Anonymisation de trajectoires d'appels mobiles à l'aide de modèles en grilles et de chaines de Markov

Françoise Fessant*, Fabrice Clérot *
Koray Ozbek *,**

*Orange Labs, Lannion
francoise.fessant, fabrice.clerot@orange.com,
**Université Limoges
koray.ozbek@hotmail.fr

Résumé. Dans cet article, nous proposons une méthodologie pour anonymiser un ensemble de trajectoires individuelles d'appels mobiles. L'objectif est de publier un ensemble de trajectoires anonymes construit à partir de l'ensemble initial, qui protège contre le risque de ré-identification. En d'autres termes, on ne doit pas pouvoir lier une trajectoire dans les données publiées à un individu présent dans la base originale. La solution proposée associe la segmentation des trajectoires d'appels à l'aide de modèles en grille, à un modèle générateur basé sur les chaines de Markov, pour produire des trajectoires d'appels synthétiques qui peuvent ensuite être publiées à la place des trajectoires originales. L'utilité des données synthétiques, ainsi que le niveau de protection offert sont évalués à travers différents indicateurs statistiques.

1 Introduction

La multiplication des équipements connectés, omniprésents dans notre vie quotidienne et l'adoption des services basés sur la géolocalisation, rendent possible le traçage numérique d'une grande partie de nos activités et déplacements. Tweeter au sujet d'un évènement, rechercher un itinéraire avec le système de navigation de son véhicule, appeler quelqu'un avec son téléphone mobile, payer avec sa carte bancaire sont des exemples de situations qui génèrent automatiquement des grandes masses de données qui sont collectées et stockées dans des bases de données. L'analyse de ces données est l'objet d'un intérêt soutenu de la part de différents acteurs : compréhension des mobilités humaines pour le transport (gestion du trafic routier) ou les collectivités locales (fréquentation des sites, (Jiang et al., 2017)), le géomarketing (Giannotti et al., 2009) ou l'aide aux politiques publiques (Blondel et al., 2012). Ces analyses reposent essentiellement sur la collecte de trajectoires de mobilité, c'est à dire l'historique horodaté des positions géographiques successives visitées par un individu. Cependant, l'exploitation ou la publication de telles données pose des questions relatives à la protection des données à caractère personnel. En effet, connaitre les déplacements d'une personne peut permettre d'inférer par exemple ses lieux de résidence ou de travail, son identité, ses centres d'intérêts, ses habitudes, ses préférences politiques et sa santé voire une déviation par rapport à son comportement habituel (Gambs et al., 2011).

Ce problème a initialement été adressé par les techniques de pseudo-anonymisation qui consistent à remplacer les champs directement identifiant (nom, numéro de téléphone, etc.) par un nouveau champ (en général au moyen d'une fonction de hachage) qui rend impossible la corrélation entre lui et l'individu. Mais différentes études ont montré que les techniques de pseudo-anonymisation ne suffisent pas à protéger contre la réidentification. Par exemple, dans une base de données de géolocalisation, quatre points spatio-temporels suffisent pour identifier une personne (Montjoye et al., 2013). Les données pseudonymisées sont toujours considérées comme des données personnelles. Elles restent donc soumises au règlement européen sur la protection des données personnelles (RGDP [1]) et 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]

Dans le cadre de cet article nous détaillons une méthodologie d'anonymisation des trajectoires d'appels mobiles, dans un objectif de publication de données individuelles. Le risque contre lequel on souhaite se protéger est le risque de ré-identification. Pour cela on va générer automatiquement des trajectoires synthétiques d'appels à partir d'un ensemble réel de trajectoires. La solution proposée associe la segmentation des trajectoires d'appels à l'aide de modèles en grille, à un modèle générateur basé sur les chaines de Markov. Ce travail représente une extension de la méthodologie présentée dans (Fessant et al., 2017) pour la publication de données multidimensionnelles. Après avoir décrit les différentes étapes de la solution, on présente des résultats expérimentaux obtenus à partir de comptes rendus d'appels mobiles (CRA) réels. On évalue notamment l'utilité des données synthétiques à travers différents indicateurs statistiques, ainsi que le niveau de protection apporté.

2 Travaux connexes

La littérature sur le domaine de la publication des trajectoires spatio-temporelles respectueuse de la vie privée s'organise essentiellement autour de deux notions de protection : la ré-identification et l'inférence de nouvelles connaissances à partir des données publiées. Les techniques de k-anonymat (Sweeney, 2002) sont très largement utilisées pour protéger contre le risque de ré-identification. Le principe du k-anonymat est de faire en sorte que chaque trajectoire, considérée dans son ensemble, ne puisse pas être distinguée de k-1 autres trajectoires. On ne peut donc pas lier un individu à un individu du fichier à protéger, mais à un groupe d'individus. La probabilité de ré-identification est au plus de 1/k. Le k-anonymat est atteint par des techniques combinant généralisation (avec la granularisation de l'espace et du temps) et suppression de trajectoires ou de parties de trajectoires. Les algorithmes se distinguent essentiellement par leur stratégie de généralisation ou de suppression. Dans GLOVE (Gramaglia et Fiore, 2015) proposent une généralisation spécifique pour chaque point spatiotemporel selon le coût que représente cette généralisation pour l'utilité des données. (Nergiz et al., 2008) suppriment les points de la trajectoire dont la généralisation est trop couteuse en termes d'utilité. Le k-anonymat peut également être atteint par des techniques de clustering, comme la micro agrégation (Domingo-Ferrer et Trujillo-Rasua, 2012) dans laquelle les trajectoires sont d'abord partitionnées en clusters de cardinalité k. Chaque cluster est ensuite représenté par

1. https ://www.cnil.fr/fr/textes-officiels-europeens-protection-donnees

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.

une trajectoire prototype. Le travail porte essentiellement sur la métrique de similarité utilisée (Abul et al., 2010), (Trujillo-Rasua et Domingo-Ferrer, 2015). (Gramaglia et Fiore, 2015) ont analysé l'impact de la forte unicité des trajectoires sur l'utilité des données anonymes produites et le coût nécessaire pour atteindre un k-anonymat avec $k > 2$.

Pour protéger contre l'inférence on fait appel aux techniques de confidentialité différentielle (ou DP). 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 calculé sur les données publiées (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). Dans le contexte de la publication de trajectoires spatio-temporelles les propositions consistent à générer des trajectoires synthétiques qui conservent les propriétés des trajectoires originales. Par exemple, dans DP-Where (Mir et al., 2013) le modèle générateur est construit à partir de l'estimation de différentes distributions de probabilités reflétant les caractéristiques principales des déplacements des individus. Celles-ci sont obtenues à partir des données originales et bruitées pour respecter la DP. Une approche similaire DP-Star est proposée dans (Roy et al., 2016) avec d'autres combinaisons de caractéristiques estimées et bruitées. Les données synthétiques produites servent au calcul de statistiques globales, la perte d'information étant trop importante pour une analyse individuelle. Pour une revue de littérature récente sur l'anonymisation de trajectoires spatio-temporelles on peut se reporter à (Fiore et al., 2019).

3 Description de la méthodologie d'anonymisation

Les différentes étapes de la solution technique proposée sont i) un clustering des trajectoires d'appels, ii) la représentation de la dynamique des trajectoires d'un cluster par une chaine de Markov, iii) la génération de données synthétiques à partir du modèle de mobilité obtenu. On rappelle les principes des modèles en grilles et des chaines de Markov, avant de décrire la solution technique. L'objectif du coclustering est de capter les différents comportements d'appels, celui de la chaine de Markov de modéliser ces comportements pour les différents clusters. On adopte ici une démarche similaire à celle de (Bondu et Dachraoui, 2015) qui se sont intéressés à la simulation de séries temporelles de consommations électriques.

3.1 Coclustering

Les modèles en grille. Le coclustering est une technique qui a pour but de réaliser une partition simultanée des lignes et des colonnes d'une matrice de données. 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,

3. www.khiops.com

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 Bayésienne 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.

Coclustering pour les trajectoires d'appels mobiles. Dans le cadre de ce travail, on s'intéresse plus spécifiquement aux trajectoires d'appels mobiles enregistrées de manière automatique par les opérateurs de téléphonie mobile pour des raisons légales ou de facturation. Chaque évènement qui se produit entre le réseau et le terminal mobile (appel entrant ou sortant, envoi ou réception d'un SMS) est collecté avec son estampille temporelle, ainsi que les coordonnées spatiales de l'antenne avec laquelle le mobile s'est connecté. La trajectoire d'appels est formée par la séquence des évènements (les comptes rendus d'appels mobiles ou CRA) ordonnés dans le temps. Chaque trajectoire d'appels mobiles peut donc contenir un nombre variable d'évènements. La base de données de facturation fournit une information à l'échelle de l'antenne, uniquement quand le téléphone est en communication. La position de l'individu n'est connue qu'en cas de communication, au niveau de l'antenne.

On notera S_i une trajectoire contenant m_i évènements d'appels caractérisés eux mêmes par 2 variables : T qui représente l'estampille temporelle (ou timestamp) et Z l'identifiant de la zone géographique à laquelle appartient l'antenne de connexion. La ième trajectoire d'appels est notée : $S_i = (t_{ij}, z_{ij})_{j=1}^{m_i}$. L'approche MODL suppose un recodage de l'ensemble des trajectoires sous la forme d'une ligne par évènement d'appel, chaque ligne étant décrite par 3 informations (C, T, Z) où C est l'identifiant de la trajectoire originale. On dispose au final pour l'analyse d'une table d'autant de lignes qu'il y a d'évènements au total contenus dans l'ensemble des trajectoires. Le triclustering de cette table produit simultanément un clustering des trajectoires, un clustering des zones géographiques et une discrétisation des timestamps, ainsi qu'une estimation de la densité jointe des variables $P(C, T, Z)$. Etant donné que $P(T, Z/C) = \frac{P(C,T,Z)}{P(C)}$, le modèle peut également être interprété comme un estimateur de la densité jointe entre T et Z, qui est constante pour chaque cluster de trajectoires. Le schéma de la figure 1 illustre le principe du triclustering appliqué aux trajectoires spatio-temporelles.

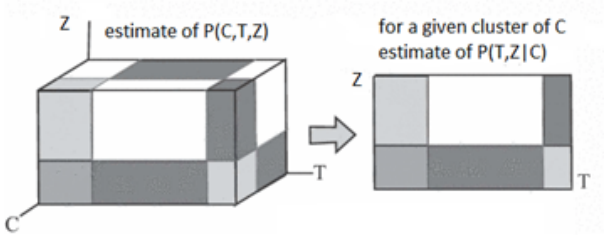


FIG. 1 – *Illustration d'un modele de triclustering appliqué aux trajectoires d'appels mobiles.*

3.2 Modélisation de la mobilité avec les chaines de Markov

Les chaines de Markov permettent de représenter de manière compacte le comportement de mobilité d'un individu (Gambs et al., 2012) ou d'un groupe d'individus (ASAH 2011). Un modèle de mobilité de Markov est composé de i) un ensemble d'états, ii) un ensemble de transitions qui représentent la probabilité de se déplacer de l'état *i* à l'état *j*. Le modèle de mobilité est représenté par la matrice des probabilités des transitions entre états. Les lignes de la matrice correspondent aux états d'origine et les colonnes aux états de destination. La valeur de la cellule correspondant à la probabilité de transition associée. La matrice peut contenir des 0 correspondant à l'absence de transition observée entre les états. La probabilité de se déplacer dans un état dépend uniquement de l'état courant et de la distribution de probabilité des transitions entre états.

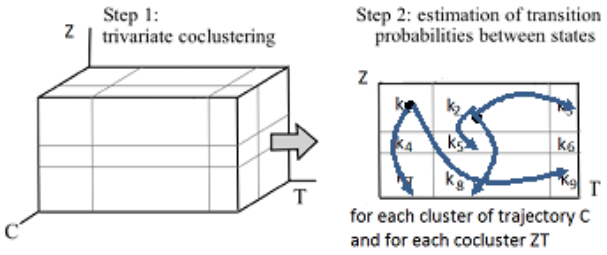


FIG. 2 – *Exploitation des chaines de Markov et des modèles en grille pour modéliser la dynamique des trajectoires. Le label d'un cocluster k_i correspond à un ensemble de zones géographiques/ tranche horaire.*

3.3 Approche pour la génération des trajectoires synthétiques

Le modèle générateur proposé associe l'approche MODL à une chaine de Markov. Les différentes phases de la méthodologie d'anonymisation sont décrites ci-dessous :

1. **Partitionnement des trajectoires originales**
   Lors de cette première phase, les trajectoires d'appels originales sont partitionnées à l'aide d'un triclustering qui capte et structure les différents comportements individuels. On obtient simultanément un clustering des trajectoires, un clustering des zones géographiques, une discrétisation des timestamps en plages horaires. Le triclustering est donc utilisé ici pour résumer les données et permettre de construire un modèle de mobilité par cluster d'individu qui reflète les différentes dynamiques d'appels, ce que ne permettrait pas un modèle global.
2. **Construction du modèle de mobilité**
   Pour chaque cluster de trajectoires, on s'intéresse plus spécifiquement à la grille bivariée, formée par les coclusters de zones géographique et d'intervalles de temps, qui estime la densité jointe entre les variables Z et T (figure 1). On note k_T et k_Z les nombres

d'intervalles de temps et de groupes de zones géographiques de la grille et k le nombre de cellules (coclusters) de la grille bivariée, $k = k_T k_Z$. On affecte un label k_j à chaque cocluster avec $j \in [1, k_T k_Z]$. Un cocluster correspond simultanément à un intervalle de temps et à un groupe de zones géographiques. Ces coclusters vont nous servir à définir les $\{k\}$ états du modèle de Markov, un pour chaque cocluster. Une trajectoire d'appels S_i appartenant au cluster de trajectoires c peut maintenant être recodée par une séquence d'états $\{k^i\}$. On affecte à chaque CRA (ou élément de la trajectoire) son cocluster dans la grille bivariée (i.e. son état). On estime ensuite une matrice des transitions entre états pour chaque cluster de trajectoires. Les dimensions de la matrice sont fonction du nombre d'états : $k \times k$. Cette matrice contient les probabilités conditionnelles $P(k_{i,t+1}/k_{j,t)})$. Selon les hypothèses de Markov, la probabilité d'observer un état particulier k_i à un instant donné dépend uniquement de l'état à l'instant précédent k_j. En pratique, les probabilités conditionnelles sont obtenues en comptant les transitions d'un cocluster de la grille bivariée à l'autre dans les données réelles (pour deux appels successifs d'une même trajectoire d'appels, on compte simplement le nombre de fois où un appel est affecté au cocluster correspondant à l'état k_j et le suivant au cocluster correspondant à l'état k_i. On incrémente les comptes pour toutes les trajectoires du cluster. Les comptes sont ensuite normalisés en ligne pour obtenir la matrice des probabilités conditionnelles. Le modèle de Markov associé à chaque cluster de trajectoires cherche à apprendre la dynamique de mobilité du cluster. La figure 2 illustre le principe de modélisation.

3. **Génération des trajectoires synthétiques**
   Le but est maintenant de générer une trajectoire synthétique à partir du modèle de mobilité. Pour cela on va initialiser le premier élément de la trajectoire synthétique à simuler avec le premier état de la trajectoire originale correspondante, ainsi qu'une longueur de trajectoire attendue similaire à celle de la trajectoire initiale. Ce premier état correspond à l'un des coclusters de la grille bivariée (ou états du modèle de mobilité, figure 2). On infère ensuite successivement les coclusters suivants à partir du modèle de Markov jusqu'à ce que la longueur de la trajectoire soit atteinte. Pour transformer la séquence d'états en séquence d'appels individuels, on tire aléatoirement pour chaque état une zone géographique d'antenne et un timestamp dans la distribution des appels du cocluster correspondant.

On procède ainsi pour l'ensemble des clusters et des trajectoires à simuler. On obtient des trajectoires synthétiques qui ne sont plus les trajectoires empiriques et qui ne correspondent donc plus à des individus réels. L'architecture de la solution est schématisée figure 3.

4 Expérimentation

La méthodologie proposée ci-dessus est déroulée pour l'anonymisation d'une base de Comptes Rendus d'Appels (CRA) mobiles réels issue du système d'information d'Orange.

4.1 Conditions de l'expérimentation

Les données de l'étude représentent une journée de CRA de la région de Lyon correspondant à 780 000 clients, répartis sur 1180 antennes. Les antennes sont regroupées en 117

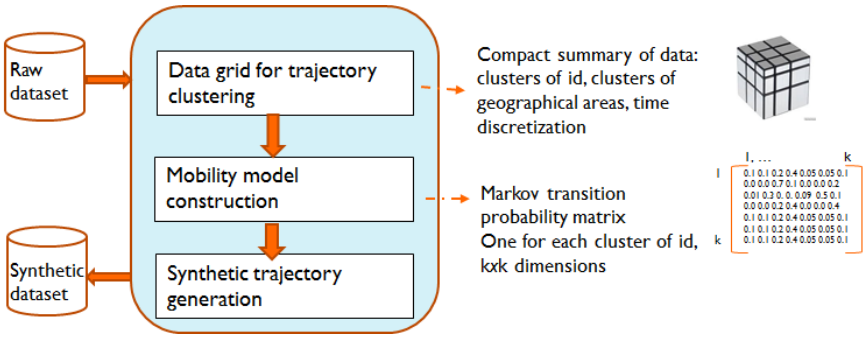


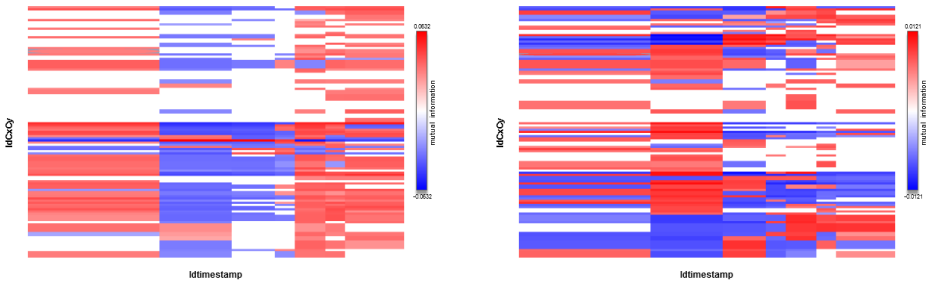
FIG. 3 – *Représentation de l'architecture de la solution de génération de trajectoires synthétiques.*

zones géographiques repérées par un identifiant ; une zone peut contenir de 1 à 4 antennes. L'affectation des antennes à leur zone géographique a été réalisée en amont au moyen d'un premier coclustering sur les coordonnées spatiales des antennes (processus non détaillé ici). Pour l'analyse on n'a conservé que les individus qui faisaient au minimum 2 CRA et le temps a été recodé en minutes. Le fichier contient au total 6 761 571 CRA. Le nombre moyen de CRA par trajectoire est de 9 (couvrant une échelle de 2 à 437), 50 % des trajectoires ont moins de 5 CRA.

4.2 Triclustering des trajectoires d'appels

On dispose donc pour l'analyse d'une table de 6 761 571 de lignes (une par CRA) décrites par 3 variables : 2 catégorielles (les identifiants de la trajectoire C et de la zone géographique de l'antenne Z) et une numérique (le temps T en minutes sur la journée d'appels).

Le triclustering le plus fin produit 336 clusters de trajectoires, 81 clusters de zones géographiques d'antennes et 7 intervalles de temps. Chaque cluster de trajectoires est représenté par une grille bivariée de 81×7 coclusters qui estime la densité jointe $P(ZT/C)$. Le cluster le moins peuplé contient 900 trajectoires. Les 7 tranches horaires sont les suivantes : $[0 - 8h30]$, $[8h30 - 13h]$, $[13h - 15h45h]$, $[15h45 - 17h]$, $[17h - 19h]$, $[19h - 20h15]$,$[20h15 - 24h]$. La figure 4 présente deux portraits de clusters de trajectoires qui montrent deux types de comportements d'appels différents, selon les localisations géographiques des antennes (sur les lignes de la grille) et les tranches horaires (sur les colonnes). On affiche la contribution de chaque cocluster à l'information mutuelle avec en rouge une contribution positive (excès d'appels), en bleu une contribution négative (déficit d'appels), en blanc une absence de contribution. Le premier portrait (figure 4.a) est essentiellement caractérisé par une absence d'appels sur les deuxième, troisième et quatrième tranches horaires pour la plupart des zones géographiques. Sur le deuxième portrait (figure 4.b) on observe une activité d'appel plus intense sur la deuxième tranche horaire $[8h30 - 13h]$ et certaines zones géographiques essentiellement actives en soirée sur les 3 dernières tranches horaires.



(a) (b)

FIG. 4 – *2 portraits a) et b) de clusters de trajectoires d'appels, avec les intervalles de temps en colonne et les regroupements de zones géographiques sur les lignes.*

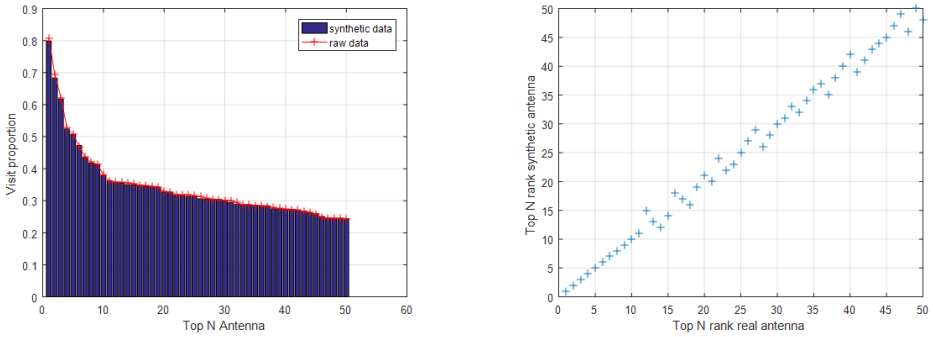
4.3 Qualité des trajectoires synthétiques

On s'intéresse maintenant à l'exploitation des données synthétiques générées. La question que l'on se pose est : les données synthétiques sont-elles réalistes et peuvent-elles être utilisées de la même manière que les données réelles ? L'utilité des données est évaluée sur différentes statistiques, globales et de mobilité.

Dans chaque cluster de trajectoires on a généré autant de trajectoires synthétiques qu'il y avait de trajectoires réelles, selon la méthodologie présentée section 2. Pour construire le modèle de mobilité du cluster, on s'est appuyé sur la grille bivariée formée des 81×7 coclusters (qui constituent les états du modèle) et les transitions entre ces coclusters pour peupler la matrice de transition. On reporte figure 5a les populations de CRA observées à partir des trajectoires synthétiques (en bleu) pour les 50 antennes les plus fréquentées (en proportion du nombre de CRA) et en rouge pour les trajectoires originales. On observe que les distributions des 2 populations sont proches. On donne également figure 5b l'ordre des antennes les plus peuplées pour les données réelles (en abscisse) et synthétiques (en ordonnée). Une diagonale indiquerait une correspondance parfaite entre les 2 populations. On observe une bonne correspondance entre données réelles et synthétiques.

La figure 6 donne l'histogramme des timestamps calculé à partir de l'ensemble des appels passés sur la journée d'analyse pour les données réelles (en rouge). On affiche également la même information pour les données synthétiques (en bleu). On a observé une légère différence dans les deux distributions pour la dernière période temporelle (timestamp >1215, ie la tranche horaire [20h15-24h]).

La figure 7 donne un premier indicateur de la conservation des informations de mouvement dans les données synthétiques. On affiche les comptes des transitions entre deux états de la matrice de transition pour les transitions les plus peuplées, pour les données réelles (en rouge) et leur correspondance pour les données synthétiques (en bleu).



(a) (b)

FIG. 5 – *Populations des antennes les plus peuplées, synthétiques (en bleu) et réelles (en rouge) a), et correspondance entre les antennes b).*

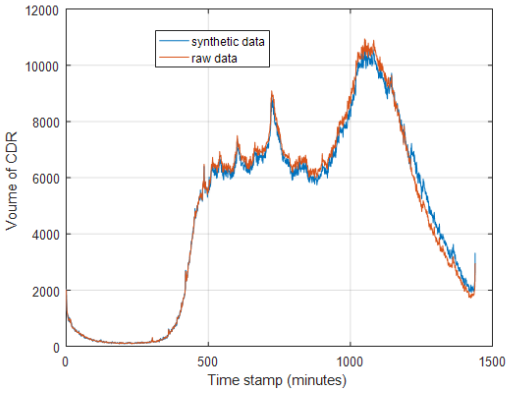


FIG. 6 – *Histogramme des timestamp pour les données réelles (en rouge) et synthétiques (en bleu). Le temps est donné en minutes sur la journée.*

4.4 Protection des individus

L'ensemble synthétique compte 477164 trajectoires différentes, l'ensemble réel 504480. Pour évaluer le niveau de protection des données synthétiques contre la ré-identification, on s'est intéressé aux trajectoires similaires dans les deux ensembles réels et synthétiques. On retrouve dans les données synthétiques 36024 trajectoires identiques á celles des données réelles. La figure 8 donne la distribution des populations de ces trajectoires dans les trajectoires réelles. Au total 17545 trajectoires sont uniques (2,25 % de l'ensemble total des trajectoires) ; la longueur moyenne de ces trajectoires est de 4 items. La préconisation est de supprimer les trajectoires uniques avant publication. Les autres trajectoires similaires sont partagées par 2 indivi-

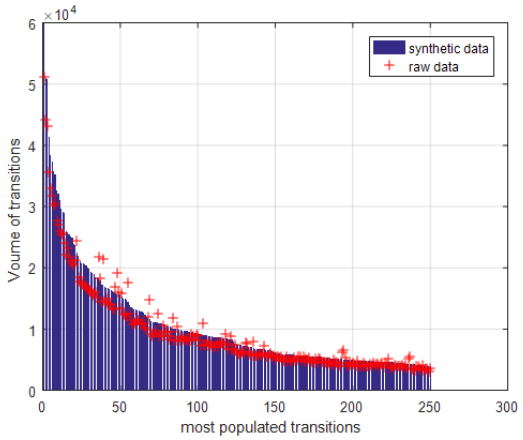


FIG. 7 – *Histogramme des comptes dans la matrice de transitions pour les transitions les plus peuplées, données réelles (en rouge) et synthétiques (en bleu).*

dus ou plus. En termes de protection, un utilisateur peut décider de l'effectif que doit atteindre chaque trajectoire appartenant simultanément aux ensembles réels et synthétiques et supprimer les trajectoires n'atteignant pas l'effectif minimal requis.

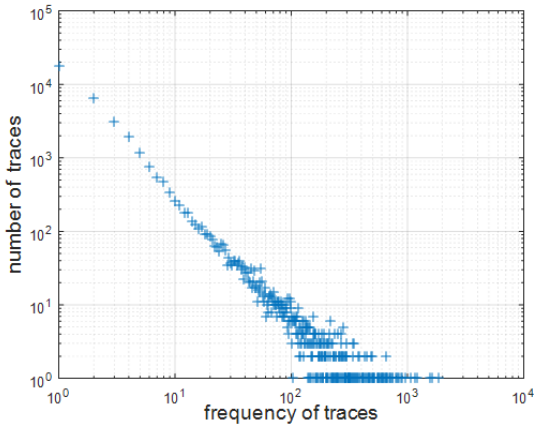


FIG. 8 – *Fréquence des trajectoires similaires des ensembles réels et synthétiques, avec en abscisse la fréquence des trajectoires, en ordonnée leur nombre (17545 trajectoires sont uniques, 1100 trajectoires ont 5 occurrences dans les données rélles, etc).*

5 Conclusion

Dans cet article nous avons proposé, une méthodologie pour anonymiser des trajectoires spatiotemporelles individuelles (trajectoires de comptes rendus d'appels mobiles), en vue de leur publication. L'approche a consisté à associer un modèle de coclustering et modèle de mobilité . Dans une première phase, un triclustering partitionne conjointement les variables descriptives des trajectoires individuelles. Obtenir le coclustering optimal ne nécessite aucun paramétrage utilisateur ni de préparation spécifique des données. Le coclustering est couplé à une chaine de Markov qui reproduit le comportement de mobilité d'un ensemble de trajectoires. On s'appuie sur ce modèle de mobilité pour générer des trajectoires synthétiques du même format que les trajectoires initiales. On a montré à travers le calcul de différents indicateurs que les données synthétiques conservent les statistiques globales des données originales et qu'il est possible d'envisager leur utilisation pour la fouille. Ces résultats nécessitent cependant d'être précisés, notamment par le calcul d'indicateurs de mobilité supplémentaires et confrontés à d'autres approches d'anonymisation de trajectoires. Le niveau de protection apporté par les trajectoires synthétiques vis à vis du risque de réidentification a également été évalué.

Références

Abul, O., F. Bonchi, et M. Nanni (2010). Anonymization of moving objects databases by clustering and perturbation. *Information Systems 35*(8), 884–910.

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*.

Bondu, A. et A. Dachraoui (2015). Realistic and very fast simulation of individual electricity consumptions. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE.

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.

Domingo-Ferrer, J. et R. Trujillo-Rasua (2012). Microaggregation-and permutation-based anonymization of movement data. *Information Sciences 208*, 55–80.

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

Fessant, F., T. Benkhelif, et F. Clérot (2017). Anonymiser des données multidimensionnelles à l'aide du coclustering. In *EGC*, pp. 153–164.

Fiore, M., P. Katsikouli, E. Zavou, M. Cunche, F. Fessant, D. L. Hello, U. M. Aivodji, B. Olivier, T. Quertier, et R. Stanica (2019). Privacy of trajectory micro-data : a survey. *arXiv preprint arXiv :1903.12211*.

Gambs, S., M.-O. Killijian, et M. N. n. del Prado Cortez (2012). Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, MPM '12, New York, NY, USA, pp. 3 :1–3 :6. ACM.

Gambs, S., M.-O. Killijian, et M. Nuñez Del Prado Cortez (2011). Show me how you move and i will tell you who you are. *Transactions on Data Privacy 4*(2), 103–126.

Giannotti, F., M. Nanni, D. Pedreschi, C. Renso, et R. Trasarti (2009). Mining mobility behavior from trajectory data. In *2009 international conference on computational science and engineering*, Volume 4, pp. 948–951. IEEE.

Gramaglia, M. et M. Fiore (2015). Hiding mobile traffic fingerprints with glove. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, CoNEXT '15, New York, NY, USA, pp. 26 :1–26 :13. ACM.

Jiang, S., J. Ferreira, et M. C. Gonzalez (2017). Activity-based human mobility patterns inferred from mobile phone data : A case study of singapore. *IEEE Transactions on Big Data 3*, 208 – 219.

Mir, D. J., S. Isaacman, R. Cáceres, M. Martonosi, et R. N. Wright (2013). Dp-where : Differentially private modeling of human mobility. In *2013 IEEE international conference on big data*, pp. 580–588. IEEE.

Montjoye, Y.-A., C. Hidalgo, M. Verleysen, et V. Blondel (2013). Unique in the crowd : The privacy bounds of human mobility. *Scientific reports 3*, 1376.

Nergiz, M. E., M. Atzori, et Y. Saygin (2008). Towards trajectory anonymization : a generalization-based approach. In *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS*, pp. 52–61. ACM.

Roy, H., M. Kantarcioglu, et L. Sweeney (2016). Practical differentially private modeling of human movement data. In S. Ranise et V. Swarup (Eds.), *Data and Applications Security and Privacy XXX*, Cham, pp. 170–178. Springer International Publishing.

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

Trujillo-Rasua, R. et J. Domingo-Ferrer (2015). *Privacy in Spatio-Temporal Databases : A Microaggregation-Based Approach*, pp. 197–214. Cham : Springer International Publishing.

Summary

In tis paper we propose a methodology to anonymize individual mobility trajectories. 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 a Markov mobility model is designed for each cluster of trajectories. The mobility model is then used to draw synthetic individual trajectories. 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.