Tendances et nouveaux défis en statistiques pour le climat, l'environnement et l'écologie

Groupe Environnement de la Société Française de Statistiques, Denis Allard, Liliane Bel, Olivier Gimenez, Philippe Naveau, Aurélien Ribes

Les statistiques sont mobilisées dans le système « Planète Terre » lorsque se posent des questions d'exploration de jeux de données, d'estimation de grandeurs ou de paramètres, de comparaison (en différents lieux, à différentes époques, à différentes échelles), de validation d'hypothèses scientifiques, ou encore de confrontation entre modèles[1] et mesures. Les thèmes couvrent l'ensemble de la triade Terre fluide – Terre vivante – Terre habitée, allant de la climatologie et la paléo-climatologie à l'échelle globale ou continentale, à l'observation environnementale et écologique, au suivi des populations végétales ou animales (démographie, déplacement, colonisation, invasion, …), mais aussi et de plus en plus à l'interaction entre plusieurs de ces thèmes. Tout effort de recherche dans ces champs disciplinaires en lien avec ces thématiques est évidemment le bienvenu, mais l'ambition de cette contribution est de dégager des tendances nouvelles et des fronts de recherche à des interfaces encore peu étudiées.

Les évolutions technologiques, environnementales et sociétales des deux dernières décennies ont profondément renouvelé les enjeux de recherche dans les champs statistiques concernés par ces domaines – et vont très certainement continuer à le faire dans un avenir prévisible. Ces évolutions ou tendances, nous pouvons les décliner en 4 axes, bien entendu articulés les uns aux autres. La première évolution est en lien avec l'explosion informatique et les progrès technologiques des instruments de mesure qui nous ont fait passer de la gestion de la rareté des données à la gestion de leur profusion. La deuxième évolution a trait aux champs étudiés : la climatologie, l'environnement, l'étude des individus, des peuplements et des communautés en écologie nécessitent d'appréhender les phénomènes étudiés à des échelles de plus en plus grandes. La troisième tendance majeure que nous percevons correspond à une demande grandissante dans notre société : explorer les possibles afin de cerner – et si possible quantifier – les incertitudes pour éclairer et rationaliser la décision dans l'incertain. En bref, passer d'une vision déterministe à une vision plus stochastique. La dernière évolution enfin, qui découle en partie des trois premières nous fait passer d'une approche analytique d'un phénomène particulier, étudié en situation plus ou moins bien contrôlée à une approche intégrée, systémique prenant en compte l'ensemble des facteurs agissant sur un système. Nous allons maintenant détailler comment, selon nous, ces grandes tendances font évoluer la science statistique dans son rapport avec l'étude de la planète Terre. Du fait de nos préoccupations de recherche, nous nous focaliserons plus particulièrement sur les deux premiers volets de l'ARP : Terre fluide et Terre vivante.

De la rareté des données à leur abondance

Les progrès de l'instrumentation, des systèmes électroniques embarqués qui permettent par exemple d'étudier les déplacements d'animaux et de l'imagerie satellite génèrent de très grandes quantités de données. Un grand nombre de variables, parfois assez fortement dépendantes, sont disponibles. Celles-ci sont souvent très hétérogènes en type (qualitative, quantitative discrète, quantitative continue), en qualité, en support spatial concerné, en quantité. Les défis posés par cette profusion de variables sont considérables. On assiste de fait, à l'éclosion d'une « éco-informatique » ou « envi-informatique », à l'image de la bio-informatique il y a une vingtaine d'années. Ces disciplines en devenir posent de nouvelles questions informatiques liées à la description, au format, au stockage et à la diffusion de ces données qui sortent du cadre de cette contribution. Elles vont également renouveler profondément le champ des statistiques.

---

[1] Une précision s'impose concernant l'usage du terme modèle, utilisé à de nombreuses reprises dans ce texte. Pour un mathématicien, il désigne les équations décrivant le système étudié ; pour un statisticien, il désigne l'ensemble des choix menant aux distributions de probabilité représentant les variables étudiées ; pour un scientifique, il désigne le modèle numérique utilisé pour simuler le comportement des systèmes biologiques physiques étudiés. En règle générale, le contexte permet de comprendre de quel modèle il s'agit. Pour éviter toute ambiguïté, nous avons ajouté un adjectif précisant le sens lorsque cela était nécessaire.

- L'hétérogénéité de ces données nécessite de définir des modèles statistiques complexes avec de multiples dépendances entre les variables. Les modèles bayésiens sont particulièrement adaptés pour décrire des schémas de dépendance complexes. Les algorithmes d'estimation pour ces modèles sont basés sur les simulations Monte Carlo par Chaînes de Markov qui nécessitent des temps de calcul très longs. Un enjeu consiste à rendre ces chaînes plus efficaces ; c'est l'objet des recherches, encore nécessaires, visant à améliorer les propriétés de mélangeance de ces chaînes ; c'est également l'objectif des méthodes basées sur l'approximation de Laplace ou, plus récemment, des approximations de Laplace imbriquées INLA (Integrated Nested Laplace Approximation).
- Le volume de données est parfois tel que les fonctions et routines numériques habituellement utilisées par les statisticiens ne peuvent plus être utilisées sans des modifications profondes. Il existe un réel besoin, non couvert à ce jour en France, pour développer des méthodes numériques efficaces utiles au statisticien, à l'image sans doute de la synergie qui existe de longue date entre mathématicien des équations différentielles et numériciens.
- Du fait de l'abondance de données et en particulier des covariables accessibles, les questions posées aux statisticiens seront (et sont déjà) moins souvent l'estimation de telle quantité ou tel paramètre et plus souvent la sélection des covariables pertinentes et le test d'hypothèses scientifiques dans ce type d'environnement. La multiplication des facteurs explicatifs amène à une variété de modélisation quasiment infinie. La sélection de modèles est un thème statistique développé depuis maintenant une vingtaine d'années avec des techniques aussi populaires que PLSR (Partial Least Square Regression) ou LASSO (Least AbsoluteShrinkage and SelectionOperator). La recherche sur ce thème continue à être très active. Elle vise à proposer de multiples variantes des méthodes pré-citées qui permettent de sélectionner parmi des familles de modèles gigantesques suivant des critères pouvant s'adapter aux configurations des données et objectifs de l'étude. Une autre façon d'aborder le choix de modèle, qui est un thème de recherche très actif en ce moment, est de les considérer tous, par exemple en les agrégeant ou en activant une procédure de sélection adaptative.
- Le maximum de vraisemblance, difficile voire impossible à calculer, n'est plus l'outil unique ni même principal des méthodes d'estimation ; il peut être remplacé par d'autres scores : pseudo-vraisemblance, quasi-vraisemblance, vraisemblance profilée, contrastes calculés sur des moments. Bien souvent on utilise une combinaison de ces scores pour les différents paramètres. Les propriétés de ces méthodes d'estimation ne sont pas toujours connues aussi bien que celles du maximum de vraisemblance habituel. Dans certains cas, et en particulier quand la vraisemblance ne peut se calculer voire ne peut s'écrire (c'est le cas pour les modèles individu-centrés par exemple), un champ de recherche récent et actif consiste, encore plus radicalement, à s'en passer en recourant à des techniques de type ABC (Approximate Bayesian Computation).
- La fréquence d'acquisition des données a également fortement augmenté, permettant la mise à jour régulière des estimations. Il serait sans doute intéressant que le rapprochement entre la communauté statistique et celle de l'assimilation de données s'amplifie, en particulier pour développer des algorithmes efficaces en assimilation de données spatialisées.
- Des domaines tels que la surveillance de la biodiversité demandent une quantité d'observations ne pouvant être obtenue par les moyens conventionnels car encore trop coûteuses. On a alors recours aux approches participatives : les données sont collectées par des citoyens bénévoles, naturalistes amateurs et/ou grand public qui alimentent des bases de données disponibles en ligne. Ces données ont l'avantage d'être massives (plusieurs dizaines de millions d'observations) mais ne peuvent être exploitées directement à cause de leur très grande hétérogénéité (en termes de qualité, de répartition spatiale et temporelle, d'effort d'observation, etc.). Les défis à surmonter sont multiples : données massives et en grande dimension, exploitation pertinente des régularités spatio-temporelles, modélisation fine des processus de récolte des données, identification des possibles biais socio-géographiques. Des recherches sont nécessaires pour affiner les techniques statistiques permettant d'extraire l'information

disponible dans ces données, notamment en s'appuyant en partie sur des données moins abondantes mais mieux standardisées.

De l'échelle locale à l'échelle régionale ou globale

Lorsqu'on évoque la planète Terre, il vient immédiatement à l'esprit l'échelle globale, ou du moins continentale ou semi-continentale. De fait, c'est bien à cette échelle que l'on rencontre de nombreux verrous méthodologiques et/ou numériques en statistiques.

- Tout d'abord un travail fondamental autour de la définition des modèles pertinents de champs aléatoires (même Gaussien!) spatio-temporels et multivariés. Des progrès significatifs ont été faits récemment, mais le champ de recherche reste très actif. En particulier, seuls des modèles spatio-temporels multivariés simplistes existent à l'heure actuelle. Il en va de même pour les modèles de champs aléatoires définis sur la sphère (ou sur une variété quelconque) et de ceux qui vérifient des lois physiques, comme la nullité de la divergence ou du rotationnel.
- Un second enjeu, véritable serpent de mer des statistiques spatiales, consiste à faire communiquer les échelles et les niveaux d'organisation. La diversité croissante des technologies d'acquisition amène à mesurer une même grandeur à des supports différents : du prélèvement ponctuel, au pixel de taille variable obtenu par imagerie satellite ou aéroportée, en passant par le transect pour des observations visuelles, mais aussi de nature différente requérant le passage par des proxys. Pour faire communiquer ces échelles et niveaux d'organisation on a souvent recours à la modélisation hiérarchique pour laquelle les besoins en algorithmes d'inférence et simulation efficaces et rapides sont criants.
- Bien entendu, travailler aux grandes échelles spatiales nécessite en général des quantités de données assez grandes, ce qui nous renvoie aux enjeux développés au paragraphe précédent. L'étude du climat à l'échelle globale en est une illustration. La dimension spatio-temporelle des jeux de données utilisés est grande, typiquement 10^5 pour un seul paramètre (ex : température), au pas de temps annuel et à une résolution spatiale de 500 km, donc relativement grossière. Dans un certain nombre de problèmes, faire des statistiques nécessite au préalable de caractériser la variabilité climatique, donc a minima d'estimer la matrice de covariance de ce vecteur, et même sa distribution si l'on se passe de l'hypothèse Gaussienne. Il s'agit d'un problème d'estimation d'une matrice de covariance en grande dimension, thématique actuellement en plein essor en probabilités / statistiques. Dans le cas du climat, le problème est difficile à simplifier, car il est difficile de faire des hypothèses paramétriques simplificatrices sur la structure de cette matrice, qui n'est pas creuse. En effet, il existe des « dépendances » spatiales et temporelles de grande échelle, comme par exemple les événements El Niño ou certains modes de variabilité océanique qui ont des constantes de temps caractéristiques de plusieurs décennies. Afin de contourner ce problème, une alternative consiste à réduire a priori la dimension des données, ce qui est souvent fait de façon empirique. Une réflexion méthodologique sur l'optimalité des techniques de réduction utilisées semble utile.

D'un paradigme déterministe à un paradigme statistique

Dans ce domaine, la demande sociétale, exprimée à travers les décideurs, les média ou les associations citoyennes, est particulièrement forte. Fournir une estimation, une prédiction unique, fut-elle le résultat du travail le plus honnête, le plus méticuleux et le plus scientifiquement établi, ne suffit plus. La société, dans toutes ses composantes, demande au minimum une quantification de l'incertitude autour de ce point, et le plus souvent une analyse par scénarios, chacun avec ses incertitudes et ses probabilités. Le statisticien est habitué de longue date à fournir un intervalle de confiance autour de son estimation. Le numéricien utilise depuis longtemps les filtres de Kalman, qui mélangent les modèles physiques et statistiques. Notons que, dans ce domaine particulier, des recherches restent nécessaires pour proposer des méthodes et algorithmes efficaces pour traiter la grande dimension et les non-linéarités.

La nouveauté est que la nécessité de quantifier les incertitudes concerne maintenant des domaines où la culture déterministe – pour de bonnes raisons liées à la physique des équations par exemple – est dominante. En un mot, il faut mettre des statistiques dans les approches déterministes et réciproquement venir compléter des modèles statistiques par des mécanismes mécanistes ou physiques, lorsque cela est possible.

- Dans cette optique, un premier champ d'activité assez actif en ce moment en mathématiques pour l'écologie et pour l'évolution – ainsi qu'en épidémiologie où les besoins sont pressants – consiste à croiser les modélisations déterministes (EDO et EDP) et la modélisation statistique. Les modèles déterministes, notamment ceux basés sur des EDP, sont généralement des modèles « pour la connaissance ». Ils peuvent aboutir à des résultats théoriques qualitatifs et permettre une meilleure compréhension de phénomènes biologiques, mais en règle générale il est difficile de comparer leur solution aux données disponibles. Ces modèles d'EDP sont également généralement peu réalistes au sens où ils sont loin des mécanismes biologiques sous-jacents. Une façon de répondre à ces lacunes repose sur la combinaison de modèles d'EDP avec des modèles statistiques / probabilistes en considérant des modèles hybrides.
  - Parmi ceux-ci, les modèles dits mécanistico-statistiques combinent un sous-modèle d'EDP décrivant la dynamique spatio-temporelle du processus étudié avec un sous-modèle statistique décrivant le processus d'observation. Cette approche combine les avantages des modèles statistiques empiriques en permettant de prendre en compte un grand nombre de types de données (binaires, censurées), et ceux des modèles déterministes du type EDP, qui créent une forte contrainte sur la dynamique et réduisent ainsi les incertitudes sur les paramètres estimés. Des efforts de recherche pour ce type d'approche sont nécessaires, pour mieux en connaître les propriétés mathématiques, ainsi que pour proposer des méthodes d'estimation de leur paramètre qui soient efficaces et robustes.
  - Un second type de modèle hybride, qui intervient très naturellement en dynamique des populations du fait de la diversité des échelles impliquées, combine des modèles discrets, comme les modèles individu-centrés et des modèles agrégés tels que les modèles d'EDP. Ce type de modèle devrait offrir un meilleur réalisme que les modèles d'EDP notamment à faible densité de population, tout en conservant certains avantages des modèles d'EDP (temps de calcul, outils analytiques). Pour ces modèles, des efforts de recherche sur les propriétés probabilistes et sur les méthodes d'estimation sont également nécessaires.
- Un second champ d'activités consiste à développer des méthodes pour explorer les modèles numériques et/ou les ensembles de modèles, hiérarchiser les sources principales d'incertitude. La quantification des incertitudes est ainsi très importante dans le cas du climat, mais d'autres domaines scientifiques sont confrontés aux mêmes questions. Lorsqu'il s'agit de projections climatiques (simulations du climat du futur), les deux stratégies les plus couramment utilisées consistent à i) utiliser une méthode d'ensemble, c'est-à-dire perturber un modèle physique donné pour évaluer la dispersion des réponses simulées, ou ii) utiliser l'ensemble des modèles disponibles dans la communauté. Se pose la question, partiellement résolue à ce jour seulement, de quantifier l'incertitude à partir de ces sorties de simulations, en utilisant par exemple les méthodes de scoring. Les modèles numériques utilisés ne pouvant pas véritablement être considérés comme indépendants (sans qu'on sache réellement mesurer leur dépendance), des recherches sont nécessaires afin de développer des méthodes de scoring adaptées.
- Parce qu'il est très coûteux de réaliser un nombre important de simulations numériques pour évaluer l'effet de la variation de ces paramètres, des approches reposant sur l'utilisation d'émulateurs ont été développées. Ceux-ci permettent de simuler, de manière très rapide, des trajectoires dont les comportements statistiques sont similaires à ceux des modèles numériques d'origine. Cette stratégie, en plein essor, offre une porte d'entrée pour explorer l'espace des paramètres d'entrées.

D'une vision analytique à une approche intégrée

La plupart des grands défis sociétaux qui structurent actuellement les schémas stratégiques de recherche nationaux ou internationaux soulèvent des questions qui appellent une approche multidisciplinaire et intégrée. Ce passage d'une approche analytique qui décrit au plus près des mécanismes élémentaires à une approche très intégrée constitue une toile de fond pour les questions de recherche qui touchent au système Terre. Ainsi par exemple, c'est aussi à l'échelle de l'écosystème que désormais les décideurs et les porteurs d'enjeux envisagent la conservation. Cette façon globale d'approcher les questions de recherche mobilise l'ensemble des questions déjà vue dans les paragraphes précédents, mais soulève également de nouvelles questions.

- Un premier ensemble de questions, assez techniques, provient du couplage des modèles physiques agissant dans des compartiments différents du système Terre, couplage présent dans toute approche intégrée. Les questions relatives à l'exploration des modèles et des ensembles de modèles, à la hiérarchisation des sources principales d'incertitude se reposent dans ce contexte, avec la difficulté supplémentaire de devoir gérer la propagation des variabilités (erreurs et incertitudes) entre modèles, et les rétro-actions possibles entre différents modèles.
- Enfin, une approche intégrée du système Terre ne peut ignorer la place de l'homme dans le système, à la fois comme agent modifiant *ex-ante* l'état du système, et comme agent tentant de s'adapter *ex-post* aux modifications de celui-ci. Ces interactions, aux rétro-actions complexes, appellent à une meilleure articulation entre les mathématiques appliquées, les géo-sciences et les sciences humaines et sociales en lien avec la Terre.