Fondements et état de l'art

Nathalie Mitton, Ludovic Brossard, Tassadit Bouadi, Frédérick Garcia, Romain Gautron, Nadine Hilgert, Dino Ienco, Christine Largouët, Evelyne Lutton, Véronique Masson, et al.

3

Fondements et état de l'art

Auteurs – *Nathalie Mitton, Ludovic Brossard, Tassadit Bouadi, Frédérick Garcia, Romain Gautron, Nadine Hilgert, Dino Ienco, Christine Largouët, Evelyne Lutton, Véronique Masson, Roger Martin-Clouaire, Marie-Laure Mugnier, Pascal Neveu, Philippe Preux, Hélène Raynal, Catherine Rousset, Alexandre Termier, Véronique Bellon-Maurel.*

Remerciements (contribution, relecture, édition) – *Isabelle Piot-Lepetit.*

Après avoir décrit au chapitre précédent les enjeux de l'agriculture et en particulier ceux de l'agroécologie et des systèmes alimentaires durables, qui constitueront la « cible » de nos questionnements, nous abordons ici les fondements des technologies numériques et de leur usage en agriculture, et les recherches en cours. L'introduction a rappelé les piliers de l'agriculture numérique, qu'on peut résumer par la donnée, les capacités de traitement, la connectivité qui permet l'échange des données et informations et enfin l'automatisation. Ainsi, les défis de l'agriculture intéressent tous les niveaux du cycle de la donnée, de sa captation à son exploitation en passant par sa collecte, sa traçabilité, son traitement, son stockage, son interprétation, sa restitution ou son usage dans des systèmes automatisés ou robotisés.

3.1 La donnée

L'usage de technologies numériques en agriculture génère une masse de données fortement hétérogènes, pouvant aller jusqu'à constituer un *"big data"* (*Bellon-Maurel et al.*, 2018) qui se caractérise par sa complexité car il comprend l'observation d'objets et milieux complexes et de natures variées à des échelles spatiotemporelles très différentes (par exemple du gène au champ) avec des interactions fortes intra et interniveaux, et l'implication de nombreux acteurs. Cette complexité amène à se poser des questions sur les données que l'on doit collecter (nature, fréquence, objectif…), afin de guider le déploiement d'une solution technique à tout niveau (matériel, logiciel, interface, etc.).

Captation de la donnée (quoi, pourquoi, où, comment)

Les défis liés à la captation de la donnée sont à la fois matériels et logiciels. Savoir à quoi cette donnée est destinée aide à déterminer le choix du matériel de mesure.

Il faut d'abord spécifier la nature de la mesure (température, taux d'humidité de l'air, du sol, état des feuilles d'une plante, poids d'un animal, etc.) et la précision recherchée. Ces prescriptions qui émanent de la définition des besoins varient fortement d'une utilisation à une autre. Il faut ensuite s'interroger sur la manière de capter cette donnée. La nature, la taille, le poids, l'encombrement et la résistance du capteur dépendront de la nature de la mesure, de l'objet auquel elle s'applique et du milieu dans lequel il sera placé : un capteur porté par un animal sera choisi en fonction du poids et de l'encombrement du matériel et de la taille de l'animal. De même, un capteur pour des mesures en champ sur le sol ou les végétaux nécessitera une protection pour le rendre résistant à l'environnement (humidité, variations de température, résistance aux chocs…). Enfin, l'utilisation de la donnée

permet de définir l'échantillonnage, en particulier le lieu de collecte, la résolution spatiale et temporelle (*Brun-Laguna et al.*, 2018) : doit-on déployer un capteur par m² ou par km² ? Dans le cas du suivi de la position des animaux, doit-on équiper tous les animaux du troupeau ou uniquement quelques-uns (*Jabbar et al.*, 2017) ?. Quelle est la fréquence temporelle requise et doit-elle être constante ? Certaines applications nécessiteront une régularité et des fréquences spatiales et temporelles élevées, ce à quoi la télédétection satellitaire peut répondre. D'autres se contenteront de mesures plus ponctuelles, par exemple de données participatives (*Minet et al.*, 2017).

Les choix techniques et matériels ainsi que des méthodologies à mettre en place pour le déploiement des capteurs ont donné lieu à de nombreux travaux durant les deux dernières décennies avec des applications en production animale ou végétale : identification et géolocalisation par RFID (*Ruiz-Garcia et Lunadei*, 2011) ou GPS ; imagerie (2D, 3D, infrarouge, hyperspectrale), accéléromètrie, acoustique, mesures biochimiques sur fluides (dont biomarqueurs), automates de mesure tels que bascules de pesée, compteurs à eau ou à lait, distributeurs d'aliments, etc. (*Chastant-Maillard et Saint-Dizier*, 2016 ; *Halachmi et al.*, 2019). Dans la plupart des cas, des compromis sont à faire entre coût, résolution, précision et praticité (*Foubert et Mitton*, 2019). Les recherches visent à limiter ces concessions, soit en réalisant des capteurs toujours plus précis, plus économes en énergie, plus petits, moins intrusifs et à moindre coût, soit en concevant des dispositifs d'acquisition massive de données (à base d'images satellitaires, drones…). Le déploiement de nouvelles constellations de satellites (*Sentinel2*[27]), dont les images à haute résolution spatiale et temporelle sont mises gratuitement à disposition, offre de nouvelles opportunités de suivi.

En conclusion, ces travaux de mise au point de systèmes d'acquisition, par nature pluridisciplinaires, doivent associer agronomes, biologistes, zootechniciens, généticiens, informaticiens, électroniciens et utilisateurs finaux pour satisfaire les attentes des utilisateurs (parfois eux-mêmes chercheurs d'un autre domaine) en combinant la connaissance des objets d'étude, de leurs spécificités et de leurs contraintes, et la connaissance des technologies numériques.

27. https://sentinel.esa.int/web/sentinel/missions/sentinel-2

Par exemple...

Plusieurs unités INRAE mettent au point de tels dispositifs d'acquisition pour le phénotypage ou le suivi des animaux ou des cultures. En voici quelques exemples. L'Unité Mixte de Recheche (UMR) **PEGASE** a développé en collaboration avec l'Institut Technique Agricole **IDELE** et la société **3D Ouest** un portique d'acquisition haut débit d'images 3D et les méthodes de traitement associées pour estimer l'état corporel et la morphologie des vaches laitières. L'Unité Expérimentale (UE) **PEAT** et l'UMR **BOA** ont développé une mangeoire automatique pour étudier les quantités ingérées et le comportement alimentaire individuel des volailles élevées en groupes. Le détecteur électronique de chevauchement « **ALPHA** » (Société Wallace), basé sur un lecteur RFID autonome porté par un bélier, a été conçu par l'UMR **SELMET** pour une détection automatisée des chaleurs chez les moutons notamment en élevage extensif. Dans le domaine végétal, l'UMR **ITAP** et l'Unité Mixte Technologique (UMT) **CAPTE** mettent au point des capteurs optiques, pour le phénotypage ou pour la détection précoce des maladies des plantes. L'UMR **TETIS** utilise la télédétection satellitaire pour détecter des dysfonctionnements dans les parcelles. L'acquisition de données phénotypiques à l'aide de capteurs fait l'objet de programmes et d'infrastructures de recherche de grande ampleur tels que **PHENOME** sur la caractérisation des plantes cultivées, en serre et en champ ou **IN SYLVA** sur la forêt. Les données qui en sont issues permettent d'améliorer les capacités prédictives des modèles et leur prise en compte des interactions entre génotypes et environnement.Plus largement, des dispositifs de phénotypage haut-débit sont aussi développés et testés dans les unités expérimentales animales et végétales d'INRAE.

Collecte et transmission de la donnée (Quelle donnée envoyer, quand, comment ?)

Une fois la donnée acquise, il s'agit de la transmettre. Si certains systèmes utilisent des liaisons filaires (Ethernet, série...), celles-ci sont parfois impossibles et les capteurs doivent alors être équipés d'un moyen de communication sans fil, ce qui soulève de nouveaux défis. La captation et la transmission de la donnée en agriculture font de plus en plus appel aux technologies de l'Internet des objets (*Zhao et al.*, 2010), en particulier la RFID et les réseaux de capteurs sans fil, avec des spécificités liées à l'agriculture.

Suivi des animaux en élevage extensif © Selmet – CIRAD.

La plupart des capteurs sans liaison filaire reposent sur une énergie limitée (par exemple batterie) et/ou variable (par exemple *via* un capteur solaire) qu'il faut donc préserver. La transmission de la donnée est souvent le facteur le plus consommateur d'énergie et l'un des plus gros défis. On cherchera donc à limiter la quantité de données à envoyer tout en maintenant une fréquence d'envoi nécessaire au bon fonctionnement de l'application. Les recherches se concentrent alors sur le traitement de la donnée dans le capteur, lui-même contraint en capacité de calcul et de mémoire : agrégation spatiale et/ou temporelle des données (*Salim et al.*, 2020) ainsi que sur des méthodes d'intelligence artificielle allégées. Par exemple, on exploite la corrélation entre deux grandeurs (comme température et humidité) afin de ne transmettre qu'une des deux valeurs et interpréter la seconde. On peut aussi prédire localement la prochaine valeur mesurée, et ne la transmettre que si elle ne correspond pas à la valeur prédite. Plus une application est exigeante en résolution temporelle ou en précision, plus on devra effectuer d'envois. Il y a donc là aussi un compromis à considérer entre efficacité, précision et coût.

Par exemple...

Les équipes-projets **FUN** et **EVA** d'Inria travaillent sur la collecte de données pour l'agriculture *via* des réseaux de capteurs sans fil. Elles travaillent à la fois sur les protocoles réseau spécifiques et sur les données à remonter afin de ne pas saturer les medias de communication et réduire la consommation énergétique de ces transmissions. En particulier, l'équipe-projet **FUN** déploie des capteurs dans des vignes en Afrique du Sud pour mieux gérer l'arrosage et la gestion de l'eau et collabore avec Sencrop qui déploie des capteurs dans les cultures de céréales et de pommes de terre. **EVA** déploie des capteurs sur des pêchers en Argentine pour prévenir du gel.

Le choix de la technologie de communication dépendra de la quantité de données à remonter, de la distance, mais également du lieu de déploiement des capteurs. La remontée à fréquence faible (par exemple une température une fois par jour) utilisera une technologie longue portée à faible débit et faible consommation alors qu'une mesure exigée à haute fréquence (le suivi des animaux par vidéo) demandera du haut débit. Les points de mesure peuvent se situer dans des zones non couvertes par une technologie cellulaire (de type 3G/4G/5G ou LPWAN – *Low Power Wide Area Network*), ce qui exigera de mettre en place des mécanismes réseaux spécifiques, tels que le routage (action de relayer les données de proche en proche jusqu'à atteindre la station de base), qui doit intégrer les contraintes et exigences des applications et les limitations matérielles et caractéristiques des technologies radio existantes (*Foubert et Mitton*, 2021) et de l'environnement dans lequel les capteurs sont déployés (*Ferreira et al.*, 2020). Une difficulté supplémentaire tient à l'hétérogénéité des technologies qui doivent coexister et parfois coopérer, et aux défis plus généraux de l'Internet des objets (IoT), traités par ailleurs dans le livre blanc Inria sur l'Internet des objets[28].

Enfin, dans les zones blanches, des solutions mobiles émergent pour moissonner la donnée, de solutions très frugales (des « puits de données » mobiles portables dans des sacs à dos comme le projet *COWSHED* en Afrique[29]) à des solutions high tech avec des « puits » volants (drones ou nanosatellites). Ces derniers peuvent collecter des données soit à partir de milliers d'objets connectés à faible débit (protocole LoRa), soit à un débit plus élevé (c'est-à-dire 100 kB par transmission) à partir d'un nombre plus restreint (une centaine) de terminaux

28. Les défis scientifiques de l'Internet des objets. Livre Blanc Inria. À paraitre.
29. https://hal.archives-ouvertes.fr/hal-03102190/document

au sol (protocole UHF). Les applications dans le domaine de l'agriculture se développent, par exemple en Australie où les agriculteurs contrôlent à distance le niveau des réservoirs d'irrigation par nanosatellites[30].

Définition

L'Internet des objets ou *IoT* (*Internet of Things*) est l'interconnexion entre l'Internet et des objets, des lieux et des environnements physiques. L'appellation désigne un nombre croissant d'objets connectés à l'Internet permettant ainsi une communication entre nos biens dits physiques et leurs existences numériques. L'IoT réunit un grand nombre de technologies hétérogènes, allant de la simple étiquette RFID aux applications sur téléphones portables en passant par les réseaux de capteurs sans fil. Les technologies de communication radio sont diverses et présentent des spécificités différentes en termes de débit, consommation, portée, etc. Les capteurs peuvent présenter des microcontrôleurs plus ou moins puissants et consommateurs d'énergie.

Stockage et échange de la donnée, traçabilité

Une fois les données captées et transmises, elles peuvent avoir différents usages. Premièrement, elles peuvent être stockées et traitées pour en extraire de la connaissance, anticiper des dysfonctionnements, etc. Ces données peuvent être très hétérogènes, de qualité variable, présenter des échantillonnages très différents car elles sont issues de diverses sources (capteurs physiques, capteurs « humains », voire résultats de simulations) et en grande quantité (grand nombre de points de captation, fréquence temporelle potentiellement élevée). Les méthodes issues de la gestion des données multivariées et aujourd'hui du *big data* permettent de répondre aux enjeux de volume, de vitesse de traitement, et de diversité des formats et des sources (*Bellon-Maurel et al.*, 2018). Le préalable pour leur bonne valorisation est que les données répondent aux principes directeurs des « données FAIR » (*Findable, Accessible, Interoperable et Reusable*)[31], qui garantissent la capacité des systèmes informatiques à trouver, interopérer et réutiliser des données avec une intervention humaine minimale. Ainsi, une nouvelle génération de systèmes d'information adaptée à l'agriculture est attendue pour

30. https://which-50.com/world-first-australian-iot-uses-satellites-to-monitor-farmers-tanks-rain-levels/
31. Wilkinson et al. *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific Data 3, 160018. doi:10.1038/sdata.2016.18.

gérer et structurer ces masses de données complexes en implémentant les principes FAIR. Métadonnées et données doivent être bien décrites, en s'appuyant sur des ressources sémantiques (ontologie, taxonomie, thésaurus), pour être comprises et pour faciliter leur accès *via* des protocoles standardisés.

Par exemple...

À INRAE, des centres automatisés de traitement de l'information **(CATI)** fédèrent et structurent les compétences, méthodologies et technologies pour faciliter la réutilisation des données, comme par exemple le centre automatisé de traitement de l'information **CATI SICPA** (Systèmes d'informations et calcul pour le phénotypage animal), le **CATI Codex** (phénotypage végétal) ou le **CATI GEDEOP** (GEstion des Données d'Expérimentations, d'Observations et de Pratiques sur les agro-socioécosystèmes).

Enfin se pose la question de la validation de la donnée, un sujet qui devient central avec l'accroissement des quantités collectées : la valeur mesurée est-elle correcte ? pour quelles applications et dans quelles conditions ?

Il existe une variété de SGBD (Systèmes de Gestion de Bases de Données), selon le modèle de données utilisé, le modèle relationnel étant aujourd'hui le plus répandu. Par exemple, *Benchini et Stöckle* (2007) ont couplé un modèle dynamique de simulation des systèmes de culture avec une base de données relationnelle afin d'obtenir un stockage et une analyse efficaces de données du modèle, à l'échelle de l'exploitation agricole. Mais lorsqu'il s'agit de gérer de très grands volumes (de l'ordre du pétaoctet), ou des données complexes et hétérogènes (graphes, documents, etc.) dans des contextes fortement distribués (serveurs distants, Internet des objets, etc.), les bases de données de type *NoSQL* exploitant un autre modèle de données – qui relâche certaines contraintes conceptuelles du modèle relationnel – sont plus pertinentes.

Dans d'autres cas, l'utilité principale de la donnée est d'être partagée par des acteurs multiples, par exemple pour limiter des fraudes ou attester certains processus (respect d'un parcours, respect de la chaîne de froid, d'une production locale ou sans insecticides). Un des outils numériques les plus prometteurs actuellement pour assurer cela est la *blockchain* (*Bermeo-Almeida et al.*, 2018). La *blockchain* implémente une base de données distribuée ne nécessitant aucune entité de contrôle, datant et garantissant l'intégrité des éléments qui la composent. En agriculture, elle permet d'enregistrer les étapes de vie d'un produit

et assure sa traçabilité (*Kamilaris et al.*, 2019). Ses avantages sont multiples (cf chapitre 4 – opportunités) : transparence, suivi des transactions entre agriculteurs, fournisseurs, acheteurs, consommateurs, etc. Dans certaines filières, la mise en place d'une *blockchain* permet de se passer de certification longue et coûteuse (*Lin et al.*, 2017).

Cependant, l'utilisation de la *blockchain* soulève de nouveaux défis numériques et organisationnels. Les risques de piratage informatique inhérents à tout système d'information existent. De plus, les *blockchains*, initialement conçues pour assurer des flux et partages de biens intangibles (monnaie, certificats, diplômes), ne sont pas infaillibles quand le flux informationnel doit être couplé à un flux physique, comme en agriculture et en agro-alimentaire : il faut s'assurer que les données numériques représentent à l'identique le flux physique. Par ailleurs, la confiance instaurée dans une *blockchain* repose sur le mécanisme de « preuve de travail » (des opérations informatiques) qui valide les nouveaux blocs à intégrer. Les *blockchains* publiques s'appuient sur de grandes quantités de « preuves de travail », d'où une très forte consommation énergétique. C'est pourquoi les recherches actuelles se penchent sur la réduction de la complexité de ces algorithmes cryptographiques pour en réduire la consommation.

3.2 Modélisation, simulation et optimisation

Si la donnée constitue un levier majeur de l'agriculture numérique, la modélisation se révèle également indispensable pour relier les mesures et observations aux interprétations et préconisations qui viennent aider les acteurs de la filière agricole à mieux comprendre, piloter et améliorer leurs systèmes de production.

Dans le domaine de l'agriculture et de la recherche agronomique, la modélisation est une démarche scientifique qui a émergé très tôt, avec comme principal objectif la prédiction des récoltes : chez les Égyptiens sous le règne de Sésostris I^{er}, avec l'utilisation des hauteurs de crues pour prédire la richesse des récoltes à venir (*Gros de Beler*, 1998), ou encore chez les paysans incas qui connaissaient plusieurs mois à l'avance les cycles agricoles en observant la nature (*Gutiérrez*, 2008). Bien plus tard, les travaux pionniers de Mendel (*Mendel*, 1907) puis Fisher (*Street*, 1990) ont définitivement légitimé l'utilisation de modèles statistiques dans les domaines de la génétique et de l'agronomie. Dans la seconde partie du XX^{e} siècle, la modélisation en agriculture s'est particulièrement développée dans l'économie rurale, pour la rationalisation et l'optimisation de la production, de l'agronomie et de la zootechnie, de la conduite des cultures et de l'alimentation animale, et enfin de la sélection génétique végétale et animale. Avec le développement de l'informatique et des premiers calculateurs, les modèles ont alors progressivement dépassé le cadre des

statistiques ou de la recherche opérationnelle, pour reposer de plus en plus sur des formalismes symboliques et algorithmiques pour une modélisation exprimée en termes mathématiques et informatiques où la simulation joue un rôle clé.

La fonction générale d'un modèle est souvent définie comme une fonction de médiation : « *Pour un observateur B, un objet A* est un modèle d'un objet A, dans la mesure où B peut utiliser A* pour répondre à des questions qui l'intéressent au sujet de A* » (*Minsky*, 1965). Cette médiation peut contribuer à différentes quêtes cognitives : faciliter l'expérience, la formulation intelligible, la théorisation, la communication et la coconstruction des savoirs, la décision et l'action (*Varenne et Silberstein*, 2013). Aujourd'hui, la modélisation appliquée à l'agriculture concerne un très large spectre d'objets, et vise principalement quatre objectifs : analyser, communiquer, prédire-contrôler l'évolution de diverses composantes d'un système agricole, concevoir-optimiser le système considéré. Dans la suite de cette section, nous présentons quelques grandes classes de modèles, et l'usage qui peut en être fait pour l'agriculture numérique grâce à la simulation et l'optimisation.

Modéliser quoi, pour quels objectifs, avec quels outils

Modéliser quoi ? En agriculture, les objets d'étude, sujets de la modélisation, sont des systèmes naturels anthropisés qui couvrent éventuellement plusieurs échelles et niveaux d'organisation. La modélisation porte sur les composants de ces systèmes, les processus qui régissent leur dynamiques, les événements qui activent ou inhibent ces processus, et les facteurs exogènes qui les influencent (par exemple les conditions météorologiques) (*Martin et al.*, 2011). Les composants sont pour partie de nature biophysique (par exemple les cultures avec les processus de croissance, les maladies) (*Kumar et Sinhg*, 2003) et une autre partie est centrée sur les rôles joués par les acteurs humains. Dans ce dernier cas, la modélisation peut concerner aussi bien un individu (*Martin-Clouaire et Rellier*, 2004 ; 2009) qu'un collectif d'individus (par exemple les membres d'une coopérative) et des processus sociaux de coordination d'activités menées par différents individus du collectif (*Drewniak et al.*, 2013 ; *Manson et al.*, 2016).

Par exemple...

De très nombreuses équipes et unités développent des modèles de processus, de flux, d'interactions. Voici quelques illustrations à différentes échelles
Plusieurs unités travaillent à INRAE sur la modélisation de cultures ou en élevage, à l'échelle de la parcelle (modèles de culture multiespèces tels que **STICS**, décrivant la croissance en fonction de variables climatiques et environnementales), de l'individu (modèles de croissance chez l'animal en fonction de l'alimentation et de l'environnement), ou à des échelles plus larges (modèles d'épidémiologie des plantes incluant la dispersion entre parcelles, ou d'épidémiologie animale décrivant les transmissions intertroupeaux, etc.). Les plates-formes de modélisation telles que **RECORD**, **OPEN ALEA** ou **OPEN FLUID** hébergent ces modèles à partir desquels des simulations peuvent être lancées sous différents scénarios climatiques et contextuels.

L'équipe-projet **STEEP** (Inria, CNRS, Université de Grenoble Alpes) développe des modèles numériques pour analyser les flux de matières (production, transformation, échanges, consommation, déchets) dans les filières agricoles et forêt-bois et ainsi 1) appréhender les vulnérabilités amont/aval des filières, 2) questionner l'usage des ressources naturelles et les éventuels problèmes de concurrence d'usage et enfin 3) estimer des empreintes environnementales. Les outils développés reposent sur une modélisation des filières en termes de produits, secteurs et flux pouvant exister entre ces produits et secteurs. Une des difficultés majeures ici est liée au caractère particulièrement lacunaire et incohérent des données.

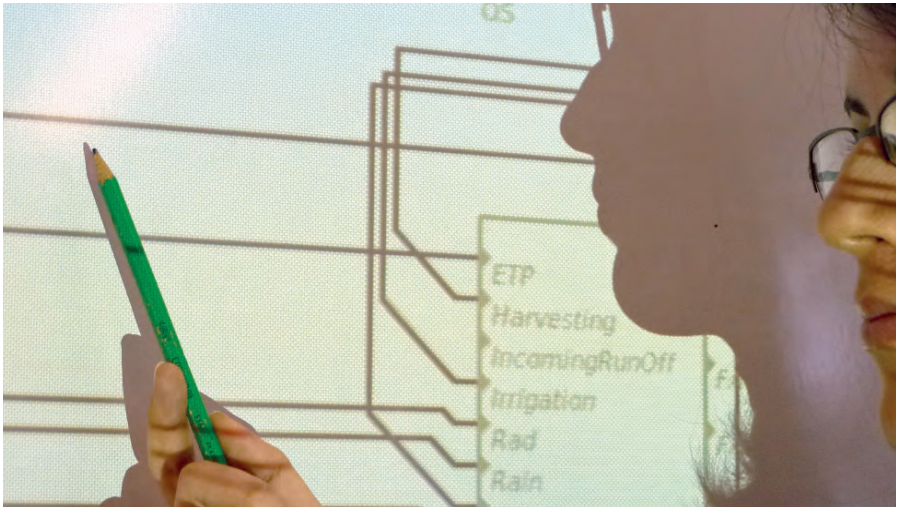
L'équipe-projet **EASE** (Inria, Ecole Nationale Supérieure Mines-Télécom Atlantique Bretagne Pays de la Loire, Université Rennes 1) développe un ensemble complet de nouveaux modèles d'interaction, des outils pour augmenter et qualifier les informations issues des systèmes complexes. Ces travaux ont été appliqués à la gestion de l'énergie dans le monde agricole. En particulier, leur modèle permet de définir comment rendre sa consommation plus verte, qu'il s'agisse d'un ancien site à optimiser ou d'un nouveau à équiper. Ils démontrent que l'optimisation d'un seul paramètre (production locale, stockage ou transfert de processus) ne peut à elle seule maximiser l'autoconsommation et minimiser les coûts d'énergie.

Modéliser pour quels objectifs ? Les systèmes au cœur d'une étude sont modélisés à des niveaux de détail déterminés par l'objectif visé et par l'outil envisagé. Les objectifs les plus communément envisagés vont de l'identification des objectifs,

et moyens mis en œuvre dans la gestion des agroécosystèmes à la prédiction des performances (*Rio et al.*, 2019) étant donné différents scénarios, en passant par l'identification des risques et l'examen critique sur le fonctionnement et la conduite des systèmes de production agricole (*Li et al.*, 2019). La modélisation peut également permettre la conception de nouveaux systèmes incluant la configuration et le dimensionnement d'une chaîne logistique (*Taghikhah et al.*, 2021).

Par exemple...

De nombreux modèles concernent le pilotage des agroécosystèmes. Ainsi, l'équipe-projet commune **Inria**/**INRAE BIOCORE** (CNRS, Sorbonne Université-UPMC) s'intéresse à la modélisation et au contrôle en épidémiologie pour l'agriculture tropicale. À INRAE, l'UR **MIAT** et l'UMR **MISTEA** développent des modèles de simulation et des méthodes d'optimisation dédiées explicitement à la gestion des agroécosystèmes à l'échelle de l'exploitation. Dans le domaine animal, des unités comme les UMR **BIOEPAR, SELMET, MoSAR, UMRH** ou **PEGASE** développent des modèles autour de la santé et de l'épidémiologie animale, des phénomènes dynamiques ingestifs, digestifs et métaboliques, et des systèmes d'élevage. Par exemple, de nouveaux modèles pour ajuster l'apport alimentaire quotidien aux besoins nutritionnels de chaque animal ont été développés à l'UMR **PEGASE** pour l'alimentation sur mesure des truies en gestation et en lactation (*Gauthier et al.*, 2019).



Construire un modèle agronomique. © INRAE.

Modéliser comment ? La modélisation informatique pour soutenir l'analyse, la conception ou la gestion d'agroécosystèmes est associée à des approches privilégiant soit la simulation soit l'optimisation (*Li et al.*, 2020). La simulation dynamique met l'accent sur le réalisme de la modélisation des phénomènes jugés importants pour les objectifs de l'étude (*Kaghazchi et al.*, 2021). L'optimisation consiste à faire une exploration algorithmique de l'espace des alternatives qui recherche efficacement une solution optimale selon un ou plusieurs critères explicitement formulés, en utilisant des modèles mathématiques réductionnistes (*Ezanno et al.*, 2020 ; *Casagli et al.*, 2020). Les deux approches ont des objectifs relativement antagonistes (réalisme de la modélisation *versus* efficacité calculatoire) et, de ce fait, font généralement appel à des modélisations différentes.

Cadres de représentation

Les agroécosystèmes sont des objets complexes dont la modélisation porte d'une part sur le fonctionnement des entités biophysiques qui les composent (le sol, les plantes, les animaux, les flux de minéraux, d'eau, etc.) et d'autre part sur les activités humaines de décision et d'action sur ces entités biophysiques (*Zabala et al.*, 2021). Les modèles véhiculent des connaissances issues principalement de disciplines scientifiques telles que l'agronomie, la zootechnie, les sciences de l'environnement, les sciences de gestion, et les sciences humaines et sociales.

Les modèles biophysiques peuvent être classés en trois grands domaines : mécaniste, empirique et hybride (*Reyniers*, 1996). La modélisation mécaniste se concentre sur les événements, les relations de causalité et les processus. Par contraste, les modèles empiriques traitent un système comme une « boîte noire » et ne décrivent généralement pas les phénomènes biophysiques sous-jacents. Ces modèles représentent la dynamique entrée-sortie d'un composant du système en termes de données d'observation. En fait, il existe peu de véritables modèles mécanistes ou empiriques. Les modèles sont plutôt hybrides ou classés dans l'une ou l'autre catégorie en fonction de la prépondérance des composantes mécanistes ou empiriques. La compréhension globale et les informations nécessaires pour construire ces modèles augmentent en allant des modèles empiriques aux modèles mécanistes. En rendant la causalité explicite, les modèles mécanistes peuvent être plus complexes, tandis que les modèles empiriques sont généralement plus simples, leur champ d'application étant toutefois limité par la disponibilité de données statistiques.

La modélisation des aspects décisionnels varie suivant l'hypothèse que fait le modélisateur sur l'agent décideur. Dans une première hypothèse, le décideur est supposé parfaitement rationnel (au sens des économistes) et détermine en situation de décision le choix qui est mathématiquement optimal conformément

à des fonctions d'utilité explicitées *a priori*. Selon une seconde hypothèse dite de rationalité limitée, l'agent élabore la décision qui conduit à un résultat qu'il juge satisfaisant compte tenu de l'information disponible et de son niveau d'aspiration. Il est souvent fait appel à des modèles mentaux plus ou moins sophistiqués, dont les modèles à base de règles de décision qui associent des situations à des décisions ou actions (*Martin-Clouaire*, 2017). Pour faciliter et normaliser le développement de ces modèles on peut s'appuyer sur des ontologies qui définissent les concepts, les relations et autres distinctions qui sont pertinents pour les domaines concernés (grande culture, élevage…) (*Roussey et al.*, 2011). Une ontologie (cf section 3.4) constitue un modèle abstrait (métamodèle) du domaine et fournit les primitives de représentation permettant d'instancier des modèles de systèmes spécifiques sous la forme de bases de connaissances (*Martin-Clouaire et Rellier*, 2004 ; *Fishwick*, 2007).

Définition

Une ontologie fixe un vocabulaire et les liens sémantiques entre les éléments du vocabulaire. Le vocabulaire est composé de noms de concepts (ou « classes »), qui sont les types d'entités connus par le système, et de noms de relations (ou « rôles ») possibles entre ces entités (par exemple, la relation « est bioagresseur de » lie deux entités de type « organisme vivant »). L'ontologie est décrite dans un langage informatique et logique plus ou moins expressif de représentation de connaissances. Elle peut ainsi aller d'une simple taxonomie (un ensemble de concepts structuré par spécialisation progressive) à des descriptions complexes des éléments du vocabulaire et de leurs liens sémantiques. Le langage utilisé permet la mise en œuvre de raisonnements automatiques.

Systèmes à événements discrets, temps discret et temps continu – Un modèle de simulation à événements discrets permet la représentation d'un système dynamique à l'aide de variables dont l'évolution dépend entièrement de l'occurrence d'événements asynchrones au cours du temps. Un cas particulier est celui où la progression temporelle se fait à incrément fixe. L'approche par événements discrets contraste (sans s'opposer toutefois) avec l'approche de type *System Dynamics* dans laquelle l'état du système est modifié en continu dans le temps sur la base d'un ensemble d'équations différentielles définissant les taux de changement des variables d'état. Dans ces différents cas de modèle de simulation, on s'intéresse à la représentation des relations causales, c'est-à-dire, pour la partie biophysique, à des modèles mécanistes. Parmi les formalismes les plus connus citons *Discrete Event System Specification – DEVS* (*Zeigler et al.*, 2000) qui s'appuie sur un cadre

générique permettant diverses particularisations à des formalismes spécifiques tels les réseaux de Petri, les automates cellulaires et plus généralement les modèles à pas de temps fixe. Le formalisme mathématique des réseaux de Petri est plus particulièrement apprécié pour sa capacité à représenter la synchronisation de processus exécutés en parallèle et pour les possibilités d'analyse rigoureuse des modèles. Un automate cellulaire est constitué d'un réseau de cellules discrètes et est tout-à-fait approprié pour représenter une dynamique spatiale (par exemple la propagation d'une infestation) et des phénomènes d'auto-organisation (par exemple les dynamiques paysagères de reboisement naturel). Certains formalismes tels que les *statecharts* (*Léger et Naud*, 2009), les automates temporisés (*Hélias et al.*, 2008) et les réseaux de Petri (*Guan et al.*, 2008) peuvent aussi se prêter à des procédures de vérification du comportement (par exemple pour s'assurer que le modèle ne peut pas se bloquer) ou des propriétés temporelles du modèle.

Détection automatique des mangues par intelligence articielle à partir d'une photo de smartphone.
© Hortsys – CIRAD.

Modèles à base d'individus ou d'agents – Ces modèles s'intéressent à des systèmes qui peuvent se décomposer en un ensemble d'entités (par exemple plantes, animaux, zones) qui agissent ou interagissent (*Daudé*, 2004 ; *DeAngelis et Diaz*, 2019). Couplé à une approche de type automate cellulaire, un modèle à base d'individus permet la représentation spatiale et la simulation de processus biophysiques sur un territoire parcellisé. Lorsque les entités modélisées sont dotées de capacités cognitives et décisionnelles plus élaborées (*Bahri et al.*, 2020), on parle de modèles d'agents qui permettent par exemple de simuler le comportement décisionnel d'un ensemble d'agents (agriculteurs par exemple) opérant sur un territoire (*Huber et al.*, 2018). La gestion des exploitations agricoles a souvent été modélisée en utilisant des mécanismes simples de déclenchement de règles de décision associées aux situations possiblement rencontrées. Il s'avère toutefois difficile avec cette approche de contrôler l'ordre dans lequel les règles sont utilisées et de maintenir la base de règles dès qu'elle atteint une certaine taille. Un enrichissement a été apporté par l'approche BDI (*Belief, Desire, Intention*) (*Georgeff et al.*, 1970 ; *Bratman*, 1987) qui permet de modéliser le processus par lequel un agent prend ses décisions sur la base d'une perception de la situation courante (*Belief*), d'objectifs déclarés (*Desire*) et d'engagements sur la façon de procéder vers les objectifs (*Intention*).

Par exemple...

Dans l'esprit de l'approche « BDI » et de la planification temporelle et dans l'incertain en intelligence artificielle, l'UR **INRAE-MIAT** a développé plusieurs formalismes pour représenter et simuler le comportement décisionnel des agriculteurs dans leurs activités de gestion de l'exploitation agricole. Par exemple, Martin-Clouaire et Rellier abordent le problème de gestion de production comme celui de la coordination d'un ensemble d'activités organisées dans des plans flexibles dont on peut simuler la mise en œuvre dans un contexte particulier (*Martin-Clouaire et Rellier*, 2009). Pour des exemples d'application voir *Martin et al.* (2011) en élevage laitier et *Martin-Clouaire et al.* (2016) en viticulture.

Modèles à base de contraintes – Les modèles à base de contraintes regroupent un ensemble de formalismes variés qui reposent principalement sur la notion de graphe modélisant des relations binaires entre variables (*Hurley et al.*, 2016). Ces relations peuvent modéliser des corrélations, des influences causales, déterministes ou probabilistes, comme c'est le cas avec les réseaux bayésiens ou les chaînes de Markov. Ces réseaux peuvent également décrire des contraintes entre variables,

portant sur les combinaisons de valeurs acceptables ou non, conduisant à une problématique de satisfaction de contrainte (CSP pour *constraint satisfaction problem*) (*Moummadi et al.*, 2011). Dans un esprit proche, les méthodes de programmation linéaire reposent sur l'optimisation d'une combinaison linéaire de plusieurs variables reliées entre elles par des relations linéaires appelées contraintes (*Maqrot et al.*, 2017).

Par exemple...

À INRAE, l'UMR **BAGAP** travaille sur la modélisation du problème d'allocation dynamique des cultures au sein d'une exploitation agricole, basé sur l'emploi de contraintes spatiales et temporelles et l'utilisation du solveur *toolbar2* (*Akplogan et al.*, 2013). Par exemple, l'équipe s'est penchée sur l'analyse du paysage bocager du pays charolais brionnais pour démontrer l'unicité de ce paysage, sa capacité à incarner les différentes structures et fonctions des haies. Cette analyse a permis l'inscription de ce bocage sur les listes de sites qui pourraient être soumises au patrimoine de l'UNESCO.

Modélisation et simulation

Le principal atout des approches de modélisation est sans doute la possibilité de modéliser et simuler des comportements complexes de systèmes agricoles et, plus largement, de systèmes socioécologiques incluant des agroécosystèmes (*Peart et Curry*, 1998). Les modèles, en particulier ceux à base d'agents, sont souvent complexes dans le nombre et l'hétérogénéité des composants et des interactions, ainsi que dans la réactivité aux aléas qui pèsent sur ces systèmes. Leur comportement est difficile à étudier parce que les phénomènes qui y interviennent sont non linéaires, avec de multiples discontinuités et rétroactions entre les niveaux d'organisation et les échelles. Certains de ces modèles représentent des agents cognitifs exhibant des comportements de rationalité limitée. De nombreuses applications relatives à l'agriculture se sont appuyées sur les plates-formes CORMAS (*Bommel et al.*, 2015) et GAMA (*Taillandier et al.*, 2010) comme par exemple sur l'étude de la gestion de l'eau, sur la réforme de la politique agricole commune ou encore sur la réduction de la consommation de pesticides et de développement de l'agriculture biologique.

Par exemple...

À INRAE, l'UMR **AGROECOLOGIE** coordonne le développement de la plate-forme MAELIA de modélisation et évaluation intégrées des systèmes socioagroécologiques et vise à produire des connaissances sur la structure, le fonctionnement et les performances de ces formes d'agriculture de l'échelle de la parcelle à celle du paysage et/ou du territoire.

En pratique, les approches de modélisation-simulation donnent lieu à des utilisations variées allant de l'analyse en laboratoire faite par des scientifiques, de l'aide à la décision (*Huber et al.*, 2018) ou de la prise de décision en temps réel par des agriculteurs ou conseillers agricoles, ou encore du support à des négociations entre acteurs (par exemple des modèles d'accompagnement pour gérer l'eau en commun sur un territoire), à la coconception par un collectif d'agriculteurs de nouveaux systèmes de production ou à des formations. L'agriculteur ou un groupe d'agriculteurs peuvent progresser dans la compréhension du fonctionnement biophysique et glaner des idées d'amélioration du système étudié sur : la qualité des produits, la vulnérabilité du système, les conséquences environnementales des pratiques mises en œuvre, la réduction de surcharge et de pénibilité du travail et, enfin, les performances économiques liées à l'application des principes de l'agroécologie.

Par exemple...

À INRAE, des compétences sont regroupées au sein de **CATI** pour modéliser des systèmes de grande échelle, par exemple les **CATI IMOTEP** (Information, Modèles et Traitement des données en Epidémiologie et dynamique de Populations) et **IUMAN** (Informatisation et Utilisation des Modèles pour les Agroécosystèmes Numériques). Les travaux portent aussi bien sur la modélisation de la propagation d'épidémies chez les végétaux ou les animaux, que sur le développement logiciel de plates-formes et preuves de concepts permettant le partage et l'informatisation de ces nouveaux modèles à de multiples échelles.

Modélisation et optimisation

Par définition, l'optimisation explore selon différentes méthodes les solutions possibles à un problème donné pour trouver un ou des optimums selon un critère ou un ensemble de critères (*Zelinka et al.*, 2013). Elle est utilisée en agriculture dans différents domaines et à différentes échelles (*Plà-Aragonés*, 2015). Au niveau de l'exploitation agricole, l'optimisation est présente, de façon explicite ou sous-jacente pour l'exploitant, que ce soit pour la formulation d'aliments, la gestion du troupeau, la planification des abattages des animaux, la planification des cultures ou de l'utilisation des terres, ou la gestion de l'eau. L'optimisation est également utilisée à des échelles de groupes de fermes, de territoires, de régions, de pays pour la gestion de l'utilisation des terres, de l'eau, des questions économiques d'échanges et de marché (*Carpentier et al.*, 2015). Dans ce cas, les modèles bioéconomiques sont utilisés dans une démarche d'analyse : l'objectif premier devient l'évaluation de l'impact de l'application des contraintes et des critères sur les solutions optimales.

Du fait de la complexité des systèmes agricoles et de l'évolution des questions liées à l'agriculture, l'optimisation a également évolué en agriculture (*Jones et al.*, 2016). Les premiers modèles économiques dans les années cinquante s'intéressaient surtout à maximiser le revenu. La formulation d'aliment à moindre coût vise encore aujourd'hui principalement à obtenir un aliment le moins cher possible en respectant des critères nutritionnels. Progressivement, l'optimisation est devenue multiobjectif pour combiner différents objectifs productifs (par exemple production animale ou végétale, temps de travail), économiques (par exemple revenus, coût) mais aussi environnementaux (bilans de nutriments, impact environnemental calculé par analyse de cycle de vie, services écosystémiques...). Pour les optimisations « sous contraintes », les contraintes sont également variées : biologiques, structurelles ou réglementaires, mais aussi environnementales et décisionnelles.

L'optimisation pour les modèles en agriculture s'est également appuyée sur l'évolution des méthodes d'optimisation, utilisant un panel varié de méthodes. Les méthodes déterministes de programmation linéaire sont toujours très présentes, avec des adaptations permettant la résolution de problèmes multiobjectifs. Les méthodes stochastiques de métaheuristique sont appliquées seules ou en combinaison avec les précédentes. Ces méthodes de métaheuristique permettent notamment d'aborder l'optimisation multicritère et d'obtenir un ensemble de solutions optimales dites admissibles dans ce cadre (appelé Front de Pareto) ; elles comprennent par exemple les algorithmes évolutionnaires (comme les algorithmes génétiques, ou l'évolution différentielle) qui travaillent sur une population de solutions, l'optimisation par essaim particulaire, la recherche tabou, le recuit simulé, etc. (*Kaim et al.*, 2018 ; *Memmah et al.*, 2015).

Les questionnements actuels autour de l'optimisation portent notamment sur l'adaptation des méthodes à des modèles de plus en plus complexes, avec en particulier la prise en compte de l'incertitude (*Crespo et al.*, 2010) et du temporel dans la formulation du problème d'optimisation (*Akplogan et al.*, 2013), rejoignant ainsi des questions historiquement traitées dans la communauté de l'automatique et du contrôle optimal. Le couplage entre optimisation et simulation (*Borodin*, 2014) constitue également un front de recherche, en lien en particulier avec les méthodes d'apprentissage par renforcement (*Gosavi*, 2015). Malgré l'évolution technologique autour de la puissance de calcul, le temps de traitement des processus d'optimisation reste toujours une dimension à considérer du fait de la complexité croissante des modèles considérés, et les développements récents autour de la métamodélisation constituent une possible stratégie de simplification pour réduire ces temps de traitement.

Par exemple...

À INRAE, les unités UMR **PEGASE** et **SMART-LERECO** développent des approches d'optimisation multicritère (performance zootechnique et économique, impact environnemental) des stratégies alimentaires en élevage porcin, sur la base d'un modèle de l'exploitation porcine.

Chez Inria, plus d'une vingtaine d'équipes-projets travaillent sur la mise au point d'algorithmes d'optimisation, de recherche opérationnelle ou de contrôle.

3.3 Apprentissage et extraction de connaissances multiéchelle

Les deux sous-chapitres précédents ont présenté les approches utilisées pour collecter les données, puis les techniques de modélisation reposant principalement sur l'analyse humaine. Dans ce sous-chapitre, nous nous intéressons aux grandes familles d'approches permettant de construire directement des modèles depuis les données, et ainsi d'en extraire automatiquement des connaissances. Les connaissances découvertes peuvent soit être présentées à des experts humains, soit rester internes à une approche d'apprentissage et permettre de réaliser des tâches de prédiction ou d'identification par exemple.

Nous montrerons d'abord que les données « brutes » envoyées par les capteurs ne peuvent en général pas être utilisées telles quelles et que leur prétraitement représente un défi en soi. Commençons par présenter les types de données fréquemment utilisées en agriculture numérique, et qui pourront constituer un *Big Data*.

Les données massives en agriculture

En agriculture, les données les plus « massives » sont issues de capteurs à grande résolution temporelle ou spatiale, comme les séries temporelles et les données de télédétection ou de cartographie issues de capteurs embarqués.

Séries temporelles. Une série temporelle est une suite de valeurs numériques représentant l'évolution d'une variable mesurée sur un individu au cours du temps. De telles suites de variables peuvent être modélisées individuellement pour en comprendre l'évolution passée et prévoir le comportement futur, à l'aide de modèles de type ARMA (*Box et al.*, 2015). Aujourd'hui les expérimentations en agronomie permettent d'observer la même variable sur des milliers d'individus (par exemple surface foliaire de milliers de plantes en serre, température d'animaux d'élevage) et sur de longues périodes. L'objectif d'analyse de ces séries temporelles a donc évolué vers la recherche de caractéristiques communes entre ces séries, de différences majeures, ou d'acquisition de connaissance plus fine quant aux mécanismes internes (par exemple effet des génotypes) ou externes (par exemple liés aux variables environnementales) qui influent sur les variables observées. Les séries temporelles sont ainsi étudiées plus généralement comme des fonctions du temps. Elles portent aussi le nom de « données fonctionnelles » ou « longitudinales ».

Données de télédétection. Les données de télédétection sont des images d'une zone donnée, prises par satellite ou par des drones. Les images satellitaires – sur lesquelles nous nous focaliserons par la suite – peuvent être enregistrées à différentes périodes, ces séquences constituant des séries temporelles. Elles peuvent aussi, pour une même période, être issues de satellites différents, chacune ayant un contenu radiométrique différent (i.e. information radar, information optique). La dynamique de la végétation peut aujourd'hui être suivie avec une résolution spatiale compatible avec la taille des objets d'intérêt et une revisite temporelle fréquente grâce aux récentes missions spatiales comme, par exemple, *Copernicus*. La mission satellitaire *Sentinel-1* permet d'acquérir des informations radar (deux bandes) tous les cinq ou six jours sur une même zone à dix mètres de résolution spatiale. Cette source de données permet d'accéder à des informations sur la structure des objets (*i.e.* biomasse forestière ou agricole) ainsi que de suivre et estimer les surfaces humides et la part de terrains qui ont été irrigués sur une certaine période. Une autre mission satellitaire avec autant d'intérêt

est la mission *Sentinel-2* qui délivre elle aussi des informations tous les cinq ou six jours, toujours à une résolution spatiale de dix mètres, dans le domaine de l'optique multispectrale (treize bandes). Ce capteur optique est particulièrement adapté pour cartographier l'occupation et l'usage des sols, suivre la biodiversité des états naturels ainsi que pour l'estimation de rendements à large échelle sur des grandes surfaces (*Lambert*, 2018).

À l'opposé de cette échelle, c'est-à-dire vers le microscopique, les données métaomiques issues du *metabarcoding*[32] permettent de mieux caractériser l'environnement biologique des cultures ou des animaux. Ces données métaomiques sont l'assemblage des « empreintes » des génomes présents ou de leurs expressions (ARN, protéines), ce qui permet d'analyser de nouvelles dimensions des écosystèmes, pouvant mieux expliquer le comportement des cultures ou des animaux. On est seulement au début de l'exploration de ces nouvelles sources de données, dont certaines restent difficiles à obtenir (protéomique, métabolomique…).

Prétraitements des données

Les challenges importants en prétraitement des données sont : i) identifier des données aberrantes ou peu fiables : les données collectées pendant les expérimentations ou sur le terrain sont nombreuses, très bruitées et sujettes à de multiples sources d'erreurs, comme un capteur défectueux. Des outils spécifiques sont donc nécessaires pour annoter ces données, identifier rapidement les capteurs défectueux, diagnostiquer l'hétérogénéité dans le champ ou la serre, afin d'améliorer la qualité des jeux de données pour les analyses futures ; ii) lier données et connaissance de l'expert dans les analyses comme par exemple mimer automatiquement le raisonnement d'un expert quand il valide un « petit » jeu de données, ou utiliser la connaissance de l'expert pour recaler des courbes (alignement des dates de stades phénologiques).

Un défi particulier est la fusion de données. Des informations difficiles à obtenir directement peuvent être retrouvées en combinant des données, soit de même nature (par exemple, la surface foliaire d'une plante (une donnée scalaire) est prédite à partir de l'analyse de quinze images de cette plante prise sous différents angles), soit de natures différentes. Pour cela, de plus en plus de données, de

32. Le *metabarcoding* est une méthode d'indentification des espèces à partir de segments d'ADN ou d'ARN. En ne ciblant pas d'espèce spécifique mais déterminant la composition d'espèces dans un échantillon, le *metabarcoding* permet l'identification de nombreux taxons dans un assemblage de populations (de bactéries ou autres micro-organismes) au sein un échantillon environnemental (par exemple échantillon de sol, sédiments, excréments…). C'est ainsi une des méthodes les plus rapides d'évaluation environnementale de la biodiversité de systèmes écologiques riches en espèces inconnues ou difficiles à déterminer.

natures variées et hétérogènes, sont collectées pour suivre un même phénomène ou une même zone d'étude (données dites « multisources »). Les connaissances contenues dans ces données représentent une vraie opportunité pour mieux comprendre les phénomènes complexes associés aux pratiques agricoles modernes afin de pouvoir, par la suite, mieux les suivre et les gérer. Dans ce cadre général, un des enjeux principaux est aujourd'hui de savoir comment exploiter au mieux ces sources d'informations hétérogènes et complémentaires pour en tirer le maximum d'information (car, en sciences de la complexité, « le tout est plus grand que la somme de ses parties »[33]). Selon la typologie des sources impliquées dans le processus, deux stratégies de fusion sont envisageables : la fusion précoce et la fusion tardive. Dans le premier cas de figure, les données sont combinées ensemble au début du processus pour constituer un unique et nouveau jeu de données homogènes. Par exemple, ramener toute l'information disponible à la même résolution spatiale, temporelle ou d'unité d'analyse. Dans ce contexte, une fois le nouveau jeu de données constitué, des techniques classiques de l'analyse monosource peuvent être utilisées. Dans le second cas de figure (fusion tardive), un processus d'analyse est mis en place par chaque source de façon spécifique et la combinaison est faite au niveau des descripteurs ou de la prise de décision. Par exemple à partir de chaque source, des descripteurs propres peuvent être extraits, qui ensuite seront combinés pour exploiter des interactions à plus haut niveau entre les différentes sources considérées. Enfin, les différentes sources peuvent être combinées dans un processus dit *"end-to-end"*, c'est à dire où les étapes classiques de traitement sont remplacées par un système unique (généralement un réseau de neurones profond) qui prend en entrée les sources brutes et renvoie en sortie les décisions attendues (*Charvat et al.*, 2018 ; *Plaisant*, 2004 ; *Tonda et al.*, 2018).

Dans le cas des séries temporelles, la fusion de séries ayant des résolutions temporelles différentes est un défi important. Par exemple, le capteur d'activité d'un collier porté par un animal envoie des informations toutes les cinq minutes, mais la pesée de ce même animal ne sera faite qu'une fois par jour. Pour comparer les individus entre eux, il peut être alors nécessaire d'interpoler les séries temporelles sur un même support de temps (par des méthodes de lissage linéaires ou polynomiales) en faisant éventuellement correspondre leurs similarités (dates de stades phénologiques, pics de croissance...). Le *Dynamic Time Warping* (*Sakoe et Chiba*, 1978) est une technique bien connue de mesure de similarités entre deux séries. Cette technique ne répond pas à toutes les questions de recalage de courbes que l'on rencontre autour du vivant, où la prise en compte du temps phénologique est fondamentale. Ces questions constituent un défi encore largement ouvert en biologie.

33. http://www.scilogs.fr/complexites/le-tout-est-il-plus-que-la-somme-des-parties/

Il existe également des méthodes capables d'extraire plusieurs modèles des séries temporelles à des échelles temporelles différentes, puis de choisir les plus pertinents grâce à des approches issues de la théorie de l'information (principe de longueur de description minimale – MDL) (*Vespier et al.*, 2012). L'intérêt de ces approches est qu'elles permettent de se focaliser sur l'échelle temporelle des phénomènes observés, et non sur celle de la valeur technique d'échantillonnage.

Dans le cas de la télédétection, avec l'explosion du nombre de missions satellitaires (*i.e Sentinel, Spot, Pléiades et PléiadesNeo, PlanetScope*, etc.), il devient possible de collecter, à moindre coût, des informations dans différents domaines spectraux (optique et radar) et à différentes échelles spatiotemporelles qui décrivent une même zone d'étude. Cette quantité massive d'information multisource requiert le développement de nouveaux outils de gestion et d'analyse de données (*Schmitt et Zhu*, 2016). Généralement, dans un processus classique de fusion multisource de données d'observation de la Terre, les sources sont exploitées à travers un processus de fusion précoce. Par exemple, dans le cas d'imagerie à différentes échelles spatiales, une étape de rééchantillonnage de l'information pour porter toutes les images à la même échelle spatiale est adoptée au préalable. Malheureusement, ce type de processus peut introduire des biais ou des erreurs dus à la génération de nouvelles informations synthétiques. C'est pour cela qu'aujourd'hui, des approches de fusion tardive sont préférées autant que possible. Des premiers exemples dans le contexte de la cartographie de l'occupation du sol commencent à apparaître mais nous sommes encore loin d'une solution générique pouvant être déployée de façon systématique sur des territoires différents et adaptée aux différents pratiques agricoles.

Par exemple...

À INRAE, l'équipe **MISCA** de l'UMR Tetis développe *Warping,* des méthodes de gestion de l'information permettant de répondre aux grands enjeux sociétaux liés à l'environnement, qu'il s'agisse de stocker, de gérer, de partager ou d'analyser de gros volumes de données. En particulier, elle contribue à la cartographie des sols en appliquant des techniques de *deep learning* sur de très gros jeux de données.

Chez Inria, plusieurs équipes-projets (**GEOSTAT**, **TITANE**, **FLUMINANCE** (Inria, INRAE, Université Rennes 1),...) et l'action exploratoire **AYANA** travaillent sur l'analyse d'images satellitaires.

Au-delà du multisource purement satellitaire, d'autres types d'information sont aujourd'hui associés à des données d'observation de la Terre. Par exemple, des informations géolocalisées « spontanées » ou issues de sciences citoyennes (*Ienco et al.*, 2019) ont une grande plus-value pour mieux étalonner et compléter les informations purement physiques des capteurs satellitaires.

Approches supervisées

L'analyse supervisée consiste principalement en deux tâches : la classification supervisée et la prédiction de futures valeurs. La classification supervisée consiste, étant donné une série temporelle et un ensemble de classes prédéterminées (par exemple « animal malade » et « animal bien portant »), à assigner l'une de ces classes à la série temporelle. En pratique, cela peut aider à déterminer l'état d'un animal ou d'une plante à partir de données de capteurs et d'informations sur les différents états possibles. Les méthodes de classification supervisée ont besoin d'être « entraînées » : pour cela, il faut leur fournir un nombre important d'exemples correctement étiquetés avec leur classe. À partir de ces exemples, l'algorithme de classification construit un (ou plusieurs) modèle(s), pour assigner une classe à une série temporelle non étiquetée en fonction des caractéristiques de celle-ci. Les grandes familles d'approche de classification supervisée diffèrent principalement par la manière dont elles construisent ces modèles. Les approches les plus simples, de type k-plus proches voisins (*k-Nearest Neighbor* ou kNN en anglais), ne construisent pas de modèle mais cherchent les k exemples de l'ensemble d'entraînement les plus proches de l'individu à étiqueter, et renvoient l'étiquette majoritaire. La difficulté est de choisir une bonne méthode de similarité (*Karlsson et al.*, 2016).

Enfin, les très populaires méthodes de réseaux neuronaux profonds peuvent être utilisées avec succès pour la classification de telles données. La méthode de ce type la plus performante actuellement, MLSTM-FCN (*Karim et al.*, 2019), combine un bloc convolutif CNN (*Convolutional Neural Network*) avec un bloc LSTM (*Long Short Term Memory*). Le bloc CNN, très utilisé en analyse d'image, sert de filtre qui parcourt la série temporelle ou le spectre et en extrait des attributs caractéristiques au temps t. Il est combiné au bloc LSTM, très utilisé en analyse de données séquentielles (en particulier des textes), et qui permet de mettre en relation des valeurs passées et présentes. Ce type d'approche peut donner d'excellents résultats (*Kamilaris et Prenafeta-Boldú*, 2018), par contre plus encore que les autres elle requiert une grande quantité de données d'entraînement étiquetées (qui peuvent être difficiles à acquérir dans certains scénarios agronomiques), et son paramétrage peut se révéler délicat (*Zhu et al.*, 2017).

Par exemple...

Chez Inria, l'équipe-projet **STATIFY** (Inria, UGA, CNRS(Inria, CNRS, Institut polytechnique de Grenoble) s'intéresse à la modélisation statistique de systèmes mettant en jeu des données ayant une structure complexe. Ses membres développent des méthodes statistiques pour capturer la variabilité des systèmes étudiés en garantissant un bon niveau de précision et en prenant en compte les valeurs extrêmes reflétant généralement des phénomènes rares. En particulier ils modélisent les événements météorologiques pour l'agroécologie.

Comme exemple de l'utilisation de méthodes de classification supervisée en agriculture, dans *Fauvel et al.* (2019), les auteurs exploitent des données de capteurs d'élevage de précision provenant de vaches laitières. Ces vaches sont équipées de thermomètres ainsi que de colliers avec un accéléromètre. L'analyse des séries temporelles de température et d'activité physique permet de détecter la période d'*œstrus* plus précisément que les méthodes existantes ou l'observation visuelle, y compris dans les cas fréquents (30 %) où les vaches n'expriment pas de comportement particulier en phase de pré-*œstrus.*

Approches non supervisées

Les approches non supervisées permettent de découvrir certaines structures dans les données, que ce soit des regroupements avec le clustering ou des motifs récurrents avec le *pattern mining*.

Clustering. Le *clustering* (ou classification non supervisée) est une méthode d'apprentissage qui a pour objectif d'identifier des classes pertinentes dans les données. À l'intérieur de chaque classe, les données sont regroupées par similarité ou par proximité. Pour obtenir une bonne classification, il faut à la fois minimiser l'inertie intraclasse (pour avoir des classes homogènes) et maximiser l'inertie interclasse (pour avoir des classes bien différenciées). Deux grandes familles de méthodes sont couramment utilisées : i) la classification ascendante hiérarchique, ou CAH, qui cherche à regrouper itérativement les individus, en commençant par le bas (les deux plus proches) et en construisant progressivement un arbre, ou dendrogramme, regroupant finalement tous les individus en une seule classe, à la racine ; ii) la classification par réallocation dynamique (dont l'algorithme des *k-means* est une version très connue) : le nombre k de classes est fixé *a priori.*

Ayant initialisé k centres de classes, tous les individus sont affectés à la classe dont le centre est le plus proche au sens de la distance choisie. L'algorithme calcule ensuite les barycentres de ces classes qui deviennent les nouveaux centres. Le procédé (affectation de chaque individu à un centre, détermination des centres) est itéré jusqu'à convergence vers un minimum (local) ou un nombre d'itérations maximum fixé.

Les principaux verrous à lever quand on fait du clustering de données multivariées sont d'identifier le « bon » nombre de classes et de définir une distance adaptée aux données, avec parfois la nécessité de réduire la dimension. Une technique courante est de faire de l'analyse en composantes principales sur les données et d'appliquer ensuite le *clustering* sur les coordonnées des données dans la base des vecteurs propres, avec toutes les difficultés de choix de dimension que cela comporte. Une façon de contourner ces difficultés est de faire du *clustering* par mélange de processus de Dirichlet (*Coquet et al.*, 2002).

Découverte de motifs *(Pattern mining).* Les motifs correspondent à des régularités/irrégularités ou des spécificités implicites des données ou de sous-parties des données. Dans les applications agronomiques, un individu peut être décrit à travers une suite de caractéristiques/événements. Par exemple, une parcelle décrite par une suite d'opérations culturales, une plante décrite par une séquence ADN, etc. Un des grands défis est l'extraction de sous-séquences fréquentes/rares dans ce type de données.

Par exemple…

À INRAE, l'UMR **TETIS** s'intéresse à l'extraction de sous-séquences fréquentes/rares dans ce type de données et de motifs fréquents sous forme d'items et de séquences (suites d'événements ordonnés dans le temps) afin de caractériser la différence de croissance de végétation entre différentes zones spatiales. En particulier, leurs travaux s'appliquent à l'estimation des surfaces humides et au suivi de la biodiversité.

D'autres approches visent à mettre en évidence des sous-parties des données dont les caractéristiques sont très différentes de celles du reste des données (par des différences de distribution sur certains attributs, etc.). Par exemple dans *Millot et al.* (2020), les auteurs ont exploité la notion de motifs discriminants afin de caractériser, à partir de données de simulations, des sous-familles de protocoles de cultures dans les fermes urbaines dont une partie des attributs (température, lumière, CO_2, etc.) montre une distribution intéressante par rapport à une mesure d'intérêt donnée.

Ces méthodes sont confrontées à un nombre de motifs trouvés trop important pour une utilisation facile par les experts. Une piste prometteuse et très étudiée actuellement est la sélection du sous-ensemble des motifs les plus pertinents. L'extraction de motifs à partir de séries temporelles peut s'effectuer après un prétraitement transformant la suite de valeurs numériques en suite de valeurs symboliques : les méthodes classiques de découverte de motifs s'appliquent alors. Quand les données numériques sont conservées, les méthodes d'extraction de sous-séries représentatives, appelées shapelets, peuvent s'appliquer.

Par exemple...

Des unités INRAE comme l'UMR **PEGASE**, **l'UMRH**, l'UMR **Toxalim** et l'UMR **GenPhyse** utilisent ces différentes approches d'apprentissage respectivement pour l'alimentation de précision, la détection précoce des anomalies de rythme d'activité de vaches laitières dans un troupeau, la détection de pathologies chez les porcelets ou l'analyse du comportement des truies.

Apprentissage par renforcement

Comme de nombreux types de données, les données agricoles sont souvent incertaines (cf. 3.1). L'apprentissage par le renforcement (en anglais *reinforcement learning*, abrégé RL par la suite) consiste à apprendre à agir dans un environnement incertain. Un exemple d'utilisation moderne du RL pour la planification de la gestion de l'ensemble des cultures a pour origine *Garcia* (1999), basé sur l'interaction avec un modèle de règles de décision pour la culture du blé, Déciblé (*Chatelin et al.*, 2005). Le simulateur empirique de culture est utilisé pour évaluer les politiques exprimées sous forme d'ensembles de règles de décision. Dans *Ndiaye*, (1999), des méthodes sans modèle – à savoir le *Q-learning* et le *R-learning* – ont été mélangées avec des algorithmes génétiques, des arbres de décision et de la logique floue pour trouver des règles de décision optimales pour la gestion des cultures couplées à Déciblé. Le résultat n'a pas été jugé aussi bon que les choix de décision auxquels un expert pourrait s'attendre. Ces premières approches étaient intéressantes dans la mesure où elles ont introduit des techniques modernes de RL pour la gestion des cultures en considérant toute une série d'actions. Elles ont également exprimé une politique optimisée de manière naturelle, c'est-à-dire sous la forme d'un ensemble de règles de décision simples qui correspondent au raisonnement des agriculteurs. Les limites des études de *Garcia* (1999), et *Ndiaye* (1999) sont que l'apprentissage est hors ligne, en utilisant un simulateur de modèle de décision empirique avec ses propres biais et domaine de validité.

L'apprentissage n'étant qu'en différé, les systèmes n'utilisent pas les commentaires des agriculteurs pour améliorer la politique apprise à l'aide du simulateur.

Ces méthodes ont ensuite été appliquées à un cadre plus complexe, en incorporant un modèle économique pour la gestion du colza, et une composante de parasites et de maladies dans la modélisation des cultures (*Trépos et al.,* 2014). Les méthodes de RL ont été appliquées avec succès à la planification de l'irrigation lorsque la disponibilité en eau est limitée (*Bergez et al.,* 2001). Néanmoins, chaque décision de gestion doit être prise en tenant compte de l'ensemble de la séquence des choix. Différentes variétés de cultures ont des besoins en eau différents ; ainsi, il y aura des coûts d'irrigation différents. *Bu et Wang* (2019) ont proposé une architecture informatique générale pour la prise de décision intelligente en agriculture, basée sur le *Q-Learning* profond. Dans la pratique, le *deep Q-Learning* nécessite des milliards d'essais et d'erreurs. En outre, il n'est pas proposé d'intégrer des connaissances spécialisées (par exemple, des connaissances sur la physiologie des plantes) dans ce système ; des approches utilisant des connaissances expertes pourraient être envisagées (apprentissage basé sur un modèle), permettant de réduire la quantité d'exemples nécessaires pour l'entraînement de plusieurs ordres de grandeur.

Par exemple...

L'équipe-projet **SCOOL** (Inria, CNRS, Université de Lille) spécialisée en apprentissage par renforcement, étudie la recommandation de pratiques dans le domaine de l'agriculture dans le contexte d'exploitations de très petite taille, en particulier dans des pays en voie de développement, et également dans le domaine du jardinage. Ces recherches sont menées dans une perspective de développement durable.

Les diverses méthodes d'apprentissage automatique et de sciences des données sont implémentées dans la bibliothèque scikit-learn, principalement développée à Inria et qui fait partie des trois bibliothèques d'intelligence artificielle les plus téléchargées au monde.

L'unité INRAE **MIAT** travaille également sur le développement de méthodes basées sur les processus décisionnels de Markov et sur l'apprentissage par renforcement appliqués à la gestion des systèmes agroécologiques, en abordant tout particulièrement les enjeux liés à la dimension spatiale des problèmes.

Les entrepôts de données et analyse OLAP

Les entrepôts de données (ED) sont apparus pour gérer de très gros volumes de données issues de sources hétérogènes (*Chandra et Gupta*, 2018). La modélisation multidimensionnelle (*i.e.* données caractérisées à travers de multiples axes d'analyse) et hiérarchique (*i.e.* un axe d'analyse peut être associé à différents niveaux de granularité) est la base des ED, et de l'analyse multidimensionnelle. Par exemple, l'analyse de la quantité de pesticides ou d'azote appliquée par les agriculteurs peut être caractérisée selon plusieurs dimensions (ou axes d'analyse) : temporelle, spatiale, et au niveau de la culture (*Bouadi et al.*, 2017). Cela permet de représenter les quantités par type de culture, par saison et par parcelle. Ces dimensions peuvent être exprimées selon différents niveaux de détail. Par exemple, l'information spatiale peut être définie à l'échelle d'une parcelle, ou à une échelle plus générale comme le bassin versant, la région, etc. En effet, chaque parcelle appartient à un bassin versant, qui à son tour appartient à une région, elle-même située dans un pays.

L'analyse multidimensionnelle fait appel à des traitements OLAP (*On-Line Analytical Processing*) permettant d'agréger, de visualiser et d'explorer de manière interactive les données. Si nous reprenons l'exemple précédent, nous pourrions alors analyser la quantité des pesticides ou d'azote à l'échelle de la parcelle ou alors à un niveau plus agrégé de la dimension spatiale comme le bassin versant. Les traitements OLAP servent à naviguer entre différentes granularités d'une ou de plusieurs dimensions, et ce de manière très efficace (*i.e.* la navigation est instantanée).

Les utilisateurs exploitent l'entrepôt de données en combinant les différentes dimensions et les différents niveaux de granularité des hiérarchies correspondantes. Pour sélectionner les données appropriées à l'échelle adéquate, les utilisateurs expriment et soumettent des requêtes à l'entrepôt de données.

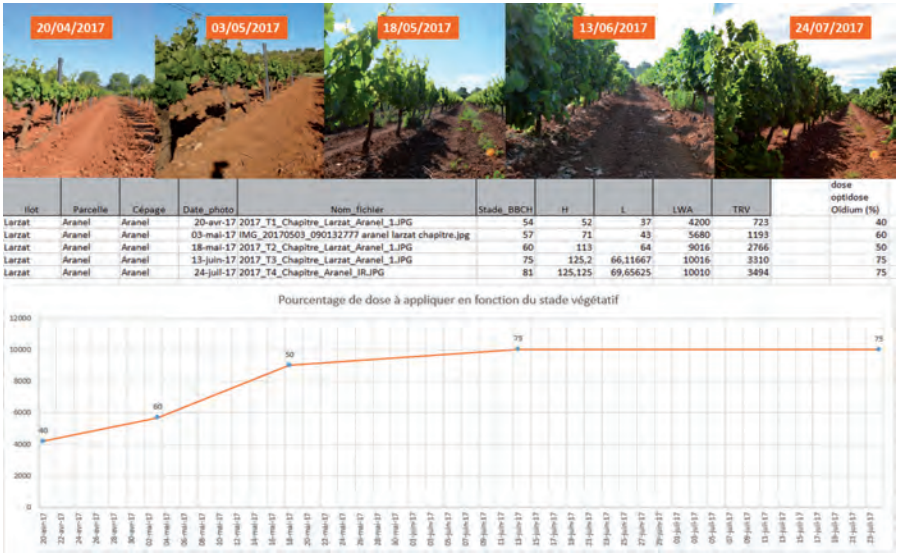
D'autres travaux (*Palpanas*, 2000) décrivent le couplage de l'analyse multidimensionnelle avec des méthodes de fouille de données (ex. *pattern mining*), l'objectif étant de proposer des méthodes hybrides associant les capacités exploratoires et analytiques de l'OLAP et les capacités descriptives de la fouille de données. Par exemple, l'outil ADSS-OLAP (*Abdullah et Hussain*, 2006), couplant OLAP et fouille de données (*clustering*) a été développé pour analyser l'incidence de la cochenille sur les cultures de coton. Pour enrichir davantage l'analyse OLAP et permettre l'exploration de données géographiques, l'idée a émergé de coupler les deux technologies OLAP et SIG (Système d'informations géographiques). Ainsi, le nouveau concept de Spatial-OLAP (SOLAP) (*Bédard et al.*, 2007) est introduit

pour pouvoir exploiter conjointement les outils OLAP (décision, graphiques, etc.) et les outils géographiques (représentation cartographique, agrégateurs géographiques, etc.).

Par exemple...

Chez Inria, l'équipe-projet **LACODAM** (Inria, Institut national des sciences appliquées de Rennes, Institut national supérieur des sciences agronomiques, agroalimentaires, horticoles et du paysage, Université Rennes 1) a modélisé et construit un entrepôt de données pour analyser/explorer, dans l'espace et dans le temps, les effets des pratiques agricoles sur les émissions d'azote dans l'eau et dans l'air (*Bouadi et al.*, 2017). Par ailleurs, l'équipe-projet étudie l'utilisation de l'apprentissage machine pour l'amélioration du bien-être animal (santé des vaches laitières et alimentation des truies).

À INRAE, l'unité **TSCF** consacre un de ses axes de recherche à l'OLAP spatial. Elle contribue entre autres à l'entreposage et l'analyse en ligne des données de biodiversité, notamment à travers le projet VGI4BIO (www.vgi4bio.fr) qui propose des méthodes d'analyse des indicateurs de biodiversité dans le contexte agricole centrés données et utilisateurs VGI.



Ilot	Parcelle	Cépage	Date_photo	Nom_fichier	Stade_BBCH	H	L	LWA	TRV	dose optidose Oïdium (%)
Larzat	Aranel	Aranel	20-avr-17	2017_T1_Chapitre_Larzat_Aranel_1.JPG	54	52	37	4200	723	40
Larzat	Aranel	Aranel	03-mai-17	IMG_20170503_090132777 aranel larzat chapitre.jpg	57	71	43	5680	1193	60
Larzat	Aranel	Aranel	18-mai-17	2017_T2_Chapitre_Larzat_Aranel_1.JPG	60	113	64	9016	2766	50
Larzat	Aranel	Aranel	13-juin-17	2017_T3_Chapitre_Larzat_Aranel_1.JPG	75	125,2	66,11667	10016	3310	75
Larzat	Aranel	Aranel	24-juil-17	2017_T4_Chapitre_Aranel_IR.JPG	81	125,125	69,65625	10010	3494	75

Optidose® un outil pour adapter la dose en fonction des paramètres de la culture et du risque épidémique. © Le Mas Numérique.

3.4 Gestion et ingénierie des connaissances pour l'aide à la décision en agriculture

Les sections précédentes ont présenté l'état de l'art sur la donnée, que ce soit sa collecte, sa gestion ou son traitement ; nous avons également vu comment la modélisation permettait de gérer et représenter de façon mathématique des connaissances à partir des mesures et observations pour aider à l'interprétation et à la préconisation. Une autre facette importante de l'agriculture numérique est la gestion de connaissances, c'est-à-dire des informations de plus haut niveau, aussi bien des connaissances scientifiques générales (par exemple processus physiologiques des plantes ou des animaux) que des méthodes spécifiques à certains acteurs des filières agricoles (par exemple conduite de troupeau d'un éleveur, méthodologie de fabrication de certains fromages, etc.). Des efforts importants ont été engagés ces dernières années pour formaliser ces connaissances et les organiser dans des ontologies permettant d'y accéder de manière structurée. Les ontologies sont une des composantes des systèmes informatiques aidant l'utilisateur à accomplir une tâche. Cette aide peut prendre plusieurs formes, de l'automatisation d'une décision d'irrigation à la découverte d'une information pour l'aider à prendre une décision. Les connaissances peuvent également être produites par les analyses présentées plus haut. Dans ce cas la difficulté est de présenter ces analyses aux acteurs humains de la manière la plus intelligible qui soit. Là encore, les développements récents en analyse des données sont d'un intérêt particulier pour l'agriculture, que ce soit *via* des approches de visualisation ou *via* les méthodes d'interprétabilité des modèles d'apprentissage. Enfin, le but de tout ce qui est présenté dans cette section est de permettre aux acteurs humains de prendre de meilleures décisions. Des outils spécifiques, les OAD (Outils d'aide à la décision), utilisant tout ou partie des techniques présentées dans ce chapitre, sont disponibles pour ces acteurs et en constante évolution : cette section se conclura par un tour d'horizon de ces outils.

Par exemple...

L'UMR **BIOEPAR** a contribué au développement de **EMULSION**, un logiciel et un OAD open source, basé sur l'intelligence artificielle. **EMULSION** permet aux modélisateurs de développer des modèles mécanistes stochastiques de systèmes complexes en épidémiologie, à différentes échelles et en utilisant différents paradigmes, tout en réduisant la quantité de code informatique à écrire (*Picault et al.*, 2019). Sur cette base, le projet **ATOM** (*Automation of decision support Tools based On epidemiological Models*) vise à développer un processus d'industrialisation de la génération d'OAD à partir de modèles épidémiologiques mécanistes (https://www6.angers-nantes.inrae.fr/bioepar/Recherche/Projets-en-cours/ATOM).

Les systèmes à base de connaissances en agriculture

Des premiers systèmes experts aux systèmes à base de connaissances – Dans les années soixante-dix sont apparus les premiers systèmes experts, issus de la recherche en intelligence artificielle. Ces systèmes sont dédiés à la résolution d'un problème précis à partir des connaissances issues d'un ou plusieurs experts pour mimer leur raisonnement dans le but de les remplacer à terme. Dans l'approche dite symbolique, les connaissances expertes sont formalisées à l'aide d'un langage de représentation des connaissances basé sur le raisonnement logique. Cela s'opposait alors à l'approche connexionniste, qui mimait le fonctionnement du cerveau humain par des réseaux de neurones.

Une particularité du secteur agricole est que les problèmes inhérents à la gestion d'une culture ou d'un troupeau nécessitent des connaissances expertes de plusieurs domaines (sciences des sols, météorologie, chimie, biologie...). Pour pallier cette demande forte de plusieurs expertises, certains systèmes experts intègrent comme composants des modèles de simulation tels que ceux décrits au 3.2. Citons par exemple le système expert *"CrOp MAnagement EXpert"* (COMAX), dédié à la culture du coton, dont l'objectif est de maximiser les rendements en minimisant les intrants (*McKinion et Lemmon*, 1985) et qui encapsule un modèle de simulation de développement du coton (GOSSYM). L'acquisition des connaissances expertes étant cruciale pour le développement des systèmes experts, l'ingénierie des connaissances se concentrait alors sur les méthodes d'acquisition de ces connaissances. Ces méthodes permettaient de guider le cogniticien dans les tâches

complexes d'identification, d'extraction et de formalisation des connaissances expertes issues de diverses sources (entretiens d'experts ou autres documents décrivant la tâche de résolution de problème).

Très prisés dans les années quatre-vingts, les systèmes experts ont été également sévèrement critiqués car non adaptables à une autre application que celle de départ, et difficilement évolutifs. Dans les années quatre-vingt-dix, les systèmes experts ont ainsi évolué vers des systèmes à base de connaissances. La notion d'ontologie apparaît alors dans le domaine informatique. Les ontologies ont pour vocation de formaliser des connaissances du domaine consensuelles et relativement stables, avec l'objectif qu'elles puissent être réutilisées dans d'autres systèmes à base de connaissances.

Un système à base de connaissances est composé de deux parties distinctes : d'une part, la base de connaissances, comprenant une ontologie structurant les connaissances du domaine, une base de faits qui instancie l'ontologie pour décrire des situations spécifiques, et éventuellement une base de règles qui enrichit l'ontologie ; d'autre part, un moteur de raisonnement qui est associé au langage de représentation des connaissances mais indépendant de toute base de connaissances particulière.

Évolution des méthodes d'acquisition et de capitalisation des connaissances – L'évolution vers les systèmes à base de connaissances s'est accompagnée d'une évolution de la conception des rapports entre l'humain et la machine. Les systèmes à base de connaissances et le système informatique intelligent associé ont pour objectif de coopérer avec leur utilisateur pour l'aider à réaliser une tâche nécessitant diverses connaissances, en complétant ses connaissances, en lui révélant les conséquences de ses choix et en lui proposant d'autres options que celles qu'il aurait imaginées. L'ingénierie des connaissances a évolué alors vers une forme de médiation de la modélisation des connaissances, produisant des « modèles de connaissances », un modèle prenant ici un sens différent plus englobant qu'en 3.2 puisque représentant des connaissances et non plus des phénomènes. Ces modèles permettent au cogniticien, en charge de l'implémentation dans un système informatique, de dialoguer avec des experts pour enrichir et valider les connaissances à représenter. Pour aider cette médiation, plusieurs méthodes ont été développées, la plus connue étant *"Knowledge acquisition and documentation structuring"* (KADS et son évolution commonKADS). Par exemple, un système préconisant des dates d'irrigation des manguiers a été développé à partir de la méthode commonKADS (*Nada et al.*, 2014).

Par exemple…

L'équipe-projet **GRAPHIK** (Inria, INRAE, CNRS et Université de Montpellier) étudie la représentation des connaissances. Elle travaille entre autres sur une méthode de collecte, de modélisation et de formalisation des connaissances visant à améliorer la qualité des fromages de terroir. La collecte est réalisée à l'aide de questionnaires ; la modélisation se fait par la création de cartes mentales (*mind maps*) pour faciliter la validation par les experts et la formalisation est réalisée dans le langage des graphes conceptuels.

Les modèles de connaissances vont se déployer sur d'autres systèmes informatiques comme les systèmes de recherche (de sources) d'information. Cette évolution se traduit par la création de « mémoires des organisations ». La mémoire d'une organisation est l'ensemble des ressources humaines et matérielles, supports de connaissances, permettant à l'organisation de réaliser ces tâches. Une mémoire peut ainsi être composée d'un ensemble de documents textuels, de vidéos, de listes de compétences des employés et d'un ou plusieurs modèles de connaissances. La formalisation de ces modèles permet leur exploitation automatique, en vue d'aider à la circulation des ressources et des connaissances au sein des membres de l'organisation. Cette formalisation prend souvent la forme d'un *thesaurus* : une liste organisée de termes normalisés, organisés par trois types de relations (équivalence, hiérarchie, association), dans le but d'indexer et d'aider à la recherche de divers contenus.

En savoir +

La FAO (organisation des nations unies pour l'alimentation et l'agriculture) a largement contribué à produire des mémoires d'organisation dans le domaine agricole (*O'Leary*, 2008). Sa base bibliographique AGRIS met à disposition différents types de ressources (documents scientifiques, jeux de données…) dans plusieurs langues. L'indexation de ces ressources est réalisée grâce au *thesaurus* AGROVOC, devenu le *thesaurus* multilingue incontournable dans le domaine agricole (*Sini et al.*, 2008). Ce *thesaurus* assure une large couverture du domaine agricole et peut être adapté en fonction des besoins d'une nouvelle organisation. Un exemple de mémoire d'organisation est le projet Agropedia réalisé par plusieurs instituts agricoles indiens avec la FAO (*Pappu et al.*, 2010), pour traduire les connaissances scientifiques issues des universités en des connaissances pragmatiques utiles pour les agriculteurs. Les modèles de connaissances utilisés dans Agropedia prennent la forme de cartes de thèmes (*Topic Maps*) agrégeant l'ensemble des connaissances nécessaires à une culture donnée.

Accès sémantique à des sources d'information – La naissance du Web sémantique au début des années 2000 a fortement impacté le domaine de la représentation des connaissances. Les technologies du Web sémantique sont un ensemble de langages, de protocoles et d'outils, standardisés sous l'égide du W3C, en vue de permettre l'exploitation automatisée des ressources du Web à partir de leur contenu. Les ressources du Web (des documents HTML par exemple, ou plus largement des données quelconques disponibles sur le Web) sont annotées avec des métadonnées décrivant le contenu de ces ressources dans un langage formel, constituant une base de faits qui peut être enrichie d'un thesaurus ou d'une ontologie qui précise sa sémantique.

Définition

Les principaux langages formels du Web sémantique sont les suivants :

- *Resource Description Framework (RDF)* : le langage de description des ressources du Web sous la forme d'un graphe constitué de triplets (sujet, prédicat, objet) ;
- RDF Schema (RDFS) : une extension de RDF permettant de définir un vocabulaire en termes de classes et propriétés (ou prédicats binaires) organisés par spécialisation ;
- *SPARQL Protocol and RDF Query Language* (SPARQL) : le langage d'interrogation de descriptions RDF(S) ;
- *Ontology Web Language (OWL)* : le langage privilégié de description des ontologies du Web Sémantique ;
- *Semantic Web Rules language* (SWRL) : un langage de règles qui permet d'enrichir des descriptions OWL ;
- *Simple Knowledge Organization System* (SKOS) : spécification en RDFS dédiée à la formalisation des terminologies, *thesaurus*, classifications et autres vocabulaires contrôlés utilisés dans les systèmes de recherche d'information.

Les ontologies du Web sont modulaires, centrées sur un besoin précis pour faciliter leur réutilisation et leur combinaison. Leur mise à disposition sur le Web améliore l'interopérabilité entre les systèmes à base de connaissances. Notons l'initiative du groupe de travail de la RDA Agri Semantic qui a étudié l'usage de ces technologies et ressources associées pour améliorer l'échange et le partage des données agricoles (*Aubin et al.*, 2017). Ces technologies ont permis la transposition des mémoires d'organisation vers le Web.

Par exemple…

INRAE développe des mémoires d'organisation concernant l'impact des changements climatiques sur les pratiques agricoles et l'agroécologie. Une archive des bulletins d'alertes agricoles français, les Bulletins de Santé du Végétal (BSV), a été constituée lors du projet VESPA qui a étudié les réseaux d'épidémiosurveillance (*Roussey et al.*, 2017).

Le portail Web collaboratif GECO (https://geco.ecophytopic.fr/) a été développé pour améliorer le partage de connaissances autour de la protection intégrée des cultures et de l'agroécologie. Ce portail gère un ensemble de fiches textuelles explicatives pour proposer des moyens de lutte contre les bioagresseurs (*Soulignac et al.*, 2017). GECO permet des recherches indépendantes du niveau d'expertise de l'utilisateur.

Les ontologies du Web et les thesaurus SKOS deviennent des ressources réutilisables. Pour retrouver l'ensemble de ces ressources, des portails spécifiques ont été développés.

Par exemple…

À INRAE, l'axe informatique de l'UMR **MISTEA** a, entre autres, développé le portail AgroPortal (http://agroportal.lirmm.fr/) qui répertorie et met à disposition de façon ouverte les ontologies et thesaurus autour de l'agronomie et l'agriculture. AgroPortal fournit également des services d'aide à l'annotation de documents textuels et à la détection de liens entre les concepts de deux ontologies (problématique d'alignement d'ontologies). Il contribue également aux avancées du phénotypage haut débit des plantes.

Intégration sémantique de données structurées – Le Web de données liées (*Linked Data*), développé dans les années 2010, marque une nouvelle étape du partage de données basée sur les technologies du Web sémantique : on considère alors un réseau d'ensembles de ressources liés entre eux. Ce réseau repose sur l'utilisation de vocabulaires partagés (*thesaurus*, ontologies, etc.) pour décrire les données. Cette évolution va de pair avec la généralisation de la notion de donnée : on parle de données quelconques, notamment de données structurées issues de diverses bases de données.

Par exemple...

Quelques exemples de données structurées :
Les données météo d'une station INRAE sont disponibles sur le Web de données (http://meteo.clermont.cemagref.fr/). Une ontologie centrale *Semantic Sensor Network* (SSN) fournit un patron (*design pattern*) de description des mesures.

Le projet européen SmartOpenData (http://www.smartopendata.eu/) a proposé une infrastructure (et un schéma de données *SmartOpenData* (smod)) pour gérer les données ouvertes et liées dans le domaine de la biodiversité et de l'environnement (par exemple en gestion des données d'agroforesterie).

Le projet *Agronomic Linked Data* (AgroLD– http://www.agrold.org) intègre dans une même base RDF 50 bases de données. Son objectif est d'interroger conjointement et de lier des points de vue différents sur les plantes cultivées (génomique, protéomique et phénomique) et formalisés par au moins une des dix ontologies du Web utilisées (*Gene Ontology, Plant Trait Ontology*, etc.).

Des architectures hybrides intégrant des ontologies, un moteur d'inférence, et des bases de données de différents formats (relationnelle, NoSQL, RDF) telles que *OpenSilex* (http://www.opensilex.org/) sont utilisées pour développer plusieurs systèmes d'information sur le phénotypage haut débit. Dans cette architecture, tout objet scientifique (plant, pot, champ...) est identifié par un identifiant web (URI) et typé par un élément d'une des ontologies associées. La base RDF stocke les métadonnées descriptives statiques et la base NoSQL stocke les flux de données brutes : photos de drones, séries temporelles de capteurs en champs, etc.

Par exemple...

Le système d'information GnpIS est en charge de stocker toutes les données structurées des expérimentations faites sur le phénotypage des plantes (*Pommier et al.*, 2019). Les ontologies proposées par le réseau *Crop Ontology* (https://www.cropontology.org/) sont utilisées comme des dictionnaires de tous les traits observables dans les expérimentations.

Concernant l'animal, un réseau d'ontologies du Web développé à INRAE est utilisé pour rendre compatibles les descriptions des expérimentations animales effectuées dans différents centres de recherche. Ce réseau se compose actuellement de trois ontologies du Web (*Salaun et al.*, 2018) : *Animal Trait Ontology for Livestock* (ATOL, sur les traits phénotypiques des animaux d'élevage), *Environnement Ontology for Livestock* (EOL, sur les paramètres environnementaux d'élevage), et AHOL (pour la santé animale du bétail).

Architectures émergentes – Au-delà du Web de données, la problématique de l'exploitation intelligente de données toujours plus volumineuses et hétérogènes suscite des recherches très actives qui combinent représentation des connaissances, gestion de données, Web sémantique, fouille de données et apprentissage, etc. C'est dans ce contexte qu'a été proposée une nouvelle architecture appelée *Ontology-Based Data Access* (OBDA) (*Xiao et al.*, 2018) qui combine une approche particulière de l'intégration de données, dite par médiation, avec la notion de système à base de connaissances. Un système OBDA est structuré en trois niveaux : le niveau conceptuel, organisé autour d'une ontologie (décrite par exemple en RDFS ou OWL) ; le niveau des données, composé de diverses bases de données préexistantes et indépendantes entre elles ; et le niveau des *"mappings"* qui traduisent les données pertinentes pour l'application visée en une base de faits utilisant le vocabulaire de l'ontologie. Les requêtes au système (par exemple en SPARQL) utilisent ce vocabulaire, l'utilisateur s'exprimant au niveau conceptuel, sans connaissance du système de stockage des données (par exemple, une requête *« quels sont les auxiliaires permettant de lutter contre le bioagresseur X et les itinéraires techniques associés qui limiteraient la compétition avec la plante principale ? »* fait abstraction des schémas des bases de données sous-jacentes).

Par exemple...

À INRAE, l'unité URFM Écologie des Forêts Méditerranéennes du département **ECODIV** mène des recherches pluridisciplinaires en écologie. En particulier, elle met en œuvre des systèmes OBDA matures tels que Ontop (https://ontop-vkg.org/) et MASTRO (https://www.obdasystems.com/mastro) pour une gestion durable des écosystèmes forestiers méditerranéens.

Dans le cadre de l'Internet des objets, certains de ces systèmes utilisent aussi les technologies Web sémantique (ontologie OWL, règles SWRL, base d'annotations RDF).

Les organismes de normalisation comme le W3C ou l'ETSI travaillent actuellement sur la validation de nouveaux standards et ontologies pour combiner l'Internet des objets et le Web sémantique : l'ontologie SAREF pour ETSI et l'ontologie *"Web of Things"* (WoT) pour le W3C. Ces ontologies n'ont pas encore atteint le niveau de maturité suffisant pour être utilisées dans des applications réelles.

Par exemple...

À INRAE, l'unité **TSCF** propose une traduction de la méthode d'irrigation manuelle IRRINOV sous forme de règles SWRL et d'un réseau d'ontologies du Web pour représenter les connaissances permettant d'automatiser les arrosages.

Des questions demeurent sur la compatibilité des ontologies construites sur des principes différents : usages différents, auteurs différents, ontologies fondationnelles différentes etc. Certaines ontologies du Web proposent des schémas de données correspondant à des patrons réutilisables (*design patterns*) centrés sur un besoin précis. D'autres ontologies proposent des classifications de référence pour qualifier les données. Les gestionnaires de données doivent ainsi construire un réseau d'ontologies pour structurer leurs données, en vérifiant que ces ontologies restent compatibles entre elles. Sont-elles basées sur les mêmes patrons ? Permettent-elles de produire des inférences correctes ? Enfin, les questions de distribution du raisonnement sur tous les composants d'un système de type « Internet des objets » sont des problématiques de recherche actuelles.

Restitution des connaissances, visualisation, interactions homme-machine en agriculture

Les approches de production de connaissance à base de données (section 3.3) ont conduit à des résultats de plus en plus précis et fiables, mais aussi de plus en plus difficiles à comprendre, au point que l'on qualifie aujourd'hui la plupart de ces approches de « boîtes noires », avec l'impossibilité pour l'utilisateur de comprendre les déterminants du résultat produit (par exemple une décision de l'itinéraire technique). Une solution à ce problème est d'utiliser des approches d'interprétabilité locale comme *LIME* (*Ribeiro et al.*, 2016) ou SHAP (*Lundberg et Lee*, 2017), qui visent non pas à expliquer le modèle appris dans sa globalité, trop complexe, mais les raisons qui ont conduit le modèle à produire cette décision dans le cas précis fourni par l'utilisateur, comme, par exemple, les attributs qui ont le plus contribué (positivement ou négativement) à la décision. Par exemple, *SHAP*, utilisé pour la détection d'*œstrus* vue plus haut (*Fauvel et al.*, 2019), fournit des explications du type: *« un œstrus a été prédit aujourd'hui en se basant sur l'évolution de la température des trois derniers jours et sur un temps de repos important il y a trois jours ».*

En parallèle à la question de l'interprétabilité, la représentation visuelle de données et d'informations est un passage obligé pour tout système informatique lorsqu'il s'agit d'interagir avec des utilisateurs. « Visualiser » consiste à produire des éléments visuels (graphes, courbes, cartes, images) pour aider les utilisateurs à comprendre, explorer, analyser, donner du sens à des données, des modèles, des informations, parfois en grandes dimensions, souvent complexes (*Kubicek et al.*, 2013). Une interface homme-machine, des visualisations fluides et efficaces, sont souvent essentiels dans le succès de systèmes numériques grand public. Le domaine de l'agriculture n'échappe pas à cette réalité.

L'exigence en matière de visualisation dans ce secteur est forte du fait de la conjonction de plusieurs facteurs : une croissance saisissante des masses de données collectées, des utilisateurs non informaticiens mais souvent technophiles, la nécessité d'une vision à différentes échelles spatiales et temporelles, sur des données privées et publiques. La visualisation est parfois même considérée comme une matière stratégique, car la maîtrise de ces techniques peut représenter un avantage compétitif, un pouvoir, pour certains acteurs de la chaîne agroalimentaire. Le domaine est d'ailleurs fortement investi par les acteurs privés (équipementiers), mais il existe des initiatives pilotées par des universités et des instituts, dont INRAE et Inria, sous licences libres (par exemple *AQUAPONY*[34], *GeoVisage*[35], *PARCHEMIN*[36]) ou participatives (*I-EKbase*[37]) (*Wachowiak et al.*, 2017).

34. http://www.atgc-montpellier.fr/aquapony/
35. http://geovisage.nipissingu.ca/
36. http://www.parchemins.bzh/index.php/outil-de-visualisation-donnees-lagriculture-littorale-bretagne/
37. http://iekbase.com/hot-spots-monitorin

Par exemple...

À INRAE, le groupe Écologie et Évolution des Zoonoses de l'UMR **CBGP** analyse la diversité virale des hantavirus et des processus évolutifs qui la façonnent. Ils ont entre autres piloté le développement de AQUAPONY, un visualiseur web qui permet de naviguer de manière interactive sur un arbre phylogénique et facilite l'interprétation objective des scénarios évolutifs.

Les outils actuellement déployés pour le monde agricole reposent sur des paradigmes classiques de visualisation : cartographie (GIS, Geographical Information Systems), interfaces de visualisation de données de capteurs, collections de visualisations liées (multifacettes) et outils réactifs (*dynamic queries*). Les interfaces sémantiques et les stratégies de visualisation liées (mise en relation de vues) sont des sujets très actifs, ainsi que les visualisations 3D, qui donnent une perception de la topologie des zones géographiques (jusqu'aux profils des champs), et l'usage de la synthèse d'images, voire de la réalité augmentée.

Par exemple...

À INRAE, l'UMR **SAS** accompagne la transition agroécologique des systèmes et territoires d'élevage. En particulier, elle a participé au projet PARCHEMINS achevé en mars 2021 sur la visualisation de données d'agriculture littorale en Bretagne. Le projet PARCHEMINS a mis en place un visualiseur cartographique, outil web de consultation de cartes, qui permet à l'utilisateur d'interagir avec des données géographiques. À destination d'utilisateurs non géomaticiens ou non spécialistes des outils informatiques, il permet de représenter et d'analyser de l'information spatiale, de façon intuitive et simple d'utilisation.

L'interactivité et la rapidité de réponse des outils de visualisation sont un défi en agriculture (et ailleurs !), car ces visualisations doivent être adaptées à des terminaux légers (*smartphones*, tablettes) ou embarqués (tracteur connecté). La fluidité des visualisations de données est étroitement liée aux solutions techniques, aux protocoles d'échanges de données et à l'architecture des systèmes. La visualisation adaptative, sujet actuellement très actif en recherche, permet d'adapter la visualisation au contexte : profession de l'utilisateur, terminaux de visualisation, nature des données disponibles.

La question de la visualisation et du partage de la connaissance est peu abordée en pratique et reste ainsi pour l'instant plus dans le champ de la recherche que de l'application. Or en agriculture, l'expertise humaine joue un rôle traditionnellement très important, ce qui crée un contexte particulièrement favorable au développement de techniques interactives : il est en effet tentant de combiner les capacités expertes humaines avec des algorithmes d'apprentissage, d'optimisation ou de modélisation (*Boukhelifa et al.*, 2018). Suivant les stratégies et les systèmes, l'interaction homme-machine (IHM) peut être explicite (l'utilisateur est régulièrement interrogé *via* une interface, un système de visualisation) ou implicite (à l'insu de l'utilisateur ou en non-verbal, la machine captant des informations qu'elle utilise comme base d'apprentissage). Le courant actuel des recherches en visualisation et IHM se focalise préférentiellement sur des questions d'interprétabilité, d'explicabilité, de causalité et de transparence des interactions.

Le tandem IHM/visualisation intervient aussi dans un cadre où l'expertise humaine est essentielle pour gérer l'incertitude des données et la prise de décision sur des questions multicritères : les approches qualitatives sont complémentaires des approches automatiques/statistiques (choix des critères, par exemple), pour gérer les ambiguïtés, les trous de connaissances, les extrapolations à d'autres types de cultures. Des exemples peuvent être trouvés dans l'agroécozonage – fondé sur des techniques de clusterisation, de segmentation – ou la Web-application *Crop-GIS* couplant modélisation et visualisation pour la gestion de la culture du maïs[38]. Mais l'évaluation des systèmes interactifs est délicate car si l'algorithme apprend et s'adapte à l'humain, l'inverse est vrai aussi : l'utilisateur apprend à utiliser un système. Comprendre ces mécanismes subtils de coadaptation et de coévolution nécessite d'employer des approches de type science expérimentale (plans de tests, fiabilité des résultats, biais), de tester sur des cohortes de volontaires.

En conclusion, la visualisation et les IHM appliquées à l'agriculture sont des sujets relativement rares à la fois dans la littérature scientifique agronomique et dans celle sur la visualisation. Or c'est un facteur essentiel à l'adoption des technologies, les agriculteurs préférant des outils simples d'usage et moins précis à des outils très performants difficiles à utiliser (*Pierpaoli et al.*, 2013).

Outils d'aide à la décision (OAD)

À partir des années quatre-vingts, les programmes informatiques et l'électronique ont commencé à être utilisés pour améliorer l'efficacité de l'agriculture et mieux raisonner les activités agricoles. C'est l'émergence des premiers OAD numériques (Outils d'aide à la décision, en anglais DSS *Decision support system*).

38. https://www.cropgis.com/

Cette nouvelle révolution est plutôt bien perçue par les agriculteurs (79 % des agriculteurs qui utilisent les nouvelles technologies reconnaissent leur utilité, source *Rapport agriculture et innovation 2025* [39]) et aussi par la société qui attend des innovations numériques un moyen pour préserver l'environnement (47 % des personnes interrogées sondage *Etude Opinion Way 2016*). Les OAD numériques sont basés sur de la programmation informatique « simple » associée à un *corpus* de données de référence relativement restreint, ils s'installent sur un ordinateur personnel ou s'utilisent depuis une interface web qui permet d'accéder à l'application. Ils sont le plus souvent développés par les instituts de recherche ou les instituts techniques. Des logiciels comme *INRAtion*[40] ou *InraPorc*[41] font partie de cette génération d'OAD. Ces logiciels conçus par l'INRA sont des références françaises en termes d'aide à la conception des rations alimentaires des ruminants et porcs d'élevage. Dans le domaine du végétal, de nombreux logiciels ont aussi été développés pour aider l'agriculteur à planifier et gérer la fertilisation des cultures, la lutte contre les ravageurs ou l'irrigation. Aujourd'hui, avec la déferlante du numérique, une nouvelle génération d'OAD a émergé, qui emploie des technologies numériques actuelles comme la télédétection, le GPS, l'Internet des objets, l'intelligence artificielle etc. Ces OAD sont conçus et produits par le secteur de l'AgriTech où sont présents les majors de l'agribusiness et de nombreuses startups (*Padhy et Satapathy*, 2020).

Définition

L'agritech est un terme générique désignant les technologies de l'agriculture. Les Agritech se divisent en quatre pôles majoritaires : 1/ le biocontrôle, 2/ le *big data* agricole, 3/ la robotique et 4/ la génétique et les biotechnologies végétales. Ces quatre éléments sont souvent très liés et de nombreuses technologies agricoles sont issues de ces pôles.

L'intégration des nouvelles technologies dans les OAD permet de démultiplier la gamme de services offerts par ces outils couramment utilisés par les agriculteurs qui cherchent à prendre les décisions les plus adéquates pour gérer leur ferme (*Spanaki et al.*, 2021). Connaître de manière précise l'état des parcelles agricoles ou des troupeaux est essentiel pour l'exploitant qui peut maintenant s'appuyer sur des données (images, mesures biophysiques…) issues de capteurs connectés pour avoir plus d'informations que ce que son œil ne peut lui permettre de collecter. Après différents traitements numériques, ces informations sont mises

39. https://agriculture.gouv.fr/sites/minagri/files/rapport-agriculture-innovation2025.pdf
40. https://www.inration-ruminal.fr/
41. https://inraporc.inra.fr/inraporc/

à disposition de l'agriculteur *via* une application dédiée accessible par le Web ou sur smartphone. Dans le secteur animal, les éleveurs n'hésitent pas à adopter ces nouveaux outils basés sur le numérique quand ils sont susceptibles de leur apporter un gain technique, économique et qu'ils peuvent diminuer la pénibilité de leur travail. Il y a tout d'abord les OAD basés sur des capteurs portés par les animaux (en externe ou interne), qui mesurent en temps réel les caractéristiques physiologiques de l'animal et son activité (température, capteurs de pression abdominale, mouvement ...). En élevage laitier, l'éleveur pourra ainsi s'appuyer sur ces outils pour piloter le cycle de reproduction de l'animal et détecter de manière fiable les chaleurs, les mises bas, ou des problèmes de santé, ceci avant même que des signes extérieurs ne soient visibles par un professionnel. On voit aussi émerger des prototypes d'OAD basés sur la reconnaissance d'images (issues de caméras installées en élevage) à l'aide de méthodes d'intelligence artificielle (*deep learning*), permettant le suivi des animaux en termes de comportement et de santé, voire la reconnaissance faciale. Si une anomalie est constatée sur un groupe ou animal, une alerte peut être envoyée sur le *smartphone* de l'agriculteur. Même si nombre d'initiatives sont en cours, différentes questions, déterminantes pour qu'un OAD soit utilisé et utilisable par les professionnels, font l'objet de recherches, de plus en plus en interaction avec ces professionnels. Cela concerne notamment la précision, la pertinence (un OAD fournissant trop de fausses alertes par exemple risque d'être abandonné), l'adéquation et la forme des informations fournies à l'agriculteur selon son expertise et ses besoins, ainsi que l'ergonomie de l'outil, en lien avec les notions vues dans la partie visualisation et IHM (*Li et al.*, 2020). La manière dont les connaissances de l'utilisateur seront utilisées fait l'objet de questionnements éthiques liés plus largement à l'innovation ouverte.

3.5 Automatisation, contrôle et robotique

Comme mis en évidence précédemment l'agriculture numérique est loin de se limiter à l'acquisition et au traitement des données. En effet, elle a vocation à exploiter ces données pour prendre des décisions et déterminer les actions à mener, tant au niveau spatial que temporel, afin d'optimiser les itinéraires culturaux capables de concilier de hauts niveaux de production, la qualité des récoltes et la préservation de l'environnement. En ce sens l'aboutissement de telles préconisations requiert un travail précis et potentiellement fréquent, qui n'est pas toujours compatible avec les ressources et capacités humaines. Ceci est d'autant plus vrai que les tâches agricoles s'avèrent bien souvent fastidieuses, et parfois dangereuses. L'exploitation du potentiel complet des principes de l'agriculture numérique peut donc mener à une automatisation des tâches. Aujourd'hui, les technologies robotiques prolongent les développements d'ores et déjà mis en œuvre dans le cadre d'outils automatisés, ou de systèmes d'aide à la conduite

d'engins agricoles. Mais au-delà de l'automatisation de certains travaux, les progrès de la robotique dans le monde agricole doivent ouvrir la voie à une évolution des pratiques pour accompagner la transition écologique.

Farmstar fait partie des OAD basés sur des images spatiales à résolution infraparcellaire. Il a été développé par Airbus en collaboration avec les instituts techniques agricoles. La chaîne de traitement est ici complexe puisqu'elle mêle l'utilisation d'images spatiales avec d'autres sources de données comme les données climatiques et qu'elle mobilise la simulation informatique de modèles agronomiques. Le résultat est mis à disposition via *des API (*Application Programming Interface*) qui sont interrogées par l'application utilisateur. L'agriculteur accède ainsi aux informations utiles sous forme de cartes et d'indicateurs « tableau de bord »* via *une application web intégrée qui masque l'architecture informatique complexe et les flux de données mobilisés.*

Figure 1 : Farmstar, de l'image spatiale haute résolution aux cartes de conseils

Par exemple...

L'équipe-projet **VALSE** (Inria, Ecole Centrale de Lille, Université de Lille) étudie les problèmes issus de l'analyse de systèmes dynamiques distribués, incertains et interconnectés. Elle vise la conception d'algorithmes d'estimation et de contrôle dans divers domaines. En particulier, dans le domaine de l'ostréiculture, ces algorithmes ont permis la conception d'un biocapteur basé sur les mesures et l'interprétation du comportement des mollusques bivalves, pour la détection à distance de la pollution des eaux côtières et des conséquences du changement climatique.

Les milieux structurés, alliés des robots

L'essor de la robotique s'est historiquement ancré dans les domaines d'application de l'industrie, notamment automobile, pour l'automatisation des chaînes de production (*Bahrin et al.*, 2016). Il est alors possible de concevoir des infrastructures permettant aux robots de se référencer et d'évoluer dans des environnements parfaitement connus et inchangés, ainsi que de maîtriser les conditions d'interactions (conditions de lumière, manipulation d'objets connus, création de zones spécifiques). Ceci aide grandement la conception d'algorithmes de perceptions et commandes robustes, basés sur des modèles d'évolution des robots qui nécessitent des hypothèses fortes (roulement sans glissement, reconnaissance d'objet ou de scène, localisation précise, etc.). Par conséquent, les applications de la robotique en agriculture se sont en premier lieu focalisées à l'intérieur des bâtiments, notamment pour la production animale (*Bergerman et al.*, 2016). En ce sens, le plus gros marché de la robotique dans le domaine de l'agriculture se situe aujourd'hui dans l'élevage, avec les robots d'affouragement ou de traite. Ceux-ci peuvent en effet exploiter un certain nombre de repères, et bénéficier d'aménagements particuliers pour conserver un haut niveau de répétabilité. Ils sont ainsi à même de remplir des missions astreignantes (comme la traite ou le nourrissage des animaux) et libérer du temps à l'agriculteur. De telles évolutions accompagnent de plus en plus les pratiques, puisqu'aujourd'hui la moitié des nouvelles installations françaises en élevage laitier s'équipent de robot de traite (*Tse et al.*, 2018) .

Dans le cadre de la production végétale, il est plus difficile de mettre en place de telles infrastructures, et la structuration apportée par les cultures est par essence changeante, posant des problèmes de détection et de référencement. Néanmoins, l'automatisation de certaines tâches, et en particulier la conduite d'engins agricoles, a grandement bénéficié de l'avènement du GPS, notamment dans sa version centimétrique, constituant de fait une référence absolue. Ainsi de nombreux dispositifs visant à automatiser le pilotage, sous la surveillance d'un « conducteur », ont vu le jour, partageant un certain nombre de problématiques avec les avancées sur le véhicule autonome.

L'exploitation du seul capteur GPS demeure toutefois limitée pour la réalisation de robots entièrement autonomes (i.e. sans surveillance humaine embarquée), pour plusieurs raisons. Tout d'abord, les pertes possibles des signaux satellitaires aux abords des bâtiments, sous serres, ou à proximité de végétation haute, impliquent une reprise en main. Ensuite, l'accomplissement de travaux agricoles exige un référencement et une interaction avec la végétation et non avec une référence absolue, même si les plantations sont réalisées avec un référencement GPS. Enfin

l'absence d'un superviseur embarqué impose de doter une machine autonome de moyens de perception pour garantir la sécurité (évitement d'obstacles, gestion de la traversabilité).

Étude d'un tracteur électrique robotisé pour l'agroécologie. © INRAE.

Ainsi, plusieurs autres stratégies – vision (*Stefas et al.*, 2019), laser (Tourrette et al., 2017) – viennent en substitution ou complément d'un référencement absolu pour réaliser la fonction de navigation autonome. Celle-ci est d'ores et déjà exploitée commercialement dans des robots, principalement pour le désherbage mécanique, la tonte, la surveillance. L'efficience de ces robots demeure néanmoins limitée pour le moment en termes de tâches et les performances sont intimement corrélées aux conditions de détectabilité.

Pour envisager des travaux plus complexes (taille, récolte en plein champ), de façon complètement autonome, il convient de pouvoir lever plusieurs verrous scientifiques et technologiques pour pouvoir appréhender la variabilité des milieux d'évolution, ainsi que la diversité et la complexité des tâches à effectuer, en conservant l'intégrité du robot ou des robots.

De l'adaptation à la reconfiguration

Contrairement à la robotique mobile en milieux industriels ou routiers, les robots mobiles évoluant en milieux naturels requièrent des capacités d'adaptation pour affronter la diversité des conditions d'interaction et leur variabilité (*Bergerman et al.*, 2016). Ceci suppose la modification en ligne des paramètres de perception et de commande (comme la modification des temps de réponse en fonction de la vitesse (*Hill et al.*, 2020) ou l'adaptation de seuil de détection en fonction des conditions de luminosité. Plusieurs mécanismes d'adaptation et d'anticipation ou de commande robuste ont été proposés pour affronter la variation de ces milieux et conserver un haut niveau de précision, tout en maintenant l'intégrité du robot (*Krid et al.*, 2017 ; *Yandun et al.*, 2017). Cette dernière fonctionnalité se définit de façon relativement binaire dans les milieux structurés : éviter les collisions avec un obstacle géométrique, ne pas évoluer dans une zone interdite. En milieux naturels, la notion d'obstacle est plus floue, et sa résolution plus complexe. D'abord, la rencontre avec un obstacle n'est pas nécessairement un cas d'échec, car les robots ne doivent pas être arrêtés lorsqu'ils enjambent la végétation ou doivent pousser une branche. Ensuite, certaines zones peuvent être franchies à certaines conditions (limitation de la vitesse ou du chargement), et le franchissement dépend aussi des conditions de sol (notamment l'adhérence) et des propriétés du robot (*Guastella*, 2018). Enfin, l'évolution dans certaines zones peut mener à une perte de contrôle ou de stabilité physique du robot (*Wolf et al.*, 2019).

Par exemple...

À INRAE, l'UR **TSCF** conçoit des systèmes reconfigurables et à autonomie partagée, pour accroître les performances et la sécurité des engins œuvrant en milieux naturels, en particulier ceux rencontrés dans l'agriculture. Elle conçoit par exemple des mécanismes d'adaptation pour affronter la diversité des conditions d'interaction et leur variabilité.

Plusieurs approches permettent de prendre en compte cette complexité par la notion de traversabilité (ensemble des conditions permettant le passage d'une zone définie devant le robot). Néanmoins, le travail autour de cette notion illustre la difficulté à définir une approche unique de perception et de commande pour réaliser des tâches agricoles complexes par un robot. Aussi, de nombreux travaux sont focalisés sur la sélection en temps réel ou la fusion de comportements types

(voir le projet INRAE *Adap2E*[42]), ce qui pose le problème de l'interprétation de scène et de l'évaluation des comportements. En outre, de nombreuses stratégies en robotique agricole sont centrées sur la coopération de robots plus basiques, capables de s'associer ou de travailler sur une même zone. Ceci permet de limiter les risques sur le fonctionnement de chaque robot (énergie cinétique limitée en cas de choc) et le coût de chaque robot, mais reporte la complexité sur l'association et la synchronisation de la flotte (*Blender et al.*, 2016).

Conclusion

Dans ce chapitre, nous avons vu les différentes recherches menées pour l'utilisation du numérique en agriculture. Elles sont principalement centrées autour de la donnée, à tous les niveaux du cycle de la donnée, de sa captation à son exploitation en passant par sa collecte, sa traçabilité, son traitement, son stockage, son interprétation, sa restitution ou son usage dans des systèmes automatisés ou robotisés. Différentes compétences sont mises en œuvre pour y apporter des solutions efficientes, sûres et sécurisées incluant le réseau, la modélisation, l'apprentissage, la gestion des connaissances, le contrôle. Les buts sont, entre autres, d'assister les agriculteurs dans des tâches difficiles, de permettre une meilleure gestion de nos ressources, de favoriser les échanges et savoir-faire, le tout dans le respect autant que possible de l'environnement.

42. https://adap2e.inrae.fr/