Une nouvelle méthodologie pour l'anonymisation des entrepôts de données spatiales : application aux données de biodiversité dans le contexte agricole

Loris Croce[*,**,***] Laetitia Lemière[*,**,***] Sandro Bimonte[*], François Pinet[*]

[*] Université Clermont, IRSTEA, TSCF, Aubière, France
prenom.nom@irstea.fr,
https://www.irstea.fr/
[**]Université Clermont Auvergne, Aubière, France
prenom.nom@etu.uca.fr,
https://www.uca.fr/
[***]Ces auteurs ont contribué de manière égale.

Résumé. Dans cet article, nous nous intéresserons au problème de l'anonymisation de données agricoles géo-référencées. C'est un sujet qui n'est que peu ou pas abordé dans la littérature mais pourtant pertinent puisque le monde agricole est une source de données très importante. En effet, ces données peuvent-être utilisées pour mesurer la biodiversité. L'objectif est de rendre accessible des données agricoles à des fins de recherches sans briser l'anonymat des personnes participant à l'étude, car ces données peuvent d'avérer sensibles. Nous tentons de répondre à cela via une technique spécifique d'agrégation.

1 Introduction

Les entrepôts de données spatiales (EDS) et les systèmes OLAP spatial (SOLAP) permettent l'analyse en ligne de grandes volumes de données géo-référencées (Malinowski et Zimányi, 2008). Les données dans les EDS sont stockées selon le modèle *spatio-multidimensionnel* qui définit les concepts de dimension spatiale (dimension qui présente des valeurs géométriques) et mesure spatiale (valeur géométrique). Aujourd'hui, de plus en plus de données sont disponibles via les nouveaux systèmes d'acquisition (capteurs, images satellites, etc.), les réseaux sociaux, les données open data et les données volontaires (i.e. données des observatoires). Dans ce contexte, la mise en place des EDS peut se faire classiquement en utilisant des données internes aux organisations/entreprises, mais aussi en utilisant ces nouvelles sources de données (Ravat et al., 2016). De nombreux systèmes d'acquisition sont installés sur des terrains agricoles ce qui fait du monde agricole une source de données importantes. L'analyse de ces données serait très instructifs. Toutefois, cela soulève des nouvelles problématiques liées aux aspects de confidentialité. En effet, pour pouvoir être utilisées, ces données doivent être anomymisées car elles peuvent ne pas appartenir aux organisations/entreprises et il faut pouvoir garantir la confidentialité des informations. C'est pourquoi il est primordial d'établir une chaîne de confiance avec les personnes volontaires sur le respect de leur confidentialité.L'anonymisation

des données a aussi pour objectif de créer un climat favorable a la participation de tiers dans la collecte de données.

Une technique couramment utilisée pour anonymiser est le k-anonymat (Sweeney, 2002; Ciriani et al., 2007; Nguyen, 2014). Il consiste à regrouper des personnes parmi k autres personnes. Un individu est décrit par une ligne dans la base de données. Certaines variantes du k-anonymat où on cache un groupe (composé d'un ou plusieurs individu), sont apparues lorsque les données traitent de ménages ou que plusieurs personnes possèdent la même adresse. Dans ce travail, nous nous intéressons à la situation inverse où le même individu à protéger est directement relié à plusieurs entités de la base. Par exemple, dans une base de données concernant des logements, un propriétaire est directement relié aux logements qu'il possède et peut être ré-identifié par ces mêmes logements. En effet, il faut d'abord et avant tout qu'il y ait suffisamment d'individus pouvant être reliés aux données. Dans le cas extrême où toutes les entités du jeu de données sont reliées à la même personne alors chaque entité est un indice supplémentaire pour remonter à la personne. De plus, avec le développement de l'open data, de nombreuses sources d'informations sont disponibles et facilitent le recoupement d'informations.

Dans ce travail, nous proposons une nouvelle méthodologie pour anonymiser les données des EDS où plusieurs entités peuvent décrire le même individu en utilisant la micro-agrégation. La micro-agrégation consiste à agréger les données sensibles à des granularités plus élevées qui permettent l'anonymisation des données. Dans le contexte des EDS, la micro-agrégation peut s'appliquer naturellement en utilisant les hiérarchies des dimensions. Nous validons notre proposition en utilisant un cas d'étude. Il utilise l'EDS développé dans le cadre du projet VGI4Bio dont le but est l'analyse de la biodiversité dans le contexte agricole avec les données open-data issues de l'Observatoire Agricole de la Biodiversité (OAB).

L'article est structuré de la façon suivante : la section 2 relate l'état actuel de la littérature dans le cadre de notre étude. Ensuite, la section 3 explique plus en détail le cas d'étude. Puis, la section 4 présente la méthode qui a été développée avec un exemple d'application sur des données réelles. Les résultats sont en section 5 et la section 6 évoque la suite du travail.

2 État de l'art

La littérature propose de nombreuses solutions pour anonymiser des points qui est une représentation spatiale possible des données. Par exemple, l'échange de position entre deux points (Zhang et al., 2017) est l'une d'entre elles. Toutefois, cette représentation ne garantit pas l'anonymisation dans notre contexte. En effet, la méthode proposée dans (Zhang et al., 2017) cache des adresses de personnes sans prendre en compte la taille réelle de l'objet qui n'est pas une information identifiante dans leur contexte. Or, dans beaucoup d'applications réelles une représentation des données spatiales sous la forme de polygones est nécessaire - la surface des polygones devient alors un attribut permettant une ré-identification. Par exemple, dans le contexte agricole, il peut y avoir entre des parcelles voisines une grande diversité de surfaces. La surface devient donc un attribut particulièrement identifiant. Ainsi, il n'est pas possible d'échanger des valeurs entre des points sans prendre le risque de créer des valeurs incohérentes.

Une autre approche proposée par la littérature consiste à utiliser l'agrégation de points. Une des difficultés de cette méthode est le choix la taille des zones à l'intérieur desquelles on agrège. La technique proposée par Chow *et al.* (Chow, 2008) (qui fait varier la taille de la zone en fonction de la densité des points) s'avère pertinente. Toutefois, si les zones obtenues

ne correspondent pas à des zones facilement utilisables par d'autres acteurs alors la méthode perd de l'intérêt. L'utilisation de masques matriciels (Armstrong et al., 1999) est une approche envisageable pour notre problème. Toutefois,elle se heurte à la même difficulté que l'échange de points c'est-à-dire la taille des parcelles dans un contexte agricole. Pour illustrer cette difficulté la figure 1 représente une zone contenant 3 parcelles voisines. Si on agrège les trois terrains en un seul et que l'on fait par exemple la moyenne des valeurs caractéristiques, alors la parcelle A est plus déterminante car elle est bien plus grande que ses voisines. Par conséquent les trois parcelles n'influent pas de manière équivalente sur les données de la zone finale. Si on sait à qui appartient la parcelle A (la plus grande parcelle de la région) et que l'on connaît un indicateur moyen sur toute la région. Alors, on sait que A a fortement influencé la valeur de l'indicateur. Cela a pour conséquence de ne pas préserver la confidentialité de la personne détentrice de la parcelle A.

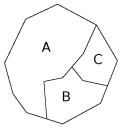


FIG. 1: *Exemple d'une parcelle prédominante.*

Puisque l'on ne peut ni échanger les valeurs ni regrouper les parcelles en plus grande zone par l'agrégation, l'utilisation de la micro-agrégation respectant la k-anonymisation (Sweeney, 2002) semble être la meilleure approche. Un exemple classique de transformation de données en données 3-anonymes est présent à la figure 2.

Définition (k-anonymat)**.** *Un ensemble d'enregistrements respecte la propriété du k-anonymat si n'importe quel enregistrement ne peut être différencié de $k-1$ autres. Pour cela il est nécessaire de supprimer les identifiants directs et de* généraliser *si besoin les* quasi-identifiants.

En effet, cette méthode permet de garder les informations au niveau des entités (appelé les microdonnées). Les microdonnées sont anonymes lorsqu'il y a k microdonnées « identiques » dans la même zone géographique. Donc, si on ne peux distinguer des microdonnées, on ne peut pas remonter de manière certaine à la personnes reliée aux microdonnées. Par exemple, en agriculture, si des parcelles de plusieurs exploitants sont indistinguables alors il ne sera par possible d'obtenir le nom correct de leur exploitant.

Les faits des EDS peuvent donc naturellement représenter les microdonnées car un fait est décrit par les membres des dimensions aux niveaux les plus fin, et les valeurs des mesures peuvent représenter les données sensibles.

Deux objets sont identiques si leurs quasi-identifiants le sont.

Définition (Quasi-identifiant (De Capitani Di Vimercati et al., 2012))**.** *Un attribut qui, combiné avec des sources d'informations externes, peut mener à une ré-identification de la personne à laquelle la donnée se réfère ou à diminuer l'incertitude à propos de son identité.*

Les quasi-identifiants sont définis en fonction des scénarios d'attaques (Kounadi et Resch, 2018). Ce sont les attributs qui dans un scénario donné aident à l'identification. Ces attributs sont généralisés afin de rendre des microdonnées identiques. Lors d'une généralisation une valeur réelle devient un intervalle ou une catégorie plus grande la contenant. Par exemple, considérons

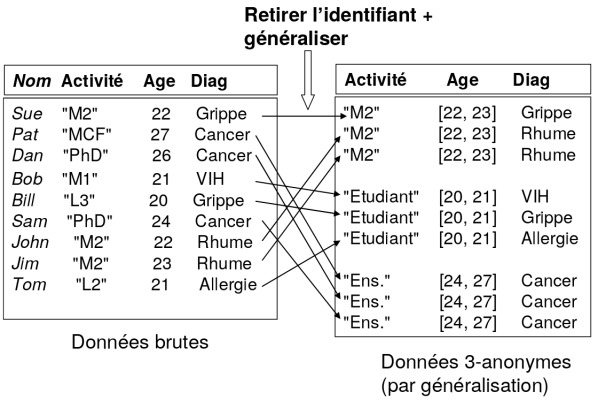


FIG. 2: *Exemple de k-anonymisation (Nguyen, 2014).*

un ensemble de données où les microdonnées (i.e. les faits) représentent les personnes vivant en France. Pour appliquer la micro-agrégation respectant le 5-anonymat, on peut, par exemple, choisir de remplacer les adresses exactes par la ville. Autrement dit, l'attribut « adresse » subit une généralisation. Ainsi, les microdonnées sont groupées par villes. Et pour respecter le 5-anonymat, dans chaque groupe, il faut que chaque individu ait au moins 4 autres individus identiques (c'est-à-dire ayant les mêmes quasi-identifiants que lui). De plus, micro-agréger garde l'information au niveau des faits contrairement à l'agrégation où l'on construit un objet résumant les faits. Ainsi, les analyses sur des micro-agrégations ont plusieurs angles possibles et peuvent être plus précises.

La micro-agrégation est donc directement applicable dans les EDS grâce à la présence des hiérarchies de dimensions. Dans notre approche, une opération de micro-agrégation n'agrège pas les valeurs des mesures, mais associe simplement le fait à un niveau moins détaillé (par exemple ville à la place d'adresse) avec une association de plusieurs à plusieurs, entre fait et dimension.

3 Cas d'étude

Le cas d'étude utilisé dans ce travail concerne l'EDS développé dans le cadre du projet VGI4bio pour l'analyse de données de biodiversité dans le contexte agricole. Cet EDS présente plusieurs dimensions : le temps (lors de l'expérimentation présenté à la partie 5, seulement les données de l'année 2017 ont été prise en compte), la localisation (avec la hiérarchie : parcelle, exploitation, ville, département et région), les cultures groupées par type de culture (dont la hiérarchie est visible à la figure 3) et la conduite. La mesure est représentée par l'abondance d'abeilles qui est agrégée avec la moyenne. Dans l'expérimentation, elle est donc la donnée à protéger. L'EDS peut permettre une mise en relation des pratiques et de la biodiversité. Afin de préserver la confidentialité des données pour un traitement à des fins de recherche, il faut pouvoir anonymiser les parcelles et, par conséquent, les agriculteurs (Bimonte et al., 2018). Donc, les microdonnées de la micro-agrégation sont les faits (comme par exemple des relevés sur des parcelles) considérés avec la valeur de la mesure et les membres de dimensions associés.

Définition (microdonnée)**.** *Une microdonnée possède $a_1..a_m$ attributs. Les attributs $a_1..a_i$ sont identifiants alors que les attributs $d_{i+1}..d_j$ sont quasi-identifiants. L'attribut a_m est l'attribut portant l'information sensible (qui ne doit pas être reliée à un individu).*

Dans notre exemple, pour une année donnée, une parcelle considérée comme un fait possède les attributs suivants :

- Identifiants :
  - un nom de parcelle : identifiant de la parcelle.
- Quasi-identifiants :
  - une culture : plante cultivée ;
  - une conduite : culture conventionnelle, biologique ou non-renseignée.
- Autre attribut :
  - un exploitant : nom de l'exploitant de la parcelles.
- L'attribut sensible : l'abondance d'abeilles.

Id parcelle	Zone (Département)	Nom exploitant	Culture (Culture)	Conduite	Abondance d'abeille
1	Allier	M. Boyer	betterave	Conventionnelle	3.979
2	Allier	M. Boyer	ble dur	Biologique	5.723
3	Isère	D. Bernard	ble dur	Biologique	2.151
4	Allier	P. Martinez	betterave	Biologique	4.96
5	Isère	D. Bernard	betterave	Conventionnelle	1.837
6	Isère	P. Draco	betterave	Conventionnelle	8.64
7	Isère	V. Clavez	ble tendre	Conventionnelle	3.52
8	Allier	M. Boyer	ble tendre	Conventionnelle	4.81
9	Isère	H. Guerchet	ble tendre	Conventionnelle	2.987
10	Isère	H. Guerchet	ble tendre	Biologique	2.299
11	Allier	M. Boyer	ble tendre	Biologique	6.8

TAB. 1: *Exemple (fictif) de microdonnées.*

Le tableau 1 montre un exemple de 11 microdonnées de notre jeu de données. L'attribut exploitant n'est pas classé comme quasi-identifiant car un quasi-identifiant est définit en fonction des attributs des données anonymes. Or, il sera supprimé des microdonnées anonymisées car il identifie directement l'exploitant qui est l'individu qu'on cherche à protéger. Chaque quasi-identifiant possède une hiérarchie de profondeur p qui lui est propre. Pour anonymiser une microdonnée, on doit généraliser des valeurs de quasi-identifiants. La hiérarchie pour les cultures contient le type de culture comme montré en figure 3. Si on souhaite faire une micro-agrégation en utilisant la valeur « noix », alors toutes les parcelles ayant cette valeur et toutes les autres parcelles avec la valeur « amande » ou « châtaigne » remplacent la valeur de l'attribut culture par « Fruit a coque ». Cette hiérarchie a été définie en utilisant le thésaurus *FrenchCropUsage* [1] (Roussey et Bernard, 2018). De la même manière, l'agrégation sur la hiérarchie spatiale consiste à passer d'un département à une région comme le montre la figure 4.

Notre algorithme va manipuler une hiérarchie (appelée par la suite « hiérarchie de données ») qui est la composition entre une hiérarchie d'un quasi-identifiant auquel on ajoute les microdonnées. Chaque microdonnée devient l'enfant du noœud ayant la même valeur que la microdonnée. Les microdonnées sont donc les feuilles de cette nouvelle hiérarchie. Dans notre exemple, on souhaite anonymiser sur l'axe spatial, la structure manipulée par l'algorithme est donc la hiérarchie du quasi-identifiant « zone » avec les microdonnées. La hiérarchie obtenue est visible à la figure 4.

1. `http://ontology.irstea.fr/pmwiki.php/Site/FrenchCropUsage`

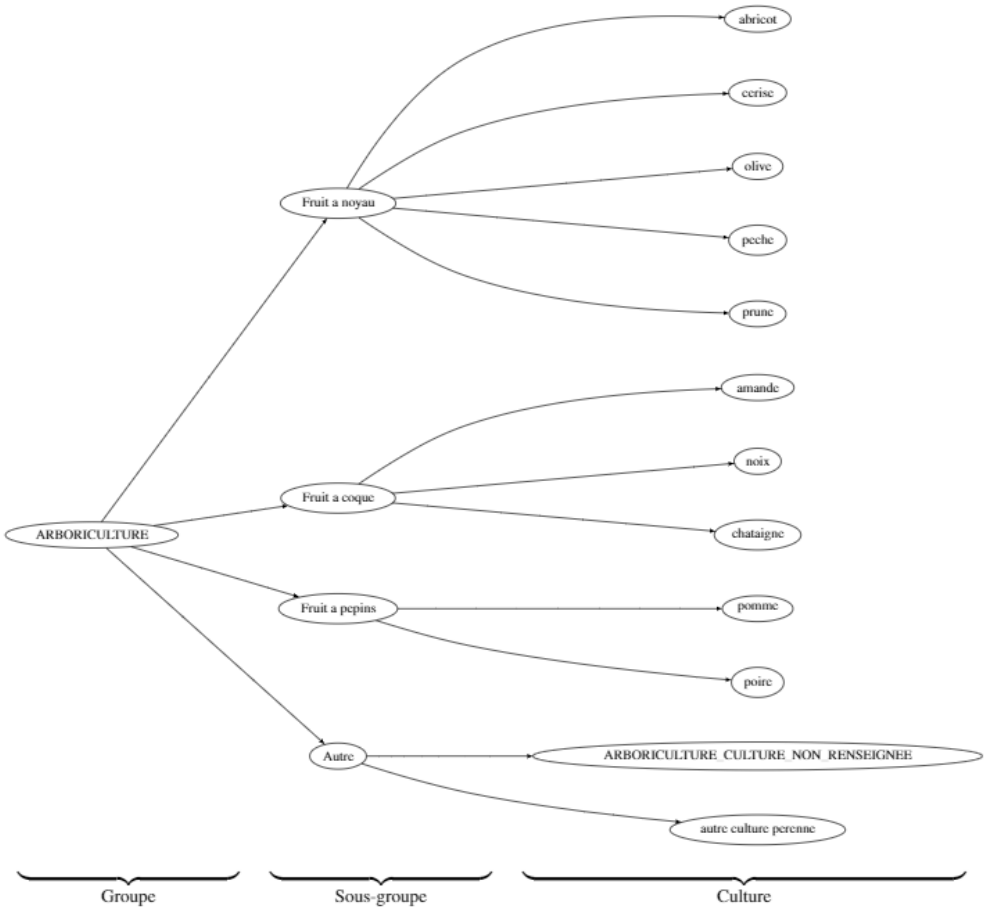


FIG. 3: *Fragment de la hiérarchie des cultures.*

Cette représentation est possible à condition qu'une microdonnée ne possède qu'une seule valeur à chaque quasi-identifiant. Cela se traduit dans la hiérarchie par les nœuds enfants ne peuvent avoir qu'un seul parent (Knuth, 2011). Ici, l'utilisation de la dimension spatiale fonctionne car une parcelle appartient à un et un seul département, comme cela est spécifié publiquement dans le registre du cadastre [2]. Lorsqu'on souhaite anonymiser des parcelles par micro-agrégation, on définit les zones en fonction de cette hiérarchie. On commence par le premier parent des feuilles puis, si besoin, on remonte dans la hiérarchie. Dans notre cas d'étude, on cherchera d'abord à anonymiser au niveau du département et si ce n'est pas possible on remontera au niveau des régions. La hiérarchie de données permet donc de regrouper les microdonnées sous un axe d'étude et d'anonymiser une microdonnée parmi l'ensemble de microdonnées ayant le même parent.

2. `https://www.cadastre.gouv.fr/scpc/accueil.do`

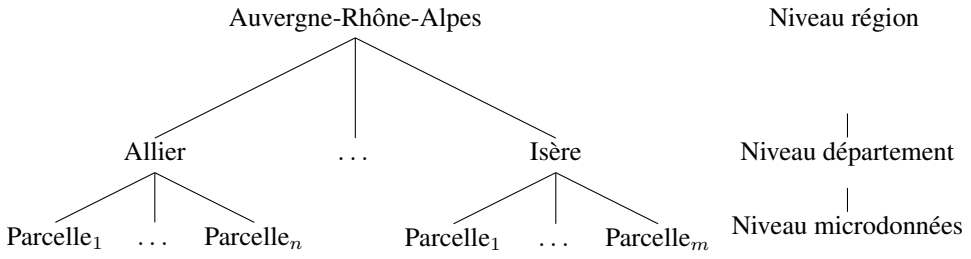


FIG. 4: *Exemple de hiérarchie sur l'axe géographique.*

Par ailleurs, il y a plusieurs microdonnées qui se rapportent au même individu. Il faut s'assurer que, dans l'ensemble de microdonnées, il y ait au moins X individus décrits par l'ensemble des microdonnées de la zone. Ce seuil sera appelé η. Dans le jeu de données agricole, il faut par exemple qu'il y ait au moins 3 agriculteurs par département.

Cette partie a permis de définir plusieurs termes comme les microdonnées, les hiérarchies et la hiérarchie de données qui seront utilisés dans le section 4. Tous ces termes ont été illustrés avec une application concrète qui est un jeu de données relatif aux parcelles agricoles.

4 Solution proposée

La solution présentée dans cette partie a pour but d'anonymiser lorsque, dans la base de données, ce ne sont pas des individus qui sont décrits mais des faits les concernant. C'est une situation qui se retrouve dans de nombreuses applications. Actuellement, il y a assez peu de méthodes pour anonymiser les données dans ce cas. Celle présentée dans cette partie possède une première étape de préparation des données et une seconde de transformation des microdonnées.

4.1 Nettoyage du jeu de données

Afin que l'anonymisation soit possible, il faut s'assurer qu'il n'y ait pas de parcelles avec des caractéristiques très particulières sur l'ensemble des parcelles de notre jeu de données. C'est la raison qui nous pousse à appliquer des pré-traitements sur les données. La liste des opérations effectuées est la suivante :

1. Suppression des identifiants ;
2. Suppression des microdonnées « uniques ou presque » : pour chaque combinaison des quasi-identifiants, sa fréquence d'apparition est calculée et nous vérifions qu'elle est supérieure à un seuil fixé appelé plus tard « seuil unique ». Si le nombre d'apparition de la combinaison des quasi-identifiants est insuffisant alors les parcelles avec cette combinaison sont considérées comme uniques et sont supprimées.

Le résultat de l'application du pré-traitement avec un seuil « uniques ou presque » à 2 sur les microdonnées présentées au tableau 1 est le tableau 2. La microdonnée 4 est supprimée car « betterave Biologique » n'est présente qu'une fois parmi toutes les microdonnées. Toutes les

autres combinaisons sont présentes au moins 2 fois. L'attribut « Id parcelle » a été enlevé car c'est un identifiant.

Zone (Département)	Nom exploitant	Culture (Culture)	Conduite	Abondance d'abeille
Allier	M. Boyer	betterave	Conventionnelle	3.979
Allier	M. Boyer	ble dur	Biologique	5.723
Isère	D. Bernard	ble dur	Biologique	2.151
~~Allier~~	~~P. Martinez~~	~~betterave~~	~~Biologique~~	~~4.96~~
Isère	D. Bernard	betterave	Conventionnelle	1.837
Isère	P. Draco	betterave	Conventionnelle	8.64
Isère	V. Clavez	ble tendre	Conventionnelle	3.52
Allier	M. Boyer	ble tendre	Conventionnelle	4.81
Isère	H. Guerchet	ble tendre	Conventionnelle	2.987
Isère	H. Guerchet	ble tendre	Biologique	2.299
Allier	M. Boyer	ble tendre	Biologique	6.8

TAB. 2: *Exemple (fictif) de microdonnées nettoyées.*

4.2 Algorithme

```
1  Function anonymisation de fait(hiérarchie de données, k, η) : hiérarchie de données
     anonymes is
2      Les microdonnées sont nettoyées.
3      for tous les premiers parents des feuilles de la hiérarchie des données do
           // Par exemple un département
4          if le parent est_k-anonymisable then
5              Transformation des microdonnées.
6          else
7              if Si le parent n'est pas la racine de la hiérarchie then
8                  Suppression du premier parent des microdonnées dans la hiérarchie.
                       // Exemple on supprime le département et donc le nouveau
                       parent des parcelles du département devient la région
9                  anonymisation de fait(hiérarchie de données, k, η)
10             else
11                 Suppression des enfants // les microdonnées contenues dans la
                       zone ne sont pas anonymisables
12     return hiérarchie de données
```

Algorithme 1 : Pseudo code de la méthode principale.

L'algorithme 1 présente la solution pour anonymiser des microdonnées géographiques. Le pré-traitement décrit dans la section 4.1 est résumé par la ligne 2. Pour expliquer, le fonctionnement de la méthode, nous l'appliquons sur les données du tableau 2 obtenues après le pré-traitement. Sa représentation sous forme de hiérarchie est la figure 5.

Le premier parent des microdonnées traité est l'Allier. La méthode « est_k-anonymisable » est appelée. Elle détermine si un ensemble de microdonnées peut être k-anonyme. Le pseudo-code est détaillé dans l'algorithme 2. Pour déterminer si la zone traitée est anonymisable, le

Auvergne-Rhône-Alpes

Allier

Nom exploitant	Culture (Culture)	Conduite	Abondance d'abeille
M. Boyer	betterave	Conventionnelle	3.979
M. Boyer	ble dur	Biologique	5.723
M. Boyer	ble tendre	Conventionnelle	4.81
M. Boyer	ble tendre	Biologique	6.8

Isère

Nom exploitant	Culture (Culture)	Conduite	Abondance d'abeille
D. Bernard	ble dur	Biologique	2.151
D. Bernard	betterave	Conventionnelle	1.837
P. Draco	betterave	Conventionnelle	8.64
V. Clavez	ble tendre	Conventionnelle	3.52
H. Guerchet	ble tendre	Conventionnelle	2.987
H. Guerchet	ble tendre	Biologique	2.299

FIG. 5: *Représentation des données (fictives) nettoyées.*

nombre d'individus présents est testé par rapport à η. Cela s'effectue en premier car cet attribut ne peux pas être généralisé. Dans l'exemple, η vaut 3 et les individus sont les exploitants des parcelles. Or, dans l'Allier, il n'y a qu'un seul exploitant : M. Boyer. Donc, la méthode « est_k-anonymisable » renverra un message indiquant que la zone n'est pas anonymisable. Dans ce cas, on remonte d'un niveau dans la hiérarchie de données en supprimant le niveau actuel. Le résultat de cette opération sur l'exemple est à la figure 6.

Auvergne-Rhône-Alpes

zone (fégion)	Nom exploitant	Culture (Culture)	Conduite	Abondance d'abeille
Auvergne-Rhône-Alpes	M. Boyer	betterave	Conventionnelle	3.979
Auvergne-Rhône-Alpes	M. Boyer	ble dur	Biologique	5.723
Auvergne-Rhône-Alpes	M. Boyer	ble tendre	Conventionnelle	4.81
Auvergne-Rhône-Alpes	M. Boyer	ble tendre	Biologique	6.8
Auvergne-Rhône-Alpes	D. Bernard	ble dur	Biologique	2.151
Auvergne-Rhône-Alpes	D. Bernard	betterave	Conventionnelle	1.837
Auvergne-Rhône-Alpes	P. Draco	betterave	Conventionnelle	8.64
Auvergne-Rhône-Alpes	V. Clavez	ble tendre	Conventionnelle	3.52
Auvergne-Rhône-Alpes	H. Guerchet	ble tendre	Conventionnelle	2.987
Auvergne-Rhône-Alpes	H. Guerchet	ble tendre	Biologique	2.299

FIG. 6: *Représentation des données (fictives) au niveau régional.*

Dans cette nouvelle zone, il y a 5 exploitants. Puisque le test sur η est validé, la méthode « généraliser » est appelée. Elle vérifie que toutes les combinaisons des quasi-identifiants présents dans la zone apparaissent au moins 3 fois car on fixe que k est égal à 3. Dans la région Auvergne-Rhône-Alples, il y a 4 combinaisons : betterave-Conventionnelle, blé dur-Biologique, blé tendre-Conventionnelle et blé tendre-Biologique qui apparaissent respectivement 3, 2, 3 et 2 fois. Dans cette situation, appliquer une micro-agrégation sur la valeur blé tendre permet de rendre les microdonnées anonymes. En effet, dans la hiérarchie des cultures visible à la figure 3, le parent de blé tendre est blé qui est aussi le parent de blé dur. La hiérarchie de données devient alors celle visible à la figure 7. Maintenant, la combinaison betterave-Conventionnelle

blé-Conventionnelle et blé-Biologique sont présentes respectivement 3, 3 et 4 fois. Toutes sont supérieures ou égales à k qui vaut 3. Donc, la zone est k-anonymisée.

Auvergne-Rhône-Alpes				
zone (région)	**Nom exploitant**	**Culture (Sous-groupe)**	**Conduite**	**Abondance d'abeille**
Auvergne-Rhône-Alpes	M. Boyer	betterave	Conventionnelle	3.979
Auvergne-Rhône-Alpes	M. Boyer	ble	Biologique	5.723
Auvergne-Rhône-Alpes	M. Boyer	ble	Conventionnelle	4.81
Auvergne-Rhône-Alpes	M. Boyer	ble	Biologique	6.8
Auvergne-Rhône-Alpes	D. Bernard	ble	Biologique	2.151
Auvergne-Rhône-Alpes	D. Bernard	betterave	Conventionnelle	1.837
Auvergne-Rhône-Alpes	P. Draco	betterave	Conventionnelle	8.64
Auvergne-Rhône-Alpes	V. Clavez	ble	Conventionnelle	3.52
Auvergne-Rhône-Alpes	H. Guerchet	ble	Conventionnelle	2.987
Auvergne-Rhône-Alpes	H. Guerchet	ble	Biologique	2.299

FIG. 7: *Représentation des données (fictives) micro-agrégées sur l'attribut culture (qui devient "type culture").*

Dans cette partie, la méthode proposée est appliquée pas à pas sur un exemple. Cet application permet, entre autre, de montrer que les données non anonymisables sont donc levées de la table de fait lors de la phase de pré-traitement. Les micro-agrégations se font donc sans ces données. De plus, la cohérence au niveau cartographique est maintenue car les micro-données sont affichées à des niveau spatiaux agregés (département, région) et non à celui le plus fin (parcelle).

5 Expérimentation

L'expérimentation est d'abord décrite par la description des différents risques de réidentification d'un individu puis par les résultats obtenus en appliquant notre méthode sur un jeu de données agricole.

5.1 Risque de ré-identification

Pour établir le risque de ré-identification avec la méthode proposée, d'abord, il faut définir les scénarios d'attaques possibles. Ici, on envisage l'identification spontanée et l'attaque utilisant des informations extérieures tel qu'ils sont définis dans le journal « ESSnet on Statistical Disclosure Control » (Hundepool et al., 2012).

L'identification spontanée est possible si les seuils fournis en entrée du programme sont mal paramétrés. Cela signifie que dans une zone, des parcelles seront mal anonymisées. Concrètement, cela peut arriver si une parcelle est particulière parmi les autres parcelles de la zone. Par exemple, si dans la méthode présentée en section 4.2, on fixe un k à 1. Alors, dans une zone (région ou département), une parcelle peut être la seule dont la culture est la pomme. Dans ce cas, l'attaquant a juste à trouver dans le « monde réel » la seule parcelle avec des pommes. Une fois la parcelle retrouvée, il peut savoir à qui elle appartient.

Le risque de ré-identification à l'aide de sources extérieures n'est pas à minimiser. Il existe de nombreuses sources d'informations en accès libre traitant de l'agriculture et, plus

```
1  Function est_k-anonymisable(microdonnées, η, k) : les microdonnées résultantes de
     l'anonymisation is
2      nombre ← compter le nombre d'individus décrits par les microdonnées de la zone.
3      if nombre < η then
4          return Zone non anonymisable
5      microdonnées ← généraliser(microdonnées, k);
6      if au moins une des combinaisons des quasi-identifiants des microdonnées < k then
7          return Zone non anonymisable
8      return réponse, microdonnées

9  Function généraliser(microdonnées, k) : microdonnées généralisés is
10     nombre ← calcul de la plus petite fréquence d'apparition des combinaisons des
       quasi-identifiants des microdonnées.
11     while critère d'arrêt do
           // Il est possible qu'une fréquence d'apparition soit toujours
              inférieure à k c'est pour cela que cette condition ne suffit
              pas à l'arrêt de la boucle while.
12         attribut , valeur à changer ← sélection de l'attribut et de la valeur dont une
           généralisation minimiserait le plus le nombre de combinaison inférieures à k.
13         Modifie les microdonnées en remplaçant « valeur à changer » et les valeurs
           sœurs sur « attribut » par le parent de « valeur à changer ».
14     return microdonnées
```

Algorithme 2 : Pseudo code des méthode « est_k-anonymisable » et « généraliser ».

particulièrement, des parcelles. Par exemple, grâce au site Koumoul [3], il est possible de récupérer l'identifiant d'une parcelle ainsi que sa superficie. Avec ces informations, on peut alors aller sur la page Géoportail [4]. Ce site donne la possibilité de rechercher la parcelle associée à l'identifiant obtenu avec Koumoul et de connaître l'historique des cultures qui ont été plantées sur celle-ci. Ces données sont publiques et facilement consultables sur le web.

Par exemple, avec l'identifiant d'une parcelle de la forme [Code Insee de la commune][Section communale][Numéro de parcelle], il est possible de recouper avec des informations externes. Ainsi, il relie une information à protéger à un individu. Ce dernier n'est donc plus sous anonymat.

À cause de la présence de ces nombreuses sources d'informations, il est crucial de mettre à l'épreuve les données en sortie du programme afin de vérifier leur réelle anonymisation. Il faut donc qu'à partir de données publiées, il ne doit pas être possible de récupérer l'identifiant d'une parcelle qui, comme expliquer ci-dessus, mène directement à un exploitant.

5.2 Résultats

On effectue une série de tests sur le jeu de données [5] en faisant varier les paramètres k (dans une zone, chaque combinaison doit au moins apparaître k fois), seuil unique (chaque combinaison doit au moins apparaître plus que ce seuil dans tout le jeu de données) et type de

3. https://koumoul.com/s/cadastre/
4. https://www.geoportail.gouv.fr/carte
5. Uniquement l'année 2017.

régions (nouvelles ou anciennes régions françaises) décrit dans la section 4. De plus, on a fixé que le nombre minimum d'exploitants dans la zone soit égal à 3.

Pour estimer l'efficacité de l'algorithme, on observe le pourcentage de microdonnées anonymes conservées, ainsi que le nombre de zones affichées. Les résultats des expérimentations se trouvent dans le tableau 3. Les tests sont classés dans l'ordre croissant de leur performance.

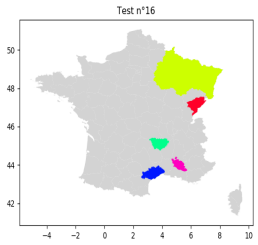
nº test	k	**Seuil unique**	**Type région**	**Nb régions**	**Nb dept**	**% de lignes anonymisées**
16	6	30	nouvelles	1	4	3.019
24	6	45	nouvelles	2	3	3.711
4	6	15	anciennes	0	3	4.214
8	6	15	nouvelles	0	3	4.214
19	5	45	anciennes	2	5	4.403
12	6	30	anciennes	0	5	5.597
20	6	45	anciennes	2	4	5.66
6	4	15	nouvelles	2	5	6.226
3	5	15	anciennes	0	5	7.61
7	5	15	nouvelles	0	5	7.61
23	5	45	nouvelles	2	5	8.428
15	5	30	nouvelles	2	7	8.868
2	4	15	anciennes	0	8	10.189
11	5	30	anciennes	2	7	10.44
14	4	30	nouvelles	4	8	13.648
13	3	30	nouvelles	5	8	14.528
5	3	15	nouvelles	4	7	14.906
10	4	30	anciennes	3	9	14.906
22	4	45	nouvelles	9	3	15.535
9	3	30	anciennes	3	12	16.415
18	4	45	anciennes	9	7	16.415
21	3	45	nouvelles	9	5	17.547
1	3	15	anciennes	2	11	19.497
17	3	45	anciennes	10	10	22.767

TAB. 3: *Résultats selon les paramètres d'entrées triés par pourcentage.*

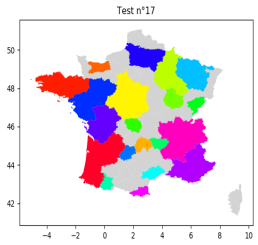
Les figures 8 et 9 montrent des cartes créées à partir du résultat de notre programme. Les cartes présentées permettent de visualiser les zones dont les parcelles sont conservées car elles sont k-anonymes. Plus spécifiquement, cela permet d'estimer la répartition et la granularité des données obtenues sur le territoire. Une zone colorée contient des microdonnées anonymisées et donc conservées tandis qu'une zone grisée ne possède pas de microdonnées car les parcelles présentes sur cette zone ne peuvent pas être k-anonymes.

À partir du tableau 3, on peut constater que le paramètre k a le plus d'impact sur les résultats. Cela s'explique car il est beaucoup plus aisé de respecter un critère de k faible. On a donc les $k = 3$ ou 4 qui apparaissent en meilleure position. D'un point de vue graphique, la figure 8 montre bien qu'un plus grande surface de la France est couverte lorsque k est faible.

De plus, l'attribut « type région » ne semble pas avoir une influence importante sur le pourcentage de lignes anonymisées. En effet, la première partie du tableau possède 7 lignes avec la valeur « nouvelles » et 5 avec « anciennes » tandis que pour l'autre moitié du tableau la répartition est l'inverse (5 pour « nouvelles » et 7 pour « anciennes »). Toutefois, on tend à conserver plus de zones au total avec les anciennes régions plutôt qu'avec les nouvelles . Il influe également sur la répartition des zones entre les régions et les départements. L'attribut « nouvelles » a, le plus souvent, plus ou autant de régions qu'avec l'attribut « anciennes » lors



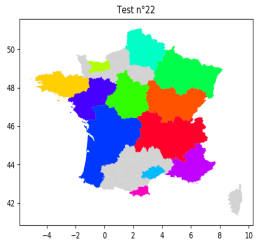
(a) *Carte résultante du test n° 16.*



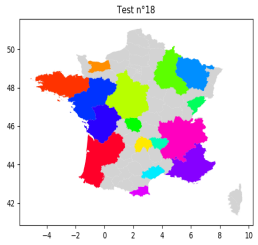
(b) *Carte résultante du test n° 17.*

FIG. 8: *Comparaison des cartes obtenues avec les tests ayant les pourcentages le plus et le moins élevé.*

d'une exécution avec des paramètres égaux. Ce fait est illustré par la figure 9. Les deux cartes ont les mêmes paramètres (sauf pour type région) en entrée du programme. Celle de gauche résulte de l'utilisation des nouvelles régions et celle de droite des anciennes régions. On peut constater que la superficie « anonymisée » de la France varie. Ce paramètre influe donc sur la précision géographique des résultats.



(a) *Test n° 22 utilisant nouvelles régions.*



(b) *Test n° 18 utilisant les anciennes régions.*

FIG. 9: *Comparaison des cartes résultats en faisant varier le paramètre zone.*

De la même façon, le critère de seuil unique ne semble pas déterminant. En effet, on a les seuils avec la valeur 15 qui sont légèrement plus présents dans la première partie du tableau et les seuils à 30 dans la seconde, partie. La répartition des seuils à 45 est homogène de chaque coté.

Par ailleurs, le nombre de zones (qui est la somme du nombre de départements et de régions) en sortie du programme est proportionnel au pourcentage de lignes conservées.

Néanmoins, il est important de rappeler que les résultats de l'algorithme sont hautement influencés par la nature des données. Le nombre de parcelles dans le jeu de données utilisé est aux alentours de 1500. Cette valeur est très faible en comparaison du nombre total de parcelles

sur le territoire français qui est d'environ 9 millions [6] . Une parcelle a donc plus de chances d'être unique dans le jeu de données car les parcelles semblables n'y sont pas.

Les risques de ré-identification peuvent provenir d'un mauvais paramétrage de la méthode mais surtout d'une mauvaise analyse des sources d'informations extérieures. La quantité de parcelles anonymes dépend principalement de l'objectif de k-anonymat. Les valeurs de k testées étaient relativement faible à cause du petit nombre de parcelles présentes dans le jeu de données. Par ailleurs, le choix des limites géographiques des zones influe sur la superficie mais assez peu sur la quantité de parcelles anonymisées.

6 Conclusions et travaux futurs

Cet article présente un premier travail autour du k-anonymat où plusieurs entités du jeu de données décrivent une même personne. D'après nos recherches, la littérature propose assez peu de solutions dans ce cas. C'est pourquoi, à partir d'un cas d'étude composé de données agricoles géo-référencées, une méthode basée sur la micro-agrégation a été proposée. Comme l'a montré la partie 5, l'efficacité de cette méthode dépend de plusieurs paramètres parmi lesquels le k du k-anonymat qui semble avoir une certaine influence sur le résultat de la méthode.

Une méthode différente est en cours de préparation. Son objectif est d'anonymiser le même type de données et où les micro-agrégations s'effectuent dans des zones les plus petites possibles ne respectant pas obligatoirement des frontières administratives.

Notre travail futur consiste à utiliser la *differential privacy* ou « confidentialité différentielle » (Dwork, 2011) pour résoudre le problème de l'anonymisation dans le contexte agricole. Cette technique plus récente est offre la possibilité de publier des données sans effectuer de suppression. Cela limite donc la perte d'information.

Mentions complémentaires

Ces travaux ont été financés par le projet ANR-17-CE04-0012 (données de biodiversité issues de ce projet). L'étude sur les techniques d'anonymisation a été financée par le projet CASDAR-MULTIPASS [7].

Références

Armstrong, M., G. Rushton, et D. Zimmerman (1999). Geographically masking health data to preserve confidentiality. *Statistics in medicine 18*, 497–525.

Bimonte, S., A. Besnard, E. Edoh-Alove, A. Hassan, C. Prince, A. Sakka, et P. Zaraé (2018). VGI users & data centered methods for the analysis of farmland biodiversity indicators open issues. In *21th AGILE International Conference on Geographic Information Science (AGILE 2018)*, Lund, Sweden, pp. 1–5.

Chow, C.-Y. (2008). *Cloaking Algorithms for Location Privacy*, pp. 93–97. Boston, MA : Springer US.

6. Selon le registre parcellaire graphique, données de 2017.
7. `https://numerique.acta.asso.fr/multipass/`

Ciriani, V., S. De Capitani di Vimercati, S. Foresti, et P. Samarati (2007). *k-Anonymity*, Volume 33, pp. 323–353.

De Capitani Di Vimercati, S., S. Foresti, G. Livraga, et P. Samarati (2012). Data privacy : definitions and techniques. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 20*(06), 793–817.

Dwork, C. (2011). *Differential Privacy*, pp. 338–340. Boston, MA : Springer US.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer, et P.-P. Wolf (2012). *Wiley Series in Survey Methodology*, pp. 287–288.

Knuth, D. E. (2011). *Art of Computer Programming, Volumes 1-4A Boxed Set*. Addison-Wesley Professional.

Kounadi, O. et B. Resch (2018). A geoprivacy by design guideline for research campaigns that use participatory sensing data. *Journal of Empirical Research on Human Research Ethics 13*(3), 203–222.

Malinowski, E. et E. Zimányi (2008). *Advanced Data Warehouse Design - From Conventional to Spatial and Temporal Applications*. Data-Centric Systems and Applications. Springer.

Nguyen, B. (2014). Techniques d'anonymisation. *Statistique et Société 2*(4), 53–60.

Ravat, F., J. Song, et O. Teste (2016). Designing multidimensional cubes from warehoused data and linked open data. In *Tenth IEEE International Conference on Research Challenges in Information Science, RCIS 2016, Grenoble, France, June 1-3, 2016*, pp. 1–12. IEEE.

Roussey, C. et S. Bernard (2018). French crop usage.

Sweeney, L. (2002). K-anonymity : A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10*(5), 557–570.

Zhang, S., S. M. Freundschuh, K. Lenzer, et P. A. Zandbergen (2017). The location swapping method for geomasking. *Cartography and Geographic Information Science 44*(1), 22–34.

Summary

In this article we will focus on the problem of anonymization of geo-referenced agricultural data. It is a subject that has received little or no mention in the literature but it is interesting because the agriculture is an important source of data. The objective is to make agricultural data available for research purposes without breaking the anonymity of the people participating in the study. We are trying to address this through a specific aggregation technique.