Journées de méthodologie statistique de l'Insee 2022

REFONTE DE LA BASE DE SONDAGE DES EXPLOITATIONS AGRICOLES : VERS UNE UTILISATION ACCRUE DES FICHIERS ADMINISTRATIFS

Thierry GUILLAUME (*)

(*) Ministère de l'Agriculture et de l'Alimentation, Service de la statistique et de la Prospective, Bureau des méthodes et de l'informatique statistiques

thierry.guillaume1@agriculture.gouv.fr

Mots-clés : Appariement, échantillonnage, base de sondage, répertoire d'entreprises

Domaine concerné : Utilisation combinée de fichiers administratifs

Résumé

La statistique agricole dispose d'une base de sondage dédiée pour les enquêtes auprès des exploitations agricoles, et ce pour deux raisons : d'une part le secteur agricole n'est pas couvert par la démographie d'entreprises de façon aussi fine que le champ de l'industrie, du commerce et des services (ICS) et par conséquent, ne bénéficie pas d'un Sirus de qualité suffisante, et d'autre part, la définition de l'exploitation agricole utilisée dans la statistique agricole européenne est spécifique (le champ du secteur agricole n'est pas défini strictement sur le code d'activité principale exercée (APE) et/ou sur la forme juridique).

La version actuelle de cette base de sondage, constituée après le recensement agricole de 2010, avait été conçue pour être actualisée principalement par les données des enquêtes agricoles, mais également par des sources statistiques et administratives du ministère de l'Agriculture et de l'Alimentation ou de ses partenaires. Ce processus, insuffisamment automatisé, générait une charge de travail conséquente pour l'administrateur de la base et les gestionnaires en région, et s'est avéré être la principale raison du peu de mises à jour par les fichiers administratifs, jusqu'à la préparation du dernier Recensement Agricole.

La multiplication, l'évolutivité et la nature même des fichiers administratifs amènent aujourd'hui le ministère de l'Agriculture et de l'Alimentation à réaliser une modernisation du dispositif actuel, par un effort de traitement sur les données administratives, des appariements et des fonctionnalités d'historisation, afin de :

- disposer d'une base de sondage plus à jour et d'une meilleure qualité,
- centraliser les informations administratives disponibles sur ces exploitations et faciliter leur utilisation dans les enquêtes, afin de diminuer la charge de réponse des enquêtés,
- améliorer la gestion des contacts (courriels, adresse et téléphone notamment),
- diminuer la charge de gestion (de l'administrateur de la base et des gestionnaires sollicités pour expertiser des unités).

L'effort mené sur la caractérisation de la situation des exploitations agricoles pourrait à terme permettre d'envisager une démographie d'entreprises annuelle sur ce champ, ce qui représente à ce jour un défi de taille.
La présentation abordera les différents points de ce projet, dont la finalisation est prévue pour le 4e trimestre 2022, après intégration des données du Recensement Agricole 2020.

Abstract

For many years, the Ministry of Agriculture and Food has been using a sampling frame of farms which is entirely renewed every ten years with the results of the agricultural census. This frame is updated meanwhile with both results of thematic surveys and administrative data. The main flaw of the current frame is the lack of automated processing of administrative data, which prevents any frequent update, due to the burden of the system. The aim of the project is to modernize the sampling frame to allow frequent updates from administrative data which become more and more accessible. Consequently, decision rules are to be implemented to integrate sometimes diverging datasets. Additionally, having a up-to-date sampling frame for farms paves the way for extending business demography statistics to agricultural farms.

1. Définition de l'exploitation agricole

Lorsqu'on s'intéresse au domaine des statistiques d'entreprises, la base de sondage est généralement constituée à partir d'un répertoire d'entreprises.

Au ministère de l'Agriculture et de l'Alimentation, la définition de l'exploitation agricole diffère de celle des entreprises du champ agricole de l'Insee[1].

Une exploitation agricole est définie comme une unité économique et de production répondant simultanément aux trois conditions suivantes :

1. avoir une activité agricole,
2. atteindre ou dépasser une certaine dimension,
3. être soumise à une gestion courante indépendante.

L'activité agricole correspond à la production de produits agricoles ou le maintien des surfaces dans un état agricole et environnemental justifiant le paiement d'aides. Les produits agricoles retenus sont ceux liés aux activités classées dans les codes A.01.1 à A.01.6 de la nomenclature statistique des activités économiques dans la Communauté européenne (NACE) avec, pour le code A.01.49Z (élevage d'autres animaux), uniquement l'élevage d'animaux semi-domestiqués et autres animaux vivants (hors insectes), et l'apiculture et la production de miel et cire d'abeille. Ainsi, par exemple, les élevages de chats et de chiens ne sont pas inclus dans le champ agricole.

L'exploitation doit également atteindre une certaine dimension économique lui permettant de jouer un rôle d'acteur économique. Cette taille assure en théorie que l'exploitation peut participer à un processus de transaction commerciale comme la vente ou l'échange. En pratique, ce critère conduit à s'assurer que l'exploitation atteint au moins un seuil minimal de surfaces cultivées, d'animaux ou de production.

La troisième condition de définition d'une exploitation stipule que la mobilisation des facteurs de production pour la conduite des travaux sur l'exploitation est indépendante de toute autre unité économique. L'existence d'un Siret est considérée comme une présomption suffisante d'autonomie. À chaque Siret ne peut donc correspondre qu'une seule exploitation agricole dans les enquêtes statistiques. Le Siret, plutôt que le Siren, a été retenu comme identifiant adéquat pour appréhender l'exploitation agricole. En effet, la notion de localisation a été jugée essentielle et des entreprises peuvent avoir plusieurs établissements avec chacune une gestion autonome (c'est notamment le cas des unités qui réalisent deux déclarations d'aides à la politique agricole commune (PAC)). Cependant, il existe des exploitations possédant plusieurs Siret (situés dans des communes identiques ou

[1] Généralement le champ de l'agriculture de l'Insee est constitué, au sens large, par les entreprises et établissements dont l'activité principale exercée appartient à la section A de la NAF rév. 2, qui regroupe l'activité agricole, la sylviculture et l'activité de pêche (cf définition sur https://www.insee.fr/fr/metadonnees/definition/c1225) .

proches) qui mobilisent des facteurs de production communs, ces unités sont alors considérées comme une seule exploitation. Une autre particularité est la présence d'une minorité d'unités pouvant remplir les conditions d'une exploitation agricole sans posséder de Siren (par exemple certaines exploitations déclarant des surfaces à la PAC).

En résumé, le fait de savoir qu'une unité économique produit des produits agricoles ne garantit pas que celle-ci soit une exploitation agricole : elle doit vendre sa production (même si la vente est limitée) et avoir une certaine dimension économique. Ainsi une APE agricole dans les répertoires Sirene/Sirus ne saurait suffire, et oblige le SSP à obtenir des informations de nombreuses sources pour considérer l'unité économique en question comme une exploitation agricole, d'autant plus que les seuils peuvent changer au fil du temps. Par exemple, au Recensement Agricole 2010, il fallait 10 ruches en production pour être considéré comme une exploitation agricole. Ce seuil est passé à 50 lors du Recensement Agricole 2020.

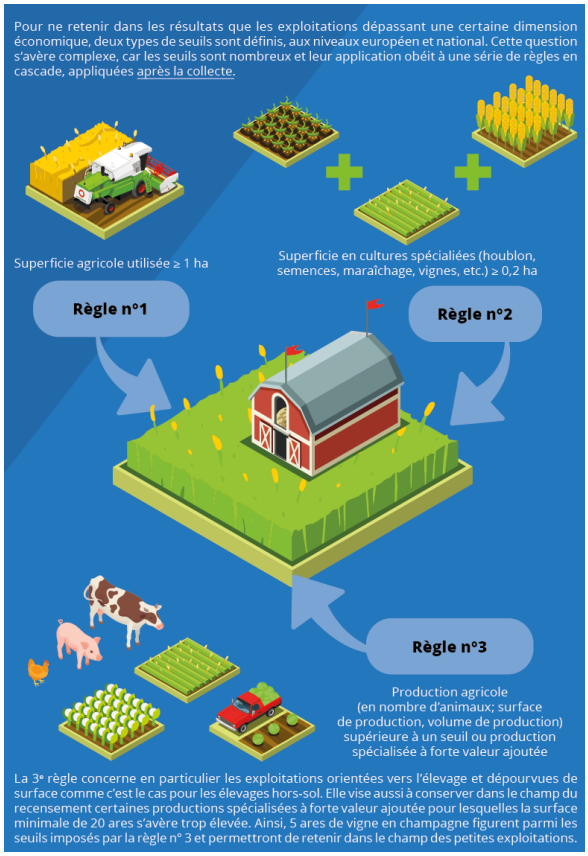

Figure 1: Les règles pour qu'une unité économique soit considérée comme une exploitation agricole

2. Les variables de la base de sondage

La base de sondage, dite BALSA et gérée par le Service de la statistique et de la prospective (SSP) au ministère de l'Agriculture et de l'Alimentation, contient plusieurs types de variables :

- d'identification : identifiant de l'exploitation agricole, raison sociale, Siret (si celui-ci existe et est connu), identifiants administratifs (numéro Pacage relatif aux dossiers de la PAC), numéro d'exploitation vitivinicole (EVV), numéro des Établissements de l'Élevage), adresse de l'établissement,
- de contacts : nom et prénom des répondants aux enquêtes, adresses de contact, téléphones et mails,

- des variables pouvant être des variables de stratification dans le cadre de l'élaboration des plans de sondage : surfaces des cultures, effectifs animaux, effectifs salariés et des variables composites comme, par exemple, la production brute standard (PBS) qui représente la valeur de la production potentielle hors tout aide (un proxy du « chiffre d'affaires » potentiel) et l'orientation technico-économique (OTEX) (calculée à partir de la PBS) qui se rapproche conceptuellement de la notion d'APE.

3. Pourquoi refondre la base de sondage ?

La version actuelle de la base de sondage, constituée après le recensement agricole de 2010, avait été conçue dans un premier temps pour être alimentée principalement par les données des enquêtes agricoles et par les liasses issues du répertoire statistique de Sirus. Du fait de leur volumétrie (plusieurs milliers de liasses quotidiennement) et des cas complexes pouvant apparaître, l'utilisation de ces avis n'a finalement jamais été implémentée.

La prise en compte de Sirus a aussi été abandonnée.

En effet, dans Sirus, une unité est mise en cessation statistique l'année N lorsque les trois critères ci-après sont réunis :

- l'unité n'a pas d'effectifs salariés deux années consécutives (en N et N-1),
- l'unité n'a pas de chiffre d'affaires trois années consécutives (en N, N-1 et N-2),
- l'unité n'a répondu à aucune enquête durant 3 années consécutives (en N, N-1 et N-2).

Ces critères posent un certain nombre de difficultés pour le champ agricole, par l'absence fréquente de chiffre d'affaires et le taux élevé d'exploitations sans salariés. En effet, les chiffres d'affaires sont obtenus par le traitement des liasses fiscales des bénéfices agricoles simplifiés et normaux transmises à l'Insee par la Direction Générale des Finances Publiques (DGFiP). Actuellement, aucun traitement n'est mis en œuvre pour extrapoler les données en cas d'absence de liasse fiscale pour le champ agricole, ce qui entraîne une absence de chiffre d'affaires pour de nombreuses exploitations. Au final, le taux de mise en cessation statistique par Sirus pour le champ agricole était au final très élevé (41 % en 2016) comparé au reste du champ Sirus (9 %). Aussi, à partir de la campagne 2016, les unités du secteur agricole ont été désormais exclues du champ de calcul de la cessation statistique de Sirus.

Pour éviter ce problème, l'univers a été établi à partir du répertoire des entreprises et des établissements Sirene, géré en continu à partir des créations, cessations et successions déclarées aux centres de formalités des entreprises.

Il a été complété avec les fichiers administratifs ayant un rapport avec l'activité agricole :

- des déclarations (surfaces de cultures et effectifs bovins, ovins et caprins) dans le cadre de la PAC,
- du casier viticole informatisé (CVI) sur les surfaces de vignes, des mouvements et des détenteurs d'animaux (bovins, ovins, caprins et porcins) provenant des bases de données de la Direction Générale de l'Alimentation (DGAL),
- de la mutualité sociale agricole (MSA) sur les salariés et les cotisants,
- des fichiers de l'AgenceBio et de l'institut national de l'origine et de la qualité (INAO),
- des fichiers de bénéfices et micro-bénéfices agricoles,
- des fichiers sur des orientations spécifiques tels l'horticulture, l'apiculture, les équidés,…

Les fichiers administratifs permettent de mettre à jour l'ensemble du champ des exploitations agricoles. Elles permettent d'enrichir la base avec des créations d'établissements, d'actualiser l'état économique de l'unité, les variables de contact et certaines variables de stratification (en fonction du fichier administratif).

La base de sondage est également actualisée avec les résultats des enquêtes thématiques réalisées par le SSP.

Ces enquêtes permettent d'actualiser l'état économique de l'unité, les variables de contact et les variables de stratification. Elles ont mis à jour sur la dernière décennie environ 15 % des unités de la base de sondage, avec une mise à jour plus importante des exploitations à fort poids économique, qui sont très souvent intégrées dans des strates exhaustives des enquêtes.

Pour mettre à jour la base et préparer au mieux le recensement agricole, des opérations qualités ont été menées. La stratégie retenue a consisté à ajouter par précaution le maximum d'unités qui pourraient faire partie du champ, puis de mener des opérations qualité ciblées pour écarter les unités hors champ[2].

Ces opérations ont été menées avec deux approches :

- élaborer avec les gestionnaires en région des règles de décision de masse, sur la base de tests sur des échantillons (par exemple, vérification des conditions d'être une exploitation agricole pour des unités présentes dans la base dont l'activité principale de l'établissement est « Location de terrains et d'autres biens immobiliers ») ;
- traitement individuel de listes par les gestionnaires, en fonction d'objectifs spécifiés a priori (par exemple, vérification de l'activité économique d'exploitations agricoles active économiquement, gestion de doublons identifiés, vérification des informations d exploitations agricoles, sirétisation manuelle d'unités...).

Tableau 1 : Evolution du nombre d'unités dans la base de sondage entre 2010 et 2020

		Nov2010		Nov2015	Nov2016	Nov2017	Nov2018	Nov2019	Nov2020
Etat économique	**Actif**	**518 024**		**513 316**	**542 652**	**547 462**	**532 629**	**510 381**	**508 128**
	Vacant			2 595	7 038	7 039	6 644	3 576	1 675
	Cessé			8 976	72 504	72 807	89 759	197 321	247 417
	Total	**518 024**		**524 887**	**622 194**	**627 308**	**629 032**	**711 278**	**757 220**
Dont Actifs Sans-Siret				40 941	41 914	44 741	43 352	20 930	39 367
Nombre Exploitations agricoles estimées		**491 400**		**442 100**	**437 400**	**431 700**	**426 000**	**420 300**	**414 600**
Taux couverture		**105,4%**		**116,1%**	**124,1%**	**126,8%**	**125,0%**	**121,4%**	**122,6%**

Le tableau 1 montre bien l'enrichissement important de la base depuis la fin de l'année 2018 et la mise en qualité qui a permis de considérer un fort taux d'exploitations inactives d'un point de vue économique.

Le processus étant faiblement automatisé/outillé, il générait une charge de travail conséquente pour l'administrateur de la base et les gestionnaires en région.

L'objectif de la refonte est donc de permettre des mises à jour automatisées à partir des données administratives, plus fréquemment, dans la mesure de ce que les sources permettent de faire (entre quotidienne et annuelle). Une mise à jour mensuelle paraît être un bon compromis entre la fraîcheur des données et la charge de traitement.

Ce nouveau dispositif permettra :

- d'améliorer la qualité des données, à la fois la qualité des échantillons d'enquêtes et donc des résultats statistiques, mais aussi les données de contact et de localisation des exploitations facilitant ainsi la préparation et la réalisation des enquêtes ;
- de diminuer la charge de réponse des répondants en exploitant directement les données administratives. Cette évolution s'inscrit dans le programme général dit « Dites le nous une fois » et répond également au 9e principe du Code des Bonnes Pratiques de la Statistique Européenne, élaboré par Eurostat ;

[2] Il a également été envisagé de réaliser une enquête par internet de l'ensemble des unités de la base possédant un courriel valide, mais cette option a finalement été écartée.

- de réduire la charge de travail de l'administrateur de la base qui intègre aujourd'hui ces données via des traitements manuels.

4. Comment refondre la base de sondage ?

4.1 Utilisation des résultats du recensement agricole 2020

Avant le recensement agricole de 2020, on considérait qu'une exploitation agricole était active si :

- elle était présente dans un fichier administratif « agricole » récent (dernier millésime disponible),
- elle disposait d'un Siret « actif » dans Sirene avec une activité principale exercée de l'établissement (APET) dit « agricole ».

Cependant, en fonction du fichier administratif, des unités ont été intégrées (ou conservées) dans la base de sondage même si elles ne respectaient pas ces règles (absence de Siret, APET « non agricole ») car on considérait que ces unités se trouvaient dans un fichier administratif « stratégique ». Les règles ont notamment fluctué au fil du temps pour éviter une sous-couverture (notamment dans la perspective du recensement agricole).

Les résultats du recensement agricole 2020 (RA 2020), dont les données seront définitives en avril 2022, vont permettre d'affiner ces hypothèses. Une analyse globale et source par source des unités considérées ou non comme une exploitation agricole est en cours.

À noter que comme la majorité des fichiers administratifs à usage non statistique, certaines données du RA 2020 sont à prendre avec prudence et nécessitent des expertises et traitements complémentaires. En effet, près de 15 000 exploitations ne déclarent tout de même pas de Siret et près de 5 000 des Siret déclarés sont fermés depuis plusieurs années dans Sirene. D'autres exploitations ne déclarent pas certains identifiants administratifs. Ces absences ou erreurs de déclarations ne facilitent pas l'appariement avec les sources administratives.

4.2 Le problème récurrent de l'appariement

4.2.1 Limite de l'appariement sur l'identifiant direct Siren

Actuellement, l'identifiant Siren ou Siret n'est pas utilisé comme identifiant principal par l'ensemble des administrations françaises ayant des contacts avec les exploitations agricoles. Ceci a pour conséquence de devoir gérer les problèmes de sous-couverture et de doublons. En effet, certaines unités dont le Siret n'est pas connu dans le fichier administratif peuvent ne pas être intégrées dans la base de sondage ou alors l'être mais sans être rattachées au Siret existant déjà dans la base, créant ainsi un doublon.

Les problèmes de sur-couverture existent aussi : certaines entreprises cessées d'un point de vue économique sont toujours présentes dans le répertoire Sirene ou dans les fichiers administratifs. C'est également le cas de certaines unités enregistrées dans Sirene qui n'ont même jamais démarré d'activité (et pour certaines ne la démarreront jamais).

Des travaux ont été menés pour estimer sur le champ agricole le nombre de ces unités actives dans Sirene mais ne remplissant pas les conditions pour être une exploitation agricole. Ce taux se situe au alentour de 35 % pour les unités dont l'APET est « agricole ». Un des déterminants de la qualité de la base de sondage est l'appariement correct des données.

Chaque fichier administratif a son propre identifiant administratif (par exemple le Pacage pour les fichiers PAC, le numéro MSA pour les fichiers des cotisants de la MSA ou encore le numéro d'exploitation vitivinicole dans le Casier Viticole). Le Siret y figure mais n'est jamais complètement renseigné (avec des taux variant de 25 % pour les fichiers sur les équidés à 98 % dans les fichiers de la

PAC, hors fichier Sirene). Il peut être aussi renseigné mais avec des erreurs manifestes (comme un nombre inexact de chiffres ou clé de contrôle incorrecte).

La base de sondage contient son propre identifiant (numéro à 7 chiffres). Cependant, l'appariement se fait, dans un premier temps, sur le seul identifiant commun à chaque fichier administratif : le Siret.

Avant tout appariement, la première approche est de qualifier la validité du Siret déclaré dans le fichier administratif. On vérifie si le Siret déclaré est actif ou fermé dans le répertoire Sirene.

S'il est actif, on s'assure que dans les cas d'exploitations agricoles possédant plusieurs Siret « actifs » mais ne correspondant en réalité qu'à une seule exploitation agricole (les moyens de productions appartiennent à la même entité juridique), on puisse apparier avec le Siret de notre base de sondage (afin de ne pas créer de « doublons »).

Pour les Siret déclarés et fermés dans le répertoire Sirene, il est nécessaire de rechercher le Siret « réel ». Pour cela, on utilise la base Sirene (le fichier Stock pour rechercher un autre établissement de l'entreprise et localisé sur la même commune et le fichier des liens de succession des établissements (prédécesseurs et successeurs))

Dans un second temps, pour les unités pour lesquelles on n'a pas de Siret déclaré ou alors un Siret fermé ou invalide, on met en place des techniques d'appariement sur identifiants indirects. Le processus utilisé actuellement se base sur un processus séquentiel de relâchement successif de clés de moins en moins robustes, décrit plus en détail ci-après.

4.2.2 Normalisation des données identiques pour la base de sondage et le fichier administratif

Les informations utilisées pour l'appariement et donc présentes conjointement dans le fichier administratif et le fichier Stock Sirene sont le nom et le prénom (pour les personnes physiques), la raison sociale (pour les personnes morales), l'adresse de l'établissement et le code Insee de la commune. En fonction des fichiers administratifs, on dispose dans certains cas du nom de naissance.

La normalisation commence par prendre en compte les caractéristiques du fichier Sirene. Dans ce fichier, l'adresse est présente sous plusieurs variables qui sont le numéro de voie, le type de voie, le libellé de voie et l'indice de répétitivité. L'adresse de chaque fichier administratif est donc découpée pour obtenir les informations indiquées ci-dessus.

Les données des fichiers administratifs sont hétérogènes : certaines sont mal ordonnées, d'autres mal orthographiées, d'autres en majuscule ou minuscule.

Il est donc nécessaire de nettoyer et normaliser ces données. Que ce soit pour les noms, les prénoms et les adresses, les caractères spéciaux, les accents, les chiffres (sauf pour le numéro de voie) sont exclus.

Tableau 2 : exemple de normalisation d'adresses

adresse	numeroVoieEtablissement	typeVoieEtablissement	libelleVoieEtablissement
35 HAMEAU DE RIBEAUVILLE	35	HAM	RIBEAUVILLE
1 RUE DE CHARLES BRUYERES	1	RUE	CHARLES BRUYERES
3 ROUTE DE WALLERS	3	RTE	WALLERS
11 CARRIERE ETREUX EN BAS	11		CARRIERE ETREUX BAS
11 LA HAIE LONGPRE	11		HAIE LONGPRE

Une autre méthode sur l'adresse en ne retenant qu'un mot directeur de l'adresse a aussi été testé. Ce mot directeur est le mot le plus long du libellé de voie ou le dernier mot du libellé de voie. Elle permet d'améliorer légèrement le taux d'appariement.

4.2.3 Processus séquentiel de relâchement successif des clés les moins robustes

Le processus est basé sur une démarche séquentielle de relâchement successif de clés de moins en moins robustes :

- Adresse complète (numéro de voie, type de voie, libellé de voie) + code commune,
- Nom complet (les prénoms et nom de famille) + code commune,
- Adresse complète sauf le numéro de voie + code commune,
- Premier prénom et le nom de famille + code commune,
- Premier prénom et le nom de famille inversé + code commune,
- Adresse sauf le type de voie + code commune.

Par ailleurs, pour réduire les temps de calculs et le nombre de Siret appariés sur le même identifiant, la comparaison avec le fichier Sirene s'est réalisé tout d'abord avec les unités avec une APET agricole (environ un million d'unités actives et cessées) et ensuite une APET non agricole (environ 30 millions d'unités actives et cessées).

À chaque étape, pour chaque unité, lorsqu'un seul Siret est apparié, on considère que ce Siret est exact et est donc validé. Si plusieurs Siret sont renvoyés, on met en concordance le statut administratif de l'établissement avec la date de validité du fichier administratif pour permettre d'éliminer certains Siret (ceux fermés) afin de se retrouver dans le cas précédent et pouvoir valider le Siret.

4.2.4 Autres méthodes d'appariement automatiques testées ou envisagées

L'approche actuellement utilisée est prudente : on privilégie le fait d'apparier sur le bon Siret plutôt que d'apparier à tort. Des méthodes comme l'utilisation de distances (Jaro Winkler) ont été testées mais n'ont pas pour le moment donné de résultats totalement satisfaisants (ces méthodes méritent cependant d'être approfondies).

Le service d'identification automatique de masse (SIAM) de l'Insee a également été utilisé par le passé puis abandonné. Le SIAM de Sirene 3 permet à partir d'informations relatives à l'identification comme la désignation et l'adresse d'implantation, de retrouver l'unité légale ou l'établissement recherché en s'appuyant notamment sur un système de scoring. Dans le cadre du projet de refonte de l'outil SIAM associée au programme Sirene 4, le ministère de l'Agriculture et de l'Alimentation a participé aux tests utilisateurs pour évaluer les premiers résultats de ce nouveau moteur d'identification et proposer des pistes d'amélioration. Les résultats montrent pour les tests sur les exploitations agricoles un taux d'appariement supérieur au SIAM de Sirene 3.

Sur un échantillon de près de 3 000 unités dont le Siret était absent de la base de sondage mais qui a été déclaré par l'exploitant agricole au Recensement Agricole, le SIAM de Sirene 4 a identifié le Siret exact dans 78 % des cas (contre 41 % pour le SIAM de Sirene 3).

Le principal inconvénient reste la sélection d'unités à partir des nombreux échos SIAM, même si le SIAM de Sirene 4 réduit aussi le nombre d'échos retenus (7 000 échos pour les 3 000 unités de l'échantillon, contre 25 300 échos pour Sirene 3). Pour cela, des travaux vont être lancés lorsque le service SIAM de Sirene 4 sera mis en production pour trouver un algorithme capable d'optimiser la validation du bon « écho » (et donc trouver le Siret exact) en prenant en compte l'information contenue dans les différents fichiers administratifs.

Des projets d'appariement de données via d'autres outils (notamment matchID[3], Rapsodie[4] et le moteur d'appariement d'InserJeunes[5]) sont aussi à l'étude.

La réflexion au sein du ministère est d'intégrer un outil de sirétisation « automatique » au sein de la base de sondage. Cette approche ne nécessite pas d'intervention humaine lorsque l'outil renvoie un Siret avec une forte probabilité d'existence (seulement fixer un seuil acceptable).

4.2.5 Validation et expertise par des gestionnaires

Une expertise humaine est nécessaire pour déterminer de la qualité ou non de l'appariement. Pour l'instant, le principal problème reste la volumétrie des validations par les gestionnaires qui est assez importante au regard de la capacité en région pour ces expertises.

Par exemple, le fichier annuel de la PAC, sirétisé à 98 %, implique l'expertise humaine d'environ 5 000 à 8 000 unités. Le fichier annuel des micro-bénéfices agricoles, sirétisé à 75 %, nécessite l'expertise d'environ 30 000 unités.

Ainsi, par le passé, l'expertise manuelle n'a pas pu être réalisée sur tous les fichiers : le choix avait été fait de privilégier des fichiers « stratégiques » et/ou en appliquant des seuils (en priorisant par exemple les unités avec un poids économique important).

Pour l'expertise manuelle, afin de réduire la volumétrie, des réflexions sont en cours pour utiliser les résultats non pas d'un seul fichier mais de plusieurs fichiers administratifs. Le principe est de considérer qu'une unité présente dans plusieurs fichiers administratifs a de meilleures chances de relever du champ des exploitations agricoles qu'une unité présente dans un seul fichier.

4.2.6 Autres variables d'appariement

La méthode d'appariement a aussi été appliquée entre les fichiers administratifs et la base de sondage, car les adresses présentes dans la base peuvent différer de celles présentes dans Sirene. En effet, un exploitant agricole renseigne parfois une adresse de résidence et non l'adresse de l'établissement pour les démarches administratives et l'on retrouve ainsi le lieu de résidence dans de nombreux fichiers administratifs (fichiers de la PAC et de la MSA par exemple) car l'exploitant reçoit son courrier à cette adresse. Cette approche par plusieurs fichiers permet d'augmenter les données appariées.

4.3 Les sources

Un important travail d'ajout de nouvelles sources a été réalisé en 2019 et 2020. Pour que la nouvelle base de sondage soit « optimale », il est nécessaire de hiérarchiser les diverses sources de mises à jour et d'établir un arbre décisionnel dans la mise à jour de chacune des variables de la base.

Les variables présentes dans plusieurs fichiers administratifs sont :

- les données de contact (adresse, courriels, téléphones),
- la surface agricole utile,
- certaines surfaces de culture (par exemple viticoles entre les déclarations de la PAC et celles issus du Casier Viticole),

[3] MatchID est un outil d'appariement, qui a été développé au ministère de l'Intérieur dans le contexte des challenges d'Entrepreneur d'intérêt général, pour identifier et retirer les personnes décédées du registre du permis de conduire.

[4] RAPSODIE (RAPprochement des données SOciales, Des Impôts et des Enquêtes) est une application développée par le Pôle Revenus Fiscaux et Sociaux (RFS) de l'Insee.

[5] Le moteur d'appariement d'InserJeunes, qui apparie des sources administratives, principalement sur identifiants indirects, a été développé par le service statistique du ministère de l'Éducation Nationale.

- certains effectifs animaux (par exemple les déclarations issues des aides bovines, caprins ou ovins de la PAC et celles issus des fichiers de la base de données nationale d'identification (BDNI) sur les bovins et du recensement ovins-caprins).

Cela entraîne des divergences pour lesquelles il faut arbitrer. Principalement deux approches sont possibles : différence de qualité entre les sources et fraîcheur de l'information.

Il est nécessaire pour cela de disposer le plus rapidement possible (sans excès, par exemple si l'information infra-journalière était disponible, elle n'aurait pas d'intérêt pour BALSA) des fichiers administratifs « externes » afin de pouvoir mettre en place les règles de décision en fonction de la date effective de l'information et de mettre à jour la base en « temps réel ». Il est aussi essentiel de travailler avec chaque fournisseur de données pour améliorer la qualité des données (notamment en les sensibilisant à l'utilisation de Siret validés). Des échanges ont lieu avec certains producteurs de données pour prendre en compte dans les fichiers producteurs la sirétisation réalisée par le ministère. Ces échanges portent aussi sur des comparaisons de champ : par exemple l'absence (respectivement la présence) de déclaration d'être engagé dans une démarche de qualité et/ou environnementale et la présence (respectivement l'absence) de cette unité dans le fichier producteur (par exemple ceux de l'INAO et de l'AgenceBio).

D'autres sources, comme les fichiers de TVA, les décès des personnes physiques ou encore les liquidations judiciaires issues du bulletin officiel des annonces civiles et commerciales (Bodacc) ont été explorées mais n'ont pas encore été intégrées à la base de sondage. Des expertises supplémentaires doivent être réalisées pour qualifier les données et définir avec les résultats du recensement agricole les règles d'intégration éventuelle de ces données.

La refonte doit permettre d'aboutir à un système évolutif pour faciliter la prise en compte des évolutions des sources administratives (notamment favoriser le recours aux API pour une intégration plus facile des données administratives) et l'intégration de nouvelles sources et règles. Certaines règles notamment sur la mise à jour de certaines variables de superficies avec les fichiers de la PAC ou les fichiers viticoles ont déjà été arrêtées.

4.4 Vers des règles de décision définies pour chaque source ?

À la lumière de l'expérience acquise au cours de la mise à jour de Balsa durant la décennie précédente et notamment de la forte montée en charge du nombre de sources utilisées et de leur accès pérenne (pour l'instant), il est manifeste que la qualité globale de la base dépend très fortement de la qualité de l'appariement ainsi que des règles de décision sur la caractérisation d'une unité comme étant une exploitation agricole dans chaque source.

Actuellement, l'intégration d'une unité dans la base de sondage reprend les règles établies (cf paragraphe 4.1). Une réflexion doit être menée sur le fait de s'appuyer ou non sur des modèles qualifiant ou non une unité d'être une exploitation agricole (et ainsi se poser la question de cesser des unités dont la probabilité d'être une exploitation agricole est inférieure à un seuil). Cette réflexion est un véritable changement de culture au sein du ministère où les unités ont longtemps été conservées dans les bases de sondages même si on considérait que la probabilité de cessation était forte (par peur de ne pas collecter les données d'une unité lors d'une enquête ou lors du recensement).

Des méthodes d'approche supervisée (régression logistique, forêts aléatoires…) doivent permettre de déterminer ces règles.

Décrivons un exemple d'approche supervisée testé : pour le fichier de la MSA, on peut mettre en place un modèle économétrique expliquant le fait d'être dans le champ du recensement agricole 2020 (unités enquêtées et répondantes) par certaines caractéristiques dans le fichier administratif.

Tableau 3 : variables utilisées par le modèle économétrique

Variable	Libellé	Valeurs prises
Unité retenue dans le champ du RA	CHAMPRA	1 si dans le champ
Nombre d'assurés	Nb_assu	1 à 6
Région du siège	region	13 régions
Champ de l'activité	naf_champ	4 modalités selon le champ SSP
Catégorie juridique	cat_jur	9 catégories
Superficie	sup	Superficie réelle en ares
Cotisant solidaire	cot_sol	Vrai si au moins un assuré est cotisant solidaire
Catégorie de risque Atexa (selon l'orientation technico-économique)	cris_ate	26 groupes

Le modèle logistique utilisé est le suivant :

CHAMPRA ~ Nb_assu + sup + region + naf_champ + cat_jur + cot_sol + cris_ate

Il comprend des variables sur la localisation de l'exploitation, la catégorie juridique de l'exploitation, le champ de l'activité selon la NAF, le nombre d'assurés, la notion de cotisant solidaire (pour cibler notamment les petites exploitations sous le seuil du recensement agricole), la superficie et la catégorie de risque Atexa qui est un « proxy » de l'orientation technico-économique (grandes cultures, viticulture, maraîchage, élevage bovins lait, élevage bovins viande, élevage porcins…).

Les estimateurs sont majoritairement significatifs au seuil de 0,1 %. Les variables sélectionnées expliquent donc effectivement, au moins en partie, l'indicatrice d'appartenance au champ du recensement.

Le tableau suivant regroupe l'ensemble des estimateurs du modèle.

Tableau 4 – Résultats de la régression logistique

Variable	Estimateur	Std. Error
Intercept	1.35***	(0.0517)
Nombre de chefs d'exploitation	0.06***	(0.0173)
Superficie de l'exploitation	0.00***	(0.0000)
Région du siège		
11 - Île-de-France	Réf.	
24 - Centre-Val de Loire	0.57***	(0.0506)
27 - Bourgogne-Franche-Comté	0.46***	(0.0498)
28 - Normandie	0.61***	(0.0491)
32 - Hauts-de-France	0.54***	(0.0429)
44 - Grand Est	0.75***	(0.0479)
52 - Pays de la Loire	0.66***	(0.0496)
53 - Bretagne	0.64***	(0.0500)
75 - Nouvelle Aquitaine	0.62***	(0.0459)
76 - Occitanie	0.47***	(0.0457)
84 - Auvergne-Rhône-Alpes	0.56***	(0.0561)
93 - Provence-Alpes-Côte d'Azur	0.33***	(0.0484)
94 - Corse	0.45***	(0.0709)
Présence dans le champ SSP		
Champ complet	Réf.	
Champ partiel	-2.98***	(0.0257)
Champ très partiel	-4.13***	(0.0143)
Hors Champ SSP	-3.44***	(0.0319)
Catégorie juridique		
1-Exploitant individuel	Réf.	
2-GAEC	0.83***	(0.0378)
3-EARL	0.63***	(0.0217)
4-SCEA	0.34***	(0.0331)
5-GFA	-0.49**	(0.0172)
6-SA/SARL	-0.66**	(0.0348)
7-Société de fait	-0.11***	(0.0776)
8-Autre société	-0.39***	(0.0747)
9-Pluralité d'exploitation	-1.02***	(0.0200)
Cotisant solidaire		
OUI	Réf.	
NON	-1.02***	(0.0147)
Catégorie de risque Atexa		
00-Non concerné	Réf.	
01-Maraîchage, floriculture	-0.44***	(0.0303)
02-Arboriculture fruitière	0.12**	(0.0384)
03-Pépinière	-0.40***	(0.0570)
04-Grandes cultures	0.40***	(0.0259)
05-Viticulture	0.27***	(0.0253)
06-Sylviculture	-1.41***	(0.0179)
07-Autres cultures spécialisées	-0.72***	(0.0418)
08-Élevage bovins-lait	0.35***	(0.0323)
09-Élevage bovins-viande	0.49***	(0.0293)
10-Élevage bovins mixte	0.37***	(0.0490)
11-Élevage ovins, caprins	0.30***	(0.0364)
12-Élevage porcin	0.13*	(0.0561)
13-Élevage de chevaux	-0.96***	(0.0356)
14-Autres élevages de gros animaux	-0.48***	(0.0918)
15-Élevage de volailles, lapins	0.19	(0.0383)
16-Autres élevages de petits animaux	0.81***	(0.0381)
17-Entraînement, dressage, haras, clubs hippiques	-0.41***	(0.0442)
18-Conchyliculture	-2.58***	(0.0198)
19- Poly-culture /élevage	0.15***	(0.0269)
20-Marais salants	-2.71***	(0.0274)
21-Exploitation de bois	-2.10***	(0.0925)
22-Scieries fixes	-1.62***	(0.0294)
23-Travaux agricoles	-0.09***	(0.0456)
24-Jardins-Paysagistes-Reboisement	-2.47***	(0.0556)
25-Mandataires-Mutuelles agricoles	-2.56***	(0.0506)

Significativité : *** significatif au seuil de 0.1% | ** significatif au seuil de 1%
* significatif au seuil de 5%

Les estimateurs pour chaque région sont significatifs. Les indicateurs sont assez semblables à ceux attendus. Toutes choses égales par ailleurs, être un GAEC, une EARL ou un GFA, plutôt qu'une société, apporte une probabilité plus importante d'être une exploitation agricole. C'est aussi le cas pour les non-cotisants solidaires qui ont une probabilité d'être dans le champ beaucoup plus importante que les cotisants solidaires (d'ailleurs la question de ne garder dans la base que les cotisants solidaires du fichier MSA avait déjà été envisagée avant le recensement agricole de 2020). En ce qui concerne la variable de catégorie de risque Atexa, les principales activités agricoles ressortent avec des estimateurs positifs (grandes cultures, viticulture, bovins, ovins, caprins, porcins, …). En revanche, cela n'était pas spécialement attendu pour l'élevage des autres petits animaux. En effet, une des raisons d'être hors champ est souvent le fait d'être présent dans la base de sondage mais réaliser de l'élevage de chiens et chats (considéré comme une activité non agricole).

Pour évaluer la qualité du modèle, il est possible de confronter la probabilité estimée P^(Y = 1) et la valeur effectivement prise par Y.

En posant la probabilité 0,5 comme seuil de décision, il est possible de définir :

$$\hat{Y} = 1 \, si \, \hat{P}(Y = 1) \geq 0{,}5$$

Effectif	***Y=1***	***Y=0***
Ŷ = 1	*306 785*	*2 197*
Ŷ = 0	*15 168*	*3 306*

Les unités classées hors champ au recensement agricole (celles qui nous intéressent) sont plus souvent mal classées que les unités dans le champ. Une grande partie des unités hors champ ont des probabilités d'être dans le champ supérieures à 0,5. La plupart ont des probabilités estimées généralement supérieures à 0,8, changer le seuil de décision n'est donc pas une solution.

Une autre approche supervisée testée est de mettre en place un modèle économétrique expliquant le fait d'être dans le champ du recensement agricole 2020 (unités enquêtées et répondantes) par la présence ou non dans un fichier administratif.

Tableau 4 : variables utilisées par le modèle économétrique « multi-sources »

Variable	Libellé	Valeurs prises
Unité retenue dans le champ du RA	CHAMPRA	1 si dans le cham
Présence dans le fichier de déclarations de surfaces de la PAC2020	PAC20	1 si présent
Présence dans le fichier de déclarations des aides bovines, ovins et caprins de la PAC2020	PAC_CHEPTEL_2020	1 si présent
Présence dans le fichier de déclarations de surfaces du CVI2020	CVI_2020	1 si présent
Présence dans le fichier de déclarations de récolte viticole	RECOLTE20	1 si présent
Présence dans le fichier des cotisants MSA en 2020	MSA_COT20	1 si présent
Présence dans le fichier des salariés MSA en 2020	MSA_SAL20	1 si présent
Présence dans le recensement ovins-caprins 2020	RECENS_OVCAP20	1 si présent
Présence dans le fichier bovins de la BDNI en 2020	BDNI2020	1 si présent
Présence dans le fichier de l'AgenceBio en 2020	AGENCEBIO_2020	1 si présent
Présence dans le fichier Resytal de 2019	RESYTAL	1 si présent
Présence dans le fichier Ruchers 2020	RUCHER2020	1 si présent
Présence dans les bénéfices-agricoles 2019	BA2019	1 si présent
Présence dans les bénéfices-agricoles 2020	MICROBA2018	1 si présent
Présnce dans les fichiers de déclaration de TVA en 2018, 2019 ou début 2020	TVA_DECLA	1 si présent
Présence dans le fichier Horticulture de FranceAgrimer	HORTICULTURE	1 si présent

Le modèle logistique utilisé est le suivant :

CHAMPRA~PAC20+PAC_CHEPTEL_2020+CVI_2020+RECOLTE20+MSA_COT20+MSA_SAL20+RECENS_OVCAP20+BDNI2020+AGENCEBIO_2020+RESYTAL+RUCHER2020+BA2019+MICROBA2018+TVA_DECLA+HORTICULTURE

En posant la probabilité 0,5 comme seuil de décision, on obtient :

Effectif	***Y=1***	***Y=0***
Ŷ = 1	*366 207*	*16 205*
Ŷ = 0	*17 216*	*38 006*

À la vue des résultats, l'idée serait plutôt d'utiliser l'approche utilisée avec le fichier de la MSA et d'élargir cette méthode aux différents fichiers disponibles. Le but est ensuite de calculer un indicateur synthétique à partir des différentes probabilités obtenus des différents fichiers. Se pose aussi la question de la pondération de ces probabilités en fonction notamment de la fiabilité des fichiers.

Les modalités de déclaration « hors-champ » au recensement doivent être analysées plus finement aussi. Les situations (cessations, auto-consommation, doublons, autres raisons,…) sont diverses et ne sont pas toujours homogènes, car le recensement est déclaratif.

4.5 Prolongements et points à approfondir

Trois grands axes peuvent être dégagés quant aux prolongements qui pourraient être donnés aux développements présentés ci-dessus :

- affiner les travaux sur les méthodes, les variables et les sources pour définir des règles permettant une maximisation de la validation automatique,
- adopter une approche globale de la validation « manuelle », prenant en compte des éléments de coût et de disponibilité des ressources humaines,
- mesurer la qualité de la base (qualité des sources utilisées, du processus et des résultats).

5. Quelques objectifs attendus

5.1 Fréquence des mises à jour

La fréquence des mises à jour dépend bien entendu de la disponibilité des sources mais aussi de l'intérêt pour le répertoire de disposer d'une information la plus récente possible. Certaines données ne sont produites qu'annuellement, à une date précise. Pour les sources disposant d'une mise à jour en continu, il est notamment prévu de réaliser des mises à jour « en temps » réel, ce qui est facilité par la mise à disposition croissante d'API.

5.2 Historisation des données et retour arrière

L'historisation des données est un enjeu majeur. En effet, actuellement les données ne sont pas historisées dans la base. On peut comparer des données à deux dates données en fonction des sauvegardes (manuelles) qui ont été réalisées. Cependant, l'historisation des variables n'est pas réalisée. Il n'est donc pas simple de connaître la source qui met à jour chaque variable.

Une historisation précise des mises à jour permettrait d'éviter de perdre une information antérieure de meilleure qualité et de réaliser des « retour-arrière » (les erreurs d'appariement pourraient ainsi être rectifiés ou un changement des règles de décision serait plus facile à réaliser).

5.3 La mesure de la qualité

Un cadre qualité de la base de sondage devra être défini et couvrir trois domaines :

- la qualité des sources utilisées (inputs),
- la qualité du processus,
- la qualité de la base de sondage (outputs).

La qualité des données en entrée concerne la possibilité d'utiliser une source administrative reçue et les traitements nécessaires à sa prise en compte dans le système de production statistique. Il est prévu de mettre en place un processus automatique de contrôle des données en entrée. Un outil comme le module ARC (acquisition, réception, contrôle) est en cours d'étude. En effet, ARC pourrait être adapté à la refonte et permettre de décrire le modèle de données cible et de définir les règles indiquant comment les données collectées vont rejoindre ce modèle.

Pour la qualité du processus, il n'existe pour l'instant aucune démarche formalisée.

Des informations seront disponibles sur la qualité des éléments contenus dans la base : nombre de présence et ancienneté dans chaque fichier administratif (information qui vient s'ajouter aux informations de Sirene).

Des indicateurs de qualité par la confrontation entre les données des fichiers administratifs et celles disponibles dans la base de sondage seront construits. Des indicateurs suite aux contrôles qualité opérées par des gestionnaires (par exemple, validation d'une proposition de Siret ou de retenir ou non une exploitation dans le champ) seront aussi définies.

Toutes ces informations seront utilisées dans une approche globale destinée à améliorer la qualité des processus de production statistique.

Cette approche globale de la qualité concernera aussi la question des arbitrages à opérer entre moyens « humains » consacrés aux opérations statistiques stricto sensu, et moyens « humains » (temps consacré à des gestionnaires pour la mise en qualité) consacrés à l'infrastructure que représente notamment la base de sondage.

La mise à jour de la base de sondage doit s'accompagner d'une mesure globale de la qualité des données. En particulier, on ne doit pas constater des hausses inexpliquées d'exploitations agricoles alors que la tendance est à la baisse (moins d'exploitations depuis plusieurs décennies mais avec des surfaces ou effectifs animaux plus importants).

L'utilisation des sources devrait permettre par comparaison entre les données de la base et celles du recensement agricole de 2020 corrigé de l'évolution annuelle du nombre de chefs d'exploitations agricoles (diffusé par la MSA) d'obtenir une estimation du nombre d'exploitations agricoles et en déduire une mesure de la couverture de la base de sondage. En ce qui concerne les données des variables utiles à la stratification, l'utilisation des données de la Statistique Annuelle Agricole pourrait permettre d'avoir une estimation de la qualité de chacune des variables. Les enquêtes du ministère (notamment pour les cheptels porcins, ovins et caprins) permettront aussi d'améliorer la base et notamment par une comparaison de la valeur des variables de la base avant et après enquête.

5.4 Vers une démographie d'entreprises ?

Un objectif complémentaire du projet est de réaliser à partir de cette nouvelle base, de la démographie d'entreprises sur le champ des exploitations agricoles.

Une démarche exploratoire a été réalisée sur entre les recensements de 2010 et de 2020. La fiabilité des recensements agricoles fait de celui de 2010 un bon point de départ et de celui de 2020 un bon

point d'arrivée pour permettre un bilan. Plusieurs approches ont été testées. L'une d'elle est de partir du stock d'une année N, d'enlever les unités qui ont fermé dans Sirene cette année N et d'ajouter dans Sirene en amont les « vraies » créations du domaine agricole (créations donnant suite à une activité économique réelle c'est-à-dire se retrouvant dans un fichier administratif agricole) et de retirer en aval ce qui semble être des « faux-actifs » (sortants d'un fichier administratif). Pour cela, le choix s'est porté sur le fichier MSA des cotisants, car ce fichier a un champ assez proche de celui du recensement agricole et qu'il est disponible chaque année depuis 2010.

La formule de récurrence est la suivante :

$$\text{stockCorrige}_{N+1} = \text{stock}_{N+1} - \text{sortantsMSA}_{N+1}$$

$$\text{stock}_{N+1} = \text{stockCorrige}_{N} + \text{creationsFA}_{N+1} - \text{fermetures}_{N+1}$$

Où : sortantsMSA_{N+1} désigne les siret présents dans le fichier MSA de l'année N mais absents de celui de l'année N+1 et creationsFA_{N+1} désigne les siret créés avec une apet agricole au sens du SSP dont le siren apparaît au moins une fois dans un fichier administratif sur la période

Cette méthode donne des résultats contrastés. Si elle permet d'être cohérent avec la réalité économique que représente la baisse continue des exploitations agricoles, une partie non négligeable des Siret récupérés en 2020 ne se retrouvent pas dans le recensement agricole. Le suivi des Siret au fil du temps dans les fichiers administratifs est ainsi compliqué. Cette méthode a aussi été réalisée avec d'autres sources administratifs que celles de la MSA mais le problème est que certains fichiers ne sont disponibles que depuis quelques années seulement, d'autres étaient d'une qualité « médiocre » en termes de fiabilité du Siret.

6. Conclusions et perspectives

En conclusion, la base de sondage va être mise à jour plus efficacement et plus régulièrement avec les différents fichiers administratifs à disposition. De plus, l'intégration de nouvelles unités sans réalité économique (inscription dans un fichier sans activité réelle par la suite) devrait être plus faible que par le passé (grâce aux travaux réalisés sur les résultats du recensement agricole et notamment le repérage des unités qui ont été déclarées hors-champ). D'autre part, les travaux pour trouver dans les sources administratives des indices suggérant l'arrêt de l'activité de l'exploitation doivent permettre de réduire le maintien des « faux-actifs » dans la base. Les fichiers administratifs et leur expertise sont donc la clef pour répondre aux problèmes décrits (même s'il peut être difficile de les manier dans la mesure où ils peuvent être parfois être contradictoires) et réaliser à court terme une approche rigoureuse de la démographie d'exploitations d'agricoles et d'entreprise agricoles au sens du ministère de l'Agriculture et de l'Alimentation qui soit comparable aux travaux de l'INSEE sur le champ de l'industrie, du commerce et des services.

Bibliographie

[1] Le Grand H., « Le recensement agricole de 2020, cinq innovations qui feront date », Courrier des Statistiques, n° 7, Insee, janvier 2022

[2] Brion P., « La mise à jour de répertoires d'entreprises », Journées de méthodologique de l'Insee 2015

[3] Tessier L., « Démographie d'entreprises sur le champ agricole : étude de faisabilité », Mémoire de Master Évaluation et Décision Publiques Études Statistiques, 2021

[4] Midy L., « Un outil d'appariement sur identifiants indirects : l'exemple du système d'information sur l'insertion des jeunes », Courrier des Statistiques, n°6, Insee, juin 2021