Réflexions sur la partie V du rapport du BAP sur l'agroécologie

Timothée Flutre

Réflexions sur la partie V du rapport du BAP sur l'agroécologie

Auteur : Timothée Flutre (INRA, BAP, UMR AGAP, équipe DAAV, UMT Géno-Vigne)

Date : 04/01/2016

Préambule

Le département Biologie et Amélioration des Plantes (BAP) de l'INRA a confié à sept de ses chercheurs une mission de réflexion sur l'implication de la génétique et l'amélioration des plantes dans la thématique de l'agroécologie. Par ailleurs, au sein de l'UMR AGAP qui rassemble des agents du CIRAD, de l'INRA et de Montpellier SupAgro, un groupe de réflexion s'est rassemblé sur ce thème, prenant comme point de départ le rapport établi entre temps pour le compte du BAP (Litrico et coll., 2014). En tant que généticien quantitatif au BAP, impliqué dans le développement d'une approche intégrée de la création variétale chez la vigne, et ayant activement participé au groupe de réflexion sur l'agroécologie d'AGAP, je retranscrit dans ce document quelques réflexions suscitées par la partie V du rapport concernant la méthodologie de la sélection. Elles peuvent être perçues comme des critiques du rapport, mais au contraire, c'est bien là toute la pertinence du rapport que d'avoir suscité une telle démarche de lecture critique. C'est aussi pour moi l'occasion de tisser des liens entre disciplines en rassemblant des références rarement discutées dans un même cadre, et de lancer des pistes de réflexion dans plusieurs directions. Il ne me reste plus qu'à espérer que celles-ci soient quelque peu pertinentes.

Éléments contextuels concernant l'INRA et le BAP

Sur son site officiel[1], l'INRA s'est donné comme objectif de « nourrir le monde de façon saine et durable d'ici 2050 ». Sa stratégie se décline en quatre priorités : « (1) améliorer toutes les performances de l'agriculture, (2) atténuer les changements du climat et s'y adapter, (3) transformer la biomasse pour la chimie, l'énergie, les bio-matériaux, et (4) assurer des systèmes alimentaires sains et durables ». Pour ce faire, deux « domaines interdisciplinaires émergents » ont été identifiés : « l'agroécologie et la biologie prédictive ».

Quant au BAP, selon son site officiel[2], il a cinq missions : « (1) produire des connaissances et explorer de nouveaux fronts de science dans les domaines du développement des plantes, de leur adaptation à l'environnement et de la qualité des produits des végétaux ; (2) comprendre et maîtriser le déterminisme génétique des caractères d'intérêt agronomique et de ceux liés à l'usage des produits des plantes ; (3) contribuer à la construction d'une biologie intégrative, basée sur des approches prédictives novatrices, allant de l'étude des gènes et du génome jusqu'à l'innovation variétale ; (4) développer la recherche translationnelle en biologie des plantes et occuper une position stratégique dans ce domaine ; et (5) contribuer à la conservation et à la valorisation des ressources génétiques végétales et à la sélection de matériels innovants. »

Bref résumé de la partie « Méthodologie de la sélection »

Dans la partie V du rapport du BAP sur l'agroécologie, les méthodologies de sélection sont catégorisées en deux approches. Dans la première, l'approche « boîte noire », il n'est pas « nécessaire de connaître les mécanismes et les caractères impliqués dans la réponse aux contraintes environnementales ». La sélection s'opère directement dans l'environnement cible et, par exemple, le rendement est utilisé comme synthèse de performance. Dans cette approche, la sélection phénotypique décentralisée et la sélection participative sont mentionnées, la dernière permettant aussi une meilleure adoption des variétés développées.

La seconde approche, qualifiée de « mécaniste », implique de « connaître les caractères impliqués dans la réponse à la contrainte environnementale ». Ceci est décrit comme revenant à optimiser un critère de sélection (voire plusieurs) à partir des héritabilités des caractères et des corrélations génétiques.

Il est précisé que ces deux approches ne sont pas mutuellement exclusive, notamment via la fixation de gènes de résistance en amont (par SAM), mais qu'elles diffèrent principalement par la structure génétique du matériel végétal utilisé. La stratégie classique est mono-génotypique. Elle vise à obtenir une variété « tout en un », les « élites » (lignées pures, hybrides F1, clones). Cette stratégie est majoritaire au BAP, et est encouragée par les règles DHS du CTPS.

L'agroécologie se fondant sur la valorisation de la diversité génétique et des interactions biologiques au sein des agrosystèmes, la stratégie pluri-génotypique est présentée, celle-ci ayant déjà montrée

1 http://institut.inra.fr/Recherches-resultats/Strategie
2 http://www.bap.inra.fr/Le-departement/presentation

son intérêt mais restant encore sous-utilisée. Au sein de peuplements mono-spécifiques, plus de recherches sont nécessaires afin d'explorer, par exemple, l'association de variétés ayant différents gènes de résistance, la sélection pour l'aptitude au mélange, la sélection directement des mélanges eux-mêmes, etc.

A cela s'ajoute la possibilité d'assemblage multi-spécifique, notamment pour répondre aux services écosystémiques autres que la seule production individuelle de chaque espèce.

Réflexions sur quelques notions-clés

Relations science-société

Rien que par son titre, « Méthodologie de la sélection », la partie V du rapport du BAP se distingue des autres par son appartenance plus explicite au registre de l'*action*, au-delà des registres plus familiers des scientifiques que sont la *description* des phénomènes naturels et la *compréhension* des mécanismes sous-jacents. Cette distinction est importante car, depuis les années 1990, le BAP s'est substantiellement désengagé des activités de sélection afin de se recentrer sur la science dite académique, notamment « l'acquisition de connaissances d'amont et le développement de méthodologies de sélection » (Lefort et Riba, 2006). A l'heure actuelle, en ce qui concerne la sélection, le BAP est surtout impliqué au niveau du *pre-breeding* : « fait référence à toutes les activités visant à identifier les caractères d'intérêt et/ou les gènes au sein de matériel végétal non-adapté qui ne peut être utilisé directement dans les schémas de croisement, et à transférer ces caractères dans du matériel intermédiaire que les sélectionneurs peuvent utiliser pour produire de nouvelles variétés » (Renard, 2015).

Dans ce contexte, le rapport du BAP rappelle que « l'agroécologie est fondée sur l'hypothèse qu'il est possible d'augmenter les productions agricoles en quantité et en qualité et de réduire l'impact de l'agriculture sur l'environnement en valorisant la diversité génétique et les interactions biologiques au sein des agrosystèmes ». Mais d'où viendrait l'inefficacité des acteurs actuels de la sélection à (suffisamment) valoriser ces aspects de diversité et d'interactions ? Le rapport semble fournir un début d'explication en indiquant que « l'approche majoritairement utilisée aujourd'hui » correspond à la stratégie monogénotypique « encouragée par les règles d'inscription des variétés au CTPS notamment orientées par le besoin de distinction variétale sur la base d'une variabilité inter-variétés supérieure à une variabilité intra- », et précisant « qu'avec de telles structures génétiques, la possibilité d'introduire de la diversité dans la parcelle reste limitée ».

La réglementation actuelle et son « paradigme fixiste » (Bonneuil *et al.*, 2006) semble donc conditionner plus qu'autre chose la valorisation des aspects de diversité et d'interactions. Le premier verrou à lever semble donc principalement réglementaire et non méthodologique. Sur ce point, les auteurs du rapport ont fait le choix de préciser dans le préambule que « l'anticipation des verrous réglementaires et socio-économiques […] sont traités dans le cadre du chantier Agroécologie-INRA et dans le Métaprogramme-EcoServ ». Mais ce choix est-il pertinent en ce qui concerne la partie V ? Certains écophysiologistes peuvent à la rigueur arguer que les aspects réglementaires sur les variétés ne les concernent pas directement, mais cela devient nettement plus difficile en ce qui concerne les généticiens impliqués dans le *pre-breeding* ou la sélection, et, de plus en plus, les

biologistes moléculaires (cf. les débats sur la brevetabilité du vivant en parallèle des progrès sur l'édition de génomes).

Il eut donc été préférable de poser crûment la question du rôle du chercheur du BAP dans une telle situation. La partie V du rapport, du fait qu'elle traite de sélection, aurait été l'endroit opportun pour introduire aux lecteurs peu familiers du domaine des *science studies* le fait que les rapports science-société ne se posent plus dans les termes de la conception moderne : « society [is] like the flesh of a peach, and science its hard pit » (Latour, 1998). Pour reprendre les mots du même auteur dans un entretien avec des collègues de l'INRA, « toute recherche est action » (Latour, 1997). Ne pas se préoccuper des aspects réglementaires sous prétexte que ce n'est pas à un généticien de le faire, c'est choisir de laisser à d'autres ce qui conditionne pourtant l'activité même du généticien.

Ce débat dépasse bien entendu de beaucoup les généticiens et le contexte du BAP. Cependant, à l'occasion de l'encadré 4 présentant la sélection participative distribuée, il est pertinent de mentionner les réflexions de chercheurs d'autres domaines mais résonnant particulièrement bien avec ce type de démarche, par exemple en économie (Dreze, 2002). De manière plus générale, lorsque l'on parle de « sélection », il est aussi important de rappeler les initiatives pour plus et mieux impliquer les agriculteurs en contact *direct* avec les chercheurs dans les processus de recherche et développement via des innovations à petite échelle (MacMillan et Benton, 2014).

Mécanisme et causalité

Le terme de mécanisme est utilisé dans la partie V du rapport du BAP pour différencier initialement deux méthodologies de sélection, « boîte noire » et « mécaniste ». La notion de mécanisme se comprend au travers de la notion de causalité. En effet, c'est une fois que l'on a pu montrer le caractère causal d'une relation que l'on peut véritablement parler de mécanisme. Or, dans sa description de l'approche mécaniste, le rapport parle de « corrélations génétiques » et de « niveau d'héritabilité ». De plus, le rapport reconnaît par la suite que les deux méthodologies de la sélection ne sont pas mutuellement exclusives. A la lecture du rapport, il n'est donc pas aisé de comprendre ce qui distingue concrètement ces deux approches, voire même s'il y a lieu de les distinguer.

Premièrement, il est bon de rappeler qu'une corrélation n'implique pas la causalité (Pearl, 2009). C'est particulièrement important avec des jeux de données multivariées, d'autant plus qu'ils sont consubstantiels à la démarche agroécologique (plusieurs caractères, plusieurs environnements, plusieurs génotypes, plusieurs espèces, etc). Deuxièmement, dans un contexte de génétique et de sélection, il est indispensable d'être clair sur le fait que l'héritabilité correspond à une proportion de variation phénotypique expliquée par une variation génotypique. Celle-ci ne devrait donc pas être interprétée en terme de causalité (Lewontin, 2006).

Mais au-delà de ces aspects un peu techniques, bien qu'essentiels, il est pertinent de réaliser que la causalité peut être abordée de deux manières différentes (Gelman, 2011) : en estimant les effets des causes (*forward causal questions*) ou en recherchant les causes des effets (*reverse causal questions*). La statistique inférentielle se préoccupe très majoritairement de la première, la deuxième nécessitant une intervention (*doing*), et non seulement une observation (*seeing*), pour conclure sans ambiguïté, de telles interventions n'étant pas toujours possibles en pratique (Lindley, 2002). Mais

certains ont récemment proposé un autre point de vue visant à incorporer la recherche de causes dans le cycle de la statistique inférentielle classique (Gelman et Imbens, 2013). Dans ce nouveau cadre, les questions de causalité vers le futur traitent d'estimation alors que les questions de causalité vers le passé traitent de vérification de modèle et de construction d'hypothèses.

Dans ce contexte, il apparaît plus clairement non seulement que les deux méthodologies de la sélection sont complémentaires, mais surtout qu'elles ne se distinguent pas par le fait d'être plus ou moins « mécanistes » mais plutôt par d'autres critères tels que l'utilisation de données génétiques et non seulement phénotypiques, l'utilisation d'essais dans plusieurs environnements, etc.

Modélisation statistique et décision en condition d'incertitude

En mettant l'accent sur la valorisation de la diversité génétique et des interactions biologiques, l'agroécologie suppose une réflexion sur l'identification des caractères d'intérêt, leur combinaison, leur déterminisme génétique, etc. Ces aspects font l'objet des parties II et III du rapport du BAP. A ce sujet, ces parties mentionnent quelques pistes de recherche à développer sur les aspects de modélisation. Mais ces aspects sont trop absents de la partie V sur la sélection. A l'heure actuelle, selon les espèces considérées, les acteurs concernés et la réglementation en cours, il n'est pas toujours possible en pratique d'utiliser un modèle formel visant à prendre la décision (la plus) optimale dans un programme de sélection. Mais en tant que chercheur du secteur public, il apparaît important d’œuvrer à la mise au point de modèles formels permettant une meilleure utilisation des données que les procédures informelles de sélection, et une meilleure compréhension de l’efficacité des différentes méthodes de sélection.

Historiquement, deux grands courants de modélisation ont avancé en plus ou moins grande ignorance l'un de l'autre. D'un côté, avec leur « modèle mixte », les généticiens ont formalisé *statistiquement* la contribution génétique à la variation phénotypique observée au sein d'un ensemble d'individus apparentés (Visscher, Hill et Wray, 2008). De l'autre, avec leurs « modèles de culture », les écophysiologistes ont décomposé le système plante-milieu en processus *physiques* (Brisson, Wery et Boote, 2006). L'intérêt de l'approche génétique réside dans le fait que des individus génétiquement différents soient analysés conjointement. L'intérêt de l'approche écophysiologique réside dans le fait que plusieurs caractères importants soient analysé conjointement. Naturellement, la force de l'un correspond à la faiblesse de l'autre : les généticiens se restreignent la plupart du temps à analyser les caractères séparément, et les écophysiologistes se restreignent la plupart du temps à analyser les individus séparément.

Mais ce n'est pas que les modèles des uns soient plus pertinent que les modèles des autres. C'est plutôt qu'ils n'ont pas les mêmes données sur les mêmes individus. Cette situation a déjà changé avec les technologies de génotypage à haut-débit, et est en train de changer avec l'introduction d'appareils de phénotypage à haut-débit. Les données évoluant, il faut maintenant que les modèles évoluent. Mais cela ne revient pas à demander aux uns d'abandonner leurs modèles au profit des autres. Non, cela consiste plutôt à reconnaître l'importance d'analyser conjointement les données tout en prenant en compte l'hétérogénéité des sources de données, ceci permettant de propager l'incertitude correctement entre les variables inconnues du modèle (Green, 2003). En un mot, c'est mettre plus de modélisation statistique dans l'analyse des données de génétique et d'écophysiologie.

Ceci est d'autant plus important qu'une fois l'idéotype définit et une fois les données collectées, c'est l'heure du choix, c'est-à-dire de la décision de garder tel individu ou telle population, que ce soit dans un but direct de production ou dans le but de poursuivre l'évolution du matériel végétal, alors qu'on n'*observe* pas la valeur génétique des individus, mais qu'on l'*estime*. Cette étape de sélection consiste donc bien à prendre une décision en condition d'incertitude, objet d'étude de la *théorie de la décision statistique* (Berger, 1985). En pratique, il est donc hautement recommandé de se préoccuper de modélisation au début des projets scientifiques et en amont des programmes de création variétale, et non une fois que les données aient été collectées.

Sélection multi-caractères

L'importance de la modélisation statistique est d'autant plus centrale que, dans une démarche agroécologique, l'idéotype se complexifie rapidement en impliquant conjointement plusieurs caractères. Il devient alors de plus en plus nécessaire d'utiliser un modèle formel, un aspect trop peu présent dans la partie V du rapport.

Le modèle classique d'indice de sélection permet d'analyser plusieurs caractères simultanément en les combinant linéairement, ceux-ci étant pondérés par leur valeur économique relative (Hazel, Dickerson et Freeman, 1994). Des extensions permettent de prendre en compte l'incertitude autour des valeurs économiques, de les combiner de manière non-linéaire, de réaliser une sélection en plusieurs étapes, de maintenir certains caractères constants, etc. Mais ces modèles ont cependant des limites aux deux étapes de leur mise en œuvre.

La première étape consiste comme d'habitude à estimer les variances génétiques de chaque caractère, mais, cette fois-ci, il faut également estimer les covariances génétiques entre caractères. Or la méthode classique d'estimation n'est plus stable au-delà d'un nombre somme toute assez restreint de caractères : par exemple, 10 caractères impliquent une matrice de variance-covariance contenant 100 paramètres. Cette limite concerne également le modèle de sélection génomique multivarié cité dans le rapport (Jia et Jannink, 2012). Non seulement il devient utile de se pencher sur l'estimation des matrices de variance-covariance (Barnard, McCulloch et Meng, 2000), notamment dans le cas de corrélations négatives, mais il peut également devenir nécessaire d'appliquer des méthodes de réduction de dimension, celles-ci devant être utilisées conjointement à l'estimation des paramètres (Runcie et Mukherjee, 2013).

Une fois les variance-covariances génétiques estimées, la deuxième étape consiste à calculer l'indice de sélection à proprement parler. Celui-ci requiert de connaître les valeurs économiques des caractères, même imparfaitement. Or les modèles existants considèrent généralement que seul le profit est à maximiser. Il serait donc nécessaire d'adapter ces modèles pour prendre en compte une multiplicité d'objectifs autres que le seul profit, tels que le maintien d'un certain niveau de diversité génétique au sein du matériel végétal en sélection, la capacité à s'affranchir des intrants extérieurs à l'environnement de culture, l'introduction de critères de durabilité, etc, critères pour lesquels il est moins évident de choisir des pondérations, notamment du fait même qu'il n'existe pas de consensus sur la signification de « durabilité » (Ostrom, 2009) ou de « résilience » (Döring *et al.*, 2015).

De plus, lorsque l'on parle de sélection multi-caractères, il est généralement implicite que chaque

caractère est lui-même univarié. Or de nombreux phénomènes biologiques n'ont de sens que s'ils sont caractérisés dans leur dimension temporelle. Bien sûr, à coût constant, cela dépend de l'augmentation du débit des technologies de phénotypage pour qu'elles soient utilisées sur du matériel végétal d'une diversité génétique suffisamment large pour permettre la sélection. Mais cela pose aussi immédiatement la question du modèle statistique permettant d'évaluer la contribution génétique à la variation phénotypique mesurée via des séries temporelles. Certains modèles de détection de QTL ont été proposés pour des caractères fonctionnels (Kwak, Moore, Spalding et Broman, 2015), mais il reste encore à les développer pour analyser conjointement plusieurs caractères fonctionnels. Cela nécessitera de ne plus travailler seulement avec des matrices de phénotypes, mais avec leur généralisation en dimension 3 et plus (Hoff, 2011).

Sélection multi-environnements

En insistant sur l'importance de tirer profit des particularités de chaque écosystème, l'agroécologie implique de sélectionner des variétés adaptées à tel ou tel environnement, c'est-à-dire à spectre moins large peut-être que les variétés actuelles. Mais bien sûr, tout dépend de ce qu'on entend par « environnement ».

L'objet désigné par le terme « environnement » est naturellement multi-dimensionnel. Habituellement, on le décrit via le climat, le sol, les pratiques culturales, etc, ces catégories étant elles-mêmes raffinées. Mais d'autres dimensions ont été proposées, notamment pour mieux tirer parti des interactions entre génotypes et « environnements » (quelle que soit la signification donnée à ce terme). Outre l'environnement biophysique et les pratiques culturales, certains auteurs ajoutent les compétences des acteurs, les circuits et lieux de distribution et vente, les réglementations et la société au sens large (Desclaux, Nolot, Chiffoleau, Gozé et Leclerc, 2008). Dans une démarche agroécologique englobante, notamment lorsqu'on se sent concerné par l'acceptabilité des nouvelles variétés issues de la sélection, il paraît donc pertinent de ne pas limiter sa réflexion à l'environnement biophysiques et aux pratiques culturales.

Sur le plan méthodologique, la sélection décentralisée et participative (que l'on peut aussi qualifiée de « distribuée et collaborative ») est décrite dans un encadré de la partie V du rapport du BAP. Une telle sélection est généralement associée à des échanges de semences entre acteurs. On parle alors de gestion dynamique de la ressource génétique au sein d'un collectif. Un tel contexte a la particularité de mettre en exergue l'importance de la structure du réseau d'échange qui influence en retour la diversité et l'adaptabilité du matériel végétal échangé (Barbillon, Thomas, Goldringer, Hospital et Robin, 2015). Cependant les modèles de génétique quantitative les plus utilisés ne tiennent pas encore compte de ces aspects sociaux.

Au-delà des raisons sociologiques et motivations politiques pouvant servir à interpréter l'implication d'acteurs dans une telle méthode de sélection (Demeulenaere, 2014), il existe aussi des raisons d'optimisation mathématique allant dans le sens d'une meilleure efficacité de ce type de sélection. Il a en effet été montré que le fait de faire varier l'environnement lors d'un processus de sélection a pour conséquence d'accélérer l'évolution par rapport à l'alternative habituelle de garder l'environnement fixe (Kashtan, Noor et Alon, 2007). Les raisons derrières ce phénomène ont trait au fait que la sélection pour un optimum donné en milieu changeant a tendance à également

sélectionner des architectures modulaires, plus faciles à faire évoluer par la suite. Le gain est d'autant plus intéressant lorsque l'optimum est multivarié, ce qui est typiquement le cas dans la démarche agroécologique comme discuté dans la section précédente.

Perspectives pour le BAP et l'UMR AGAP

Parmi ces réflexions éparses suscitées par la partie V du rapport du BAP, une notion transversale ressort particulièrement, celle de « modélisation statistique ». Je choisis d'employer ici le terme « modélisation » plutôt que « modèle » afin de mettre l'accent sur le processus à l'œuvre, c'est-à-dire sur le fait qu'il n'y a pas « un modèle » mais plutôt une succession de modèles, que l'on étend et raffine au fil des expériences et des questions posées. Je préfère aussi le terme « statistique » à « mathématique » pour insister sur l'importance de quantifier l'incertitude.

Dans la démarche agroécologique telle que présentée dans le rapport du BAP, notamment la partie V, il apparaît clairement que les questions posées se complexifient en impliquant des objets hétérogènes de grandes dimensions et de multiples échelles spatio-temporelles. A ce stade, il est indispensable de réaliser que la modélisation statistique n'est pas qu'un simple outil d'analyse de tableaux de données. Bien au contraire, elle sert de « liant », non seulement pour les données mais aussi pour les acteurs ayant généré ces données ou intéressé par ce que l'on peut en tirer. On peut même aller jusqu'à interpréter la modélisation statistique comme un véritable moyen de communication (Hennig, 2010).

Les modélisateurs-statisticiens intéressés par la méthodologie de la sélection, donc intéressés par les aspects appliqués plus que théoriques, ont de fait une position centrale. Ils développent et utilisent les modèles qui relient les données génotypiques, les données phénotypiques, les contextes environnementaux et les hypothèses scientifiques à évaluer. Ils naviguent également entre les différents acteurs de la sélection, qu'ils soient chercheurs, par exemple en génétique et écophysiologie, ou membres d'instituts techniques, agriculteurs, responsables syndicaux, militants associatifs, etc. Afin de permettre à ce « liant » de « prendre », un équilibre est à trouver : (1) les différents acteurs devraient rendre les données accessibles *facilement* aux modélisateurs-statisticiens ; (2) en retour, les modélisateurs-statisticiens se doivent de rendre leur travail *facilement* reproductible et compréhensible par les acteurs impliqués dans la sélection.

Ces aspects sont absents de la partie V du rapport du BAP, or de nombreuses difficultés surgissent dès la mise en pratique de ces recommandations. Mettre l'accent sur l'importance de la modélisation statistique a donc pour but de mettre en lumière les freins à lever et les écueils à éviter. Les générateurs de données du monde académique ne rendront accessibles leurs données brutes, ainsi que les méta-données, que s'ils sont reconnus en tant que tel, donc cités. Par exemple, il serait nécessaire que le système d'information du BAP (GnpIS[3]) censé stocker les données importantes générées par des agents de ce département mette en place la possibilité de faire des requêtes par identifiant de publication (DOI, PMID, etc), et signale de manière visible la nécessité de citer les générateurs des données.

De plus, de nombreuses données générées dans le cadre de projets de recherche ne sont pas

3 https://urgi.versailles.inra.fr/gnpis/

analysées par manque de temps ou de compétences suffisantes en modélisation statistique de la part des générateurs de données. Le BAP pourrait donc demander à ces agents la publication « d'articles de données » (*data papers*) au maximum 6 mois après la fin du projet. Comme les comportements ne changent pas sans incitation, ce type d'action pourrait être pris en compte dans l'attribution de financement additionnel aux équipes. Une telle incitation pourrait également être mise en place, bien qu'à plus petite échelle, au sein de l'UMR AGAP.

Dans le cadre de projets collaboratifs avec d'autres acteurs hors du monde académique, ce qui est souvent le cas dans les programmes de sélection, il est indispensable de mettre en place un contrat préalablement à la collaboration. C'est maintenant chose courante en ce qui concerne le matériel végétal produit dans le cadre du projet. Dans certains cas, la possibilité est donnée aux chercheurs de publier le résultat de leurs travaux, mais les données brutes ne sont rarement voire jamais explicitement mentionnées. De manière plus générale, il aurait été pertinent d'aborder dans le rapport du BAP la question des droits d'usage et des licences telles celles développées par Creative Commons, d'autant plus que l'INRA a un groupe de travail sur ces questions. Il est aussi intéressant de voir que certains acteurs non-académiques sont assez avancés sur le sujet (Réseau Semences Paysannes, 2015).

Au sein de l'UMR AGAP, la situation se complique du fait que l'INRA est un EPST, le CIRAD un EPIC et Montpellier SupAgro un EPSCP. Il y a donc peu de chance qu'un consensus soit trouvé. Mais *a minima*, il serait pertinent de sensibiliser chaque génération de doctorants à l'importance de la modélisation participative et des enjeux sous-jacents (Voinov et Bousquet, 2010).

En retour de l'accès aux données et méta-données, il est du ressort des modélisateurs-statisticiens appliqués d'assumer pleinement leur rôle de « liant ». La reproductibilité des analyses via l'utilisation de logiciels « libres » (à distinguer des logiciels seulement « ouverts ») est, en quelque sorte, le pendant de l'accessibilité aux données. Il est dommage cependant de l'avoir ignorée dans le rapport du BAP. Concrètement, de manière analogue aux « articles de données » qui obligent à organiser les données et méta-données de manière claire, les analyses reproductibles ont des impacts avantageux en terme de clarté pour les auteurs eux-mêmes, mais aussi en terme de supports pédagogiques facilement ré-utilisables. Rendre les analyses reproductibles via des logiciels libres contribuent aussi à les rendre plus facilement améliorables par d'autres. Ignorer ces deux aspects mène déjà à des situations inefficaces, voire absurdes, qu'il serait grand temps de corriger. Le BAP et l'UMR AGAP ont un rôle à jouer dans la sensibilisation de leurs agents à cette question, ne serait-ce qu'en proposant une formation au logiciel git de gestion décentralisée de versions[4].

De manière plus générale, étant donné la complexité des objets d'intérêt pour l'agroécologie, l'implication de modélisateurs-statisticiens va vraisemblablement augmenter en importance dans le futur. Il semble donc pertinent de rapprocher l'agroécologie de l'autre « domaine interdisciplinaire émergent » de l'INRA, la biologie prédictive. L'importance de la prédiction en sélection est déjà largement reconnue par les généticiens avec historiquement le BLUP et plus récemment la sélection génomique. Mais en s'engageant dans cette voie, il est fréquent (voire inéluctable ?) que l'interprétabilité des modèles se dégrade au bénéfice de leur performance prédictive. Pour s'en

4 http://www.git-scm.com/ et http://swcarpentry.github.io/git-novice/

convaincre, il suffit de prendre connaissance de l'état de l'art dans d'autres domaines d'application (LeCun, Bengio et Hinton, 2015).

Avec l'acquisition de plus en plus massive des données et la sophistication grandissante des modèles, les modélisateurs-statisticiens, particulièrement ceux du secteur public, acquièrent une responsabilité accrue envers les commanditaires de leurs recherches. Ce nouveau contexte de « *big data* » pose déjà de nombreuses questions de gouvernementalité (Lascoumes, 2004). Qu'en sera-t-il pour l'agriculture, l'environnement et l'alimentation ? Une quelconque vigilance a-t-elle sa place dans les missions des modélisateurs-statisticiens travaillant sur les méthodologie de la sélection ?

Références

Barbillon, P., Thomas, M., Goldringer, I., Hospital, F. et Robin, S. (2015). Network impact on persistence in a finite population dynamic diffusion model: Application to an emergent seed exchange network. *Journal of Theoretical Biology*, *365*, 365–376. doi:10.1016/j.jtbi.2014.10.032

Barnard, J., McCulloch, R. et Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*, 1281–1311.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd éd.). Springer. Repéré à http://dx.doi.org/10.1007/978-1-4757-4286-2

Bonneuil, C., Demeulenaere, E., Thomas, F., Joly, P.-B., Allaire, G. et Goldringer, I. (2006). Innover autrement ? La recherche face à l'avènement d'un nouveau régime de production et de régulation des savoirs en génétique végétale. *Dossier de l'environnement de l'INRA*, *30*, 29–52.

Brisson, N., Wery, J. et Boote, K. (2006). Fundamental concepts of crop models illustrated by a comparative approach. Dans D. Wallach, D. Makowski et J. W. Jones (dir.), *Working with dynamic crop models : evaluation, analysis, parameterization, and applications* (p. 257–279). Elsevier.

Demeulenaere, E. (2014). A political ontology of seeds: The transformative frictions of a farmers' movement in Europe. *Focaal*, 45–61. doi:10.3167/fcl.2014.690104

Desclaux, D., Nolot, J. M., Chiffoleau, Y., Gozé, E. et Leclerc, C. (2008). Changes in the concept of genotype × environment interactions to fit agriculture diversification and decentralized participatory plant breeding: pluridisciplinary point of view. *Euphytica*, *163*(3), 533–546. doi:10.1007/s10681-008-9717-2

Döring, T. F., Vieweger, A., Pautasso, M., Vaarst, M., Finckh, M. R. et Wolfe, M. S. (2015). Resilience as a universal criterion of health. *Journal of the Science of Food and Agriculture*, *95*(3), 455–465. doi:10.1002/jsfa.6539

Dreze, J. (2002). On research and action. *Economic and Political Weekly*, *37*(9), 817–819.

Gelman, A. (2011). Causality and statistical learning. *American Journal of Sociology*, *117*(3), 955–966.

Gelman, A. et Imbens, G. (2013). Why ask why? Forward causal inference and reverse causal questions. *National Bureau of Economic Research*, (19614). doi:10.3386/w19614

Green, P. J. (2003). Diversities of gifts, but the same spirit. *Journal of the Royal Statistical Society:*

Series D (The Statistican), *52*(4), 423–438. doi:10.1046/j.1467-9884.2003.02060.x

Hazel, L. N., Dickerson, G. E. et Freeman, A. E. (1994). The selection index: then, now, and for the future. *Journal of Dairy Science*, *77*(10), 3236–3251. doi:10.3168/jds.s0022-0302(94)77265-9

Hennig, C. (2010). Mathematical models and reality: a constructivist perspective. *Foundations of Science*, *15*(1), 29–48. doi:10.1007/s10699-009-9167-x

Hoff, P. D. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Analysis*, *6*(2), 179–196. doi:10.1214/11-BA606

Jia, Y. et Jannink, J.-L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, *192*(4), 1513–1522. doi:10.1534/genetics.112.144246

Kashtan, N., Noor, E. et Alon, U. (2007). Varying environments can speed up evolution. *Proceedings of the National Academy of Sciences*, *104*(34), 13711–13716. doi:10.1073/pnas.0611630104

Kwak, I.-Y., Moore, C. R., Spalding, E. P. et Broman, K. W. (2015). Mapping Quantitative Trait Loci Underlying Function-Valued Traits Using Functional Principal Component Analysis and Multi-trait Mapping. *G3: Genes|Genomes|Genetics*, g3.115.024133+. doi:10.1534/g3.115.024133

Lascoumes, P. (2004). La Gouvernementalité : de la critique de l'État aux technologies du pouvoir. *Le Portique*, *13*.

Latour, B. (1997). Toute recherche est action! *Etudes et Recherches sur les Systèmes Agraires et le Développement*, *30*, 197–208.

Latour, B. (1998). From the world of science to the world of research? *Science*, *280*(5361), 208–209. doi:10.1126/science.280.5361.208

LeCun, Y., Bengio, Y. et Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi:10.1038/nature14539

Lefort, M. et Riba, G. (2006). Quelles perspectives pour l'innovation variétale à l'INRA ? *Dossier de l'environnement de l'INRA*, *30*, 57–64.

Lewontin, R. C. (2006). The analysis of variance and the analysis of causes. *International Journal of Epidemiology*, *35*(3), 520–525. doi:10.1093/ije/dyl062

Lindley, D. V. (2002). Seeing and doing: the concept of causation. *International Statistical Review*, *70*(2), 191–197. doi:10.1111/j.1751-5823.2002.tb00355.x

MacMillan, T. et Benton, T. G. (2014). Agriculture: Engage farmers in research. *Nature*, *509*(7498), 25–27. doi:10.1038/509025a

Ostrom, E. (2009). A general framework for analyzing sustainability of social-ecological systems. *Science*, *325*(5939), 419–422. doi:10.1126/science.1172133

Pearl, J. (2009). Causal inference in statistics: an overview. *Statistics Surveys*, *3*(0), 96–146. doi:10.1214/09-ss057

Renard, M. (2015). Public-private partnerships in pre-breeding within France. Dans *Public-Private Partnerships for Pre-Breeding*. Montpellier, France : CIRAD, INRA, Montpellier SupAgro. Repéré à http://umr-agap.cirad.fr/actualites/international-workshop-on-the-promotion-of-public-private-partnerships-for-pre-breeding-2-4-february-2015

Réseau Semences Paysannes. (2015). *Éléments de réflexion sur la gestion des données dans les*

Maisons des Semences Paysannes : Risques de biopiraterie et droits d'usage collectifs. Repéré à http://www.semencespaysannes.org/bdf/bip/fiche-bip-212.html

Runcie, D. E. et Mukherjee, S. (2013). Dissecting high-dimensional phenotypes with Bayesian sparse factor analysis of genetic covariance matrices. *Genetics*, *194*(3), 753–767. doi:10.1534/genetics.113.151217

Visscher, P. M., Hill, W. G. et Wray, N. R. (2008). Heritability in the genomics era: concepts and misconceptions. *Nature Reviews Genetics*, *9*(4), 255–266. doi:10.1038/nrg2322

Voinov, A. et Bousquet, F. (2010). Modelling with stakeholders. *Environmental Modelling & Software*, *25*(11), 1268–1281. doi:10.1016/j.envsoft.2010.03.007