PMA
PERFORMANCE MONITORING FOR ACTION

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

BILL & MELINDA GATES INSTITUTE for
POPULATION and REPRODUCTIVE HEALTH

Protocol to Aggregate and Displace GPS Data in Performance Monitoring for Action Surveys

Background

Research involving geospatial analysis had grown rapidly in recent years as geo-referenced data have become publicly available. Population-based survey projects (e.g., Demographic and Health Survey (DHS) Program and Living Standards Measurement Study) routinely make GPS coordinate data publicly accessible. However, because GPS data can be used to identify individuals, raw GPS data from confidential surveys cannot be released publicly. The DHS Program has developed an approach to degrade the accuracy of the GPS coordinates so that the true location cannot be derived. This procedure nearly eliminates the likelihood of identifying individuals with GPS data, yet retains the locational detail for spatial analysis. Performance Monitoring for Action (PMA) uses the DHS approach to randomly displace the GPS latitude and longitude positions of PMA survey respondents. PMA data can be used to perform research using location information while respondent confidentiality is maintained.

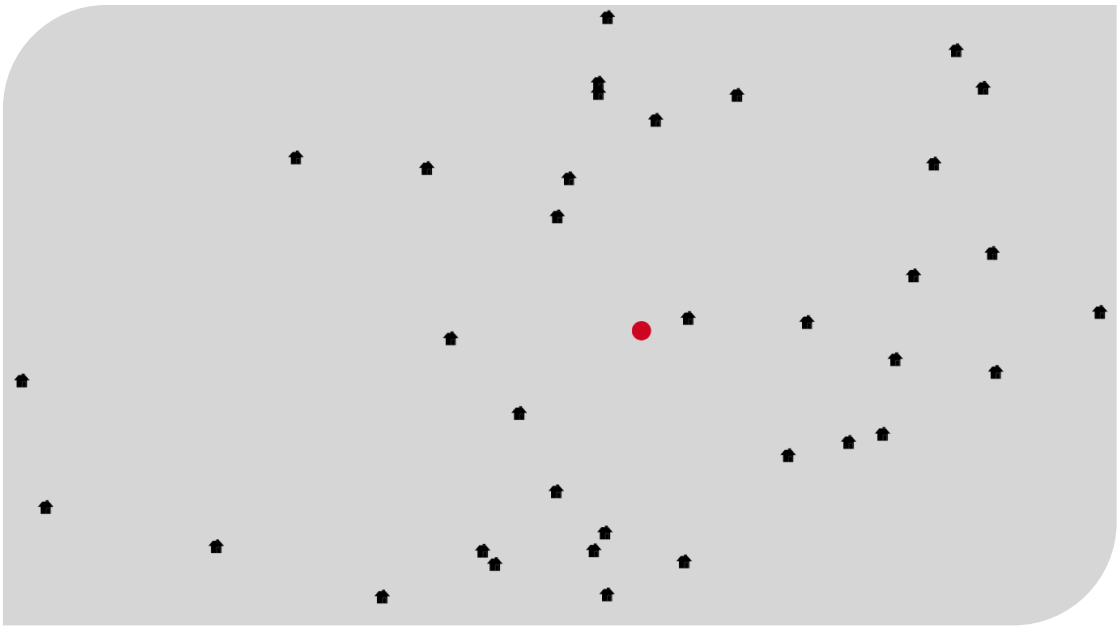
GPS data collected in PMA household surveys

In PMA survey rounds, GPS coordinate data are collected at the household level during the household listing process and after the completion of every household interview. The household listing process involves listing every household in a sampled enumeration area (EA). EAs usually include about 200 households. Within each EA, between 35 and 42 households are randomly selected for the household interview per PMA survey round. The GPS data are recorded as geographic coordinates (i.e. degrees in latitude and longitude). During ideal GPS data collection situations (i.e. flat horizon, no obstructions from vegetation canopy or buildings), the level of accuracy of the coordinates is typically within six meters.

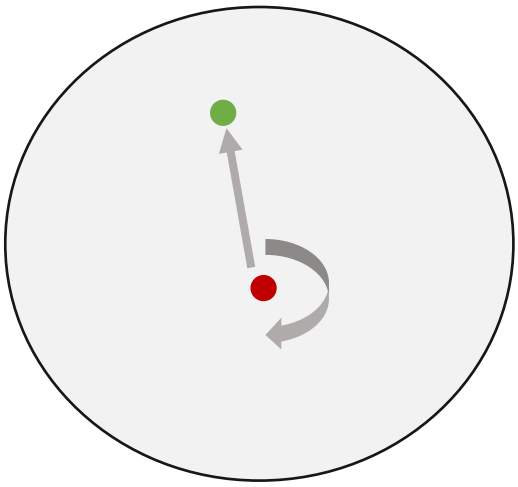
Protocol to aggregate and displace GPS data and ensure confidentiality

1) Aggregation:
   - The centroid in each EA (red circle below) is found using GPS coordinate data from all listed households in the EA (black dots below). Data are used from the first household listing conducted in each EA (i.e., first round for the original EAs and first listing for any replacement or additional EAs[1]).
   - Each EA has one centroid GPS coordinate data point.
   - No household-specific GPS coordinate data are available for researchers requesting PMA data. Only the displaced coordinates of the EA centroids will be made accessible.

[1] EAs are occasionally replaced or added to PMA surveys. Reasons for replacing EAs include accessibility issues related to security or weather/ natural disasters. Reasons for adding EAs include the need for a larger sample size or increased representation of geographic areas.

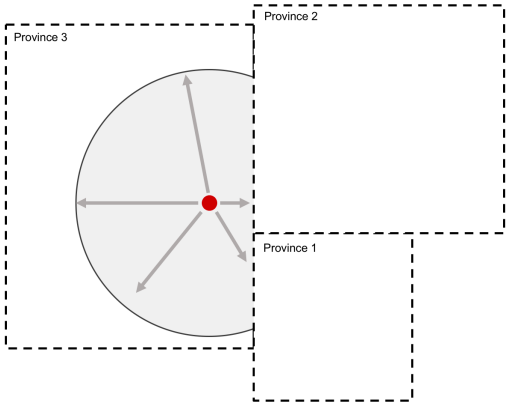


2) Displacement:
   - The displacement protocol[2] described below is the same protocol used by the DHS Program[3].
   - The EA centroid is displaced **randomly** – by angle and distance. Specifically:
     a. **Displacement direction** is randomly selected between 0 and 360 degrees.
     b. **Displacement distance** is randomly selected. The distance parameter is different between urban and rural EAs considering the lower population density in rural areas.
        - Urban EAs are displaced from their true location up to 2 km.
        - Rural EAs are displaced from their true location up to 5 km. Additionally, a random sample of 1% of rural EAs[4] will be displaced up to 10km.



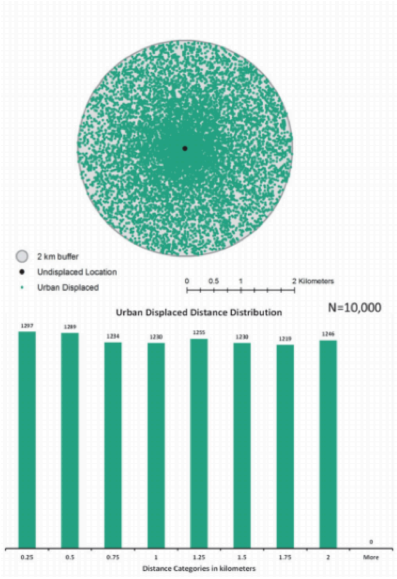
[2] The protocol description provided by DHS is available in Annex I.
[3] The protocol is available here: http://dhsprogram.com/pubs/pdf/SAR7/SAR7.pdf.
[4] In countries where the number of rural EAs is less than 100, one rural EA will be randomly selected.

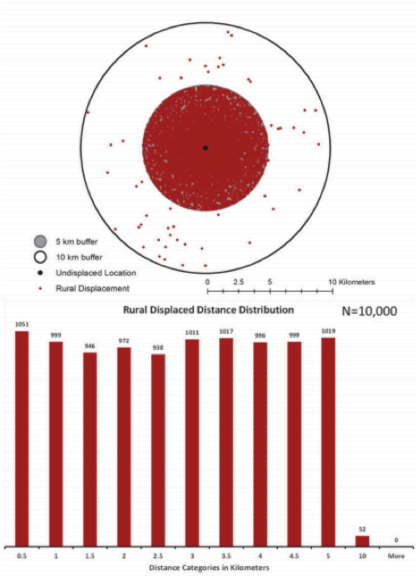
c. **<u>Displacement Restriction</u>**: For some EAs near an administrative boundary (e.g., rural EAs located within 5km from national, regional, and sub-regional boundaries), it is possible that the random procedure could displace them across those boundaries. To avoid misclassification in analysis using administrative-level data, the displacement does not cross the administrative boundary specified as the "Displacement Restriction" in the title of this document. It is possible for a point's coordinates to cross a lower level administrative boundary during the displacement process.



- These displacement errors are randomly and blindly applied to each original GPS point. The new list of coordinates can be understood as each new GPS point having a circular error buffer zone (of 10km, 5km, or 2km depending on whether it is urban or rural) within which the original GPS point falls.

3) Illustrative examples
   - A study[5] (see figures below) simulated 10,000 random displacements for a given DHS cluster. The below figures show that the displaced location (green dots in urban and red dots in rural) is truly random and nearly evenly spread in the potential displacement location (circle).
   - The study demonstrates that, for a given displaced cluster in the figure (green dots in urban and red dots in rural), if one applies the same random-displacement in attempt to identify the true location (black dot), the probability would be extremely small–or nearly impossible.



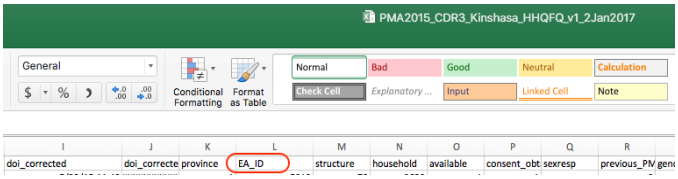


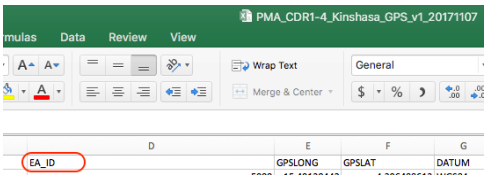
[5] Source : Burgert et al. 2013

Instructions to use displaced PMA GPS data

1) Compatible datasets:
   - GPS data files are compatible with all Household/Female and SDP datasets for a given group of rounds, from the same country.
   - PMA will select a new sample of EAs after a certain number of round of data collection.

2) Linking the GPS data file to other PMA datasets:
   - GPS data files contains a variable/column named "EA_ID" that is present in every other PMA dataset.
   - You can link the GPS dataset with another PMA dataset by using the “EA_ID” variable/column. In your GIS software, you can perform a **table join**, using "EA_ID" as the join and target fields. In STATA, you can merge the datasets based on “EA_ID”.





3) When using the GPS data for analysis users should beware of issues related to common mistakes:
   - GPS locations are based on the center of EAs, NOT household-specific locations.
   - The GPS location of every EA is displaced.
   - EA sizes can vary greatly.
   - Measuring the distance from a GPS point to another site (facility, school, etc.) will NOT capture the true distance since the GPS point has been displaced. Using categories of distances or buffers is a better approach.

GPS data access policy

1) All registered PMA users must submit a separate request for cluster level displaced GPS data. Being granted access to the publicly available HQ/FQ and SDPQ data does not give automatic access to GPS data.
2) In their application, they must include (1) specific research questions, and (2) justification to use the GPS data in the analysis. Clear guidelines will be available for those who wish to request the data.
3) Each application will be reviewed by PMA staff, for additional level security and by the in-country Principal Investigator, at country discretion.

If you have any questions, please contact: datamanagement@pma2020.org.

Annex I: DHS's protocol[6] to displace GPS data

'The geographic displacement methodology has been revised to the following set of steps which are conducted using a Python script. The Python script allows for a polygon layer to be specified as a displacement restrictor:

1) Convert the coordinates from decimal degrees to meters using a fixed conversion factor from degrees to radians and a scalar to correct for differences in the number of meters in a degree of latitude across the earth.
2) Generate a random direction by generating angle between 0 and 360 and converting the angle from degrees to radians.
3) Generate a random distance in meters of 0-2,000 meters for Urban points, and 0-5,000 meters for Rural points with 1% of rural points being given 0-10,000 meters distance.
4) Generate the offset by applying trigonometry formulas (law of cosines) using the distance as the hypotenuse and the radians calculated in step 2.
   xOffset = math.sin(angle_radian) * distance
   yOffset = math.cos(angle_radian) * distance
5) Add the offset to the original coordinate (in meters) to return the displaced coordinates.'
6) Re-convert the coordinates from meters to decimal degrees using a fixed conversion factor from radians to degrees and a scalar to correct for differences in the number of meters in a degree of latitude across the earth.
7) Determines whether the displaced coordinates are within the same polygon feature as the un-displaced coordinates. Repeats steps 1-6 as many times as necessary to generate displaced coordinates within the same polygon feature as the un-displaced coordinates.

[6] Source: *GPS_Displacement_README.txt* provided with DHS GPS dataset.