile=Q3-Q1

Pour ce graphique, la sélection peut porter soit sur les points atypiques, soit sur un ou plusieurs intervalles inter-quartiles.

5.3. Résultat

Intéressons nous à la répartition spatiale des sites scolaires qui affichent les plus forts taux de réussite. Le taux de réussite de chacun des sites correspond au rapport entre le nombre d'élève n'ayant pas redoublé et le nombre total d'élèves. Il est évident que d'autres facteurs rentrent en jeu dans l'observation d'un taux de réussite faible, comme par exemple les exigences d'un établissement pour maintenir sa réputation, les conditions d'enseignement…

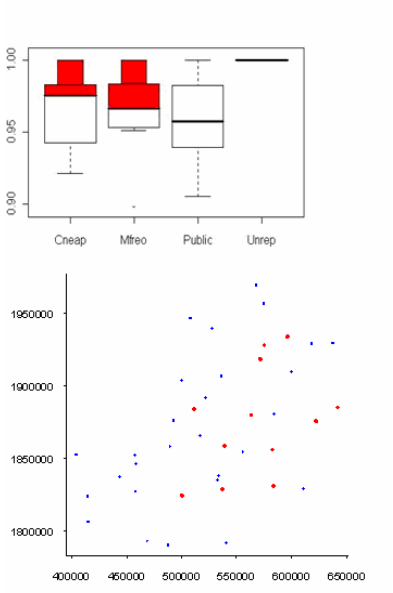

Figure 3 : Analyse du taux de réussite par site.

Le secteur privé est composé des MFREO (Maisons Familiales Rurales d'Education et d'Orientation), des sites de l'UNREP (Union Nationale Rurale d'Education et de Promotion) et des sites du CNEAP (Conseil National de l'Enseignement Agricole Privé). Les sélections faites sur le graphique (en rouge) montrent que 50% des sites privés ont un taux de réussite de plus de 97% (figure 3). Il n'est pas improbable que le taux de réussite soit un déterminant du choix du site de formation pour les élèves ou les familles.

Ici, la sélection permet d'extraire de nouvelles connaissances, l'information est visible immédiatement grâce aux couleurs et symboles employés. De nombreuses autres données ont été étudiées et d'autres types de graphique utilisés afin de construire des indicateurs d'attractivités. Des indicateurs ont été mis en évidence : le temps d'accès des élèves au lieu de formation, la capacité d'accueil des sites, l'attractivité des sites (degré de diversité des formations, indice d'attractivité des communes des sites).

Cet exemple simple montre que GeoXp est adapté à l'analyse de données géoréférencées et que les interfaces proposées sont simples et propices à la visualisation complexe et à l'extraction de connaissances nouvelles.

Ces premiers résultats ont été déterminants lors de la résolution d'un problème de géolocalisation plus complexe : déterminer le site de formation optimal pour l'ouverture d'une formation. Ce problème nécessite des outils d'optimisation évolués. Parmi ces outils d'optimisation, notre choix s'est porté sur les algorithmes évolutionnistes et en particulier sur les algorithmes génétiques.

6. AGMC

Les algorithmes évolutionnistes définis de la façon suivante : « the capability of a system to adapt its behaviour to meet its goals in a range of environments » [19], regroupe trois types : les Algorithmes Génétiques, les Stratégies d'Evolution, et la Programmation Evolutionnaire. Les Algorithmes Génétiques sont des méthodes adaptatives pour des problèmes d'optimisation, ils apportent des réponses satisfaisantes là où les méthodes classiques d'optimisation échouent face à la dimension de l'espace de recherche des solutions. Ils permettent un bon compromis entre l'exploitation et l'exploration [20].

Dès 1975, John Holland [21] les introduit en s'inspirant des capacités d'adaptation et d'évolution des espèces. Ce sont les travaux de David Goldberg [22] qui ont permis leur développement.

Nous avons intégré la bibliothèque des Algorithmes Génétiques Multicritères, « AGMC », qui regroupe les algorithmes évolutionnistes de notre équipe de recherche depuis plusieurs années. Nous ne cherchons pas ici à justifier ou valider ces méthodes d'optimisation mais à montrer l'intérêt de les coupler à un SIG.

Nous présentons dans une première partie un simple rappel du principe des algorithmes génétiques et du multi-objectifs, puis un exemple d'optimisation.

6.1. Principe

Les algorithmes génétiques sont inspirés du concept de sélection naturelle élaboré par Charles Darwin [23]. Il montre que l'apparition des espèces distinctes est fondée sur la lutte pour la vie. Il en résulte que les individus les plus adaptés, « fittest » en anglais, tendent à survivre et à se reproduire plus facilement.

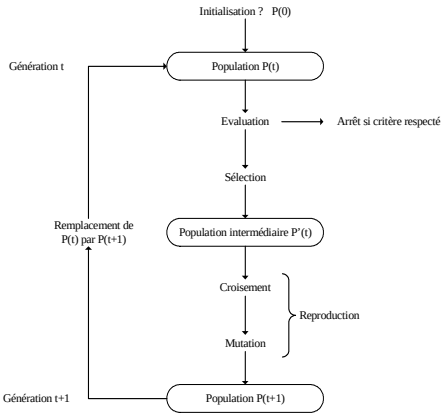

Figure 4 : Principe de fonctionnement d'un algorithme génétique.

Le principe (figure 4) est de faire évoluer une population d'individus constitués d'un génotype (son code génétique). Ce code représente les paramètres d'un problème à optimiser et est une solution potentielle de ce problème.

Cette évolution se fait à travers plusieurs étapes qui constituent une génération et qui tendent à faire survivre les individus les plus adaptés. Les individus sont d'abord évalués à l'aide d'une fonction d'évaluation qui mesure le degré d'adaptation d'un individu, appelée fitness. Certain de ces individus sont ensuite sélectionnés et on leur applique des opérateurs. Il existe deux types d'opérateur, l'opérateur de croisement et de mutation.

L'opérateur de croisement consiste à partir de deux chromosomes, à appliquer une fonction de mélange pour conserver une partie du patrimoine génétique des parents tout en créant de la diversité. Deux chromosomes enfants sont ainsi créés et ils conservent une partie du patrimoine génétique de leurs parents. Il est alors possible de créer un individu mieux adapté [24].

Le second opérateur est un opérateur de mutation qui consiste à substituer un gène par un autre. Cela correspond à modifier aléatoirement la valeur d'un élément. Cet opérateur permet d'éviter que l'évolution se fige en assurant une recherche aussi bien locale que globale selon le nombre de gènes mutés.

Une nouvelle génération d'individus est ainsi créée. Au fur et à mesure des générations successives, les individus les plus adaptés aux contraintes de leur environnement émergent.

6.2. L'optimisation multi-objectifs

Le but d'une optimisation mono-objectif est de trouver la solution optimale globale d'une fonction. Pour une optimisation multi-objectifs, il faut trouver un bon équilibre entre les solutions optimales satisfaisant au mieux, chacune, un objectif.

Soit le vecteur de fonctions objectifs noté, $f : f(x)= (f_1(x), f_2(x), \ldots, f_k(x))$ avec f_i les objectifs et k le nombre d'objectif, tel que les m contraintes notées $h_j(x)\leq 0$; j=1,...,m soient satisfaites. Le vecteur de décisions est noté $x = (x_1, x_2, \ldots, x_n)$, avec x_i; i=1,..., n les variables du problème et n le nombre de variable.

L'optimisation multi-objectifs recherche le vecteur x' telle que les contraintes $h_j(x')$ soient satisfaites et qu'il optimise f(x'). Aucune amélioration ne doit pouvoir être faite sur un critère de cette solution sans dégrader au moins la valeur d'un autre critère. Ces solutions optimales forment l'ensemble des solutions de « Pareto optimales » [25].

Notons qu'une solution A domine une solution B si et seulement si : pour tout $i \in \{1, 2, 3, \ldots, k\}$: $f_i(A) \leq f_i(B)$ et il existe $j \in \{1, 2, 3, \ldots, k\}$: $f_i(A) < f_j(B)$.

Les solutions Pareto optimales sont l'ensemble des solutions non dominées, dans l'espace des objectifs il représente le « front de pareto » (figure 5).

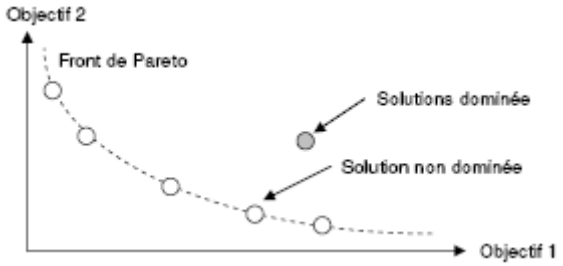

Figure 5 : Exemple de « front de Pareto » pour le problème de minimisation de deux objectifs.

6.3. Le multi-objectifs dans l'AGMC

Le premier algorithme évolutionnaire d'optimisation multi-objectifs développé est une extension d'un algorithme génétique simple, il s'appelle VEGA : Vector Genetic Algorithm. Il est présenté par Schaffer en 1985 [26]. Si on considère k objectifs et une population de n individus, une sélection de n/k meilleurs individus est effectuée pour chaque objectif. Ces k sous-populations sont mélangées pour former une nouvelle population de n individus et des opérateurs de croisement et de mutations sont appliqués.

La dominance de Pareto est utilisée dans de nombreuses méthodes multi-objectifs pour rechercher les solutions d'un problème. Certaines d'entre elles sont intégrées à l'AGMC : MOGA (Multiple Objective Generic Algorithm), NSGA (Non Dominated Sorting Genetic Algorithm), NPGA (Niched Pareto Genetic Algorithm) [27]. Elles sont dites non élitistes car :

- elles ne conservent pas les individus Pareto-optimaux trouvés au cours du temps,
- elles maintiennent difficilement la diversité sur la frontière de Pareto,
- la convergence des solutions vers la frontière de Pareto est lente.

Pour résoudre ces difficultés, nous envisageons d'intégrer les méthodes NSGA II et PESA qui :

- introduisent une population externe ou une archive permettant de stocker les individus Pareto-optimaux,
- utilisent des techniques de « niching », « clustering » et « grid-based » pour répartir efficacement les solutions sur les frontières de Pareto.

L'application présentée au paragraphe 7 est une optimisation mono-objectif. Dans cette partie nous avons préféré présenter un exemple d'optimisation issu de questions concrètes que se posent des décideurs afin de montrer les potentialités de la plateforme, au détriment d'un exemple plus complexe qui montrerait les qualités de l'AGMC [28] [29].

6.4. Les applications

Les applications des AG sont multiples : résolution numérique, économétrie, finance, traitement d'image, synthèse de formes complexes, optimisation de pièces pour l'industrie… et bien d'autres encore. Dans ce paragraphe nous en présentons quelques-unes.

Ils sont souvent utilisés comme outil de résolution numérique. Les économistes face à la complexité croissante des problèmes se sont tournés vers ces méthodes. Dorsey et Mayer [30] ont montré leur potentialité pour résoudre des problèmes d'optimisation difficile présentant des non-différentialités, des multi-modalités ou encore des discontinuités. Dans une même logique, Beaumont et Bradshaw [31] proposent de résoudre des problèmes non-linéaires tels que les modèles à croissance optimale. Les auteurs développent une version des algorithmes génétiques dite « distribuée parallèlement » afin d'éviter les problèmes de convergence prématurée vers un minimum local. Toujours dans le cadre de la résolution de modèle de croissance optimale, mais cette fois-ci stochastique, Duffy et McNellis [32] arrivent eux aussi à la conclusion de l'efficacité des AG.

Les AG peuvent être utilisés pour contrôler un système évoluant dans le temps (chaîne de production, centrale nucléaire…) car la population peut s'adapter à des conditions changeantes. En particulier, ils supportent bien l'existence de bruit dans la fonction à optimiser.

Les AG sont également utilisés pour optimiser des réseaux (câbles, fibres optiques, mais aussi eau, gaz…), des circuits VLSI [33], des antennes [34]… Il est envisagé l'intégration d'AG dans certaines puces électroniques afin qu'elles soient capables de se reconfigurer automatiquement en fonction de leur environnement *(Evolving Hardware* en anglais*).*

Ces dernières années, les livres de finances intègrent des applications utilisant les AG. La raison est simple comme le souligne Pereira [35]: « Genetic algorithms are a valid approach to many practical problems in finance which can be complex and thus requiere the use of an efficient and robust optimisation technique. Some applications of genetic algorithms to complex problems in financial markets include: forecasting returns, portofolio optimisation, trading rule discovery, and optimisation of trading rules".

L'ensemble des applications que nous avons abordés n'est pas exhaustif, il montre simplement combien les AG sont utilisés et adaptés pour des résolutions d'optimisation complexes. Dans le paragraphe suivant nous exposons les résultats d'une application utilisant l'AGMC.

7. Problème de l'ouverture d'une formation

Il s'agit d'apporter une solution aux problèmes de placement des formations agricoles au sein des établissements. La problématique correspond au scénario de localisation d'une nouvelle formation qui est à envisager à chaque rentrée scolaire. Le processus d'optimisation permettra de fournir des éléments de réponse en apportant, d'une part la localisation de la formation sur un lieu d'étude, et d'autre part une masse d'apprenants potentiellement réaffectés à cette formation. Le résultat obtenu devrait permettre aux décideurs de juger de l'ouverture ou non de cette formation en s'appuyant notamment sur le nombre d'élèves potentiels.

Soit *f* la formation que les décideurs veulent ouvrir. L'ensemble des élèves potentiellement réaffectés pour la formation *f* est l'ensemble des élèves qui suivent la formation *f* et pour qui le trajet entre son lieu d'étude et son lieu d'habitation n'est pas minimal.

7.1. Représentation du problème

Le chromosome établissement représente l'indice de l'établissement *Ind* accueillant la formation *f.* La liste des élèves potentiellement réaffectés est un chromosome composé d'autant de gènes qu'il n'y a d'élèves. Chaque gène élève possède deux modalités :

- si la valeur du gène est 1, l'élève est réaffecté à l'établissement *Ind,*
- si la valeur du gène est 0, l'élève n'est pas réaffecté et reste dans son établissement initial. Il n'entre alors pas en compte dans le calcul de la fitness.

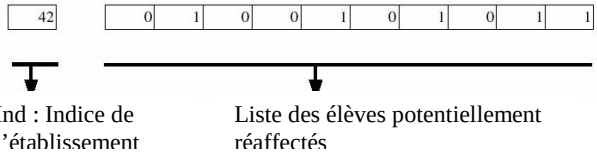

Figure 6 : Représentation des chromosomes.

7.2. Fitness

Pour l'indicateur des temps d'accès, on cherche à minimiser les temps d'accès Apprenant – Etablissements. Il faut donc calculer la moyenne des temps d'accès *Moy* pour chaque élève réaffecté (gène = 1). La note *Note1* s'obtient en rationalisant cette moyenne par la pire moyenne des temps d'accès possible (*MoyMax* = *N*TAMax*).

Où : *N* = le nombre total des élèves de la liste,

TAMax = temps d'accès maximum, c'est-à-dire entre les 2 communes les plus éloignées.

Ainsi la note renvoyée pour l'indicateur des temps d'accès est égale à :

$$Note1 = 1 - (\sum TA) / TAMax$$

où *TA* est le temps d'accès de l'élève à l'établissement considéré.

Si *Note1* est proche de 1 alors l'hypothèse est très proche de la solution, mais si *Note1* est proche de 0 l'hypothèse n'est pas adaptée.

De plus si un indicateur de capacité d'accueil des sites est ajouté, la note *Note2* est égale à 1 si le nombre d'élèves est inférieur à la capacité maximum d'accueil des établissements, 0 sinon.

La note *N* globale de l'hypothèse est une combinaison linéaire des notes pour chaque indicateur. Ainsi,

$$N = alpha*Note1 + beta*Note2$$

7.3. Résultats

Dans cet exemple, le problème est le suivant : dans quel établissement ouvrir une formation en particulier celle du « service aux personnes ». La figure 7 montre la situation actuelle des établissements (carrés de couleur) qui dispensent cette formation, ainsi que les élèves qui leurs sont affectés. Les points de couleur représentent les élèves localisés à la commune et leur couleur correspond à la couleur du site d'affectation.

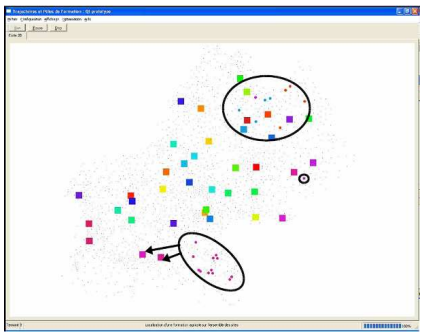

Figure 7 : Carte d'affectation réelle.

La figure 8 présente les résultats après optimisation.

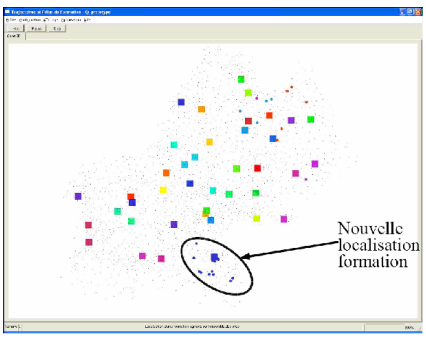

Figure 8 : Résultat de l'optimisation.

On peut observer un changement d'affectation des élèves situés dans le cercle, ils sont associés à un établissement plus proche, ici le carré bleu.

Après analyse des résultats, il en a été conclu qu'il fallait modifier la localisation de la formation « service aux personnes » du lycée violet, vers le lycée bleu.

Le couplage de GEOXP et de l'AGMC permet ainsi de fournir un puissant mécanisme d'aide à la décision pour ce type de problème très complexe.

8. Conclusion et perspectives

La plateforme à travers de nombreuses applications a suscité beaucoup d'intérêt. Les qualités et potentialités des outils ont été démontrées sur des applications réelles. Cet environnement propose les fonctionnalités d'un SIG ainsi que celles d'outils de traitement évolués, cependant nous envisageons des changements fondamentaux.

L'utilisation d'un SIG comme Géoconcept n'est pas un choix judicieux car le coût de la licence de cet outil reste un frein important au déploiement de la plate-forme. C'est pourquoi la première évolution que nous envisageons est l'intégration d'un SIG Open source permettant une visualisation WEB et une navigation 3D. Il sera couplé dans un second temps aux graphiques de la bibliothèque GeoXp afin de ne pas dupliquer les cartes. Puisque nous souhaitons une utilisation WEB du SIG, il est primordial de faire évoluer la bibliothèque GeoXp vers un interfaçage WEB afin de conserver une certaine cohérence et faciliter d'utilisation. Enfin, il est évident que l'interaction entre les graphiques et la représentation des unités spatiales doit être conservée.

Cette évolution permettra à une communauté d'utilisateur de collaborer et de partager des données mais aussi une plus large diffusion de cet outil.[3]

[1] J. Bertin, Sémiologie graphique : Les diagrammes, les réseaux, les cartes. Editions de l'Ecole des Hautes Etudes en Science, collection les Réimpressions, 431 p, 1999.

[2] H. Pornon, « La cartographie assistée par ordinateur », Ed. Hermès, Paris, 1989, 64 p.

[3] M. Worboys, M. Duckham, GIS A Computing Perspective, CRC Press, publié 2004.

[4] T. Bernhardsen, Geographic Information SystemsAn Introduction, John Wiley and Sons, Science, publié 2002.

[5] T.W Foresman, The History of Geographic Information Systems: Perspectives from the Pioneers, Upper Saddle River, NJ: Prentice Hall, (ed) 1998.

[6] P. A. Longley, M. F. Goodchild, D. J. Maguireand, D. W. Rhind, Geographical Information Systems: Principles,

[3] Des remerciements sont adressés à l'ensemble des membres du projet DYNASPAT et plus particulièrement à Christine Thomas, professeur à l'UT1, Basmah Fassi-Fehri et Mathieu Deltorre, stagiaires respectivement en statistique et informatique, qui ont travaillé sur les applications présentées dans cet article.

Techniques, Management and Applications. New York: John Wiley, 1998.

[7] M. Ilgn, « Open GIS Consortium, Aperçu et perspectives de l'Open GIS dans le domaine du Web Mapping », Lausanne, Octobre 2001.

[8] M. Neteler, H. Mitasova, OPEN SOURCE GIS: A GRASS GIS Approach, publié 2002, Springer, 434 pages.

[9] R. S. Bivand ,Using the R statistical data analysis language on GRASS 5.0 GIS data base files, Department of Geography, Norwegian School of Economics and Business Administration, Breiviksveien 40, N-5045 Bergen, Norway, 2000.

[10] A. A. Vachon, Open-source GIS, Literature Review, GIST 5120-Project Planning, October 2002.

[11] X. Song, Y. Kono and M. Shibayama, Environmental Cambodia: An Open Source GIS Approach to Web Mapping, International Journal of Geoinformatics, 2005.

[12] P. Baumard, Stratégie et surveillance des environnements concurrentiels, Masson, Paris, 1991.

[13] M. J. Ungerer; M. F. Goodchild, Integrating spatial data analysis and GIS: a new implementation using the Component Object Model (COM), International Journal of Geographical Information Science, Volume 16, Issue 1 January 2002, pages 41 – 53.

[14] I. Héba, E. Malin and C. Thomas-Agnan], "Exploratory Spatial Data Analysis with GeoXp", submitted to Geographical Analysis, 2003.

[15] A. Ruiz-Gazen and C. Thomas-Agnan, GeoXp: an R package for exploratory spatial analysis, 2006.

[16] J-M. Zaninetti, Statistique spatiale : méthodes et applications géomatiques (Coll. Applications des SIG), 2005.

[17] L. Anselin, A. Getis , Spatial statistical analysis and geographic information systems, The Annals of Regional Science, Springer Berlin / Heidelberg, Volume 26, Number 1 / March, 1992.

[18] M. Magrini, « Interaction des caractéristiques individuelles et territoriales dans le processus migratoire des jeunes en phase d'insertion professionnelle », 54ème congrès de l'AFSE, Paris, septembre 2005

[19] Fogel D.B. Evolutionary Computation (Third Edition), Series: IEEE Press Series on Computational Intelligence, Published Online: 25 Apr 2005, 1995.

[20] D. Beslay, D.R. Bull and R.R. Martin, An Overview of Genetic Algorithms: Part 1. Fundamentals, University Computing, vol 15, n°2. p 58-59, 1993.

[21] J.H. Holland, Adaptation in Natural and Artificial Systems.The University of Michigan Press. 1975.

[22] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, Reading, Mass., 1989.

[23] C. R. Darwin, On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London: John Murray. 1st edition, 1st issue, 1859.

[24] J.-L. Dessales, L'ordinateur génétique. Paris : Hermès, 1996.

[25] C Coello et al., Evolutionary Algorithms for Solving Multi-Objective Problems. Kluwer. Academic Publishers, New York 576 p, 2002.

[26] D. Schaffer., Multiple Objective Optimisation with Vector Evaluated Genetic Algorithm, In genetic Algorithm and their Applications: Proceedings of the First International Conference on Genetic Algorithm, p. 93-100, 1985.

[27] A. Berro, Algorithmes évolutionnaires pour l'optimisation multi-objectif. Dans: Technique et Science Informatiques, Hermès Science Publications, Numéro spécial: Evolution artificielle, V. 25, N. 8-9/2006, p. 991-1021, 2006.

[28] A. Berro, S. Sanchez, Autonomous agents or multiobjective optimisation. Dans: Genetic and Evolutionary Computation Conference GECCO'2004, Seatlle, 26/06/2004-30/06/2004, Springer, p. 51-54, 2004.

[29] A. Berro, S. Sanchez, Y. Duthen. Optimisation par algorithme génétique du placement des succursales d'une entreprise. Dans : 3ième congrès de la société Française de Recherche Opérationnelle et d'Aide à la Décision, ROADEF'2000, Nantes, France, 2000.

[30] R. Dorsey, W. Mayer, Genetic algorithms for estimation problems with multiple optima, non-differentiality, and other irregular features, Journal of Business and Economic Statistics, 1995.

[31] P. Beaumont, P. Bradshaw, A distributed parallel genetic algorithm for solving optimal growth models, Computational Economics, 159-180, 1995.

[32] J. Duffy, P.M. Nelis, Approximating and simulating the stochastic growth model: Parameterized expectations, neural networks, and genetic algorithm, Journal of Economic Dynamics and Control, 1273-1303, 2001.

[33] D. Beasley, D.R. Bull et R.R. Martin, An Overview of Genetic Algorithms: Part 1, Fundamentals, University Computing, vol. 15, n°2, p. 58-59, 1993a.

[34] A. Reineix, D. Eclercy et B. Jecko, FDTD/genetic algoritm coupling for antennas optimization, Annales de Télécommunications , vol. 52, n°9-10, 1997.

[35] Pereira, Genetic algorithm optimisation for finance and investment, Technical report, La Trobe University, 2000.