Restitution du séminaire *Optimisation - Modèle Complexe* 4 au 6 novembre 2014 - La Rochelle

Comité scientifique : Hilaire Drouineau, Nicolas Dumoulin, Patrick Lambert, Stéphanie Mahévas, Victor Picheny, Lauriane Rouan, Jean-Christophe Soulié

Introduction

La modélisation est fortement utilisée pour renforcer la compréhension du fonctionnement des systèmes complexes et aider à la prise de décision dans un environnement dynamique incertain. Le bon usage de ces modèles, le plus souvent complexes pour reproduire au plus juste le système étudié, requiert le respect de bonnes pratiques adhoc de modélisation à défaut d'avoir une méthodologie formalisée. Il s'agit en tout premier lieu de la transparence dans le processus de construction et de la validation du modèle. Dans le cadre du réseau Mexico, nous avons largement exploré les approches basées sur les analyses de sensibilité globale pour faciliter le développement, l'utilisation des modèles et la validation de ces derniers dans un processus de co-construction avec les experts thématiciens. Cette famille de méthodes permet d'identifier et hiérarchiser les paramètres clés qui, dès lors, doivent être estimés (calibrés) avec précision pour garantir un bon contrôle des sorties du modèle. L'optimisation est un champ de méthodes mathématiques cherchant à analyser et à résoudre analytiquement ou numériquement les problèmes qui consistent à déterminer le meilleur élément d'un ensemble, au sens d'un ou plusieurs critères quantitatifs donnés et de contraintes. Elle permet d'appréhender la calibration des modèles, la recherche d'une décision optimale selon un ou plusieurs critères prédits par le modèle, etc. En fait toute classe de problèmes modélisables peut conduire à un problème d'optimisation, pourvu que l'on y introduise des paramètres ou variables à optimiser. La principale difficulté de l'exercice d'optimisation vient du fait que les modèles complexes que nous appréhendons sont bien souvent non linéaires et non dérivables.
La problématique de ce séminaire se concentre autour de l'optimisation de la paramétrisation d'un modèle complexe qui présente au moins une des caractéristiques suivantes (une boîte noire, un code de calcul long ou coûteux, de nombreux paramètres continus ou discrets, de multiples variables de sortie et des processus déterministes ou stochastiques). L'optimisation visera à trouver i) les valeurs des paramètres permettant de reproduire au mieux les observations du système modélisé ou ii) les valeurs de variables décisionnelles (discrètes ou continues) permettant de remplir certains critères, le tout en respectant un ensemble de contraintes préalablement définies. Comme le synthétise la Figure 1, les spécificités des modèles complexes doivent être clairement prises en compte pour trouver la méthode mathématique adaptée à chacun des problèmes.

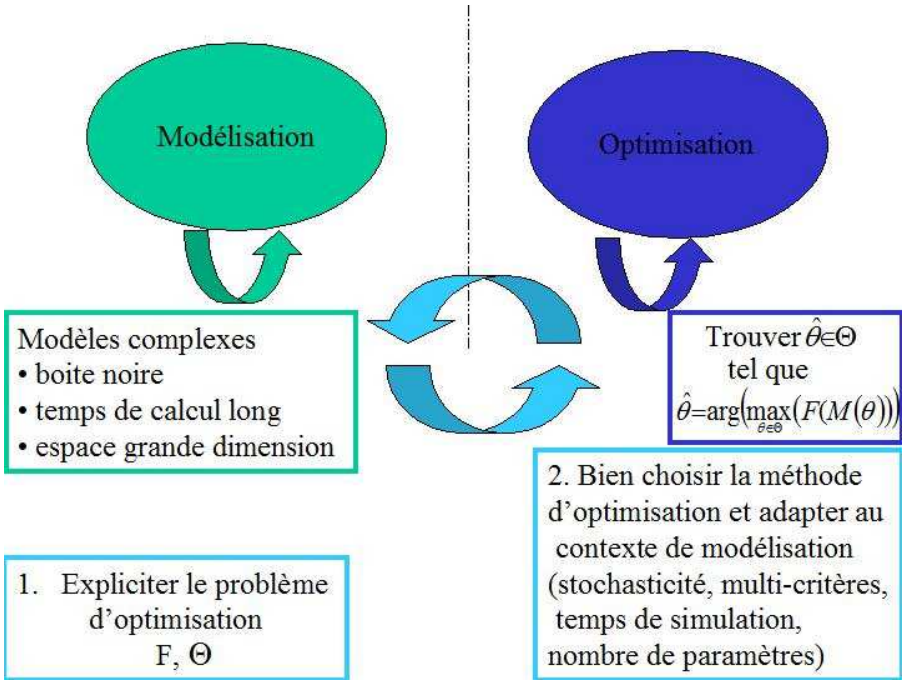

Figure 1: Optimisation dans un contexte de modélisation de système complexe

L'objectif de ce séminaire est de proposer aux membres du réseau Mexico un cadre d'échange autour de l'optimisation en modélisation de systèmes complexes. Il sera structuré autour des trois points suivants :

1. une revue des expériences d'optimisation des participants en mettant en avant les succès et les limites des méthodes utilisées pour calibrer, ajuster et exploiter leurs modèles complexes sous forme d'exposés (1/2h)
2. l'intervention des spécialistes du domaine pour des présentations sous forme de tutoriels (1h)
3. une période d'échange sous forme de tables rondes pour discuter des suites à donner dans le cadre du réseau mexico (identifier des champs de recherche pour débloquer des situations d'optimisation sans techniques adaptées, répondre à un appel d'offre sur la base d'applications concrètes dans différents domaines et de développement ou adaptation méthodologique pour déverrouiller les points bloquants actuels, écrire un position paper qui ferrait la synthèse des problématiques-approches-limites en optimisation et voies de recherche à emprunter par les participants au séminaire, autres)

Déroulement du séminaire

Tutoriel : Revue des méthodes d'optimisation et calage de modèles complexes (Rodolphe Le Riche)

This is a one hour class on the basics of numerical optimization for scientists who tune models based on experiments. It contains a short and hopefully practical classification of optimization algorithms. Some details about non-linear least squares are provided.

Exposés de modélisateurs

1. **David Odeja** - Analyse et estimation d'un modèle de la stimulation vagale pour le traitement de la insuffisance cardiaque

Les maladies cardiovasculaires représentent la principale cause de mortalité chez les adultes dans l'ensemble des pays membres de l'Organisation Mondiale de la Santé. Les processus impliqués dans les maladies cardiovasculaires sont le plus souvent complexes et multifactoriels. L'étude de telles pathologies multifactorielles nécessite l'acquisition et interprétation de données cliniques susceptibles de pouvoir fournir des indicateurs de l'état du patient. Dans ce contexte, une approche à base de modèle est utile à l'analyse de données cliniques et à la compréhension des événements impliqués dans un état pathologique.
Au Laboratoire Traitement du Signal et de l'Image~(LTSI -- Rennes), une méthodologie concernant l'utilisation de modèles multi-résolution en physiologie est appliquée sur différents cas cliniques. Dans ce travail, on s'est concentré sur deux: la modélisation spécifique-patient des coronaires pour l'étude de la circulation collatérale, et l'analyse spécifique-patient de modèles cardiovasculaires pour l'optimisation de thérapies de resynchronisation cardiaque.
Pour les deux cas, un approche d'estimation des paramètres est nécessaire pour analyser la situation physiologique des patients. Notre approche consiste en: 1) la définition d'un problème d'optimisation comme la minimisation de la différence entre les valeurs observées en chirurgie et les sorties des simulations; 2) la sélection des paramètres importants à identifier à l'aide d'une analyse de sensibilité; 3) l'application des algorithmes d'optimisation mono ou multi-objectif, et l'analyse ou interprétation physiologiques des paramètres identifies. Dans le cadre de la circulation coronarienne, nous avons fait une estimation multi-objectif sur des données cliniques obtenues durant des procédures de pontage coronarien. L'analyse et l'estimation des paramètres du modèle a permis de mettre en évidence l'importance de la circulation collatérale qui est un réseau de vaisseaux alternatifs se développant pour compenser la diminution du flux sanguin du réseau coronaire en cas de sténoses significatives. Les données cliniques obtenues durant les pontages de dix patients ont pu être reproduites de manière satisfaisante avec le modèle et l'hémodynamique coronarienne a pu être évalué. Concernant l'étude sur la thérapie de resynchronisation cardiaque (CRT), qui consiste en l'implantation d'un pacemaker, nous avons fait une estimation multi-objectif sur les données cliniques obtenues pendant un session d'optimisation du délai atrio-ventriculaire et intra-ventriculaire. Si bien les résultats de cette estimation sont relativement satisfaisants, les analyses à posteriori des paramètres identifié démontrent une difficulté pour construire des conclusions ou interprétations physiologiques utiles pour les cliniciens. Dans ce cas, l'utilisation des algorithmes d'optimisation multi-objectif est sévèrement limité par le budget de simulation et la qualité des variables observables.

2. **Audric Vigier - Calibration d'ISIS-Fish avec le Recuit simulé : choix des meta-paramètres et perrformances**

Une méthode d'optimisation, le recuit simulé, a été implémentée de manière générique pour calibrer le modèle ISIS-Fish au travers d'une approche de simulation-estimation avec une pêcherie simplifiée. Le problème de calibration se décline en plusieurs configurations variant le nombre de paramètres à calibrer, les valeurs recherchées de ces paramètres, la forme de la fonction d'objectif et la paramétrisation du recuit simulé. Le recuit simulé s'avère adapté pour calibrer peu de paramètres (ici 3), quelle que soit la fonction d'objectif : il converge vers l'optimum global en moins d'un millier d'itérations et renvoie une paramétrisation conforme conditionnellement au modèle. Par contre, en 10 paramètres, la convergence n'est pas atteinte en un nombre d'itérations volontairement limité à 2000 pour des questions de durée de simulation. Les raisons principales semblent être : les gammes de valeurs des paramètres recherchés, la fonction objectif, le voisinage dans le recuit simulé et le point initial. Une fonction d'objectif multi-critères aurait probablement pu favoriser la convergence.

3. **Diariétou Sambakhe - Définition d'un idéotype de sorgho d'après un modèle de culture**

Pour donner un objectif aux généticiens qui sélectionnent de nouvelles variétés de sorgho pour le Sahel, l'objectif de la thèse est de trouver des idéotypes de sorgho adaptés au climat présent et futur en utilisant un modèle de culture. Un idéotype est un jeu de paramètres variétaux de ce modèle qui maximise une espérance ou un quantile de la production, sous un climat donné. Celui-ci est caractérisé par un autre modèle, ajusté sur des données météorologiques, et dont les paramètres varient dans l'espace suivant un gradient approximativement Nord-Sud. Nous sommes donc devant un problème d'optimisation, celui de trouver les paramètres variétaux optimaux en fonction des paramètres du climat. La méta-modélisation semble un moyen de le résoudre.

4. **Thibaud Rougier - Calibration de mon modèle à partir de données observées avec probablement une méthode ABC d'optimisation**

La grande alose Alosa alosa est une espèce migratrice amphihaline européenne dont l'aire de distribution historique s'est considérablement restreinte au cours du dernier siècle. Afin d'améliorer notre compréhension de l'évolution observée de l'aire de distribution de cette espèce entre 1900 et 2000 et dans l'optique de formuler certaines hypothèses sur son évolution au cours du prochain siècle, nous utilisons le récent modèle mécaniste GR3D (Global Repositioning Dynamics of Diadromous fish Distribution). Pour reproduire au mieux par des simulations la distribution observée autour de 1900 de l'espèce, nous avons utilisé une méthode d'optimisation bayesienne (ABC) avec un algorithme récent (méthode Lenormand du package R EasyABC) particulièrement adapté aux modèles complexes stochastiques. Plusieurs essais de calibration ont été réalisés en variant le choix et le nombre des paramètres à calibrer ainsi que les statistiques à optimiser. Finalement, par l'intermédiaire d'une analyse de sensibilité globale, nous avons retenu trois paramètres du modèle à calibrer et trois statistiques à optimiser. Malgré quelques incohérences, les résultats sont encourageants et laissent à penser que la calibration du modèle sera satisfaisante.

<u>Tutoriel :</u> Approche par méta-modèles (Victor Picheny)

L'utilisation de méta-modèles pour faciliter l'optimisation est une solution classique dans le cadre des modèles complexes et coûteux à évaluer. Tout d'abord, nous mettons en lumière les « pièges » classiques de ces approches, et la nécessaire notion d'enrichissement séquentiel. Puis, nous décrivons les algorithmes classiques pour deux méta-modèles : les modèles de régression polynomiale (méthodes de région de confiance) et les modèles de krigeage (algorithme EGO). Enfin, différentes perspectives de recherche sont données dans le cas du krigeage afin d'aborder des problèmes complexes (contraintes, multi-objectifs, etc.).

<u>Tutoriel :</u> Prise en compte de la stochasticité (Victor Picheny et Rodolphe Le Riche)

Dans une deuxième partie, nous abordons le problème de l'optimisation sous incertitude, en particulier quand le modèle considéré est stochastique ou quand il possède des entrées stochastiques. Nous abordons tout d'abord les aspects formulation, puis nous proposons quelques solutions pour différents contextes (méthodes des moments, méta-modélisation, stratégies évolutionnaires).

<u>Exposés de modélisateurs</u>

5. **Franck Jabot - utilisation de méta-modèles pour accélérer les inférences par ABC : premiers essais.**

Les methodes ABC (Approximate Bayesian Computation) permettent d'étendre les méthodes Bayésiennes standards aux cas où le calcul analytique de la fonction de vraisemblance n'est pas possible. Ce sont donc des méthodes particulièrement bien adaptées pour calibrer des modèles complexes dans un cadre Bayésien, ce qui permet de documenter l'incertitude sur les valeurs de paramètres calibrées. Cependant, ces méthodes sont très gourmandes en temps de calcul et nécessitent typiquement plusieurs dizaines de milliers à plusieurs millions de simulations du modèle à calibrer, ce qui n'est pas accessible pour nombre de modèles

complexes utilisés dans les sciences de l'environnement. Afin d'étendre les méthodes ABC à ce type de modèles, l'utilisation de méta-modèles a été suggérée dans la littérature. Cette présentation vise à expliquer comment méta-modélisation et méthodes ABC peuvent être articulées, et à présenter quelques résultats préliminaires basés sur des méta-modèles de type régression locale. Ces résultats préliminaires permettent de pointer le principal challenge de l'approche, à savoir reproduire fidèlement la structure de la stochasticité du modèle complexe qu'on cherche à émuler.

6. **Morgane Travers - Calibration du modèle stochastique OSMOSE avec un algorithme évolutionnaire**

Pour comprendre le fonctionnement des écosystèmes marins soumis à différentes pressions de pêche et environnementales, un modèle individu-centré stochastique OSMOSE a été appliqué à la communauté de poissons de Manche Orientale. Ce modèle représentant le cycle de vie de 14 espèces de poisson utilise des paramètres liés à la structure du modèle, des paramètres biologiques fiables issus d'un grand nombre d'observations, des paramètres peu fiables et quelques paramètres inconnus dont les valeurs ne sont pas estimables via l'expérimentation. Pour estimer les 23 paramètres inconnus et les 26 peu fiables, une méthode de calibration par algorithme évolutionnaire a été développée. L'algorithme fait intervenir deux phases de calibration pour être plus efficace dans l'exploration de l'espace des paramètres pour les paramètres inconnus puis en incluant les paramètres incertains proches de leur valeur initiale. La fonction objectif est construite à partir d'objectifs partiels relatifs à la proximité des simulations avec la biomasse et les captures observées pour chaque espèce. Afin d'améliorer l'efficacité de la calibration, une analyse de sensibilité sur l'ensemble des paramètres permettrait d'identifier les paramètres susceptibles de produire de grandes variations dans les simulations.

7. **Ronan Trepos - Optimisation de variétés de tournesol sous incertitude climatique**

Le modèle dynamique de simulation de la croissance de tournesol SUNFLO calcule en sortie un rendement de la culture à l'échelle de la parcelle et de l'année culturale (d'avril à octobre). Les entrées de ce modèle sont les huit traits phénotypiques de la variété (tel que le nombre potentiel de feuilles), une conduite de culture (date de semis, fertilisations et irrigations) et ne série climatique (données journalières de pluie, evapotranspiration, températures et radiation). Dans cette étude nous ne nous intéressons pas à la conduite de culture qui est fixée a priori. En terme de données nous avons à disposition un nombre conséquent de données climatiques: 190 séries sur les sites d'Avignon, Blagnac, Dijon, Poitiers et Reims pour les années allant de 1975 à 2012 et une petite centaine de variétés réelles de tournesol. L'objectif de ce travail est de rechercher des variétés virtuelles (combinaison de traits phénotypiques) qui permettent d'optimiser le rendement en sortie de simulation. Ces variétés doivent être évaluées sur l'ensemble des profils climatiques disponibles et être robustes à l'aléa climatique. Evaluer une variété virtuelle nécessite donc de faire 190 simulations et de trouver un moyen de comparer les distributions de sorties. Dans ce travail nous proposons de sélectionner par apprentissage automatique un ensemble réduit de séries climatiques, au nombre de 10, représentatives de l'ensemble des profils climatiques. Il s'agit alors d'estimer des mesures statistiques (moyenne, quantiles etc..) à partir de ces 10 simulations qui soient le plus proche possible des mesures estimées sur les 190 séries climatiques. Ainsi, dans le cadre de la procédure d'optimisation multi-critère des mesures statistiques, l'évaluation des variétés virtuelles est beaucoup moins coûteuse. Je présenterai les approches que nous avons adoptées ainsi que des résultats préliminaires.

<u>Tutoriel :</u> Optimisation multi-objectifs (Dimo Brockhoff)

Multiobjective optimization problems occur frequently in practice when it is interesting to investigate the trade-offs among different, conflicting objective functions. In such a case, one

is interested in finding a set of solutions which approximate the set of so-called Pareto-optimal or efficient solutions of the multiobjective problem. Evolutionary Multiobjective Optimization (EMO) algorithms are often a good choice to tackle this problems, especially in cases where the objective functions are computed via expensive, highly non-linear simulations (i.e. in more general in a black box scenario). One reason is that they are stochastic algorithms that can handle noisy and uncertain objective functions; the other reason is their internal population which allows to search naturally for solution sets. This tutorial will briefly introduce the main concepts of multiobjective optimization and the basic ideas behind several common EMO algorithms such as NSGA-II and the MO-CMA-ES. In a third part, I will talk about how multiobjective optimization algorithms can be compared and benchmarked and for the latter make a short detour to the single-objective framework Coco (Comparing Continuous Optimizers, http://coco.gforge.inria.fr) and the results obtained with it in the previous Blackbox Optimization Benchmarking (BBOB) workshops.

Exposés de modélisateurs

8. **Romain Lardy - Calibration du modèle multi-agent MAELIA pour l'évaluation des normes de gestion d'étiage**

MAELIA est un modèle multi-agent visant à reproduire les impacts socio-environnementaux des normes de gestion et de gouvernance de ressources naturelles renouvelables et de l'environnement. Compte-tenu du nombre de processus nécessaires dans cette modélisation et des dimensions de la zone d'étude (bassin Garonne Amont), les simulations sont couteuses en temps de calcul. La calibration a pour but de permettre à la fois de reproduire les débits mais également de détecter les dates et durées pendant lesquels les débits descendent en dessous d'un certain seuil. Cet objectif a été traduit en quatre critères numériques utilisés dans une calibration basée sur la méthode EGO. De façon à éviter le choix du métamodèle, un mélange de krigeages a été utilisé et une pondération par l'erreur de prédiction a été utilisée. De plus, de façon à réduire l'attente utilisation, à chaque étape de l'algorithme un plan d'expérience de 36 points a été recherché. Un résultat satisfaisant a été obtenu au bout de 1200 simulations (dont 160 pour l'initialisation)

9. **Hilaire Drouineau – Hake growth : ajustement multi-criteres d'un modèle complexe par maximum vraisemblance**

Un modèle matriciel à espace d'états, structuré en taille et spatialisé, a été développé afin d'estimer les taux de croissance, de migration et d'évaluer le stock de merlu d'Atlantique Nord-Est. Ce modèle est ajusté par maximum de vraisemblance à 4 grands types de données: (i) des séries temporelles de débarquement provenant de 18 groupes de pêcheurs, (ii) cinq séries d'indices d'abondance scientifiques, (iii) des compositions en taille de débarquement et (iv) des compositions en taille provenant d'échantillonnage scientifique. Un algorithme à stratégie évolutive a été couplé à un algorithme de quasi-Newton utilisant l'auto-différentiation afin d'estimer les 86 paramètres inconnus du modèle. Différentes formulations de fonction de vraisemblance ont été testées en comparant notamment des pondérations expertes et des pondérations objectives des jeux de données. Les résultats montrent que la robustesse de la fonction de vraisemblance pourrait être améliorée afin d'éviter que certaines séries de données soient "surajustées" au détriment d'autres. Toutefois, ce travail a permis d'obtenir des estimations de paramètres satisfaisantes.

10. **Amina Ameur - Analyse de sensibilité en présence de paramètres corrélés en vue de l'optimisation d'un modèle de représentation de la structure des sols**

La représentation virtuelle de la structure des sols, c'est-à-dire l'agencement des pores et des solides, représente un enjeu important pour les problématiques environnementales impliquant la prise en compte de la qualité du sol telles que le cycle du carbone et la gestion de l'eau. Cela permet de simuler l'évolution de la structure, révélatrice du fonctionnement des sols, sous l'effet de l'activité des vers de terre, des transferts hydriques ou du développement

racinaire. Dans ce cadre, l'APSF (Arborescence Pores Solides Fractals) est un modèle particulièrement intéressant car il permet de générer une représentation synthétique multi-échelle en 3D de la structure d'un sol sous la forme d'un arbre composé de plusieurs motifs appelés canevas (arbre de canevas). Il s'agit d'un modèle descriptif qui fait appel à un algorithme récursif afin d'obtenir une répartition spatiale des cellules représentant les différents éléments de la structure d'un sol tels que les pores, les matières organiques et les matières minérales. Ce modèle a besoin de paramètres: les proportions de cellules pour chaque niveau d'échelle spatial décrite et le nombre de pores voisins. Certains de ces paramètres sont corrélés entre eux. Par conséquence, l'évaluation de leur influence sur les sorties du modèle nécessite d'adapter les méthodes « classiques » d'analyse de sensibilité avec une approche multidimensionnelle pour mieux interpréter les indices de sensibilité calculés. Dans ce travail, seront présentés la méthode d'analyse de sensibilité retenue ainsi que les résultats obtenus. D'autre part, se pose également la question du choix de l'algorithme d'optimisation, qu'il soit déterministe ou stochastique, compte tenu de la présence des contraintes reliant les paramètres clés de l'APSF, des incertitudes des mesures expérimentales ainsi que de la nature des fonctions objectif. Une fois le modèle calibré, l'APSF constituera un outil précieux dans le cadre de l'échantillonnage et de la réhabilitation des sols dégradés.

Quelle suite à ce séminaire ?

Synthèse à chaud du séminaire

Balayage d'un vaste panel de configurations d'utilisation de l'optimisation dans un contexte de modèle complexe (le plus souvent boite noire avec des temps de calcul longs) : configuration multi-modales, un nombre variable de paramètres, des coûts/temps de simulation pouvant être très limitant, des modèles déterministes et stochastiques, des paramètres discret et continus, des optimisations avec ou sans contraintes, des optimisations multi-objectifs ou mono-objectif,...

En reprenant des caractéristiques discriminantes des méthodes d'optimisation pour la modélisation complexe, peut-on cartographier l'espace des méthodes d'optimisation utilisées par les modélisateurs?

Les deux axes majeurs permettant de structurer la complexité des situations d'optimisation sont : le nombre de paramètres (synonyme de la dimension de l'espace à explorer) et le nombre d'appels au modèle (simulations) pouvant être réalisés pour rechercher l'optimum. Quatre autres caractéristiques apparaissent comme influentes dans le la formulation du problème d'optimisation ou dans le choix de la méthode d'optimisation : fonction à optimiser (mono ou multi-critères), modèle stochastique ou déterministe, paramètres continus/ discrets/mixtes, contraintes sur la fonction ou les paramètres, données disponibles.

La cartographie des expériences (Figure 2) présentées lors du séminaire fait ressortir les points suivants :

- aucune optimisation n'a été faite avec des paramètres discrets (alors que certains modèles ont des paramètres discrets: auto-censure pour absence ou non maîtrise des méthodes ?)
- peu de contraintes
- des multi-objectifs approchés parfois par du mono-critère en pondérant tous les objectifs pour n'en faire qu'un
- en dépit du vaste panel de configurations mentionnés précédemment, les cas d'études sont finalement concentrés dans la même portion du graphique. Le nombre réduit de paramètres inclus dans l'optimisation est souvent le résultat d'une analyse a priori de

la sensibilité des paramètres pour se concentrer sur ceux qui ont le plus d'influence (AS).

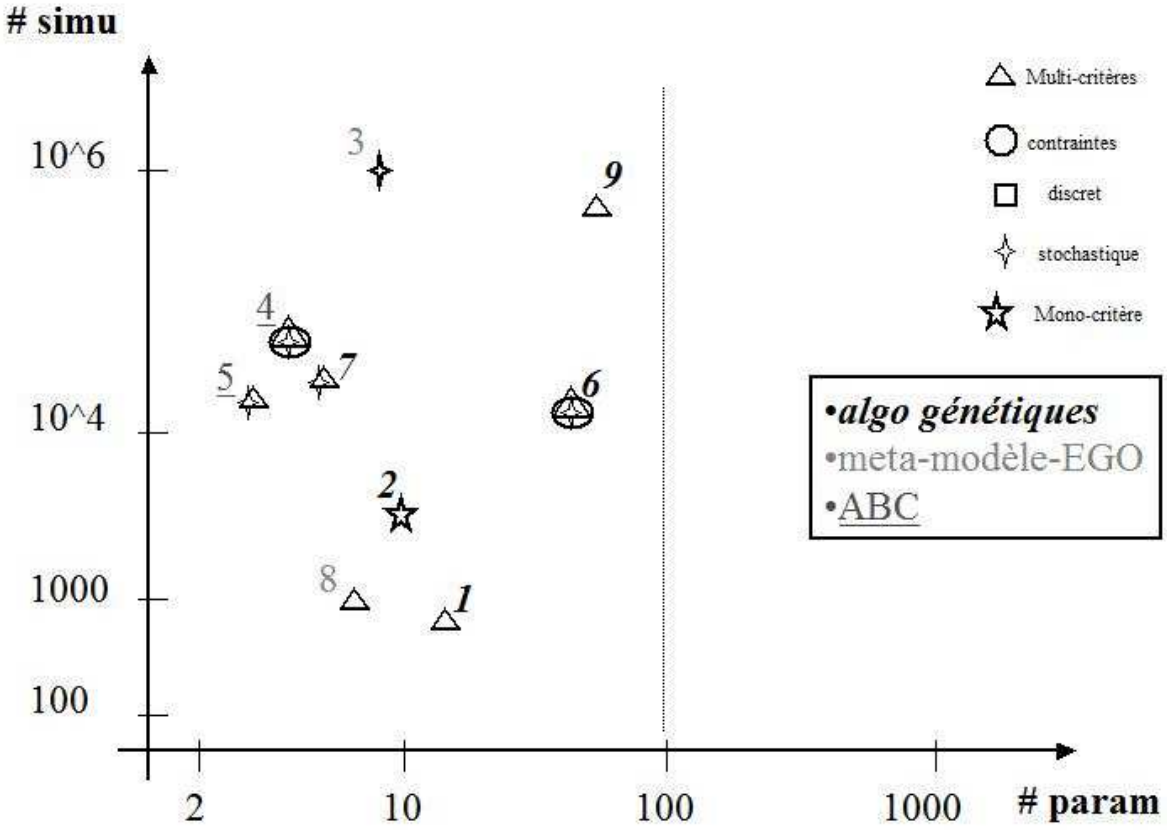

Figure 2 : Cartographies des expériences d'optimisation présentées pendant le séminaire selon différents critères (nombres de paramètres, nombres de simulations -appels au modèle-, caractéristiques du modèle, caractéristiques de l'optimisation, algorithme d'optimisation)

- seules 3 grandes familles de méthodes d'optimisation ont été utilisées : algo génétiques, approche meta-modèle-EGO, ABC. Aucun algorithme de recherche locale (en tout cas seul)

A partir des tutoriels méthodologiques, peut-on cartographier l'espace des méthodes d'optimisation proposées par les mathématiciens?

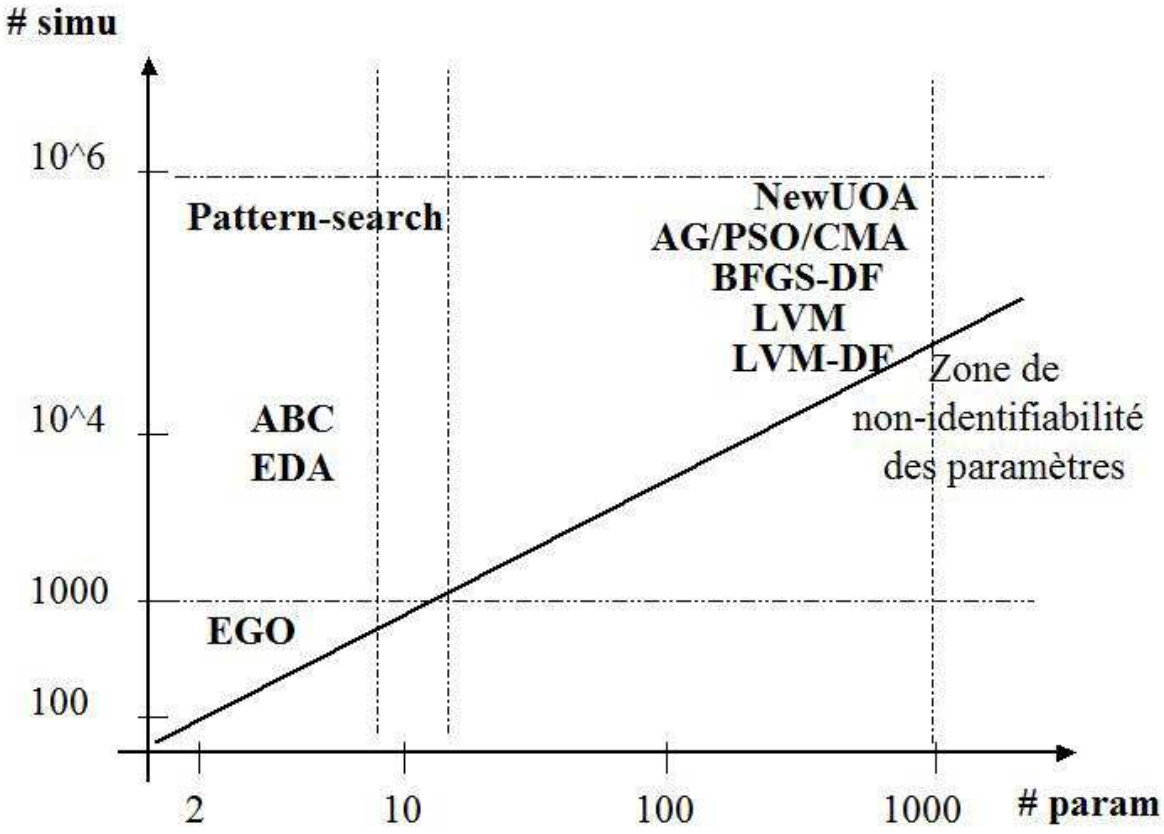

Figure 3 : Cartographie des méthodes d'optimisation présentées dans les tutoriels en fonction du nombres de paramètres à optimiser et du nombre de simulations à réaliser.

A noter que sur ce graphique (Figure 3), les méthodes pertinentes pour de nombreux paramètres le sont aussi pour moins de paramètres (qui peut le plus peut le moins).
Y a t-il concordance ? Les méthodes utilisées semblent avoir été sélectionnées de manière pertinente. On observe une frilosité et aussi beaucoup d'attentes des modélisateurs à explorer certaines méthodes récentes.

Peut-on déduire de ces deux représentations (Figures 2 et 3) un guide de choix d'une méthode pertinente d'optimisation au regard des objectifs et contraintes du modélisateur?
Les manques qui ressortent des présentations et discussions à l'issue des présentations:

- Il faut que les objectifs soient bien identifiés : pourquoi optimiser? L'objectif du modèle est-il d'estimer des valeurs de paramètres ou cette calibration n'est-elle qu'une étape intermédiaire ?
- Il faut des outils pour valider le résultat de l'optimiseur. Quelle est la robustesse du résultat? Que peut-on faire avec toutes les sorties de l'optimisation?
- il faut des outils pour évaluer la pertinence de la formulation du problème d'optimisation et notamment son adéquation face aux données disponibles (données pour l'estimation : quelle confiance? quantité d'observations ? pondération en fonction de l'incertitude), objectifs, contraintes

Restitution des discussions

Groupe méthodes

Le résultat des discussions de ce groupe peut être structuré en trois problématiques complémentaires. On distingue des problématiques liés aux aspects postérieurs (post-traitement) ou antérieurs (pré-traitement) au déroulement d'un algorithme d'optimisation, ainsi qu'un ensemble de défis méthodologiques plus généraux.

Post-traitement

La notion d'identifiabilité du problème (unicité, infinité de solutions, etc.) est apparue comme importante au cours des exposés. Ce concept est relativement classique en optimisation théorique mais a été peu traité par les intervenants, alors que ce type d'étude apporte potentiellement beaucoup de connaissance sur le modèle. Un ensemble de solutions classiques sont déjà potentiellement à la disposition des modélisateurs (linéarisation, étude des matrices hessiennes, etc.). Le cas d'étude de modèles de type boîte noire coûteux semble moins connu et a été identifié comme un potentiel sujet de recherche par les participants.

Dans le cas de la calibration de modèles, la sensibilité de l'optimum aux données est apparu comme un point critique et souvent problématique, en particulier quand ces données sont bruitées. L'idée générale est de pouvoir quantifier si une petite modification sur les données affecte les résultats d'optimisation.

Des solutions simples ont été évoquées, comme la validation croisée (on enlève une donnée, on observe la variation sur l'optimum). Il pourrait être intéressant de mettre à profit la « culture » Mexico sur l'analyse de sensibilité. D'un point de vue méthodologique, lier l'analyse de sensibilité et l'optimisation pose un certain nombre de défis (comment éviter la double boucle ?) et de pistes de recherche intéressants.

Il a été noté que la notion d'identifiabilité s'exprime également dans ce cas, mais entraîne des questionnements complexes (que signifie un gradient par rapport aux données, par exemple de type fonctionnel ?).

Des méthodes de réduction de dimension ont été systématiquement utilisées avant optimisation. Doit-on revenir dessus a posteriori? Il pourrait être intéressant de regarder la sensibilité de l'optimum à la réduction de dimension (que se passe-t-il si on rajoute des paramètres enlevés?). Cependant, est-ce accessible en temps de calcul

Formulation - prétraitement

Dans le cadre de la calibration de modèles comme dans celui de l'optimisation sous incertitudes, la description des sources comme des formes d'incertitudes est nécessaire pour que puissent travailler les métholodogistes. Plus généralement, un accès maximal aux connaissances métier est une composante critique pour une formulation efficace du problème. Un trop grand formatage (i.e. une variance pour décrire une variable aléatoire) peut faire perdre une connaissance expert (par ex. valeur impossible de cette variable en deçà d'un seuil).

Un apport méthodologique concerne l'utilisation d'autres outils statistiques que les moindres carrés ou la vraisemblance pour la calibration. Cependant, mesurer si on améliore la formulation d'un point de vue de l'optimisation semble être un problème ouvert.

La tâche de sélectionner un algorithme adéquat au problème est un problème central et complexe. Une solution peut venir des comparateurs (e.g. COCO), cependant, comment savoir quelle fonction test se rapproche du problème que l'on souhaite traiter ? Peut-on diagnostiquer a priori le « type » de problème auquel on a affaire ?

Autres défis / pistes de développement

Au cours des exposés, un certain nombre de pistes de développement méthodologique ont été identifiées, qu'on listera simplement ici :

1- Contrôle partiel de l'aléa simulé

2- Méthodes bayésiennes & recherche de distributions : ce domaine de recherche semble encore très ouvert
3- Quel algorithme utiliser dans un scénario coûteux ? Peu de réponses à ce jour (COCO : seulement 2 algorithmes testés). Il semble nécessaire de motiver les développeurs d'algorithmes afin qu'ils implémentent leurs méthodes sur des benchmarks.
4 – Prédiction fonctionnelle et méta-modélisation : la plupart des modèles ont des sorties fonctionnelles (par ex., trajectoire). On passe souvent par une statistique résumante pour la métamodélisation et l'optimisation ; devrait-on directement travailler avec des prédicteurs fonctionnels ? (connnu en mécanique des fluides, plus nouveau en statistiques)

Groupe modélisateurs

L'objectif de ce groupe de travail était de réfléchir entre modélisateurs utilisateurs de techniques d'optimisation, aux limites et souhaits quant aux outils et méthodes d'optimisation à notre disposition.

Choix et utilisation d'un algorithme d'optimisation ?

Nous avons identifié la différence entre les algorithmes locaux et globaux. Le modélisateur est naturellement tenté d'utiliser un algorithme global, mais est-ce toujours pertinent étant donné un budget de simulations disponible parfois limité ? On serait alors tenté d'utiliser un algorithme global pour dégrossir puis un local pour raffiner.

On constate que souvent, le choix d'un algorithme peut être guidé par le conseil d'un collègue, la proximité d'autres utilisateurs et par la facilité d'accès à une bibliothèque logicielle. En effet, le modélisateur cherche aussi un outil logiciel efficace qui ne va pas être trop compliqué et coûteux à utiliser, notamment lorsqu'il va s'agir de coupler le logiciel d'optimisation et son logiciel de simulation.

Reformulation du problème d'optimisation

Un autre point discuté est l'adaptation du modèle qui permet souvent de faciliter son ajustement. L'expérience montre qu'il y a des astuces utiles comme la mise à l'échelle pour centrer une variable, ou bien la reformulation pour calibrer l'exponentielle d'une variable, estimer les covariances des variables. Plus généralement, il serait très utile d'avoir des méthodes de diagnostic afin d'améliorer la conduite de l'optimisation en fonction des propriétés pouvant être déduites avant et pendant l'optimisation. De plus, le choix d'un algorithme peut introduire des métaparamètres (paramètres propres à l'algorithme d'optimisation) à calibrer, qui ne sont pas évidents à trouver dans la littérature. Ces métaparamètres étant souvent corrélés, il conviendrait de les calibrer séquentiellement. Ainsi, il arrive qu'on choisisse un algorithme pensant qu'il sera efficace, mais on parvient difficilement à l'utiliser correctement.

Formulation de la fonction objectif

Les modélisateurs sont intéressés par enrichir les fonctions utilisées pour formuler leur fonction objectif. Par exemple, l'utilisation de série temporelle peut faire appel a des distances du type "Dynamic time warping”. Un autre exemple cité concerne l'utilisation de classes d'âge dans les données. Là encore, il existe de nombreuses manières de prendre en compte cette donnée dans la fonction objectif. On peut se contenter des moindres carrés, mais on peut par exemple vouloir exprimer une tolérance sur la distance de certaines classes d'âges, mais moins sur d'autres classes d'âges. Il faudrait alors utiliser une fonction pour formuler cette distance, et éventuellement introduire de nouveaux métaparamètres. Plus généralement, des méthodes pour améliorer la robustesse des fonctions cibles aux incertitudes des données semblent nécessaires. En outre, on constate un manque de méthodes objectives pour choisir le niveau pertinent d’agrégation (spatiale, temporelle ou autre) des données dans la fonction objectif.

Nous sentons bien que beaucoup de méthodes existent et que nous ratons parfois de bonnes candidates. Nous sommes bien conscients que la formulation de cette fonction objectif peut introduire un biais important dans l'optimisation et mener parfois à l'échec. Comment diagnostiquer pendant l'optimisation un biais de la fonction objectif ? Comment diagnostiquer une incompatibilité entre plusieurs objectifs ?

Nous avons peu abordé la question des paramètres discrets, mais c'est un sujet préoccupant.

Outils de diagnostic

Nous avons déjà identifié notre besoin d'outils de diagnostic pour obtenir des propriétés des variables pendant l'optimisation et pour identifier les biais de la fonction objectif. Nous avons aussi bien compris à travers les exposés l'importance d'éviter les biais *a priori* en offrant par exemple des outils pour l'analyse des fronts de Pareto. Enfin, l'analyse des incertitudes et de l'identifiabilité demande à être mieux maîtrisées.

Groupes mélangés

Le principal constat réalisé par ces groupes mixtes est que les modélisateurs ont peu mis « en difficulté » les spécialistes de l'optimisation. Les besoins en ajustement qu'ils ont pour leur modèle, de par la nature et la dimension de l'espace de recherche et par les temps de simulation, doivent pouvoir trouver une réponse avec les outils existants. C'est donc plus des méthodes pour choisir les bons algorithmes que de nouveaux algorithmes qui sont nécessaires. Des outils, type arbre de décision, pourraient déjà aider à orienter le modélisateur. Au-delà de ça, des plates-formes permettant de tester et comparer différents algorithmes sur des modèles complexes seraient particulièrement pertinentes. Reste toutefois à vérifier si ces constats sont réels ou liés à une autocensure des modélisateurs qui cherchent à rester dans leur domaine de confort.Plus généralement, il est constaté un manque en termes d'outils de diagnostics pré et post-ajustement, autour de trois volets :

- prédiagnostic du modèle en lui-même : la formulation du modèle est-elle optimale ou non ? Diagnostic de l'autocorrélation intrinsèque (liée à l'écriture mathématique du modèle, indépendamment des données) du modèle.
- prédiagnostic et traitement de l'incertitude des données utilisées pour l'ajustement : comment juger de l'adéquation entre données disponibles et le problème d'ajustement ? Comment savoir à quelles échelles d'agrégation (temporelles, spatiales, etc.) doivent être utilisées les données pour avoir un ajustement optimal ? Comment traiter l'incertitude autour des données ?
- post-diagnostic de l'ajustement : comment vérifier l'identifiabilité extrinsèque (combinaison de l'autocorrélation liée à la formulation du modèle et celle liée aux données disponibles) des paramètres estimés ? Quelle est la sensibilité de l'estimation des paramètres aux différentes données utilisées ? Comment exploiter à fond les résultats de l'ajustement (pas uniquement la valeur à l'optimum) pour : (i) quantifier l'incertitude, (ii) arriver au front de Pareto

Sur les volets un et trois, les méthodes semblent exister (simulation-réajustement, analyse de sensibilité, analyse de la stabilité de la matrice d'Hessian, analyse de la population des algorithmes globaux). C'est autour du second point que le plus de développement méthodologique semble nécessaire.

Discussions sur les perspectives

On a identifié deux types de perspectives. Celles qui peuvent directement se réaliser **dans le cadre du réseau**

i) animation : faire en sorte que cette communauté mixte continue à exister et à discuter

ii) séminaire : quand ? prochaines rencontres mexico juin 2015 ou novembre 2015?

- diagnostic de pre et post -- trouver les experts dans ces domaines ou les bonnes volontés
- qu'est-ce que j'aimerais et que je n'ose pas - explorer les possibilités et identifier le limites

iii) préparation EC - optimisation : 2016/17

iv) écriture d'un article de positionnement - premier draft pour avril 2015 pour plos one, scope : revue des approches des modélisateurs - positionnement des méthodes dans un espace aux dimensions - identification de limites, blocages, - perspectives

dans le cadre d'un projet : projet 2016

Développement logiciel : librairie d'outils de diagnostics,

Banc d'essai type coco pour modèles complexes existant