Anonymiser des données multidimensionnelles à l'aide du coclustering

Françoise Fessant*, Tarek Benkhelif*,**, Fabrice Clérot *

*Orange Labs, Lannion
francoise.fessant, tarek.benkhelif, fabrice.clerot@orange.com,
**DUKe, LINA, Nantes
http://duke.univ-nantes.fr/

Résumé. Dans cet article, nous proposons une méthodologie pour anonymiser une table de données multidimensionnelles contenant des données individuelles (soit *n* individus décrits par *m* variables). L'objectif est de publier une table anonyme construite à partir d'une table initiale qui protège contre le risque de réidentification. En d'autres termes, on ne doit pas pouvoir retrouver dans les données publiées un individu présent dans la table originale. La solution proposée consite à agréger les données à l'aide d'une technique de coclustering, puis à utiliser le modèle produit pour générer une table de données synthétiques du même format que les données initiales. Les données synthétiques, qui contiennent des individus fictifs, peuvent maintenant être publiées. Les données produites sont évaluées en termes d'utilité pour différentes tâches de fouille (analyse exploratoire, classification) et de niveau de protection.

1 Introduction

Il y a une très forte demande économique et citoyenne pour l'ouverture des données que ce soit pour la recherche ou le marketing. Le secteur public en particulier, à travers ses instituts de statistique nationaux, de santé ou de transport est soumis à des pressions pour mettre à disposition du public autant d'informations que possible au nom de la transparence [1]. Les entreprises privées sont également concernées par la valorisation de leur données à travers l'échange ou la publication. Orange a ainsi récemment mis à disposition de la communauté scientifique différents jeux de communications mobiles collectées sur ses réseaux de Côte d'Ivoire ou du Sénégal, dans le cadre des challenges D4D (Data for Developement) dans un objectif de service à des projets de développement ou d'amélioration de politiques publiques (Blondel et al., 2012).

Quand les données publiées concernent des individus et comportent des données à caractère personnel, elles doivent être anonymisées. On peut définir l'anonymisation comme le processus par lequel des données sont rendues anonymes et à l'issue duquel elles ne peuvent plus être affectées ou rattachées à une personne en particulier. [2]

1. https ://www.republique-numerique.fr
2. Définition de l'Association Française des Correspondants à la protection des Données à caractère Personnel. Glossaire anonymisation de données de l'AFCDP du 23 mai 2007.

Le G29 qui regroupe les autorités de protection des données européennes propose trois critères pour évaluer une solution d'anonymisation i) l'individualisation : est-il toujours possible d'isoler un individu ? ii) la corrélation : est-il possible de relier entre eux des ensembles de données distincts concernant un même individu ? iii) l'inférence : peut-on déduire de l'information sur un individu ? Un ensemble de données pour lequel il n'est possible ni d'individualiser, ni de corréler, ni d'inférer est a priori anonyme. Un ensemble de données pour lequel au moins un des trois critères n'est pas respecté ne pourra être considéré comme anonyme qu'à la suite d'une analyse détaillée des risques (G29, 2014). Les deux premiers items font référence à la ré-identification d'un individu qui se produit quand un attaquant peut reconnaître l'individu dans les enregistrements publiés. Le dernier item concerne une violation de la confidentialité dans laquelle un attaquant n'a pas besoin d'identifier précisément un individu pour déduire une information sensible le concernant, information dont il n'aurait pas eu connaissance sans la publication des données (on peut prendre l'exemple de l'hôpital qui publie l'information que toutes les patientes femmes de 50-60 ans qui ont fréquenté l'hopital ont le cancer).

La littérature sur le domaine de la publication respectueuse de la vie privée s'organise principalement autour de ces deux notions de protection des données personnelles : l'individualisation avec le concept de k-anonymat et la confidentialité avec le concept de confidentialité différentielle (ou Differential Privacy, DP). Le principe du k-anonymat est de cacher un individu dans un groupe de k, en réduisant le niveau de détail des données, de manière à empêcher sa ré-identification (Sweeney, 2002). La notion de protection défendue par la DP est la suivante : faire en sorte qu'on ne puisse pas savoir si un individu contribue à un résultat agrégé (que l'individu soit présent ou non dans les données ne doit pas avoir d'impact significatif sur le résultat du calcul de l'agrégat). On atteint la DP en rajoutant un bruit aléatoire à la valeur recherchée (Dwork, 2008). La publication de données respectant la DP est appropriée à la fourniture de statistiques agrégées (comptes, histogrammes, tables de contingence) plutôt qu'à la publication de données individuelles détaillées (Cormode, 2015). Une des solutions utilisées, notamment par les organismes publics d'enquêtes, pour pouvoir fournir des données individuelles détaillées, tout en protégeant la vie privée des individus (Vilhuber et al., 2016) est la génération de données synthétiques. Une voie qui se dessine actuellement consiste à coupler modèles statistiques génératifs DP et données synthétiques. La plupart des méthodes proposent d'approximer la distribution jointe globale des données à l'aide de distributions jointes estimées de manière DP sur des sous-ensembles de variables, puis de les combiner pour construire le modèle génératif final qui sert à produire les données synthétiques (Zhang et al., 2014) ou (Chen et al., 2015). Pour une revue des avancées récentes sur le thème de la publication respectueuse de la vie privée, on peut se reporter à (Fung et al., 2010).

Dans le cadre de cet article, nous nous intéressons à la protection de tables multidimensionnelles contenant les enregistrements détaillés d'un ensemble d'individus. Les données protégées sont générées à l'aide d'une méthode qui modifie les données originales et seul l'ensemble protégé est publié. Le risque contre lequel on cherche à se prémunir ici est le risque de ré-identification de l'individu dans les données publiées. On propose d'utiliser la technique du coclustering pour construire des groupes de k individus (comme dans une technique de k-anonymat). Le coclustering produit une vue agrégée des données initiales qui peut être plus ou moins fine selon le niveau de protection désiré et optimale à chaque niveau au sens de l'information mutuelle (Boullé, 2011). Le modèle de coclustering peut également être considéré comme un modèle générateur qui permet de générer une table de données synthétiques qui

préservent les propriétés statistiques des données originales et du même format que celles-ci.

Quelle que soit la méthode de protection, elle a en général pour effet de diminuer l'utilité des données. L'enjeu est donc de publier une version anonymisée de la table qui reste exploitable pour la fouille, tout en offrant des garanties de protection (Hundepool et al., 2012).

Après avoir rappelé le principe du coclustering et détaillé le k-anonymat, on décrit la méthodologie d'anonymisation proposée. L'utilité des données synthétiques est évaluée sur différentes tâches de fouille (analyse exploratoire, classification), tâches qui ne sont pas connues au moment de la mise en œuvre du processus d'anonymisation. On évalue également le niveau de protection offert par les données publiées.

2 Description de la méthodologie d'anonymisation

2.1 Coclustering

Le coclustering est une technique qui a pour but de réaliser une partition simultanée des lignes et des colonnes d'une matrice de données. Dans le cas où les dimensions de la matrice sont les observations et les variables, on réalise simultanément une partition des individus et des variables descriptives des individus (qui peuvent être aussi bien catégorielles que numériques).

On utilise la méthode de coclustering KHC de (Boullé, 2012) utilisable via le logiciel Khiops [3]. KHC est libre de tout paramétrage utilisateur, robuste (évite le sur-apprentissage), supporte des bases volumineuses et permet de réaliser une partition de plusieurs variables, continues ou catégorielles.

KHC suit l'approche MODL (Boullé, 2006) qui permet d'estimer la densité jointe d'un ensemble de variables, sur la base de modèles en grille. Les modèles en grille réalisent cette estimation de densité de façon non paramétrique, en partitionnant chaque variable, en intervalles dans le cas numérique et en groupes de valeurs dans le cas catégoriel. Le produit cartésien de ces partitions univariées forme une partition multivariée de l'espace de représentation, i.e., une grille ou matrice de cellules et il représente aussi un estimateur de densité jointe des variables. La granularité optimale de la grille est établie au moyen d'une approche Bayesienne MAP (Maximum A Posteriori) de la sélection de modèles, et la meilleure grille est recherchée au moyen d'algorithmes d'optimisation combinatoire.

La construction du critère permettant de générer la structure du coclustering, ainsi que l'algorithme d'optimisation et les propriétés asymptotiques de l'approche sont détaillés dans (Boullé, 2011) pour le cas d'un coclustering à deux dimensions catégorielles et dans (Boullé, 2012) pour le cas de données mixtes, i.e numériques et catégorielles. (Boullé, 2012) a démontré que l'approche se comporte comme un estimateur universel de densité jointe convergeant asymptotiquement vers la vraie distribution.

L'approche MODL permet de réaliser un coclustering en d dimensions. Le problème est que le nombre d'observations nécessaires pour peupler une grille de coclustering croît exponentiellement avec le nombre de dimensions (ainsi, il faut plusieurs millions d'observations pour inférer une structure dans le cas d'un 5-clustering (Guigourès, 2013)). On peut cependant se ramener à un coclustering de deux variables catégorielles au moyen d'un recodage simple de la table. Chaque variable numérique est recodée comme une variable catégorielle (par exemple

3. www.khiops.com

en partiles). Un individu est décrit par la liste des modalités qu'il prend sur chaque variable catégorielle. Chaque individu peut maintenant être vu comme un texte décrit par un vocabulaire formé par les modalités. L'analyse peut se dérouler via un coclustering entre l'ensemble des individus d'une part et l'ensemble de ces modalités d'autre part. Le problème que l'on résout maintenant se ramène au clustering simultané des individus et des modalités de variables.

2.2 K-anonymat

Le k-anonymat est un modèle de protection dédié à la prévention de la ré-identification et bien adapté à la publication des tables multidimensionnelles (Samarati, 2001). Il permet de produire une table protégée qui respecte le même format que les données de départ, mais avec un appauvrissement de la distribution des données. Il suppose de distinguer les attributs d'une table selon :

— les identifiants qui identifient directement et de manière unique un individu (numéro de sécurité sociale, de téléphone, adresse mac d'un terminal). La préconisation est de ne pas diffuser ces informations ;
— le quasi identifiant (QI) est l'ensemble des attributs qui, combinés entre eux, et associés à des données externes permettent d'identifier un individu (par exemple {date de naissance, sexe, code postal} croisés avec des données démographiques publiques) ;
— les attributs sensibles sont ceux qui touchent à l'intimité de la personne et donc à sa vie privée et qui utilisés peuvent conduire à des discriminations (origine ethnique, croyances religieuses, opinion politique, santé).

Le k-anonymat garantit que pour chaque combinaison du QI il y a au moins k enregistrements qui partagent la même combinaison de valeurs. On ne peut donc pas lier un individu à un individu du fichier protégé mais à un groupe d'individus. La probabilité de ré-identification est au plus de $1/k$. Les algorithmes du k-anonymat sont principalement basés sur des combinaisons de suppressions et de généralisations qui consistent à remplacer les valeurs précises des différents attributs du QI par des ensembles de valeurs (intervalles numériques ou catégories) de façon à ce que la combinaison résultante soit partagée par k individus. Le groupe de k individus décrit par ses variables généralisées est appelé une classe d'équivalence. De nombreux algorithmes permettant d'atteindre le k-anonymat ont été publiés (Ciriani et al., 2007). Ils fonctionnent principalement par partitionnement (LeFevre et al., 2005) et cherchent à obtenir l'algorithme satisfaisant le k-anonymat avec un nombre minimum de généralisations. Des solutions basées sur le clustering ont également été proposées (Torra et al., 2016).

Les tables 1 a) et b) illustrent le principe du k-anonymat. On donne un exemple de table dans laquelle les atributs *code zip*, *age* et *nationalité* forment le QI et l'attribut *diagnostic* est l'attribut sensible. On donne la table originale et la table transformée pour respecter le 4-anonymat. L'*age* a été généralisé en 3 groupes de valeurs. Une partie de l'information *code zip* a été masquée. Les modalités de *nationalité* ont été regroupées en une seule et supprimées.

Le k-anonymat protège contre la ré-identification mais ne résiste pas à la divulgation d'attributs sensibles (ainsi dans l'exemple de la table 1b) 4-anonyme, l'attaquant peut observer que tous les individus de la dernière classe d'équivalence ont le cancer). Différentes techniques ont été développées pour pallier ce défaut du k-anonymat. Elles visent à imposer une diversité minimale aux valeurs sensibles des classes d'équivalence, afin de limiter ce que peut apprendre l'attaquant ayant différentes connaissances sur sa cible (l-diversité, (Machanavajjhala et al., 2007)) ou encore à créer des classes d'équivalence au sein desquelles la distribution des don-

nées sensibles est à peu près la même que dans la population globale (t-proximité, (Li et al., 2007)).

Il existe maintenant des outils open source pour l'anonymisation tels qu'ARX [4] qui implémente différentes méthodes (k-anonymat, l-diversité, t-proximté, etc). Une phase de préparation des données est nécessaire. Il faut en effet spécifier, pour chaque attribut du QI, la hiérarchie de généralisation appropriée, ce qui suppose une bonne connaissance métier. Cette phase peut s'avèrer très couteuse pour l'utilisateur. On trouve peu d'exemples d'application du k-anonymat au delà de k=10 (Prasser et al., 2016). Quand le nombre de dimensions augmente, l'agrégation détruit l'information contenue dans les données qui deviennent peu exploitables.

(a) table originale

	code zip	age	nationalité	diagnostic
1	13053	28	russe	trouble cardiaque
2	13068	29	américain	infection virale
3	13068	21	japonais	trouble cardiaque
4	13053	23	américain	infection virale
5	14853	50	indien	cancer
6	14853	55	russe	trouble cardiaque
7	14850	47	américain	infection virale
8	14850	49	américain	infection virale
9	13053	31	américain	cancer
10	13053	37	indien	cancer
11	13068	36	japonais	cancer
12	13068	35	américain	cancer

(b) table 4-anonyme

	code zip	age	nationalité	diagnostic
1	130**	<30	*	trouble cardiaque
2	130**	<30	*	infection virale
3	130**	<30	*	trouble cardiaque
4	130**	<30	*	infection virale
5	1485*	>40	*	cancer
6	1485*	>40	*	trouble cardiaque
7	1485*	>40	*	infection virale
8	1485*	>40	*	infection virale
9	130**	[30-40]	*	cancer
10	130**	[30-40]	*	cancer
11	130**	[30-40]	*	cancer
12	130**	[30-40]	*	cancer

TAB. 1 – *Table originale a) et table 4-anonyme correspondante b). Le QI est formé par les atributs {code zip,age,nationalité}, diagnostic est l'attribut sensible. Le k-anonymat de la table est obtenu par des combinaisons de généralisations et de suppressions.*

4. http ://arx.deidentifier.org développé à l'université de Munick

2.3 Principe technique de la solution

La solution proposée consiste à s'appuyer sur la technique du coclustering rappelée ci-dessus pour obtenir le k-anonymat d'une table multidimensionnelle individus x variables. On notera que pour produire un coclustering informatif sur un jeu de données, il faut suffisamment d'exemples. La méthodologie d'anonymisation que l'on décrit dans cet article est dédiée à des ensembles de données plutôt grands. Les différentes phases de la méthodologie sont décrites ci-dessous :

1. **préparation des données**

   La phase de préparation consiste à recoder toutes les variables de la table comme des variables catégorielles.
   — s'il s'agit déjà d'une variable catégorielle, on peut la conserver en l'état ou limiter le nombre de ses modalités,
   — si c'est une variable numérique, on recode en partiles (déciles, centiles, etc) en fonction de la quantité de données.

   Un individu est décrit par la liste des modalités qu'il prend sur chaque variable catégorielle.

2. **coclustering**

   Le coclustering est appliqué sur la nouvelle représentation des données entre l'ensemble des individus d'une part et l'ensemble des modalités d'autre part. On obtient une hiérarchie de clusters d'individus dont chaque niveau vise à maximiser l'information retenue. On obtient également une hiérarchie duale des modalités.

3. **simplification du coclustering**

   On remonte dans cette hiérarchie jusqu'à ce que tous les clusters d'individus soient peuplés de plus de k individus.
   — chaque individu est représenté par son cluster d'appartenance,
   — un cluster d'individus est représenté par une densité de probabilité sur les clusters de modalités

   A ce stade, on a une représentation de clusters d'individus k-anonymes (en ce sens qu'un individu est fondu dans un groupe de k et n'est plus décrit que par une densité de probabilité sur les clusters de modalités).

4. **génération des données synthétiques**

   Le cocustering est également un modèle dont on peut se servir notamment pour affecter un nouvel individu à son cluster d'appartenance. Le coclustering peut aussi être vu comme un modèle générateur qui permet de générer des individus du même format que les individus de départ en s'appuyant sur les densités de probabilités. Pour chaque individu que l'on veut construire, on a adopté la stratégie suivante :
   — sélectionner un cluster d'individus
      — pour simuler un individu correspondant à ce cluster d'individus, tirer une modalité pour une première variable selon la distribution des populations dans les coclusters. Dans chaque cocluster on dispose de la distribution des comptes sur chacune des modalités de variable du cocluster). On respecte la contrainte qui est qu'un individu simulé ne peut recevoir qu'une modalité par variable,

— continuer le tirage des modalités de variables tant que l'individu n'a pas peuplé toutes ses variables. Ce sont des tirages avec remise dans les effectifs des modalités de variables.
— procéder de même pour l'ensemble des clusters et des individus à simuler.

On obtient ainsi des individus synthétiques qui ne sont plus les individus empiriques.

3 Expérimentation

La méthodologie proposée est déroulée pour l'anonymisation d'une table de données multidimensionnelles. On évalue l'utilité des données synthétiques produites sur deux tâches de fouille : une tâche de classification supervisée et une tâche d'analyse exploratoire. Le niveau de protection offert contre la ré-identification des individus est également évalué.

3.1 Conditions de l'expérimentation

On expérimente avec la base de données Adult [5] qui contient 48842 observations décrites par 14 variables numériques et catégorielles, parmi lesquelles on retient les variables {age, workclass, education, education num, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country}. On réserve 20% des observations (choisies de manière aléatoire) pour les évaluations et on utilise les 80% restants pour la mise en œuvre du coclustering et la génération des données synthétiques.

Préparation des données. Pour la phase de préparation des données, les variables numériques sont recodées en déciles, les variables catégorielles ne sont pas modifiées.

Coclustering. La grille la plus fine du coclustering est obtenue pour 34 clusters d'individus et 58 clusters de modalités. A ce niveau, le cluster d'individus le moins peuplé compte 500 individus, le cluster le plus peuplé 950. On peut remonter dans la hiérarchie du coclustering jusqu'à ce que tous les clusters soient peuplés du nombre k d'individus désiré ; ainsi pour obtenir $k = 1300$ il faut remonter jusqu'à 15 clusters d'individus. On obtient une représentation plus grossière des clusters d'individus et des clusters de modalités. La figure 1 donne les populations min et max des clusters d'individus obtenus à un niveau donné de la hiérarchie (avec en abscisse le nombre de clusters d'individus et en ordonnée les populations correpondantes).

Génération des ensembles de données synthétiques. Pour les besoins de l'expérimentation on génère une table d'individus synthétiques à différents niveaux d'agrégation du coclustering. Les niveaux choisis : {34, 30, 25, 20, 15, 10, 5, 1} clusters d'individus. A chaque niveau, la table générée contient autant d'individus qu'il y en avait dans la table initiale.

3.2 Evaluation de l'utilité des données synthétiques

On s'intéresse maintenant à l'exploitation des données synthétiques générées. La question que l'on se pose est : les données synthétiques sont elles représentatives des données réelles et peuvent elles être utilisées pour la fouille de la même manière que celles-ci ?

5. https ://archive.ics.uci.edu/ml/

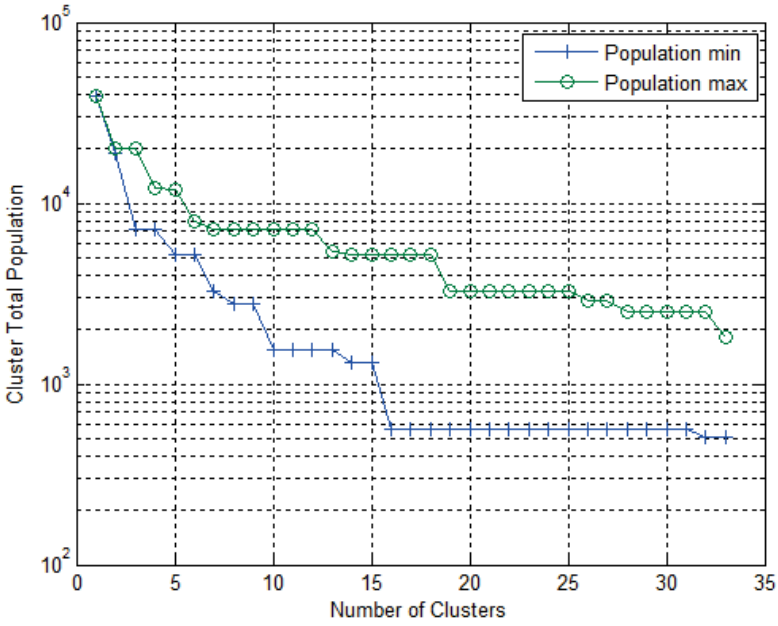


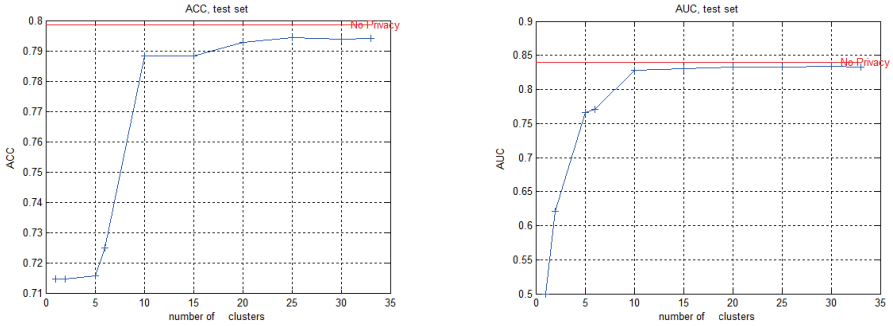
FIG. 1 – *Population min et max des clusters d'individus pour différents niveaux d'agrégation du coclustering.*

Classification supervisée La première tâche de fouille à laquelle on s'intéresse est une tâche de classification. On sélectionne une variable qui sera la cible, les autres variables constituant les variables explicatives. On a utilisé l'outil d'apprentissage supervisé de la suite logicielle khiops qui implémente un classifieur bayésien naïf avec sélection de variables et moyennage de modèles. Khiops supervisé et khiops coclustering sont téléchargeables [6].

Deux classifieurs sont appris : le premier avec l'ensemble des données synthétiques générées à un niveau donné du coclustering, le second avec les données réelles. Puis les deux modèles sont déployés successivement sur les données de test réelles qui avaient été mises de coté précédemment. On évalue ainsi les performances des classifieurs sur les «vrais individus ». Les critères qui sont évalués sont le taux de bonne classification (ACC) et l'aire sous la courbe de ROC (AUC). On présente figure 2 les résultats expérimentaux obtenus avec la variable cible *education* dont les différentes modalités ont été réparties en 2 classes {études supérieures ou non}. 11 variables explicatives ont été retenues pour l'apprentissage des modèles {age, workclass, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country}. On donne en abscisse les différents niveaux de coclustering expérimentés, et en ordonnée l'ACC figure 2 a) et l'AUC figure 2 b) obtenus sur le fichier de test. On indique également les performances obtenues quand ce sont les données réelles qui sont utilisées pour l'apprentissage du modèle (no privacy). La figure 1 fait le lien entre le

6. https ://khiops.predicsis.com/

nombre de clusters d'individus retenus pour construire un ensemble synthétique, à un niveau de la hiérarchie du coclustering et le niveau de k-anonymat correspondant.



(a) ACC (b) AUC

FIG. 2 – *ACC a) et AUC b) obtenus sur les données de test réelles. Pour un niveau donné du coclustering (donné par le nombre de clusters d'individus) le modèle a été appris à partir des données synthétiques générées à ce niveau.*

On observe que les performances de classification obtenues sur les données de test réelles, à partir des modèles appris sur les ensembles synthétiques sont proches de celles que l'on obtient quand ce sont les données réelles qui sont utilisées pour l'apprentissage du modèle. Les performances se dégradent quand le niveau d'agrégation du coclustering devient élevé.

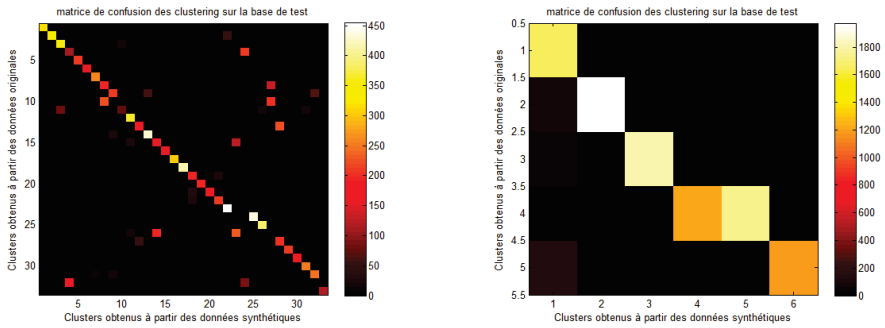
Différentes variables cibles ont été évaluées de la même manière (genre et état marital). On a observé un comportement de classification similaire à celui présenté ci-dessus. On en conclut que les données synthétiques générées conservent bien les propriétés des données réelles.

Analyse exploratoire On s'intéresse maintenant à une seconde tâche de fouille, une tâche d'analyse exploratoire. La question que l'on se pose ici est : quelle connaissance peut on extraire d'un ensemble de données synthétiques et cette connaissance se compare t'elle à celle que l'on obtiendrait sur les données réelles ? Le protocole expérimental est le suivant :

— construction d'un coclustering à partir des données synthétiques,
— déploiement du coclustering sur les données de test, le déploiement consiste à affecter à chaque individu de l'ensemble de test son cluster d'appartenance,
— déploiement du coclustering ayant servi à construire l'ensemble de données synthétiques sur les même données de test.

On peut maintenant comparer les 2 coclusterings, celui construit à partir des données réelles et celui construit à partir des données synthétiques, en traçant la matrice de confusion croisant les 2 coclusterings. La figure 3 donne les tables de confusion obtenues pour 2 jeux synthétiques obtenus à 2 niveaux d'agrégation du coclustering (pour 33 clusters d'indivus 3 a) et 5 clusters d'individus 3 b)) avec le code couleur suivant : une cellule de la matrice de confusion est d'autant plus claire qu'elle est peuplée.

Quel que soit le niveau auquel on s'intéresse, les 2 coclusterings produits «synthétique» et «réel» sont cohérents, avec des tables de confusion relativement diagonales. On retrouve dans



(a) 33 clusters d'individus (b) 5 clusters d'individus

FIG. 3 – *Tables de confusion des coclusterings synthétique et réel.*

les données synthétiques le niveau d'information qu'il y avait dans les données originales. Cependant, les données synthétiques ne conservent pas d'information à un niveau plus fin que celui auquel elles ont été générées. Elles ne permettent pas de construire un coclustering de niveau plus fin. On n'apprend donc pas de nouvelle information à travers l'analyse car il n'est pas possible de descendre en dessous du niveau prédéfini par le jeu de données synthétiques.

3.3 Protection des individus

Pour évaluer le niveau de protection contre la ré-identification d'un ensemble de données synthétiques. On se place dans le contexte où un attaquant qui dispose des données synthétiques connaît également toute la base initiale, sauf un individu, et on évalue la capacité de retrouver cet individu à partir des informations dont dispose l'attaquant. On se demande ici dans quelle mesure l'individu est caché à l'attaquant ? Le protocole expérimental est le suivant :

— sélectionner une ligne correspondant à un individu dans le fichier original ; on fait l'hypothèse que l'observation sélectionnée est inconnue de l'attaquant,
— l'attaquant peut faire la correspondance entre les 2 bases, réelle et synthétique, en fonction de la connaissance dont il dispose,
— évaluer le niveau d'incertitude de l'attaquant en comptant à combien d'individus synthétiques il peut attribuer l'observation réelle inconnue,

procéder ainsi pour tous les individus du fichier réel.

Pour un ensemble synthétique, généré à partir du niveau le plus fin du coclustering, en moyenne, une observation réelle inconnue peut être attribuée à 80,6% des exemples synthétiques. Pour un ensemble synthétique généré à partir du coclustering à 10 clusters d'individus, la valeur moyenne est de 86,4%.

4 Conclusion

Cet article a proposé une méthodologie pour anonymiser des données multidimensionnelles individuelles en vue de leur publication. L'approche consiste à coupler k-anonymat et

génération de données synthétiques. Dans une première phase, un coclustering des données partitionne conjointement les individus et les variables descriptives et permet de constituer des groupes d'individus k-anonymes. Obtenir le coclustering optimal ne nécessite aucun paramétrage utilisateur. De plus, pour atteindre le k-anonymat, il n'est pas nécessaire de passer par une étape coûteuse de préparation des données comme dans la plupart des méthodes classiques. Le coclustering est également un modèle générateur sur lequel on s'appuie pour construire des individus synthétiques du même format que les individus initiaux.

On a montré que les données synthétiques conservent les propriétés des données originales et qu'il est donc possible d'envisager leur utilisation pour la fouille.

En termes de protection, on s'est intéressé au risque de ré-identification des individus. On va maintenant étendre la technique pour répondre au risque de divulgation d'attribut sensible en intervenant sur les clusters de modalités de manière à controler l'information divulguée.

Références

Blondel, V. D., M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, et C. Ziemlicki (2012). Data for development : the d4d challenge on mobile phone data. *arXiv preprint arXiv :1210.0137*.

Boullé, M. (2006). An enhanced selective naive bayes method with optimal discretization. In *Feature Extraction*, pp. 499–507. Springer.

Boullé, M. (2011). Estimation de la densité d'arcs dans les graphes de grande taille : une alternative à la détection de clusters. In *EGC*, pp. 353–364.

Boullé, M. (2012). Functional data clustering via piecewise constant nonparametric density estimation. *Pattern Recognition 45*(12), 4389–4401.

Chen, R., Q. Xiao, Y. Zhang, et J. Xu (2015). Differentially private high-dimensional data publication via sampling-based inference. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, New York, NY, USA, pp. 129–138. ACM.

Ciriani, V., S. D. C. di Vimercati, S. Foresti, et P. Samarati (2007). κ-anonymity. In *Secure data management in decentralized systems*, pp. 323–353. Springer.

Cormode, G. (2015). The Confounding Problem of Private Data Release (Invited Talk). In M. Arenas et M. Ugarte (Eds.), *18th International Conference on Database Theory (ICDT 2015)*, Volume 31 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Dagstuhl, Germany, pp. 1–12. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Dwork, C. (2008). Differential privacy : A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pp. 1–19. Springer.

Fung, B., K. Wang, R. Chen, et P. S. Yu (2010). Privacy-preserving data publishing : A survey of recent developments. *ACM Computing Surveys (CSUR) 42*(4), 14.

G29 (2014). Article 29 data protection working party, opinion 05/2014 on anonymization techniques). Technical report, EC (www.http://ec.europa.eu/).

Guigourès, R. (2013). *Utilisation des modèles de co-clustering pour l'analyse exploratoire des données*. Thèse de doctorat, Université Paris 1 Panthéon-Sorbonne.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, et P.-P. De Wolf (2012). *Statistical disclosure control*. John Wiley & Sons.

LeFevre, K., D. J. DeWitt, et R. Ramakrishnan (2005). Incognito : Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 49–60. ACM.

Li, N., T. Li, et S. Venkatasubramanian (2007). t-closeness : Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pp. 106–115. IEEE.

Machanavajjhala, A., D. Kifer, J. Gehrke, et M. Venkitasubramaniam (2007). l-diversity : Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD) 1*(1), 3.

Prasser, F., R. Bild, J. Eicher, H. Spengler, F. Kohlmayer, et K. A. Kuhn (2016). Lightning : Utility-driven anonymization of high-dimensional data. *Transactions on Data Privacy 9*(2), 161–185.

Samarati, P. (2001). Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering 13*(6), 1010–1027.

Sweeney, L. (2002). K-anonymity : A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10*(05), 557–570.

Torra, V., G. Navarro-Arribas, et K. Stokes (2016). An overview of the use of clustering for data privacy. In *Unsupervised Learning Algorithms*, pp. 237–251. Springer.

Vilhuber, L., J. M. Abowd, et J. P. Reiter (2016). Synthetic establishment microdata around the world. *Statistical Journal of the IAOS 32*(1), 65–68.

Zhang, J., G. Cormode, C. M. Procopiuc, D. Srivastava, et X. Xiao (2014). Privbayes : Private data release via bayesian networks. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pp. 1423–1434. ACM.

Summary

In tis paper we propose a methodology to anonymize multidimensional individual data. The goal is to be able to protect data against the reidentification risk. The proposed solution is based on a coclustering method. The coclustering is used to build an aggregated representation of the data, then the model is used to draw synthetic individual data. We show that these synthetic data preserve sufficient information to be used in place of the real data. Finally the protection against the reidentification risk is evaluated.