Devenir des données brutes stockées : extraction et traitement

Emilie Cobo[1], Yann Labrune[1], Pierre-Emmanuel Robert[2], Vincent Furstoss[3]

Résumé. Le stockage des données de phénotypage n'est pas une fin en soi. Leur valorisation constitue un réel outil pour le suivi quotidien d'un troupeau d'animaux sur le terrain et est également indispensable pour répondre à des questions scientifiques émergentes. Des outils, plus ou moins complexes d'utilisation, sont disponibles pour, dans un premier temps, extraire les données brutes des bases de données puis assurer leur traitement de façon à transmettre au scientifique des variables d'intérêt. Des bibliothèques logicielles permettent également la représentation graphique de ces données pour une meilleure visibilité.

Mots clés : donnée brute, extraction, traitement statistique, représentation graphique, variable d'intérêt

Méthodes d'extraction des données brutes depuis les différentes bases de données d'élevage et d'expérimentation

Selon le besoin (conduite d'élevage, suivi du troupeau expérimental ou questions scientifiques) et la base de stockage des données, différents types d'extraction sont possibles en fonction de la complexité de la requête.

Ainsi, l'application client/serveur (logiciel Windows) des systèmes d'informations, présentés dans le chapitre 3, garantit la récupération simple et préétablie de mesures d'élevage et d'expérimentation via un module d'extraction dédié. En se fondant sur le modèle conceptuel de la base de données, ce module permet de choisir une ou plusieurs tables et d'affiner l'extraction en fonction du choix de colonnes, de critères de sélection et de tri. Les requêtes régulières peuvent être enregistrées de façon à automatiser certaines exportations. Les fichiers extraits peuvent être récupérés sous plusieurs formats texte. Certains systèmes d'informations proposent également la récupération de statistiques simples. L'extraction des données à haut débit est assurée par Sicpa Expérimentations grâce à un programme python appelé SIDExTRACT qui utilise un cœur de la machine DGA20 (serveur de calcul du Centre de Traitement de l'Information Génétique) et se connecte sur le serveur Sicpa Expérimentations situé sur le même réseau. L'extraction est protégée par une clé de 40 caractères, créée aléatoirement et associée à chaque expérimentation au moment de sa déclaration. Typiquement, 45 min sont nécessaire pour récupérer 1,7 Go de données comptant 10,3 millions d'enregistrements soit 62 millions de mesures. Un programme SAS (statistical analysis system) présent dans la base Sicpa Expérimentations permet ensuite la conversion du fichier d'extraction généré en un format exploitable sous SAS. Cette procédure est utilisée notamment pour les données de phénotypage à haut débit collectées par des outils de mesure tels que les distributeurs présentés en chapitre 4 (Weisbecker et al., 2018).

Cependant, les questions scientifiques diverses et complexes nécessitent une souplesse d'interrogation qui ne peut être totalement couverte par des modules d'extraction préétablis. L'utilisation de la procédure SQL (structured query language) du logiciel SAS devient alors nécessaire. Dans de nombreux cas, l'utilisation de cette procédure est privilégiée par les statisticiens ou les techniciens en charge de réunir l'ensemble des données brutes sous forme d'un tableau. Ainsi, l'ouverture de la base nationale caprine (20 années de données) et la puissance du langage SQL ont permis l'étude rétrospective portant sur les sources de variation génétique et non-génétique de la réussite à l'insémination artificielle. Un

1 *UMR GenPhySE, Inra, 31326 Castanet-Tolosan Cedex, France*
2 *UMR MoSAR, Inra, AgroParisTech, Université Paris-Saclay, 75005 Paris, France*
3 *UE FERLUS, 86600 Lusignan, France*
emilie.cobo@inra.fr

autre exemple, la valorisation scientifique de l'expérimentation système Patuchev de par la diversité des questions auxquelles elle permet de répondre ne pourrait se faire sans l'usage de SAS et de sa procédure SQL pour extraire les informations de la base de données Sicpa Ovins/Caprins. La mise en œuvre de ces outils nécessite de multiples compétences : comprendre la question scientifique et le modèle de données mais aussi connaître le domaine « métier » que couvre la base de données (élevage caprin par exemple). À défaut il est nécessaire, en particulier pour le domaine « métier », de pouvoir disposer d'experts dont le rôle est crucial lors de la phase d'épuration des données. Il est illusoire de penser que le dictionnaire des données et les contraintes d'intégrité d'une base de données sont garants d'une véritable cohérence « métier » des informations extraites. La levée des incohérences nécessite également les explications d'experts connaissant parfaitement la chaîne de recueil des données. Ces prérequis ne sont pas tous spécifiques à l'utilisation de la procédure SQL, mais ils prennent ici tout leur sens du fait même de la souplesse de cet outil qui apporte une réelle gratification lorsque la réponse à la question scientifique est au bout de cette chaîne de compétences.

Traitement des données brutes pour la mise à disposition de variables d'intérêt fiables

L'utilisation d'automates de mesure dans le cadre de l'élevage de précision peut générer de grandes quantités de données de phénotypage. Ainsi, une collaboration accrue est nécessaire entre la personne gérant ces données et les chercheurs pour assurer une valorisation cohérente. Cette valorisation permet un meilleur suivi des animaux du troupeau et la mise à disposition de variables d'intérêt pour répondre à une question scientifique.

En élevage expérimental, ces données peuvent être exploitées via des outils de visualisation des variables biologiques d'intérêt. L'exemple illustré en **Figure 1** présente, pour un animal donné, les courbes de pesée et de traite, de même que les périodes en expérimentation et les événements sanitaires. Un algorithme simple est utilisé pour détecter automatiquement les valeurs de pesée atypiques, indiquées par un losange. Ce calcul identifie les valeurs qui s'écartent de la moyenne de la semaine précédente au-delà d'un seuil paramétrable, fixé ici à 5 %, et prenant en compte également la variabilité des pesées de chaque animal. Le calcul peut être appliqué sur les enregistrements passés mais également aux données quotidiennes, ce qui permet au responsable du troupeau de recevoir une alerte mail face à des situations pouvant demander une action de sa part. Les bibliothèques logicielles graphiques disponibles pour le logiciel R telles que plotly (**Figure 1**) ou dygraph, permettent de zoomer facilement et d'inspecter de longues séries temporelles rapidement et aisément. Aujourd'hui, cet algorithme de détection des valeurs atypiques n'est appliqué qu'aux données de pesée. Mais il pourrait l'être aussi pour d'autres variables, le poids de lait par exemple ou bien les mesures de l'activité de l'animal au moyen d'accéléromètres, et générer ainsi un indice numérique résumé de la santé et performance de chaque animal. Cet exemple illustre l'intérêt également de développer l'application en interne, plutôt que de recourir à un produit commercial, de manière à valoriser ces données biologiques de la façon désirée par leurs utilisateurs.

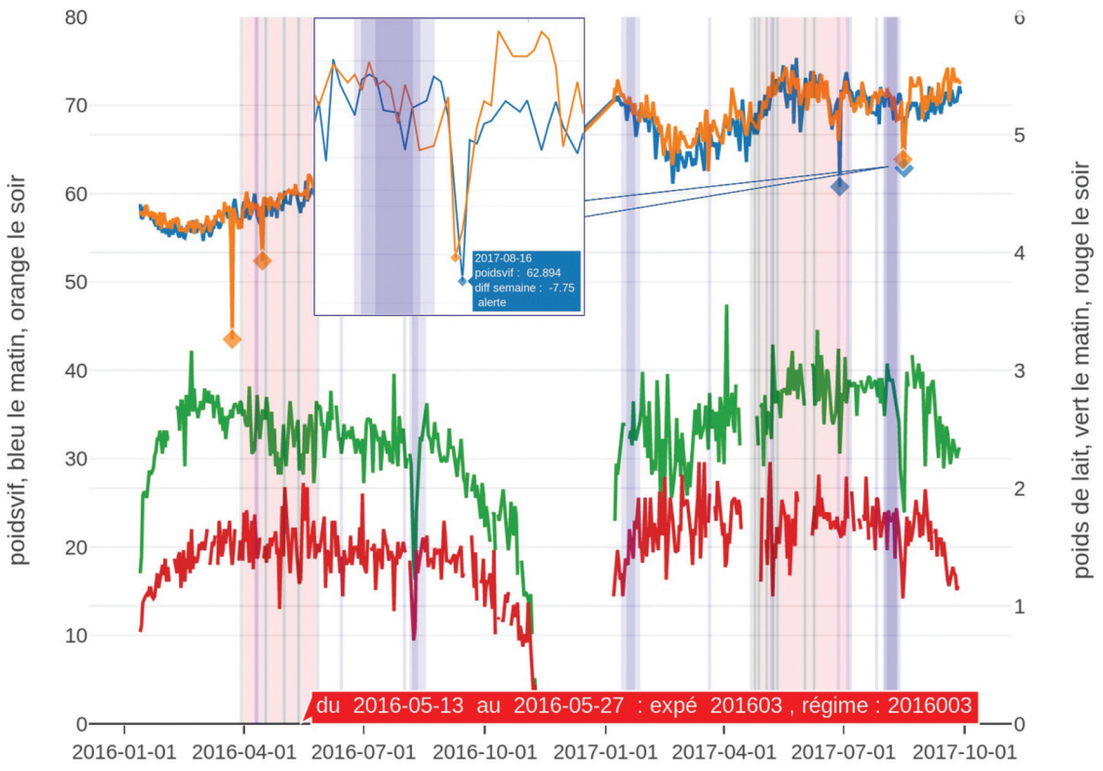

Figure 1. *Interface graphique interactive de visualisation des données du troupeau.*

En Unité de Recherche, les données de phénotypage stockées sont ce que les scientifiques appellent des données brutes. C'est à partir de ces données que sont faites les analyses statistiques. Souvent les données brutes ne sont que des variables intermédiaires dans le flux d'analyses. Afin de clarifier ce concept, voici un exemple concret. Dans le cadre du projet ADAPTMAT, l'activité de plusieurs truies bloquées en maternité a été étudiée afin d'évaluer la capacité d'adaptation des truies à ce système d'élevage, le dispositif ainsi que les résultats préliminaires sont détaillés dans l'article de Canario et al., 2018. L'expérimentation a été effectuée par l'Unité Expérimentale GenESI. Un capteur a été posé sur les truies afin de connaître leur position spatiale toutes les 10 secondes. Les données brutes sont ici la position de la truie sur les axes X, Y et Z, ce sont ces données qui sont stockées dans les bases de données. Elles ne sont pas directement utilisées par le scientifique. Il faut réaliser un traitement de la donnée brute afin d'obtenir la variable d'intérêt (ici la position physique de la truie : assis, debout ou coucher). Pour ce faire, la position physique de la truie déterminée par la visualisation des enregistrements vidéos a été mise en lien avec sa position spatiale mesurée à l'aide d'accéléromètre. À partir de ces données, un modèle est construit de façon à obtenir la position physique de la truie sur des périodes où aucun enregistrement vidéo n'a été effectué. Le modèle peut être appréhendé comme un arbre tel que sur la **Figure 2**.

Il est important de noter que bien que les données brutes ne soient pas directement intéressantes et soient bien plus volumineuses que les données d'intérêt, ce sont bien elles qui doivent être stockées dans les bases de données. En effet, le modèle est amené à évoluer en fonction des données, comme c'est le cas pour notre exemple et aussi dans la majorité des cas. De plus, avec la mise en place de l'Open Data et les évolutions des usages en matière de publications scientifiques, les jeux de données utilisés pour l'analyse ainsi que le modèle devront être fournis.

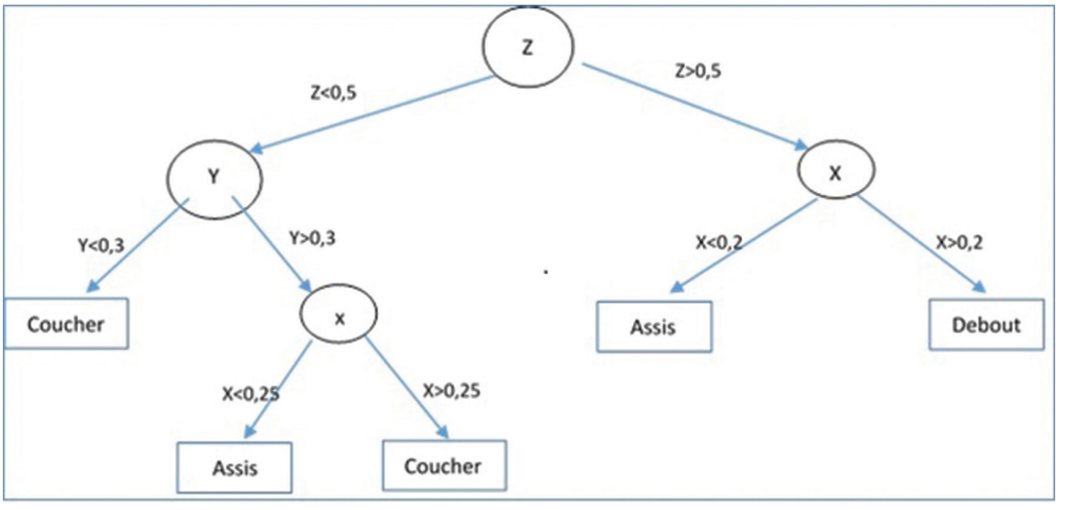

Figure 2. *Exemple d'un arbre de décision issu d'un modèle de forêt aléatoire.*

Perspectives

Plusieurs réflexions sont en cours autour de nouvelles technologies pour améliorer la facilité et la rapidité des extractions de données et leur partage tout en renforçant le contrôle d'accès aux données stockées afin d'assurer leur confidentialité. La mise en place de webservices permettrait d'accéder à plusieurs bases de données via les logiciels de statistique tels que R ou SAS, sans que le demandeur ne connaisse la structure des bases de données interrogées. Le développement d'interfaces web assurerait un accès aux données simplifiées sans devoir installer le programme client des systèmes d'informations. De plus, l'utilisation des technologies NoSQL garantirait le stockage et une extraction facile et rapide des données à haut débit notamment. Et enfin, l'instauration de l'Open Data permettrait la mise à disposition des données collectées sur nos Unités et Installations Expérimentales pour un meilleur partage et une valorisation accrue. Au vu de ces changements prochains, le rôle des biostatisticiens et des agents pivots à l'interface entre les Unités Expérimentales et les Unités de Recherche ne va cesser de prendre de l'importance, leurs compétences vont évoluer et leur position centrale sera renforcée pour une valorisation optimale des données brutes par les scientifiques.

Références bibliographiques

Canario L, Knudsen C, Delagarde R (2018) Mise au point de méthodes et d'outils spécifiques pour répondre à des questions de recherche autour du comportement animal. *Le Cahier des Techniques de l'INRA*, N° spécial Phénotypage animal, pp. 159-163.

Weisbecker JL, Lagüe M, Marcon D, Huau C, Trainini C, Ruesche J, Débat F, Portes D, Cobo E (2018) Les dispositifs de mesures individuelles de la consommation d'aliments. *Le Cahier des Techniques de l'INRA*, N° spécial Phénotypage animal, pp. 117-128.