IMPORTANCE DE L'EVALUATION QUANTITATIVE DE LA QUALITE DES DONNEES D'UN DISPOSITIF DE SURVEILLANCE : EXEMPLE DU PROGRAMME RESABEILLE *

Pauline Quéré[1], **Anne Bronner**[2], **Fayçal Meziani**[3] et **Pascal Hendrikx**[1]

RESUME

Le programme pilote d'épidémiosurveillance apicole Résabeille, composante française du dispositif européen Epilobee, a été mis en place au cours de deux campagnes consécutives (2012-2013 et 2013-2014) dans six départements. Cette surveillance programmée devait permettre d'estimer la mortalité hivernale et en saison ainsi que d'estimer la prévalence de certaines maladies infectieuses et parasitaires de l'abeille.

L'objectif de notre étude était d'évaluer de manière quantitative la qualité des données collectées dans le cadre de ce programme de surveillance en France, et d'identifier les difficultés qui avaient pu survenir au moment de la collecte et de la centralisation des données.

Plusieurs dimensions de la qualité des données (la complétude, le format, la plausibilité et la cohérence) ont pu être analysées de manière quantitative. En complément, afin d'évaluer la qualité de renseignement des fiches de visite ainsi que de leur saisie, 60 fiches de visite de printemps 2014 ont été tirées au sort et ont fait l'objet d'une analyse approfondie et d'une double saisie.

Dans le cadre de l'évaluation des données relatives à la mortalité hivernale, la plausibilité variait entre 98 et 100 % selon la variable étudiée, et la cohérence était de 82 %. A partir d'un échantillon aléatoire de 60 fiches de visite, entre 77 et 97 % des données étaient renseignées. La saisie était possible pour toutes les données à renseigner sur la fiche de visite, mais entre 12 et 27 % de ces données nécessitaient une interprétation avant la saisie (en recherchant la valeur correcte par croisement avec d'autres informations de la fiche de visite). En comparant la saisie de ces fiches réalisée par les acteurs dans les départements avec notre double saisie, nous avons estimé qu'entre 36 et 73 % des données mal renseignées sur les fiches de visites (et dont l'information correcte pouvait être retrouvée en croisant avec d'autres informations de la fiche de visite) avaient été mal interprétées au moment de la saisie, et qu'entre 0 et 2 % des données correctement renseignées sur les fiches de visites étaient mal saisies.

Compte tenu des résultats de l'évaluation, des axes d'amélioration ont été proposés pour les campagnes à venir. Ils relèvent de l'évolution du questionnaire de visite, du renforcement de la formation, de l'assistance et de la coordination des acteurs ainsi que de la mise en place d'une procédure de vérification des données avant leur saisie dans la base.

Mots-clés : abeilles, mortalité, surveillance, qualité des données, Epilobee, Résabeille.

…/..

---

* Texte de la communication orale présentée au cours de la Journée scientifique AEEMA, 20 mars 2015

[1] Anses, Direction des laboratoires, Unité de coordination et d'appui à la surveillance, Lyon, France

[2] Anses, Laboratoire de Lyon, Unité épidémiologie, Lyon, France

[3] Direction générale de l'alimentation, Sous-direction de la santé et de la protection animale, de la qualité et de la protection des végétaux, Paris, France

…/..

ABSTRACT

The "Résabeille" pilot epidemiological surveillance program is the French component of the European Epilobee surveillance system. It was implemented during two consecutive seasons (2012-2013 and 2013-2014) in six areas (French "départements") in the country. The objectives of this active surveillance system were to estimate the winter and in-season mortality as well as the prevalence of certain infectious and parasitic diseases of honey bees.

Our study was specifically designed to evaluate quantitatively the quality of the data collected under the French surveillance program and to identify the difficulties experienced in the collection and compilation of data. Several dimensions of data quality (completeness, format, plausibility and consistency) were analyzed quantitatively. Furthermore, in order to assess the quality of entries into the visit forms and the quality of their recording into the database, 60 Spring 2014 visits forms were thoroughly analyzed and double entered into the database.

In the evaluation of winter mortality data, plausibility varied between 98 and 100% depending on the variable considered and the consistency was 82%. In a random sample of 60 visit forms, between 77 and 97% of the expected data were recorded. The recording was feasible for all the items to be filled in the visit form, but between 12 and 27% of the data required an interpretation prior to the recoding (such as finding the correct value by crossing with other data in the visit form). Comparing the entry of data into the forms made by actors in the field with our double entry, led us to estimate that between 36 and 73% of the data in the visit forms were improperly recorded at the time of data entry (mostly in cases where the estimation of the correct value required crossing with other information from the visit form), which led to misinterpretation. Between 0 and 2% of data properly entered into the visit forms were incorrectly entered into the compilation.

The results of this investigation led us to propose various improvements for future campaigns, in the design of the visit forms, in the training of actors, in coordination and assistance and in data verification prior to entry into the database.

Keywords: Honeybees, Mortality, Surveillance, Data quality, Epilobee, Resabeille.

I - INTRODUCTION

Le dispositif de surveillance programmée de la mortalité des abeilles Résabeille, composante française du dispositif de surveillance européen Epilobee, a été mis en place au cours de deux campagnes consécutives (2012-2013 et 2013-2014) dans six départements pilotes. L'objectif principal de ce programme était d'estimer les taux de mortalité hivernale et en saison des colonies d'abeilles en France. Le protocole s'est appuyé sur une surveillance programmée de 66 ruchers tirés au sort dans chaque département représentant un total de 396 ruchers pour l'ensemble du programme. Dans chaque rucher, selon sa taille, de une à quatorze colonies ont été sélectionnées de manière aléatoire (permettant ainsi de détecter une maladie ou un trouble caractérisé par une prévalence minimale de 20 % au sein de chaque rucher) au cours d'une 1ère visite (V1), qui avait lieu au moment de l'entrée en hivernage (entre août et octobre). Ces colonies ont été suivies au cours de deux autres visites couvrant l'ensemble de l'année apicole, correspondant à la visite de sortie d'hivernage (visite de printemps (V2), de février à mai), et à la visite en saison apicole (visite d'été (V3), de juin à juillet).

La DGAl est responsable de la mise en œuvre de ce dispositif, qui constitue l'une des thématiques prioritaires de la Plateforme d'épidé-

miosurveillance en santé animale (Plateforme ESA). Au cours de chacune de ces visites, les colonies étaient examinées par les intervenants sanitaires apicoles qui renseignaient également une fiche de visite. Les fiches étaient ensuite saisies dans la base de données européenne développée dans le cadre du dispositif européen Epilobee. Cette saisie était réalisée par la section apicole du groupement de défense sanitaire de chaque département. Cette base de données en ligne était pilotée par le Laboratoire de référence de l'Union européenne (LRUE) sur la santé de l'abeille (laboratoire Anses de Sophia-Antipolis), en charge également de l'analyse et de l'interprétation des résultats de surveillance [Anonyme, 2013]. Toutefois, préalablement à cette analyse, de nombreux allers-retours entre le LRUE et chacun des 17 pays participants ont été nécessaires afin de corriger des données incohérentes ou de compléter des données manquantes.

Il semblait donc indispensable de procéder à une évaluation approfondie et rigoureuse de la qualité de ces données, des données de haute qualité étant définies comme des données convenables pour l'usage que nous souhaitons en faire [Kerr *et al.,* 2007]. En santé animale, aucune méthodologie d'évaluation de la qualité des données issues de programmes de surveillance n'existait [Bronner *et al*., 2015] avant le développement de la méthode Palussière *et al*. [2013b] qui vise à évaluer la qualité des données saisies dans une base de données.

Dans ce cadre, les objectifs de notre étude étaient :

1. D'évaluer de manière quantitative la qualité des données collectées par le programme de surveillance Résabeille en France au cours de la campagne 2013-2014,

2. D'étudier l'influence des étapes de collecte et de saisie sur la qualité des données,

3. De proposer des axes d'amélioration du dispositif de surveillance.

Pour répondre au premier objectif, la démarche d'évaluation quantitative de la qualité des données développée par Palussière *et al.* [2013b] et appliquée initialement au dispositif de déclaration des avortements chez les bovins a été utilisée. Cette démarche a été complétée par l'analyse approfondie d'un échantillon aléatoire de fiches de visite réalisées au cours du printemps 2014 et de leur double saisie, afin d'atteindre le 2ème objectif. Compte tenu du nombre de variables collectées et de l'objectif principal de l'étude, le présent article se focalise sur les variables utilisées pour estimer le taux de mortalité hivernale.

II - MATÉRIELS

1. SOURCE DE DONNEES

1.1. DONNEES DE LA BASE DE DONNEES EPILOBEE

Les données recueillies dans les fiches de visite sont enregistrées de façon standardisée dans la base de données européenne Epilobee via un site Internet. Cette base de données en ligne permet de centraliser rapidement les informations issues des dix-sept pays participants. L'enregistrement des données suit un ordre précis et obligatoire car la base de données est de type relationnel. A chaque sous-partie de la fiche de visite correspond un formulaire sur le site Internet. L'extraction des données à partir du site internet permet d'obtenir douze tables différentes en format Excel® portant sur les généralités sur l'apiculteur et son rucher, la gestion de son cheptel, les traitements effectués au cours de la saison apicole, les antécédents sanitaires du rucher, les transhumances, les visites réalisées (V1, V2 ou V3), les colonies examinées, la force des colonies, les symptômes observés lors de la visite ainsi que les maladies suspectées, les prélèvements effectués et les résultats d'analyse. Les tables au format Excel® peuvent ensuite être liées grâce aux champs d'identification (clés primaires et étrangères).

L'analyse de la qualité de l'ensemble des données disponibles dans la base de données a été conduite aux niveaux national et départemental pour la campagne 2013-2014. Les données utilisées ont été extraites de la base de données Epilobee en juillet 2014, date à laquelle aucune demande de correction concernant les données permettant l'estimation des taux de mortalité n'avait été encore formulée par le laboratoire de référence auprès des départements pour cette campagne. L'analyse a ensuite été réalisée sur l'ensemble des données disponibles sur la période d'étude.

1.2. SELECTION DES FICHES DE VISITE

Pour l'analyse de l'influence des étapes de collecte des données et de saisie sur la qualité des données renseignées dans la base de données, nous avons tiré au sort 60 fiches de visite de printemps 2014 (10 par département) parmi les 331 visites de printemps réalisées au cours de la campagne 2013-2014. Chacune de ces fiches avait déjà été saisie dans la base de données ce qui nous a permis d'effectuer une comparaison avec la double saisie que nous avons réalisée.

Le choix des fiches de visite de printemps a été fait car cette visite nécessite de renseigner l'ensemble des données relatives à l'estimation de la mortalité hivernale, données essentielles pour atteindre les objectifs de la surveillance de la mortalité du programme Epilobee.

2. PRESENTATION DES DONNEES CHOISIES POUR L'ETUDE

Les variables relatives à la mortalité hivernale sont présentées dans le tableau 1.

Pour l'ensemble des variables, seuls des entiers naturels pouvaient être saisis. Les variables étaient toutes pré-remplies par la valeur «9999» représentative, par convention, d'une donnée manquante. Si cette variable était effacée au moment de la saisie, la valeur nulle était entrée par défaut lors de la validation de la saisie. Ces contraintes impliquent que l'ensemble des données étaient renseignées en totalité et au bon format.

Tableau 1

Définitions des variables choisies

Nom	Définition
Random_v1_rap	Rappel du nombre de colonies tirées au sort à la visite d'entrée en hivernage (V1) qui composent l'échantillon aléatoire
Alive_v2	Nombre de colonies de l'échantillon aléatoire vivantes au cours de la visite de sortie d'hivernage (V2)
Dead_v1_v2	Nombre de colonies de l'échantillon aléatoire victimes de mortalité hivernale entre la visite V1 et la visite V2
Sold_v1_v2	Nombre de colonies de l'échantillon aléatoire cédées ou vendues entre la visite V1 et la visite V2
Merged_v1_v2	Nombre de colonies de l'échantillon aléatoire ayant été fusionnées entre la visite V1 et la visite V2
Produce_v1_v2	Nombre de colonies de l'échantillon aléatoire ayant été utilisées pour produire un ou plusieurs essaims entre la visite V1 et la visite V2

III - METHODES

1. ANALYSE DE LA QUALITE DES DONNEES RELATIVES A LA MORTALITE HIVERNALE DISPONIBLES DANS LA BASE DE DONNEES

L'évaluation quantitative de la qualité des données a été réalisée à deux échelles :

- À l'échelle des données : chaque variable pouvait être évaluée selon trois dimensions : la complétude, le format et la plausibilité (tableau 2). Dans le cadre de ces variables, la complétude et le format n'ont pas été étudiés car les contraintes

informatiques imposent une complétude et un format de 100 %. Pour l'étude de la plausibilité, une liste de valeurs plausibles pour chaque variable était établie au préalable (tableau 3) ;

- À l'échelle des visites de sortie d'hivernage (V2) : pour l'évaluation de la mortalité hivernale au sein des colonies d'abeilles, il est nécessaire que la cohérence (tableau 2) suivante soit respectée :

Random_v1_rap = Alive_v2 + Dead_v1_v2 + Sold_v1_v2

Pour chaque variable, l'étude de chaque dimension s'est faite progressivement : 1) la plausibilité a été calculée pour l'ensemble des données qui sont renseignées et au bon format et 2) pour les données collectées au cours de la visite V2, la cohérence a été étudiée pour l'ensemble des données plausibles.

Tableau 2

Définitions des dimensions sélectionnées

Niveau d'étude	Dimensions étudiées	Définitions
À l'échelle des données	Complétude	Donnée renseignée dans sa totalité [Pipino *et al.,* 2012]
	Format	Absence d'erreur de syntaxe telle qu'un format non conforme ou une orthographe incorrecte [Akoka *et al.*, 2007]
	Plausibilité	Possibilité pour une donnée d'être exacte en regard de l'élément étudié et d'autres sources d'informations [Weiskopf *et al.,* 2013]
Pour l'ensemble des visites de sortie d'hivernage (V2)	Cohérence	Existence d'une concordance entre les éléments de la base de données étudiée ou entre la base de données et une autre source [Weiskopf *et al.,* 2013]

Tableau 3

Condition de respect de la plausibilité

Variables	Liste des valeurs plausibles
Random_v1_rap	0 < x ≤ 14 ou x = 9999
Alive_v2	0 ≤ x ≤ 14 ou x = 9999
Dead_v1_v2	
Sold_v1_v2	
Merged_v1_v2	
Produce_v1_v2	

2. ANALYSE DE L'INFLUENCE DES ETAPES DE COLLECTE DES DONNEES ET DE SAISIE SUR LA QUALITE DES DONNEES RENSEIGNEES DANS LA BASE DE DONNEES

Nous avons effectué une nouvelle saisie de chacune des 60 fiches de visite tirées au sort dans le but de faire la synthèse de la qualité de renseignement des données contenues dans les fiches et de comparer notre saisie avec celle réalisée dans les départements.

L'influence de l'étape de collecte a été étudiée en analysant la qualité de renseignement des fiches de visite sélectionnées. Plus précisément, pour chaque variable étudiée, les critères suivants ont été calculés (figure 1 et tableau 4) :

- Parmi l'ensemble des données à renseigner sur la fiche de visite, nous avons évalué dans un premier temps la proportion de données
- Parmi les données non renseignées sur la fiche de visite, la proportion de données non renseignées pour lesquelles la saisie était

impossible (critère 1) a été calculée. Ceci correspondait aux données pour lesquelles la valeur ne pouvait pas être déduite de l'analyse de l'ensemble de la fiche de visite.

Une vérification de chaque valeur prise par les données a été effectuée à l'aide notamment de l'examen clinique des colonies renseigné par l'intervenant sanitaire apicole. Certaines valeurs ont ainsi été modifiées et nous avons évalué parmi les données pouvant être saisies, la proportion de données nécessitant une interprétation avant la saisie (critère 2) : en effet, soit la donnée n'était pas renseignée sur la fiche de visite mais sa valeur pouvait être retrouvée, soit la donnée était renseignée sur la fiche de visite mais la valeur indiquée n'était pas cohérente par rapport à l'examen clinique des colonies et cette valeur devait être modifiée avant la saisie.

L'influence de l'étape de saisie a été étudiée en comparant les données disponibles dans la base de données (saisies par les acteurs locaux) et les données que nous avons saisies. Pour chaque type de donnée étudié, l'évaluation de la qualité de saisie a porté sur les valeurs discordantes, en distinguant les saisies erronées constatées parmi les données nécessitant une interprétation (critère 3) des erreurs de saisie à proprement parler constatées parmi les données correctement renseignées dans la fiche de visite (critère 4).

Figure 1

Schéma illustrant la méthodologie

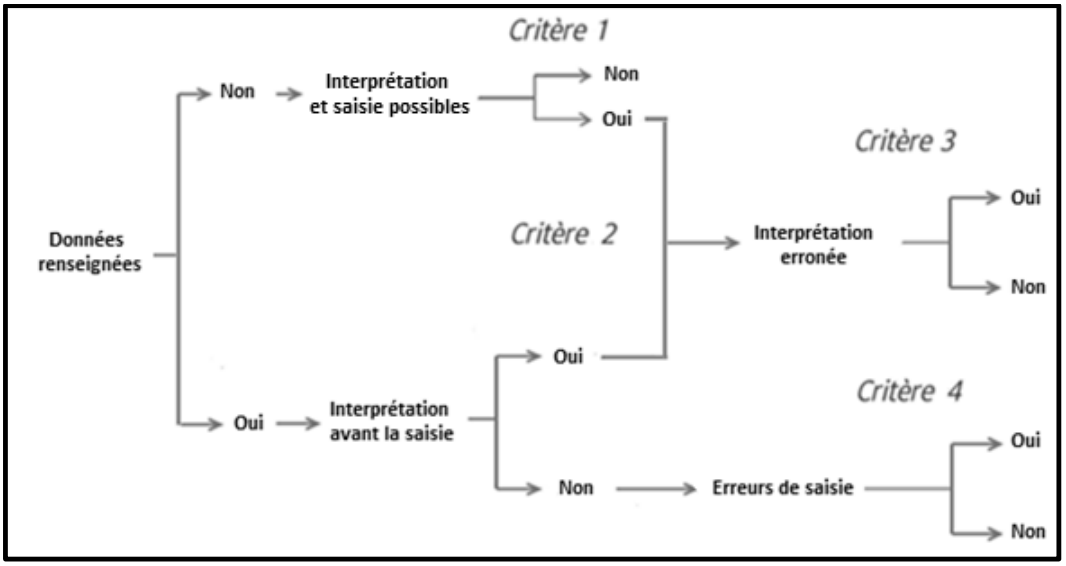

Tableau 4

Description des critères

Critères		Définitions /Indicateurs
Critères relatifs à la qualité de renseignement des fiches de visite	Critère 1	Proportion de données non renseignées pour lesquelles la saisie est impossible
		Indicateur = Nombre de données non renseignées pour lesquelles la saisie est impossible/ Nombre de données non renseignées
	Critère 2	Proportion de données pour lesquelles la saisie est possible mais nécessitant une interprétation avant la saisie
		Indicateur = Nombre de données nécessitant une interprétation avant la saisie/ Nombre de données pour lesquelles la saisie est possible
Critères relatifs à la qualité de saisie des données	Critère 3	Proportion de données nécessitant une interprétation avant la saisie et pour lesquelles l'interprétation est erronée
		Indicateur = Nombre de données pour lesquelles l'interprétation est erronée/ Nombre de données nécessitant une interprétation avant la saisie
	Critère 4	Proportion de données correctement renseignées sur la fiche de visite pour lesquelles une erreur de saisie a été identifiée
		Indicateur = Nombre de données correctement renseignées sur la fiche de visite pour lesquelles une erreur de saisie a été identifiée/ Nombre de données correctement renseignées sur le questionnaire

IV - RESULTATS ET INTERPRETATIONS

1. ANALYSE DE LA QUALITE DES DONNEES RELATIVES A LA MORTALITE HIVERNALE DISPONIBLES DANS LA BASE DE DONNEES

L'étude a été réalisée pour l'ensemble des 331 visites de sortie d'hivernage (V2) réalisées au cours de la campagne 2013-2014. Les résultats sont présentés dans le tableau 5.

La cohérence pour le calcul de la mortalité hivernale a pu être étudiée pour 326 visites de printemps sur les 331 visites V2 réalisées au printemps 2014 et était de 82 %.

Tableau 5

Résultats de l'analyse quantitative de la qualité des données

Variables	Plausibilité (en %) (n = 331)
Random_v1_rap	98 %
Alive_v2	100 %
Dead_v1_v2	100 %
Sold_v1_v2	100 %
Merged_v1_v2	100 %
Produce_v1_v2	100 %

2. ANALYSE DE L'INFLUENCE DES ETAPES DE COLLECTE DES DONNEES ET DE SAISIE SUR LA QUALITE DES DONNEES RENSEIGNEES DANS LA BASE DE DONNEES

2.1. ÉVALUATION DES CRITERES RELATIFS A LA COLLECTE DES DONNEES

La proportion de données renseignées dans les fiches de visite variait entre 77 et 97 % pour l'ensemble des variables. Le critère n°1 est de 0 % pour l'ensemble des variables étudiées et le critère n°2 variait entre 12 et 27 % (tableau 6).

2.1. ÉVALUATION DES CRITERES RELATIFS A LA SAISIE DES DONNEES

La proportion de données pour lesquelles l'interprétation est erronée variait de 36 à 73 % pour l'ensemble des variables. La proportion de données correctement renseignées sur la fiche de visite pour lesquelles une erreur de saisie a été identifiée est très faible (de 0 à 2 %) (tableau 7).

Tableau 6

Proportion de données renseignées, proportion de données non renseignées pour lesquelles la saisie est impossible (critère 1) et proportion de données nécessitant une interprétation avant la saisie (critère 2)

Variables	Proportion de données renseignées (en %) (n = 60) [IC à 95 %]	Critère 1 (en %) [IC à 95 %]	Critère 2 (en %) (n = 60) [IC à 95 %]
Random_v1_rap	97 [93-100]	0 [0-84] (n = 2)	12 [5 - 23]
Alive_v2	93 [87- 99]	0 [0-60] (n = 4)	27 [16-40]
Dead_v1_v2	88 [80-96]	0 [0-41] (n = 7)	25 [15-38]
Sold_v1_v2	78 [68-88]	0 [0-25] (n = 13)	23 [13-36]
Merged_v1_v2	77 [66-88]	0 [0-26] (n = 14)	23 [13-36]
Produce_v1_v2	80 [70-90]	0 [0-23] (n = 12)	22 [12-34]

Tableau 7

Proportion de données pour lesquelles l'interprétation est erronée (critère 3) et proportion de données correctement renseignées sur la fiche de visite pour lesquelles une erreur de saisie a été identifiée (critère 4)

Variables	Critère 3 (en %) [IC à 95 %]	Critère 4 (en %) [IC à 95 %]
Random_v1_rap	57 [18-90] (n = 7)	2 [0-10] (n = 53)
Alive_v2	69 [41-89] (n = 16)	2 [0-12] (n = 44)
Dead_v1_v2	73 [45-92] (n = 15)	2 [0-12] (n = 45)
Sold_v1_v2	43 [18-71] (n = 14)	0 [0-8] (n = 46)
Merged_v1_v2	36 [13-65] (n = 14)	0 [0-8] (n = 46)
Produce_v1_v2	54 [25-81] (n = 13)	0 [0-8] (n = 47)

IV - DISCUSSION

Nous pouvons résumer les résultats de notre étude de la manière suivante :

- La plausibilité des données variait entre 98 et 100 % selon la variable étudiée et la cohérence était de 82 % ;
- À partir d'un échantillon aléatoire de 60 fiches de visite, entre 77 et 97 % des données étaient renseignées sur les fiches ;
- La saisie était possible pour toutes les données à renseigner sur la fiche de visite, mais entre 12 et 27 % de ces données nécessitaient une interprétation avant la saisie (en recherchant la valeur correcte par croisement avec d'autres informations de la fiche de visite) ;
- En comparant la saisie de ces fiches réalisées par les acteurs dans les départements avec notre double saisie, nous avons estimé qu'entre 36 et 73 % des données mal renseignées sur les fiches de visites (et dont l'information correcte pouvait être retrouvée en croisant avec d'autres informations de la fiche de visite) avaient été mal interprétées au moment de la saisie ;
- Entre 0 et 2 % des données correctement renseignées sur les fiches de visites étaient mal saisies.

Ces résultats illustrent la nécessité de nettoyer les données avant de les analyser. Notre analyse de la qualité des données a été réalisée sur les données « brutes » juste après leur saisie par les acteurs départementaux et avant toute correction. En effet, les défauts de qualité constatés pour les données « brutes », en l'absence de correction, empêcheraient d'estimer correctement les différents taux de mortalité des colonies d'abeilles.

1. METHODE UTILISEE

L'analyse de l'influence des étapes de collecte des données puis de saisie sur la qualité des données renseignées dans la base de données se révèle complémentaire de l'évaluation quantitative de la qualité des données d'un dispositif de surveillance. Dans l'idéal, l'objectif serait d'évaluer la concordance entre les données présentes dans la base de données et la réalité. Dans le cadre de la 1ère étude, cette validité n'est étudiée que de façon approchée, en s'assurant uniquement que la valeur prise par cette donnée appartient bien à une liste de valeurs considérées « plausibles » établies au préalable [Palussière *et al.*, 2013a], L'analyse des fiches de visite permet quant à elle de se rapprocher de la valeur réellement prise sur le terrain, d'une part, grâce à l'accès aux données directement collectées et, d'autre part, en ayant accès à l'ensemble de la fiche de visite, les valeurs prises par les variables pouvant être confrontées. Bien sûr, ce type d'étude nécessite d'avoir accès aux fiches de visite, et de passer du temps à réaliser la double saisie, ce qui n'est pas toujours possible.

Cette méthode d'évaluation permet également d'analyser finement les résultats de plausibilité et de cohérence. Il a ainsi été possible de mettre en évidence que les défauts de plausibilité mais surtout de cohérence, identifiés à partir de l'analyse des données saisies dans la base de données, sont principalement attribuables au problème de collecte de ces données (données manquantes ou erronées) et à un manque d'interprétation (notamment pour les données manquantes).

2. INTERPRETATION DES RESULTATS

Pour l'étude quantitative de la qualité des données relatives à la mortalité hivernale, la plausibilité est bonne pour l'ensemble des variables malgré l'absence de contraintes informatiques imposant une liste de valeurs possibles à saisir. Il faut noter que les valeurs « 9999 » représentant des données manquantes ont été considérées plausibles. En effet, il est probable que les données à saisir dans la base de données n'aient pas été renseignées dans la fiche de visite. Ces valeurs saisies ont tout de même une répercussion sur la qualité des données entraînant une mauvaise cohérence dans 62 % des visites pour lesquelles la cohérence n'est pas respectée. Sans une étape de nettoyage des données, l'analyse du taux de mortalité hivernale des abeilles ne peut être effectuée pour 60 visites de printemps.

L'analyse de l'influence de l'étape de collecte des données sur la qualité des données renseignées dans la base de données a révélé tout d'abord un défaut de qualité de renseignement des fiches de

visite. En effet, malgré l'importance de ces données, le renseignement des fiches n'est que partiel, ce qui se traduit par une proportion de données renseignées allant de 80 % à 97 %. Dans l'ensemble, elles ont pu être interprétées grâce à l'analyse des données associées à l'examen des colonies, riche en information. Il s'agit ici d'un cas particulier lié aux données que nous avons choisies pour illustrer ce travail, l'interprétation des données n'étant pas possible pour d'autres variables du programme qui ne peuvent être inférées à partir des autres données collectées dans la fiche de visite. La proportion de données nécessitant une interprétation avant de pouvoir être saisies était particulièrement élevée pour certaines variables, ce qui peut s'expliquer par les raisons suivantes :

- Pour les variables *Alive_v2* et *Dead_v1_v2,* la majorité des cas nécessitant une interprétation relève du fait que les valeurs attribuées à ces deux variables ne correspondent pas à ce qui est observé sur le terrain, illustrant un problème de compréhension des définitions par les intervenants de terrain : par exemple, des colonies considérées comme non-valeurs ainsi que des colonies bourdonneuses n'ont pas été comptabilisées en tant que colonies mortes alors qu'elles auraient dû l'être selon les règles retenues par le protocole de surveillance, nécessitant ainsi de modifier les valeurs des données renseignées ;
- Pour les variables *Sold_v1_v2*, *Merged_v1_v2* et *Produce_v1_v2*, les données à interpréter sont, dans la majorité des cas, des données qui ne sont pas renseignées dans les fiches de visite. Cette absence de renseignement illustre des situations différentes. Ces variables ne sont souvent pas renseignées par les responsables des visites lorsqu'elles prennent la valeur zéro, mais, dans certains cas, l'absence de renseignement illustre également un problème de compréhension des définitions. En effet, certains intervenants sanitaires ont comptabilisé par exemple le nombre d'essaims produits à partir d'une ou plusieurs colonies de l'échantillon aléatoire dans la variable *Produce_v1_v2*, au lieu de mentionner le nombre de colonies de l'échantillon aléatoire ayant été utilisées pour produire un ou plusieurs essaims.

En ce qui concerne la proportion de données manquantes pour lesquelles l'interprétation est erronée (critère 3), les résultats sont également mauvais avec une valeur très élevée du critère pour toutes les variables. Ceci révèle que les personnes en charge de la saisie n'ont que très rarement procédé à l'interprétation des données avant la saisie ou que quand ils l'ont faite, ils ne l'ont pas faite de manière appropriée. Cette situation illustre manifestement un manque de formation et de préparation des personnes en charge de la saisie, et éventuellement un manque de temps passé à la saisie. En effet, des entretiens semi-directifs conduits avec les acteurs de la surveillance dans le cadre d'une autre partie de notre étude ont montré que certaines personnes en charge de la saisie des données n'avaient pas de compétence spécifique en apiculture alors que ces connaissances sont indispensables pour interpréter les données manquantes des questionnaires. La saisie des données correctement renseignées dans les fiches de visite engendre elle peu d'erreurs.

A travers cette méthode, nous ne pouvons pas exclure que des erreurs de saisie identifiées soient attribuables au fait que nous n'ayons accès qu'à une partie des sources de données par rapport aux personnes en charge de la saisie. En effet, ces dernières peuvent consulter les autres fiches de visites relatives au rucher concerné ou encore contacter directement l'intervenant sanitaire apicole en question pour connaître les valeurs prises par les différentes variables.

L'analyse de l'influence des étapes de collecte et de saisie des données permet donc de relativiser la bonne complétude observée à partir des données saisies dans la base. En effet, l'analyse des fiches de visite a permis de révéler un manque de renseignement de certaines données par les acteurs du terrain dans les fiches de visite qui était passé totalement inaperçu à la seule analyse quantitative en raison des contraintes informatiques imposant une complétude de 100 %. La mauvaise cohérence illustre ce manque de renseignement. En effet, la majorité des visites pour lesquelles la cohérence n'était pas respectée (62 %) impliquent des données manquantes renseignées par la valeur « 9999 ».

L'évaluation de la cohérence montre aussi ces limites car, pour un couple de données, la cohérence des données saisies peut être respectée sans pour autant révéler les problèmes de définition entrainant une mauvaise attribution des valeurs pour les différentes variables.

D'un point de vue méthodologique, la méthode d'évaluation de la qualité des données développée par Palussière *et al.* [2013b] est apparue tout à fait adaptée à notre problématique, ce qui apporte une expérience complémentaire pour la validation de cette méthode. La mise en œuvre d'une double

saisie sur un échantillon de visites tirées au sort est un volet complémentaire riche d'enseignements que nous recommandons d'ajouter à la méthode générique utilisée.

3. PERSPECTIVES D'AMELIORATION DU DISPOSITIF

Nous avons pu identifier certains facteurs influençant les défauts de qualité des données que nous avons constatés, ce qui nous permet ainsi de proposer plusieurs axes d'amélioration.

Tout d'abord, il faut rappeler l'importance de la formation des intervenants sanitaires apicoles et des personnes en charge de la saisie. Le nombre de données à collecter sur le terrain au cours d'une visite est très important et la compréhension du contenu de chaque champ à remplir peut être parfois complexe. Ainsi, si l'on ne s'assure pas de la qualité de la formation, cela a un retentissement manifeste sur la qualité des données de mortalité qui sont prioritaires pour atteindre les objectifs de la surveillance. En complément, un renforcement de l'animation du dispositif est à mettre en place sur l'aspect de la qualité des données afin de sensibiliser les acteurs du terrain et d'insister sur l'importance de leur rôle en termes de qualité des données. Ce message est à transmettre lors de la formation mais doit être réitéré tout au long de l'exécution de la surveillance, notamment à la faveur de réunions régulières dédiées.

Il convient ensuite de faciliter les modalités de collecte des données. Le pré-remplissage de certaines données dans les fiches de visite permettrait d'assurer un meilleur suivi des ruchers au cours de la campagne, en mentionnant par exemple au préalable dans la fiche de visite V2 la taille de l'échantillon aléatoire effectué en visite d'hivernage (V1). Cette étape de pré-remplissage nécessiterait une organisation dédiée entre les acteurs de la surveillance à l'échelon départemental. Il apparaît également important de réévaluer la partie de la fiche de visite correspondant à la mortalité des colonies dans le but d'améliorer la qualité des données collectées. Certaines données, notamment celles nécessitant une interprétation préalable, se révèlent en effet difficiles à renseigner par les agents sanitaires apicoles. Cette difficulté est liée à l'ambiguïté potentielle des définitions de mortalité, d'essaimage ou encore de fusion des colonies. Il conviendrait de préciser ces définitions par des exemples concrets illustrant la diversité des cas rencontrés sur le terrain et permettant aux acteurs de renseigner les données convenablement sans approximation.

La saisie se révèle complexe et longue dans le cas où une interprétation des données est nécessaire avant la saisie. L'analyse de la double saisie que nous avons réalisée montre qu'un contrôle préalable de chaque variable relative à la mortalité est nécessaire, qu'une donnée ait été inscrite ou non dans le questionnaire de visite. Cependant, cette analyse préalable nécessite des connaissances en apiculture et une connaissance du programme de surveillance, pour vérifier l'exactitude des données et également leur cohérence, ce qui n'est pas obligatoirement le cas pour tous les départements impliqués dans le programme. Cette analyse préalable peut également être effectuée par une autre personne de manière à ce que les données à saisir soient vérifiées avant leur saisie. Notre évaluation a en effet permis de montrer qu'il y avait très peu d'erreurs de saisie « vraie », à savoir une mauvaise retranscription d'une donnée correctement indiquée dans la fiche de visite. Enfin, au niveau de la base de données, des contraintes de saisie relatives aux relations entre les données pourraient être ajoutées dans un but d'améliorer la cohérence des données collectées dès le stade de la saisie.

VI - CONCLUSION

L'évaluation de la qualité des données issues du dispositif de surveillance programmée de la mortalité des abeilles en France montre que cette qualité pourrait être améliorée. Compte tenu des résultats de cette évaluation, il semble important d'intégrer des modalités d'amélioration de la qualité des données collectées en définissant au préalable les objectifs attendus en termes de qualité optimale des données tout en prenant en compte les différentes contraintes du terrain. Ces améliorations relèvent de l'évolution du questionnaire de visite, du renforcement de la formation des acteurs et de la mise en place d'une procédure de vérification des données avant leur

saisie dans la base. Ces aspects sont d'autant plus importants à prendre en compte qu'il est envisagé de poursuivre le programme Epilobee à l'échelon européen en ajoutant la recherche d'autres facteurs de risque de la mortalité des abeilles tels que les pesticides. Il convient donc de tirer tous les enseignements des deux premières années du programme pour faciliter la collecte et l'analyse des données pour le programme à venir.

D'un point de vue méthodologique, la méthode d'évaluation de la qualité des données que nous avons adaptée et mise en place dans cette étude mériterait sans doute d'être utilisée de manière plus systématique pour l'évaluation de la qualité des données des programmes de surveillance. Par ailleurs, le guide générique d'évaluation de la qualité des données développé à l'attention des partenaires de la Plateforme ESA par Palussière *et al*. [2013b] mériterait d'être mis à jour au vu de cette étude.

BIBLIOGRAPHIE

Anonyme - Note de Service DGAL/SDSPA/N2013-8139 du 14 août 2013 relative à la deuxième année de mise en place du réseau pilote d'épidémiosurveillance apicole 2013-2014.

Akoka J., Berti-Equille L., Boucelma O., Bouzeghoub M., Comyn-Wattiau I., Cosquer M., Goasdoué-Thion V., Kedad Z., Nugier S., Peralta V.,Sisaid-Cherfi S. - A framework for quality evaluation in data integration systems. In: 9th International Conference on Enterprise Information Systems, Madeira, Portugal, 2007.

Bronner A., Gay E., Fortané N., Palussière M., Hendrikx P., Hénaux V., Calavas D. - Quantitative and qualitative assessment of the bovine abortion surveillance system in France. *Prev. Vet. Med*., 2015, **02**, 019.

Kerr K., Norris T., Stockdale R. - Data quality information and decision making: a healthcare case study. 18th Australasian Conference on Information Systems, Toowoomba, 2007. *Citeseer*, 2007, 5-7.

Palussière M., Calavas D., Bronner A. - Évaluation de la qualité des données collectées dans le cadre du dispositif de déclaration obligatoire des avortements chez les bovins en France. *Bull. Epid. Santé Anim. Alim*., 2013a, **58**, 17-20.

Palussière M., Bronner A., Hendrikx P., Calavas D. - Guide d'évaluation de la qualité des données d'un dispositif de surveillance épidémiologique en santé animale, 2013b. Adresse URL : http://www.plateformeesa.fr/index.php?option=com_content&view=section&layout=blog&id=51&Itemid=302

Pipino L., Lee Y.W., Wang R.Y. - Data Quality Assessment. Communications of the ACM, 2002, 211-218.

Weiskopf N., Weng C. - Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J. Am. Med. Inform. Assoc*., 2013, **20**, 144-151.