Protocole d'agrégation et de déplacement des coordonnées GPS des enquêtes PMA

Contexte

A mesure que les données géo-référencées sont rendues publiques, les travaux de recherche faisant appel à la Géomatique ont connu une croissance rapide au cours des dernières années. Les projets d'enquêtes basées sur les populations (comme par exemple les Enquêtes Démographiques et de Santé, et les Etudes sur la Mesure des Niveaux de Vie) partagent régulièrement des fichiers contenant des coordonnées GPS au grand public. Mais étant donné que les localisations géographiques pourraient être utilisées afin d'identifier des individus, les coordonnées GPS brutes provenant d'enquêtes confidentielles ne peuvent pas être partagées. Le programme EDS a mis au point une approche visant à altérer l'exactitude des coordonnées GPS, de façon à ce que l'emplacement réel ne puisse être déterminé. Cette procédure élimine presque toutes les chances d'identifier précisément des individus, mais conserve le détail de la localisation, essentielle dans le cadre d'analyses spatiales. PMA2020 utilise la même approche pour modifier de façon aléatoire les positions (latitude et longitude) des répondants aux enquêtes PMA, ce qui permet aux futurs utilisateurs d'effectuer des recherches à l'aide de données de localisation tout en préservant la confidentialité des répondants.

Coordonnées GPS collectées lors des enquêtes PMA sur les ménages

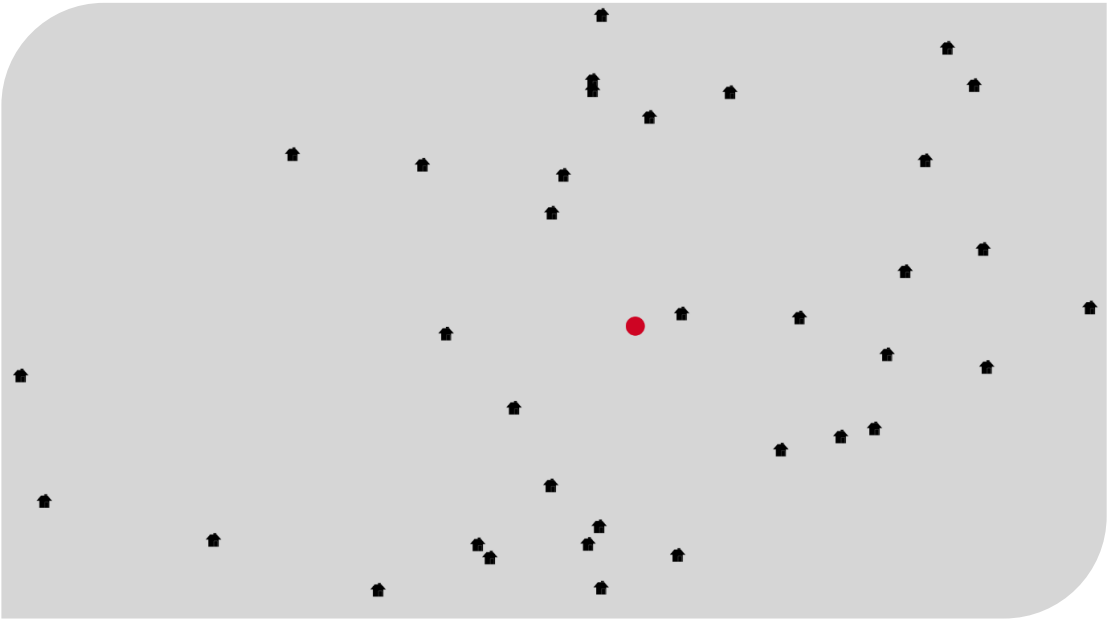
Dans les vagues de collecte PMA2020, les coordonnées GPS sont d'abord recueillies au niveau des ménages pendant le processus de recensement, puis à la fin de l'entretien de chaque ménage. Le processus de recensement consiste à inscrire chaque ménage dans une Zone de Dénombrement échantillonnée (ZD ou EA). Une ZD comprend habituellement environ 200 ménages. Au sein d'une ZD, entre 35 et 42 ménages sont sélectionnés au hasard, pour ensuite être enquêtés lors d'une vague de collecte PMA2020. Les coordonnées GPS sont enregistrées sous forme de coordonnées géographiques (en degrés en latitude et longitude). Dans des conditions idéales de collecte (horizon plat, aucune obstruction du couvert végétal ou de bâtiments), le niveau de précision des coordonnées est généralement inférieur à 6 mètres.

Protocole pour agglomérer et déplacer les coordonnées GPS afin d'en assurer la confidentialité

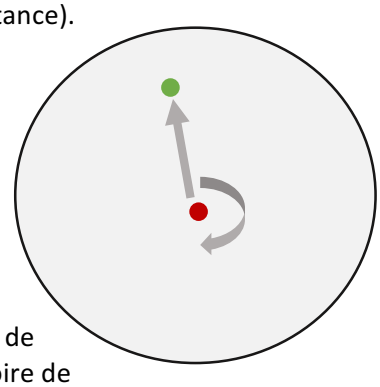
1) Agglomération :
    - Le centroïde d'une ZD donnée (cercle rouge ci-dessous) est déterminé à partir des coordonnées GPS de tous les ménages répertoriés dans cette même ZD (symboles noirs ci-dessous). Les coordonnées géographiques sont tirées du premier recensement des ménages effectué dans chaque ZD (venant de la première vague pour les ZD initiales et du premier recensement pour chaque ZD[1] supplémentaire ou de remplacement).
    - Chaque ZD possède un centroïde géographiquement référencé.

[1] Les ZD sont parfois remplacées ou ajoutées aux enquêtes PMA2020. Les raisons du remplacement de ZD sont dues aux questions d'accessibilité liées à la sécurité ou aux catastrophes naturelles. L'ajout de ZD s'explique par la nécessité d'élargir l'échantillon ou d'accroître la représentation des régions géographiques.

- Aucunes coordonnées GPS spécifiques aux ménages ne sont rendues disponibles aux chercheurs sollicitant les données de PMA2020. Seules les coordonnées déplacées des centroïdes des ZD seront rendues accessibles.

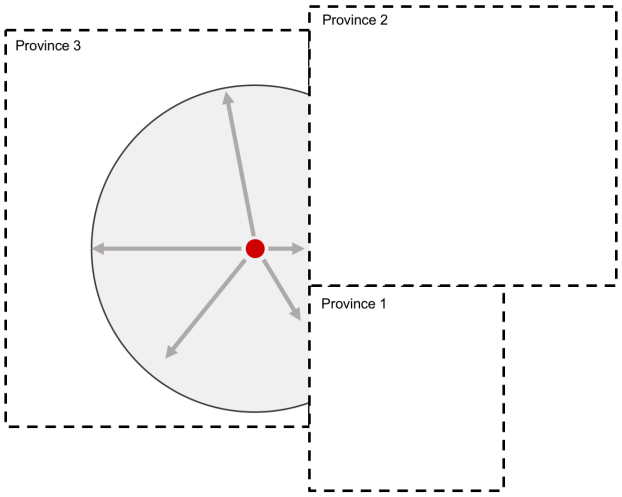


2) Déplacement :
   - Le protocole[2] de déplacement décrit plus bas est exactement le même que celui utilisé par le programme EDS/DHS[3].
   - Le centroïde d'une ZD est déplacé de manière aléatoire (direction et distance). Plus précisément :
     a. **La direction de déplacement** est choisie au hasard entre 0 et 360 degrés.
     b. **La distance de déplacement** est choisie au hasard, mais sera différente pour les ZD urbaines et les ZD rurales, compte tenu de la densité de population plus faible dans les zones rurales.
        - Les centroïdes des ZD urbaines pourront être déplacés jusqu'à 2km de leurs positions réelles.
        - Les centroïdes des ZD rurales pourront être déplacés jusqu'à 5 km de leurs positions réelles, et jusqu'à 10 km pour un échantillon aléatoire de 1 % de ZD rurales[4].



[2] La description du protocole fournie par DHS se trouve en Annexe I.
[3] Le protocole complet est disponible ici : http://dhsprogram.com/pubs/pdf/SAR7/SAR7.pdf.
[4] Dans les pays où le nombre de ZD rurales est inférieur à 100, une ZD rurale sera choisie au hasard.

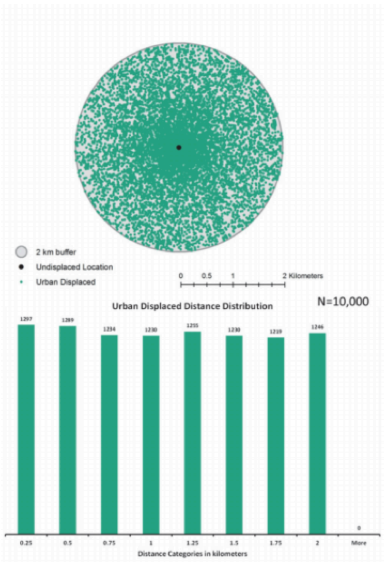
c. **<u>Limitation de déplacement</u>** : Pour certaines ZD situées près d'une frontière administrative (par exemple, les ZD rurales situées à moins de 5 km des limites nationales, régionales et sous régionales), il est possible que la procédure aléatoire puisse les déplacer au-delà de ces limites. Afin d'éviter les erreurs de classification lors d'analyses à l'aide de données administratives, le déplacement ne franchira pas la division administrative définie comme étant la "limitation de déplacement". Cependant, il est possible qu'un point puisse être déplacé au-delà d'une division administrative inférieure pendant le processus.

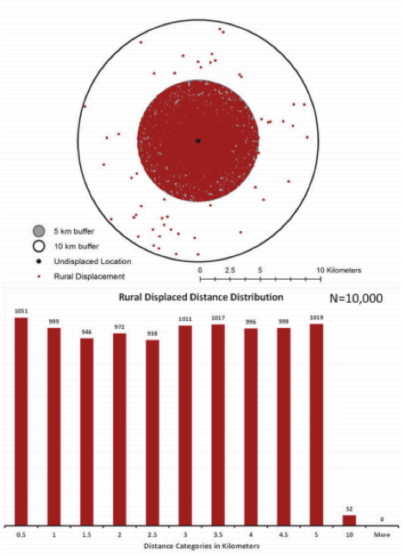


- Ces erreurs de déplacement sont appliquées aléatoirement et aveuglément à chaque point d'origine. Le résultat en sortie peut être vu comme un nouveau point (avec les coordonnées altérées) possédant une zone tampon circulaire (d'un rayon de 2, 5 ou 10 km en fonction du paramètre urbain ou rural) à l'intérieur de laquelle se trouve le point d'origine (ayant les coordonnées non-altérées).

3) Exemples :
   - Une étude[5] (voir figures ci-dessous) a simulé 10 000 déplacements aléatoires de points d'une grappe provenant du programme EDS/DHS. Les figures ci-dessous montrent que le déplacement des positions (points verts dans les zones urbaines et points rouges dans les zones rurales) est réellement aléatoire et presque uniformément réparti dans la zone de déplacement potentielle (cercle).
   - L'étude démontre également que, pour une grappe donnée et déplacée dans la figure (points verts dans les zones urbaines et points rouges dans les zones rurales), si l'on applique le même déplacement aléatoire pour tenter d'identifier l'emplacement réel (point noir), la probabilité serait extrêmement faible - ou quasiment impossible.

[5] Source : Burgert et al. 2013





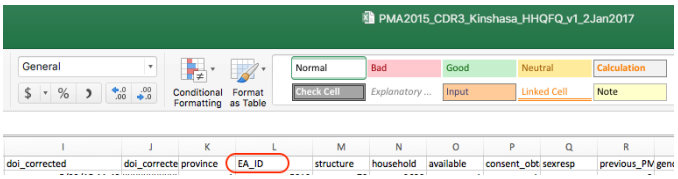
Instructions pour utiliser les coordonnées GPS déplacées de PMA2020

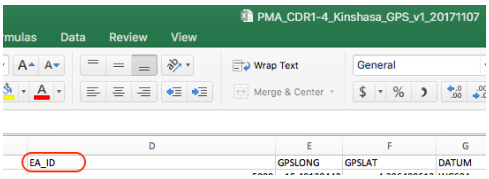
1) Ensembles de données compatibles :

- Les fichiers sont compatibles avec tous les ensembles de données provenant des Questionnaires **Ménage** et **Femme,** et **Sites de Prestation de Santé** de PMA, pour un même pays et pour un groupe de vagues donné.
- PMA2020 sélectionne un nouvel échantillon de ZD au bout d'un certain nombre de vagues de collecte.

2) Joindre le fichier de coordonnées GPS à d'autres ensembles de données PMA2020 :

- Le fichier de coordonnées GPS contient une variable/colonne nommée "**EA_ID**" qui est également présente dans tous les autres ensembles de données PMA2020 (voir figures ci-dessous).
- Vous pouvez lier les coordonnées GPS avec d'autres données PMA2020 en utilisant la variable/colonne "EA_ID" : dans votre logiciel SIG, vous pouvez effectuer une jointure de table en utilisant "EA_ID" comme champs de jointure et comme champs cible ; dans STATA, vous pouvez fusionner les ensembles de données en vous basant sur "EA_ID".





3) Afin d'éviter les erreurs courantes durant une analyse utilisant des coordonnées GPS déplacées, les utilisateurs doivent se rappeler que :

- Les positions géographiques sont basées sur le centre des ZD, et NON sur les positions spécifiques des ménages.
- La position géographique de chaque ZD a été déplacée.
- La taille des ZD peut varier fortement.

- La mesure de la distance d'un point par rapport à un autre lieu (installation, école, etc.) ne sera pas la distance réelle puisque le point a été déplacé. L'utilisation de catégories de distance ou de tampon représente une meilleure approche.

Politique d'accès aux coordonnées GPS

1) Le fait d'avoir accès aux données des Ménages/Femmes et des SPS ne donne pas l'accès automatique aux coordonnées GPS. Chaque utilisateur enregistré devra formuler une demande spécifique afin d'accéder aux fichiers contenant les coordonnées GPS déplacées.
2) La demande doit inclure (1) des problématiques de recherche spécifiques ainsi (2) qu'une explication justifiant l'utilisation des coordonnées GPS pour analyse. Des lignes directrices claires seront fournies pour ceux qui sollicitent l'utilisation des données.
3) Chaque demande sera examinée par le personnel de PMA, et/ou selon le choix du pays, possiblement par l'IP local.

Si vous avez d'autres questions, veuillez contacter datamanagement@pma2020.org.

Annexe I : Protocole[6] du programme EDS pour le déplacement des coordonnées GPS

'La méthodologie de déplacement géographique a été révisée et comprend les étapes suivantes, qui sont exécutées à l'aide d'un script Python. Le script Python permet de définir une couche contenant des polygones agissant comme limitation de déplacement :

1) Convertir les coordonnées (degrés décimaux en mètres) en utilisant un facteur de conversion fixe (degrés en radians) et un scalaire pour corriger les différences dans le nombre de mètres à chaque latitude du Globe.

2) Générer une direction aléatoire en créant un angle compris entre 0 et 360°, et en convertissant l'angle (de degrés en radians).

3) Générer une distance aléatoire de 0 à 2 000 mètres pour les points urbains et de 0 à 5 000 mètres pour les points ruraux, 1 % des points ruraux ayant une distance de 0 à 10 000 mètres.

4) Générer le déport en appliquant des formules trigonométriques (loi des cosinus) en utilisant la distance comme hypoténuse et les radians calculés à l'étape 2.

*xOffset = math.sin(angle_radian) * distance*

*yOffset = math.cos(angle_radian) * distance*

5) Ajouter le déport aux coordonnées originales (en mètres) pour obtenir les coordonnées déplacées.

6) Reconvertir les coordonnées (de mètres en degrés décimaux) à l'aide d'un facteur de conversion fixe (des radians en degrés) et d'un scalaire pour corriger les différences dans le nombre de mètres à chaque latitude du Globe

7) Déterminer si les coordonnées déplacées se trouvent dans la même entité surfacique (ici, divisions administratives= limitation de déplacement) que les coordonnées non déplacées. Répéter les étapes 1 à 6 autant de fois que nécessaire pour générer des coordonnées déplacées à l'intérieur de la même entité surfacique que les coordonnées non déplacées. '

[6] Source : *GPS_Displacement_README.txt* fourni avec le jeu de données GPS du programme EDS.