Household sampling through geocoded points and satellite view: A step-by-step approach to implement a spatial sampling method for demographic and health surveys in areas without population sampling frame and with limited resource settings

E. Apetoh, F. Roquet, F. Palstra, Carine Baxerres, J.-Y. Le Hesran

TITLE PAGE

Household sampling through geocoded points and satellite view: A step-by-step approach to implement a spatial sampling method for demographic and health surveys in areas without population sampling frame and with limited resource settings

Échantillonnage de ménages à travers les points géocodés et la vue satellitaire : une approche détaillée pour mettre en place une méthode d'échantillonnage spatial pour les enquêtes démographiques et sanitaires dans les zones d'étude sans base de sondage de la population et dans des environnements à ressources limitées

E APETOH, F ROQUET, F PALSTRA, C BAXERRES, J-Y Le HESRAN

Correspondence to Edwige APETOH

Institut de recherche pour le développement / Development research institute, Unité mixte de recherche 216 / Mixed research unit 216 : Mères et enfants face aux infections tropicale / Mother and child face to tropical infection , Faculté de pharmacie Paris-Descartes, 4 Avenue de l'observatoire 75006 Paris, France. École doctorale Pierre Louis de santé publique / Pierre Louis doctoral public health school, ED 393 Épidémiologie et Sciences de l'Information Biomédicale / Epidemiology and biomedical information sciences, Paris, France.

E-mail: eapetoh@gmail.com, Phone number: +33 (0)1-53-73-15-06

Florian ROQUET

(1) Hôpital européen Georges-Pompidou, AP-HP, 20 rue Leblanc 75015 Paris, France.
(2) ECSTRRA, CRESS - Unité mixte de recherche 1153, 1 Avenue Claude Vellefaux, Hôpital Saint Louis, 75010, Paris.

E-mail: florian.roquet@aphp.com, Phone number: +33 (0)1-56-09-31-22

Friso PALSTRA

Institut de recherche pour le développement, Unité mixte de recherche 216 : Mères et enfants face aux infections tropicales, Faculté de pharmacie Paris-Descartes, 4 Avenue de l'observatoire 75006 Paris, France.

E-mail: friso.palstra@ird.fr. Phone number: +33 (0)1-53-73-15-07

Carine BAXERRES

Institut de recherche pour le développement, Unité mixte de recherche 216 : Mères et enfants face aux infections tropicales, Faculté de pharmacie Paris-Descartes, 4 Avenue de l'observatoire 75006 Paris, France.

Centre Norbert Elias EHESS - Campus Marseille La Vieille Charité 2, rue de la Charité 13002 Marseille France

E-mail: carine.baxerres@ird.fr. Phone number: +33 (0)4-91-14-07-27

Jean-Yves Le HESRAN

Institut de recherche pour le développement, Unité mixte de recherche 216 : Mères et enfants face aux infections tropicales, Faculté de pharmacie Paris-Descartes, 4 Avenue de l'observatoire 75006 Paris, France.

E-mail: jean-yves.lehesran@ird.fr. Phone number: +33 (0)1-53-73-96-20

ABSTRACT

Introduction

Spatial sampling is increasingly used in health surveys as it provides a simple way to randomly select target populations on sites where reliable and complete data on the general population are not available. However, the previously implemented protocols have been poorly detailed, making replication difficult or even impossible. To our knowledge, ours is the first document describing step-by-step an efficient spatial sampling method for health surveys. Our objective is to facilitate the rapid acquisition of the technical skills and know-how necessary for its deployment.

Methods

The spatial sampling design is based on the random generation of geocoded points in the study area. Afterwards, these points were projected on the satellite view of Google Earth Pro™ software and the identified buildings were selected for field visits. A detailed formula of the number of points required, considering non-responses, is proposed. Density of buildings was determined by drawing circles around points and by using a replacement strategy when interviewing was unachievable. The method was implemented for a cross-sectional study during the April-May 2016 period in Cotonou (Bénin). The accuracy of the collected data was assessed by comparing them to those of the Cotonou national census.

Result

This approach does not require prior displacement in the study area and only 1% of identified buildings with Google Earth Pro™ were no longer extant. Most of the measurements resulting from the general census were within the confidence intervals of those calculated with the sample data. Furthermore, the range of measurements resulting from the general census was similar to those calculated with the sample data. These include, for example, the proportion of

the foreign population (unweighted 8.9% / weighted 9% versus 8.5% in census data), the proportion of adults over 17 years of age (56.7% versus 57% in census data), the proportion of households whose head is not educated (unweighted 21.9% / weighted 22.8% versus 21.1% in census data).

Conclusion

This article illustrates how an epidemiological field survey based on spatial sampling can be successfully implemented at low cost, quickly and with little technical and theoretical knowledge. While statistically similar to simple random sampling, this survey method greatly simplifies its implementation.

Keywords: Global Positioning System (GPS). Geographic Information System (GIS). Geographical coordinates, Households. Satellite view. Spatial sampling.

RÉSUMÉ

Introduction

L'échantillonnage spatial est de plus en plus utilisé dans les enquêtes de santé car il offre un moyen simple de sélectionner au hasard des populations cibles sur place lorsque des données fiables et complètes sur la population ne sont pas disponibles. Toutefois, les protocoles mis en œuvre jusqu'à présent sont peu détaillés, ce qui rend leur reproduction difficile, voire impossible. L'objectif de notre étude était de décrire étape par étape une méthode d'échantillonnage spatial efficace pour les enquêtes de santé afin de faciliter l'acquisition rapide des compétences techniques et du savoir-faire nécessaires à son déploiement.

Méthodes

La méthode d'échantillonnage spatial présentée a été mise en œuvre dans le cadre d'une étude transversale sur la période avril-mai 2016 à Cotonou au Bénin. Elle est basée sur la génération aléatoire de points géocodés dans la zone d'étude. Ces points ont ensuite été projetés sur la vue satellite du logiciel Google Earth Pro™ et les bâtiments identifiés ont été sélectionnés pour être visités sur le terrain. Une formule détaillée du nombre de points requis, tenant compte des non-réponses, est proposée. La densité des bâtiments a été prise en compte en considérant des cercles autours des points et en utilisant une stratégie de remplacement. L'exactitude des données collectées a été évaluée en les comparant aux données du recensement national de Cotonou.

Résultat

Cette approche ne nécessite pas de déplacement préalable dans la zone d'étude et seulement 1% des bâtiments identifiés avec Google Earth Pro™ n'existaient plus sur le terrain. La plupart des mesures résultant du recensement général se situaient dans les intervalles de confiance des mesures calculées avec les données de l'échantillon. En outre, les mesures résultant du

recensement général étaient proches de celles calculées sur les données de l'échantillon. Il s'agissait, par exemple, de la proportion de la population étrangère (non-pondérée 8,9%, pondérée 9% versus 8,5% pour les données du recensement), de la proportion d'adultes de plus de 17 ans (56,7% versus 57% pour les données du recensement), de la proportion de ménages dont le chef de famille n'est pas éduqué (non-pondérée 21,9%, pondérée 22,8% versus 21,1% pour les données du recensement).

Conclusion

Cet article décrit comment une enquête épidémiologique de terrain basée sur l'échantillonnage spatial peut être mise en œuvre avec succès à faible coût, rapidement et avec peu de connaissances techniques et théoriques. Cette méthode d'enquête est statistiquement similaire à l'échantillonnage aléatoire simple et simplifie grandement sa mise en œuvre.

Mots-clés : Système de positionnement global (GPS). Système d'information géographique (SIG). Coordonnées géographiques. Ménages. Vue satellite. Échantillonnage spatial.

INTRODUCTION

In low-income countries, administrative records are often incomplete due to the unavailability of a population census, long time intervals (sometimes greater than 10 years) between two consecutive censuses, and/or lack of addressing systems [1]. It may also be that these records simply do not exist. Carrying out an epidemiological survey based on data representative of the target population in these areas is a challenge and the best method to collect accurate data would obviously be to study all individuals in the target population, as in a general census. However, organization of censuses is costly, time-consuming and requires significant human resources [2].

In Work Package 3 of the GLOBALMED project (funded by the European Research Council) it was decided to implement a simple, efficient and reproducible protocol of random sampling to conduct four field surveys in rural and urban areas in Ghana and Benin. Problems related to counterfeit or dummy drugs are increasing internationally [3] and the global health problems associated with them are becoming more evident [4,5]. In this respect, the GLOBALMED project aims to study the realities of that market, in terms of both supply (circulation, distribution) and demand (use, consumption) in Ghana and Benin [6]. These two West African countries were chosen because of the well-known differences between their pharmaceutical legislations and drug distribution methods. While Benin's pharmaceutical legislation provides for largely state-controlled distribution, Ghana's more free market legislation leaves considerable room for manoeuvre for those involved in the importation and distribution of medicines. For example, in Ghana there are “chemical stores”, which do not exist in Benin, where a pharmacists’ monopoly is in force. The chemical stores are managed by non-pharmacists and can sell a specific list of medicines that are not subject to medical prescription. In Benin, the state does not sell drugs through non-pharmacists. However, there exists an informal market in which non-pharmacists sell drugs that are not under state control

[7]. One of the project's hypotheses is that drug distribution channels have an impact on use and consumption, which tends to increase and exceed the scope of medical recommendations. Our project aims to build on recently recommended drugs for malaria management in Benin and Ghana. Artemisinin-based combination therapies (ACTs) are a case study for evaluation of the global drug market, in terms of both supply and demand. In 2004, following emergence of resistance of the malaria parasite to the antimalarial drugs (chloroquine, sulfadoxine/pyrimethamine) used up until then, Benin and Ghana recommended ACTs for treatment of uncomplicated malaria [8,9]. Malaria is a life-threatening, preventable, and curable disease. In 2017, malaria deaths reached 435,000 and the World Health Organization (WHO) African Region was home to 93% of malaria deaths [10].

To achieve the objectives of Work Package 3 of the drug consumption project, particularly with regard to ACTs and drug sales outlets, it was decided to collect data on household malaria management through cross-sectional study. For project purposes, a household is defined as a group of people who live in the same building and eat their together. At the household level, data were to be collected by a questionnaire addressed to two randomly selected target populations, an adult over 17 years of age and a child under 12 years of age. The two age groups were chosen to determine whether adults, who are logically independent from a health standpoint, and children, who depend on their parents to manage their health problems [11], use drugs in the same way. Inability to identify target populations through basic sample frames, which were missing, led to use of a spatial sampling method.

Spatial sampling methods are widely applied to the sampling of environmental phenomena, such as invasive pathogen of plants [12], soil, water or air pollution [13–15]. By dint of recent advances in Global Positioning System (GPS) and Geographic Information System (GIS) technologies, spatial sampling methods are also increasingly used in demographic and health surveys [16–18]. These technologies facilitate the building of an effective spatial sampling

design that can accurately identify residential areas, which represent an indirect measure of household sampling frames [19].

In demographic and health surveys, different designs have been applied. The most widely used methods are based on random selection or random generation of geographic coordinates over the study area. Latitude and longitude are the most commonly used geographical coordinates. Their intersection or combination specifies the position of a place or point on the Earth's surface. In this document, the term “geocoded point” will be used to refer to the point at the intersection of the two components. In a spatial sampling design, a sample of geocoded points is initially randomly generated in the study area, after which the households identified by these geocoded points are selected for inclusion in the surveys [16,17,20,21]. In another spatial sampling design, a complete household census is initially conducted in the study area to collect the houses’ geographical coordinates, after which a sample of households is selected for the survey [18,22,23]. Census implementation may require prior displacement in the field before the fieldwork. It entails significant financial and human expenses. In any event, the creation of spatial data sets requires specific knowledge and skills [17] integrating GPS and GIS technologies.

In previously published literature reviews, little or no detail was provided on the steps and skills to be taken into consideration in the protocols through which spatial sampling methods were to be implemented [16–18,21]. Their replication is time-consuming or even impossible, especially when researchers are not familiar with GPS and GIS technologies. That is why, with regard to Cotonou, the theoretical approach and preliminary steps to implementation of spatial sampling for the purposes of the GLOBLAMED project were protracted.

The present paper aims to address these problems detailing protocol of the spatial sampling design adopted for the Cotonou survey as a surrogate to simple random sampling. First, a step-

by-step illustration of the implementation of a valid and reproducible spatial sampling process in low-income environments is presented. Easy-to-use and freely accessible tools are introduced; as concerns their operation, each step is detailed. Second, protocol effectiveness is investigated by providing descriptive results of the sample and evaluating the time required for implementation. The accuracy of the collected data is assessed by comparing them to those of the Cotonou national census.

Parts of the Cotonou study results presented in this article were previously published [24].

MATERIALS AND METHODS

Setting

The method was implemented for a cross-sectional study in April-May 2016 in Cotonou. Cotonou is a major city located in southern Benin. It is the country's economic capital and is composed of 13 districts. In 2013, the city's estimated population was 679,012 inhabitants, with 166,433 households. The city of Cotonou constitutes the department known as Littoral, with a surface area of 79 Km² [25]. It is located at the intersection of 6°20 parallel north and 2°20 meridian east. The Atlantic Ocean forms the southern limit of the city while Lake Nokoué borders it to the north. The departments of Abomey-Calavi and Oémé are located at the western and eastern limits of the city, respectively. On an urban plan, Cotonou contains very few multi-storey buildings, and may be considered as a flat city.

Spatial sampling design

Spatial sampling design is based on the random generation of a determined number of geocoded points in the study area. After which, the points are projected on Google Earth Pro™ software (Google Inc. Mountain View, CA, USA). The software offers a satellite view, which is also a geocoded map. The map's resolution directly identifies buildings but not households. For the purposes of this study, the buildings identified by the generated geocoded points constituted the sample size. Probability selection method is based on the 25-meter circle drawn around each geocoded point. If no construction was located within the perimeter of the circle, the geocoded point was deleted, which meant that in areas of lower density in human construction, there were fewer geocoded points. All sampling steps were carried out using computers. As some of the tools are hosted on internet sites, it was essential to have an internet connection. Once researchers were in the field, the identified buildings were visited. If there were several households in a building, one of them was randomly selected. Lastly, the target populations were randomly selected from within the household.

The protocol consists in the four steps described below. The manipulations performed at each step are detailed in the supplementary material.

Step 1: Definition of the study area using the random point generator (RPG) tool

The RPG tool generates one or more geocoded points at a given location on the surface of the earth. A sampling method requires definition of the study area and its geographical boundaries, which can be administrative, as in the case with Cotonou, or be determined more specifically, according to the context of the study. In the latter scenario, the study area can be defined as a perimeter around a point or a place of interest, such as the epicentre of a city, hospital, or health facility. In spatial sampling, the geographical coordinates of the boundaries are required to specify the study area in the RPG tool. In this survey, the RPG tool is available on the Geomidpoin.com website (http://www.geomidpoint.com/random/).With this tool, the study area can be specified by a rectangle or circle, but not a polygon. For Cotonou, the rectangular configuration was more practical than the circle (Figure 1).

The latitudes of the northern and southern boundaries of the study area and the longitudes of the western and eastern boundaries allow the study area to be specified in the RPG tool. With Google Earth Pro™, determination of these coordinates requires manipulation with the “Add a path” and “Add mark” features. “Add a path” draws up a specific path that defines an area of interest by reproducing its boundaries. “Add mark” collects the geographical coordinates of a point on the path having been drawn.

A paper map of Cotonou provided by the National Geographic Institute of Benin was used to identify the streets through which the city's administrative boundaries pass. After identifying these streets, the "Add a path" feature enabled these boundaries to be reproduced on the Google Earth Pro™ satellite view. In addition, the most extreme cardinal points of the boundaries were identified, and their coordinates were collected using the "Place mark"

function. The actual geographical shape of the city of Cotonou is not a perfect rectangle. As a result, the study area specified in the RPG tool was larger than the actual study area (Figure 1).

The geographical coordinates of the study area boundaries can be obtained in other ways. In some studies, researchers have previously visited the field to collect the coordinates of the study area boundaries using a GPS device [16,17]. Another "remote" approach is to use the limits given on the Global Administrative Areas website [26], which delineates administrative boundaries for almost all countries in the world in formats readable by GIS technology. Google Maps and the Openstreetmap.org website also provide administrative boundaries of cities or countries. In the Cotonou survey, when compared to data from the National Geographic Institute of Benin, the administrative delimitation proposed by these sites turned out to be incorrect. The boundaries provided by the aforementioned websites consequently had to be validated before any use by an investigator familiar with the area of interest.

Step 2: Observation of the buildings identified by the geocoded points on Google Earth Pro™

Once the study area was determined in the RPG tool (Geomidpoint.com website), geocoded points were generated (see additional documentation for more details). After which, these points were projected in the satellite view of Google Earth Pro™ to identify buildings. As the boundaries of Cotonou were previously drawn, the geocoded points generated outside (the specified zone in the RPG tool is larger than the study area) and inside the study area were identifiable. Projection of the generated points on Google Earth Pro™ requires the generation of a Keyhole Markup Language (KML) format containing the geographical coordinates (latitude and longitude) of the geocoded points. For this purpose, the geographical coordinates were saved in a Microsoft® Excel file, which was converted to the KML format using the free conversion tool available on the GPSVisualizer.com website

(http://www.gpsvisualizer.com/map_input?form=googleearth). Buildings identified by geocoded points were visualized by opening the KML file in Google Earth Pro™.

Step 3: Calculation of the number of geocoded points to generate

In this design, the number of geocoded points to be generated must take into consideration the following realities: (1) The specified zone in the RPG tool is larger than the study area (Figure 1); (2) Not all generated points necessarily identify a building; (3) Not all buildings necessarily contain a household; (4) Finally, the number of generated points must be based on the target population.

Since the area specified in the RPG tool is larger than the study area, geocoded points are generated outside the latter. In addition, urban and rural areas have heterogeneous environments, occupied by residential (households), commercial (markets, administrative buildings, offices) or unconstructed (e.g., parks, lakes, or marshes) spaces. Therefore, not all geocoded points identify a building [17]. If at each generated point the nearest building is assigned, there is a risk of over-representing low-density neighbourhoods. To take these considerations into account, a circle with a 25-meter radius was drawn around each generated geocoded point (Figure 2). This circle approximates the estimated mean surface area of buildings and their associated properties (e.g. gardens, courtyards) in Cotonou.

Calculation of the number of geocoded points to be generated for fieldwork presupposes estimation of the proportion of points likely to identify buildings within the study area. Aside from the geocoded points generated outside the study area, others will be generated inside the study area, but will not overlap with any buildings. To estimate relevant proportions, two hundred geocoded points were generated five times and projected each time on the satellite view of Google Earth Pro™. In each case, the proportion of points generated within the study area and containing at least one building within the circle was calculated. The average

proportion of geocoded points that could be exploited subsequent to the five randomizations was used to calculate the number of geocoded points required for the study.

More generally, the number of geocoded points to be generated depends on the size of the study area. The larger the area, the greater the number of simulated geocoded points will have to be, and a more sizable variation between two simulations implies a larger number of simulations.

Thanks to the survey design, there was no need for prior travel, but the occupancy status of the buildings (market, administrative buildings, offices, households) was not known prior to the survey. It was assumed that not all of the buildings identified by geocoded points would contain a household. In addition, not all of the households selected in the buildings would accept to be included in the survey. Consequently, the number of buildings to be surveyed was increased by 10 per cent [1] and an original replacement strategy was implemented for non-response and uninhabited buildings.

The target population selection plan aimed at random selection of an adult over 17 years of age and a child under 12 years of age from the chosen households. A household was included in the survey if it housed at least one of the two targeted populations. It was expected that all households would contain at least one adult, but not necessarily a child under 12 years of age. Consequently, calculation of the sample size was based on the most poorly represented target population in the study area [1]. The number n of children under 12 years of age to be included in the survey was obtained by the Schwartz formula [27] : $n = \frac{(z^2)\times(\hat{p})\times(1-\hat{p})}{e^2}$.

z represents the z-score set at 1.96; e is the desired margin of error set at 0.05 and $\hat{p}$, the proportion of children under 12 years of age having previously contracted malaria. The latter proportion was unknown, and was set at 50% in order to maximize the sample size [2]. Three

hundred and eighty-four included households had to contain at least one child. No prior information was available on the proportion of households in Cotonou containing at least one child aged under 12 years. Studies in a demographically similar city, Ouagadougou (Burkina Faso), reported 72% of households containing a child under 12 years. However, the crude birth rate in Ouagadougou is 34.7% [28] which is higher than that reported in Cotonou (32.2%) [29]. Therefore, the hypothesis was that one out of two households contained at least one child under 12 years old. Consequently, twice as many households were to be visited in order to include 384 children in the survey.

Considering the sample size n, the proportion $P_{building}$ of geocoded points, which identify a building, the proportion P_{child} of households, which contain a child under 12 years of age, 10% of non-response, the number N of geocoded points to randomize was calculated as:

$$N = n * {}^{1}/_{p_{building}} * {}^{1}/_{P_{chld}} * 1{,}1 \ .$$

To fulfil the requirements of the Cotonou survey, 1300 geocoded points had to be generated.

Step 4: On-field access to the buildings identified by geocoded points

After calculation of the number of geocoded points to generate, and following projection of these points on the satellite view of Google Earth Pro™, the geocoded points generated outside the study area and those generated within the study area but without building overlap were deleted. Buildings to visit in the field were those within the circle around the generated geocoded points in the study area. If there were several buildings in the circle, the building closest to the geocoded point (i.e. the centre of the circle) was selected for visit during the fieldwork phase.

Due to the random design, the locations designated by the geocoded points may be found on the roof of a building or inside gardens. Investigators are not allowed to enter private property to determine the exact position of a geocoded point. Moreover, utilization of the geocoded points would have complicated the identification of buildings in the field. For this reason,

problematic points were relocated and set up on the street side next to the building walls and in front of their intended entrance (Figure 2). Relocations were processed on Google Earth Pro™.

On the field, buildings were found using the GPS unit, Garmin® eTrex® 10 (Garmin International Inc., Olathe, KS, USA.). Additionally, to facilitate identification, pictures of the satellite view containing the geocoded points were given to investigators. The manipulations aimed at transferring the geographical coordinates into the eTrex® 10 unit are detailed in the supplementary material.

Replacement strategy and access to the target population: an additional step

A replacement strategy was established for situations where a building was uninhabited or in the event of refusal or absence. After all, socio-economic, demographic, economic and health outcomes can be largely influenced by the local or neighbourhood environment [30,31]. In the event of an unachievable interview, neighbouring buildings replaced the identified building. The replacement strategy (Figure 3: Image 1) was to consecutively visit up to five adjacent buildings until an interview was conducted: four on the same side of the street as the identified building (i.e. two adjacent buildings on the right and two on the left) and one facing the building on the other side of the street. A 5-level replacement strategy as therefore adopted, and if no interview was conducted after visiting the other five buildings, the geocoded point was abandoned. When there were several households in a single building, one of them was randomly selected and then, within that household, an adult and a child were randomly selected.

Statistics

All analyses were compiled using Stata® (StataCorp. 2017. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC). The data of the survey are presented as a proportion

(%) with their 95% confidence interval. Two analysis strategies were set up for comparison with Beninese national census data: (1) an analysis using design weights (i.e. normalised inverse of the inclusion probabilities), (2) unweighted raw analysis assuming that the study design was self-weighted by the diameter of the circles around the geocoded points. Confidence intervals were estimated using exact binomial confidence interval. The exact confidence interval is between p_{LB} and p_{UB} and meets the following conditions:

$$\sum_{k=0}^{k} \binom{n}{k} {p_{UB}}^{k} (1 - p_{UB})^{n-k} = \frac{\alpha}{2}$$

$$\sum_{k=x}^{n} \binom{n}{k} {p_{LB}}^{k} (1 - p_{LB})^{n-k} = \frac{\alpha}{2}$$

Where: p_{LB} is the lower bound of the confidence interval and p_{UB} is the upper bound of the confidence interval, n is the number of trials (sample size), k is the number of successes in n trials, α is the probability of a Type I error and $1 - \alpha$ is the desired 95% confidence interval.

RESULTS

Building identification on satellite view versus building identification on reality in the field

One thousand three hundred geocoded points were generated and projected on the satellite view of Google Earth Pro™. Among them, 424 (33%) were generated outside the study area and deleted. Consequently, 876 (67%) geocoded points were generated within the study area, out of which 748 (58%) points identified a building. In the field, out of the 748 buildings identified on the satellite view, 740 (99%) buildings were found. For eight geocoded points (1%), there was no building at the location. Of the 740 buildings found on the site, 516 (70%) were inhabited (containing at least one household) and 131 (18%) were uninhabited (commercial, business, or administrative buildings). In addition to these, there were 93 (12%) unoccupied buildings (under construction or abandoned).

Among the 516 inhabited buildings, 462 (89%) buildings were surveyed, resulting in the inclusion of 462 households in the survey without the replacement strategy. The investigators encountered refusals or absences in 37 (7%) and 17 (4%) buildings, respectively. Taking into account buildings that no longer existed in the field (n=8), uninhabited (n=131) or unoccupied (n=93) buildings, refusals (n=37), and absences (n=17), interviews did not occur at 286 of the generated geocoded points.

Replacement strategy efficacy

Where appropriate, the replacement strategy was implemented. It was applied in the above-mentioned cases of unachievable interview, i.e. in 286/748 (38%) geocoded points, enabling inclusion of an additional 179 households. Six hundred and forty-one households were thereby included in the survey. The percentage of households included at each level of the replacement strategy is calculated by dividing the number of buildings in which a household was surveyed

by the number of the geocoded points to investigate at that level. The results are presented as a flow chart on Figure 4. The highest percentage of successful replacement (31%) was obtained at the first level of the replacement strategy, after which the percentage decreased from the second level to the fourth level and increased at the fifth level (second (25%), third (8%), fourth (4%) and fifth levels (18%)).

Representativeness of the sample

It was assumed that one in two households had a child under 12 years of age. In the final sample, 56% of households contained this target population. The representativeness of the sample was evaluated by comparing the socio-demographic data of the sample with those of the Cotonou National Census [25].

A confidence interval provides an estimated range of values that is likely to encompass an unknown population parameter. The estimated interval is calculated from a given set of sample data from the population according to the selected level of confidence. The representative aspect of the sample resulting from the sampling design was tested through direct comparison of socio-demographic variable of sample households to those of the 2013 national census of Cotonou [25] (Table 1). Most of the values resulting from the general census are within the 95% confidence intervals of the values calculated with the sample data. Furthermore, the range of values resulting from the general census was similar to those calculated with the sample data. They include the proportion of the foreign population (unweighted 8.9% / weighted 9% versus 8.5% for census data), the proportion of adults over 17 years of age (56.7% versus 57% for census data), the proportion of households whose head is not educated (unweighted 21.9% / weighted 22.8% versus 21.1% for census data).

Time spent to set up the study

The time spent to implement this design can be considered at several levels. Drawing the boundaries of Cotonou with the "add path" tool on satellite imagery took up less than an hour. The time spent to perform the simulation and calculate the percentage of geocoded points that identifies buildings depends on the number of points to be generated and the number of simulations to be performed. In the case of the Cotonou study, the procedure took about a day (five simulations of 200 generated geocoded points). The final randomization step, including (1) building identification on the satellite view (1300 generated geocoded points), (2) adjustment of points at the supposed building entrances and (3) recording of the new coordinates of moved geocoded points, took two days. Four investigators were hired for the survey, all of whom had a degree in sociology and satisfactory knowledge of epidemiological field surveys. Training on the protocol (questionnaire, replacement strategy), use of the eTrex® 10 unit and geocoded point localisation on the field took anothet three days. Assignment of the geocoded points to the four investigators and entry of geographical coordinates on eTrex® 10 were completed in two days.

On the field, time spent on the survey depends on data collection method, number of recruited investigators and sample size. In our study, data were collected using a 16-page questionnaire. The interviews lasted on average 30 minutes (from 15 to 60 minutes). On average, per investigator, five interviews would be carried out per day, meaning 35 per week. Each investigator was consequently allocated a weekly total of 35 geocoded points. To prevent investigators from traveling long distances between buildings, a set of the nearest geocoded points (or buildings) was provided to each of them. Starting from the location of one geocoded point, the time taken to identify a building varied between two and ten minutes. Fieldwork

investigating the 748 geocoded points and covering the 641 households lasted five weeks,.

DISCUSSION

With a protocol based on spatial sampling, a large-scale survey was implemented without prior field travel in a limited resource setting. The resulting demographic data show excellent representativeness compared to existing Cotonou census data. The method applied uses only free software and is reproducible even by teams with limited financial or human resources. The objective of communicating the details of this method is to facilitate the rapid acquisition of the technical skills and know-how necessary for its deployment.

Spatial sampling enables random sampling of households in communities without the household or people identifiers (address lists, telephone numbers…) commonly used in simple sampling methods. Additionally, in any country, whether low-income or developed, it may happen that a part of the population is not included in records. This may be due to squatting or to significant migration phenomena resulting from rural exodus or exceptional events such as natural disasters. An approach by satellite provides an exhaustive view of dwellings and ensures coverage of the entire population, even in isolated areas.

Since the satellite view of Google Earth Pro™ was taken at a specific time, it could not reflect the current situation with 100% accuracy. Between the date of the satellite image and the date of fieldwork, some buildings were demolished, perhaps because they were temporary cabins set up for private or public circumstances. In general, the Google Earth Pro™ satellite view is regularly updated. For the Cotonou survey, in less than 1% of cases, buildings designated by geocoded points on a year-old image no longer existed. That said, urban areas such as Cotonou are densely populated and unlikely to experience major demographic changes over short time spans. In some areas, “street view” modes of Google Earth Pro™ allow prior inspection of the

building facades; uninhabited buildings are identified, and unnecessary field trips reduced. Unfortunately, the “street view” mode was not available in Cotonou.

As this design does not require prior displacement, actual occupation status is unknown before the fieldwork, during which adaptations need to be carried out. In the protocol by Grais et al., in cases where there was no building at the position of the generated point, the closest compound to the right (with the teams facing north) was the first to be sampled. According to the authors, this alternative strategy favours the selection of households in low-density urban areas [16]. However, assignment of the building nearest to a given geocoded points may result in overrepresentation of buildings close to empty spaces, interfering with the randomness of the sample [16,17]. Some authors have integrated the notion of density in their alternative strategies. In the design by Michelle Kondo et al., if a point was located far from a structure (a field, a wooded area…), investigators were requested to locate the nearest residence within 1000 feet to the right of the point, facing north [17]. In the protocol by Kolbe and Huston, when the geocoded point was not a residence, all households within 20 yards were identified and the location to be surveyed was randomly chosen from among them [32]. Harry Shannon and colleagues counted buildings within circles of 20 meters of radius and randomly selected one of them for survey [20]. However, counting buildings or households within a determined perimeter is complicated, time-consuming, and even impossible when property access is restricted or when the area is difficult to reach (close quarters, marshlands, slums…). In these environments, buildings that are more accessible are likely to be chosen more often. In their study, moreover, these authors did not offer a rationale for the designated perimeters. In our survey, the 25-metre radius corresponds to the estimated perimeter of a building and its property (e.g. gardens, courtyards) in Cotonou. Even if rough, this estimate was satisfactory in a city where building density is high, and the circle prevents over-representation of buildings near empty spaces. Moreover, sampling of the adjacent buildings when the identified building

is unoccupied maintains the representativeness of the environment close to where the geocoded point was generated.

Applying our strategy, percentage of replacement decreased from the second to the fourth levels and increased at the fifth level [first (31%), second (25%), third (8%), fourth (4%) and fifth level (18%)]. This is explained by the fact that at times, there were fewer than five adjacent buildings on the same side of the street as the building identified in the satellite view. In addition, at the position of the geocoded point, there could exist a complex of several buildings belonging to the same structure (administrative offices or commercial quarters). As a result, there may not have been a replacement building between the second and fourth levels of the replacement strategy. The investigator was consequently compelled to move directly to the fifth level, i.e. the building opposite the one identified in the satellite view. The increased percentage of households included in the sixth-level replacement strategy is thereby explained. It would have been preferable that the replacement strategy include two buildings on the same side as the identified building (for the first two replacement levels) and three buildings opposite the identified one, i.e. across the street (for the third to fifth replacement levels) (Figure 3: Image 2).

This survey method is not intended to replace existing statistically robust strategies based on stratification or clustering when conditions allow their use. In fact, ours is a pragmatic method using spatial sampling for probability selection, the objective being to adapt to the actual conditions of field research under circumstances where funding, available time and human resources are frequently insufficient to gather data of sufficient quality. Stratified sampling requires identification of every enumeration unit by stratum prior to sampling. If such information is not directly available, the method is not feasible. Cotonou is an urban area with urban divisions consisting in outdated and unrevised boroughs, which complexify breakdown into enumeration units. Moreover, clustered random household samples can be obtained only

from an existing sampling frame with a complete list of statistical units covering the target population. In sub-Saharan Africa census and sampling lists are often lacking, and surveys with adequate designs may require a household listing operation, which would need to be conducted before the main survey.

In other situations, the population could be stratified first, and the spatial sampling method could be used to select samples within each of the strata. Similarly, clusters could be established, and a random sample of them could be selected, within which the GIS method could be used. Stratification of the sampling [17] or weighting during data analysis [20] have previously been implemented to take into account the distribution density of the buildings in the study area. While these types of surveys are not impossible to carry out in sub-Saharan Africa under the methodological conditions of more complex survey designs, they are more expensive, more time-consuming, more labour-intensive, more often dependent on a third party database for population density analysis (e.g. WorldPop [33], Facebook [34]), and less accessible in terms of the methodological and technical knowledge required by researchers whose objective is to efficiently gather a set of reliable data.

More specifically, the spatial sampling strategy we present in our manuscript makes it possible (1) to dispense with a time-consuming and costly pre-survey, (2) to avoid depending on a previous census or survey of which the results may be unreliable because they are outdated due to demographic changes or difficult to obtain, (3) to avoid depending on a third-party database.

Probability selection methods are based on the 25-meter circle drawn around each geocoded point. If no construction is found within the perimeter of the circle, the geocoded point is deleted, which means that in areas of lower density in human construction, there are fewer geocoded points. In the city of Cotonou, as in most cities in Sub-Saharan Africa, individual or

collective residential buildings are mostly single or double story, so there is a direct correlation between human density and building density, which can be taken into account by conserving or not conserving the geocoded points. Our spatial sampling strategy is consequently directly correlated with population densities. And since it assigns geocoded points at random and independently of any other data and interviews households also selected at random among those residing at this point, all households in Cotonou had a non-zero probability of being interviewed. We do not believe that our strategy represents sampling bias, especially insofar as there were no specific household characteristics that predicted that a given household would not be included among those in a specific building; indeed, our form of spatial sampling might be considered as a self-weighting design. However, we cannot rule out the risk of under-sampling, which is why we carried out an analysis using household weights and presented the results alongside the unweighted results in Table 1; they show few noticeable differences, and remain close to the national census.

To identify buildings in the field, investigators used eTrex® 10. When available financial resources do not allow for the purchase of equipment, an alternative solution is to address satellite images to investigators who are familiar with the area under study. There also exist applications that can be installed on investigators' mobile phones to track locations based on their geographical coordinates.

Lastly, this design renders it possible to monitor compliance with household selection procedures and to verify the veracity of the visit. For example, in case of replacement, the geographical coordinates of newly selected buildings can be recorded *in situ*. As a result, the household can be located by another investigator, for supervision or for the collection of additional information.

CONCLUSION

In this paper, we present, as an alternative to simple random sampling, a step-by-step method for sampling of a population from geocoded points. This method appeared easy to apply, reproducible, and efficient. We have shown that an epidemiological field survey can be implemented inexpensively, rapidly, and with relatively little technical and theoretical knowledge. The accuracy of the collected data was assessed by comparing it to the data of Cotonou National Census and found to produce similar results. The final objective of this paper is to provide field epidemiologists with a framework for reproducibility of the design, i.e. spatial sampling, in future studies. This protocol resulted in the implementation of efficient spatial sampling not only in the Cotonou survey, but also in the three other surveys of the WP3 of the GLOBLAMED project.

Acknowledgements: The authors would like to thank Ms Gildas APETOH for her contribution to this study.

DECLARATION

Conflict of interest: None.

Ethics approval and consent to participate: The study has been approved by the Ministry of Higher Education and Scientific Research of Benin Ethics Committee CER-ISBA – FAVORABLE ADVICE N° 30. Each interviewed person was informed about the objectives of the study and the types of collected data. They all signed a document of informed consent.

Patient and public involvement: There is no patient involved in the study

Consent for publication: Not applicable

Availability of data and material: All data generated or analysed during this study are included in this published article [and its supplementary information files].

Funding: This work was supported by the European Research Council under the European Union's Seventh Framework Programme [ERC Grant agreement n°337372, FP7/2007-2013]

Authors' contributions: All authors contributed significantly to the work. EA and JYLH designed the process and supervised its implementation in the field. EA analysed and interpreted data. FP and FR participated in the writing and critical revision of the article. CB is the main investigator of the GLOBALMED project. In addition, all authors certify their approval of the final version to be published.

REFERENCES

1 United Nations. *Designing Household Survey Samples: Practical Guidelines*. New York: United Nations Publications 2008.

2 Ardilly P. *Les Techniques de Sondage*. Paris: Editions TECHNIP 2006.

3 Baxerres C. Contrefaçon pharmaceutique : la construction sociale d'un problème de santé publique. In: *Anthropologie du médicament au sud. La pharmaceuticalisation à ses marges*. 2015.

4 Janes CR, Corbett KK. Anthropology and Global Health. *Annu Rev Anthropol* 2009;**38**:167–83.

5 Nguyen V-K, Lock M. *An Anthropology of Biomedicine*. Wiley-Blackwell. 2011.

6 IRD. Les Combinaisons Thérapeutiques à base d'artémisinine : Une illustration du marché global du médicament, de l'Asie à l'Afrique. 2014.

7 Baxerres C. Du médicament informel au médicament libéralisé : les offres et les usages du médicament pharmaceutique à Cotonou (Bénin).

8 Ghana Health Service. Treatment of Uncomplicated Malaria- National Malaria Control Programme. 2015.

9 Ogouyemi-Hounto A, Kinde-Gazard D, Nahum A, *et al.* Management of malaria in Benin: evaluation of the practices of healthcare professionals following the introduction of artemisinin derivatives. *Med Trop Rev Corps Sante Colon* 2009;**69**:561–4.

10 WHO. Fact sheet about Malaria. 2017.

11 Geissler PW, Nokes K, Prince RJ, *et al.* Children and medicines: self-treatment of common illnesses among Luo schoolchildren in western Kenya. *Soc Sci Med* 2000;**50**:1771–83.

12 Demon I, Cunniffe NJ, Marchant BP, *et al.* Spatial sampling to detect an invasive pathogen outside of an eradication zone. *Phytopathology* 2011;**101**:725–31.

13 Arbia G, Lafratta G, Simeoni C. Spatial sampling plans to monitor the 3-D spatial distribution of extremes in soil pollution surveys. *Comput Stat Data Anal* 2007;**51**:4069–82.

14 Arbia G, Lafratta G. Anisotropic spatial sampling designs for urban pollution. *J R Stat Soc Ser C Appl Stat* 2002;**51**:223–34.

15 Zahid E, Hussain I, Spöck G, *et al.* Spatial Prediction and Optimized Sampling Design for Sodium Concentration in Groundwater. *PLOS ONE* 2016;**11**:e0161810.

16 Grais RF, Rose AM, Guthmann J-P. Don't spin the pen: two alternative methods for second-stage sampling in urban cluster surveys. *Emerg Themes Epidemiol* 2007;**4**:8.

17 Kondo MC, Bream KD, Barg FK, *et al.* A random spatial sampling method in a rural developing nation. *BMC Public Health* 2014;**14**:338.

18 Pearson AL, Rzotkiewicz A, Zwickle A. Using remote, spatial techniques to select a random household sample in a dispersed, semi-nomadic pastoral community: utility for a longitudinal health and demographic surveillance system. *Int J Health Geogr* 2015;**14**:33.

19 Kumar N. Spatial Sampling Design for a Demographic and Health Survey. *Popul Res Policy Rev* 2007;**26**:581–99.

20 Shannon HS, Hutson R, Kolbe A, *et al.* Choosing a survey sample when data on the population are limited: a method using Global Positioning Systems and aerial and satellite photographs. *Emerg Themes Epidemiol* 2012;**9**:5.

21 Siri JG, Lindblade KA, Rosen DH, *et al.* A census-weighted, spatially-stratified household sampling strategy for urban malaria epidemiology. *Malar J* 2008;**7**:39.

22 Kabaghe AN, Chipeta MG, McCann RS, *et al.* Adaptive geostatistical sampling enables efficient identification of malaria hotspots in repeated cross-sectional surveys in rural Malawi. *PLOS ONE* 2017;**12**:e0172266.

23 Kassié D, Roudot A, Dessay N, *et al.* Development of a spatial sampling protocol using GIS to measure health disparities in Bobo-Dioulasso, Burkina Faso, a medium-sized African city. *Int J Health Geogr* 2017;**16**.

24 Apetoh E, Tilly M, Baxerres C, *et al.* Home treatment and use of informal market of pharmaceutical drugs for the management of paediatric malaria in Cotonou, Benin. *Malar J* 2018;**17**:354.

25 INSAE. Principaux indicateurs sociodémographiques et économiques du département du Littoral. 2016.

26 Galway L, Bell N, Sae AS, *et al.* A two-stage cluster sampling method using gridded population data, a GIS, and Google Earth(TM) imagery in a population-based mortality survey in Iraq. *Int J Health Geogr* 2012;**11**:12.

27 Schwartz D. *Méthodes statistiques à l'usage des médecins et des biologistes*. Paris: : Médecine sciences publications 1993.

28 OPO. Ouaga Focus. 2012.

29 INSAE. Enquête Démographique et de Santé 2011-2012. 2013.

30 Diez Roux AV, Mair C. Neighborhoods and health. *Ann N Y Acad Sci* 2010;**1186**:125–45.

31 Molina-García J, Queralt A, Adams MA, *et al.* Neighborhood built environment and socio-economic status in relation to multiple health outcomes in adolescents. *Prev Med* 2017;**105**:88–94.

32 Kolbe AR, Hutson RA. Human rights abuse and other criminal violations in Port-au-Prince, Haiti: a random survey of households. *The Lancet* 2006;**368**:864–73.

33 Tatem AJ. WorldPop, open data for spatial demography. *Sci Data* 2017;**4**:170004.

34 Population Density Maps. Facebook Data Good.

TABLES

Table 1: Presentation of socio-demographic data on sample households collected during the national census of Cotonou.

		Census data	Sample data	
Category	Subcategory	Percent	Percent	95% confidence interval
Nature of household walls	Brick/Stone/Cement	85.8	90	87, 92
	Wood/Board/Bamboo	10*	7	5, 9
	Clay	0.2	0.3	0.07, 1
	Other	4	2.7	2, 5
Nature of household roofs	Sheet metal	81	78	74.6, 81
	Slab	12.8	13	10.6, 16
	Other	6.2	9	6.5, 10.8
Nature of household floors	Cement/Tiles	93	95.5	93, 97
	Clay/Sand/Other	7*	4.5	0.3, 6.4
Lighting	Electricity and source producing electricity	89	89.5	87, 91.6
	Other	11	10.5	8.3, 13
Drinking water supply	Tap water	97	95.4	94, 97
	Well-protected	0.6	0.2	0.2, 1.1
	Other	2.4*	4.4	3, 6.3
Household comfort	WC in the household	89	90	87, 92
	Other	11	10	8, 13
Percentage of persons in the household by age category	Under 6 years old	17.1	16.8	15.6, 18.2
	6 to 11 years old	14.1	14.3	13, 15.5
	12 to 17 years old	11.8	12.2	11.1, 13.3
	Over17 years old	57	56.7	55, 58.4
HH gender	Female	52*	66	63, 70
	Male	48*	34	30, 37

Percentage of households with HH under 20 years old		1.2	1.8	1, 3
Percentage of households whose HH has no education		21.1	21.9	18.9, 25.4
Religion of the HH	Traditional	1.9	2.2	1.3, 3.7
	Catholic	51.2	52.1	48.3, 56
	Protestant	5.8	4	3, 6
	Celestial	5.7	5.6	4, 7.7
	Islam	16.9	14.4	11.8, 17.3
	Other Christians	12.2*	16.4	13.7, 19.5
	Other religions	2.7	2.9	1.8, 4.6
	Other	2.8	2.5	1.5, 4
Percentage of foreign population		8.5	8.9	7, 11.4

*Census data not covered by the confidence interval.

HH: head of household.

FIGURES

Figure 1 : Cotonou boundaries and zone specified on the random point generator (RPG) tool.

This image is from the satellite view of Google Earth Pro™. Based on the Cotonou paper map provided by the National Geographic Institute of Benin, the boundaries of Cotonou city in red were drawn with the "Add path" tool in Google Earth Pro™. The latitudes of the northern and southern boundaries and the longitudes of the western and eastern boundaries of Cotonou enable the study area to be specified in the RPG tool of Geomidpoint.com. The places marked yellow on the figure specify these coordinates. The yellow polygon is the zone specified on the RPG tool. The specified yellow zone is larger than Cotonou. This explains the fact that certain geocoded points were generated outside the study area.

Figure 2: Illustration of the repositioned geocoded points

This image is taken from the satellite view of Google Earth Pro™. The white square represents the position of the generated geocoded points. In red, the 25m radius circle is drawn from the GPSVisualizer.com website. The geocoded points which were randomized on the roof of buildings or their associated properties (e.g. gardens, courtyards) were repositioned on the street in front of the expected entrance areas. New localizations of these points are represented by the yellow place mark (39 and 324). The geocoded points generated in front of buildings were not moved (92 and 25). Those with no buildings inside the circle were deleted (1236).

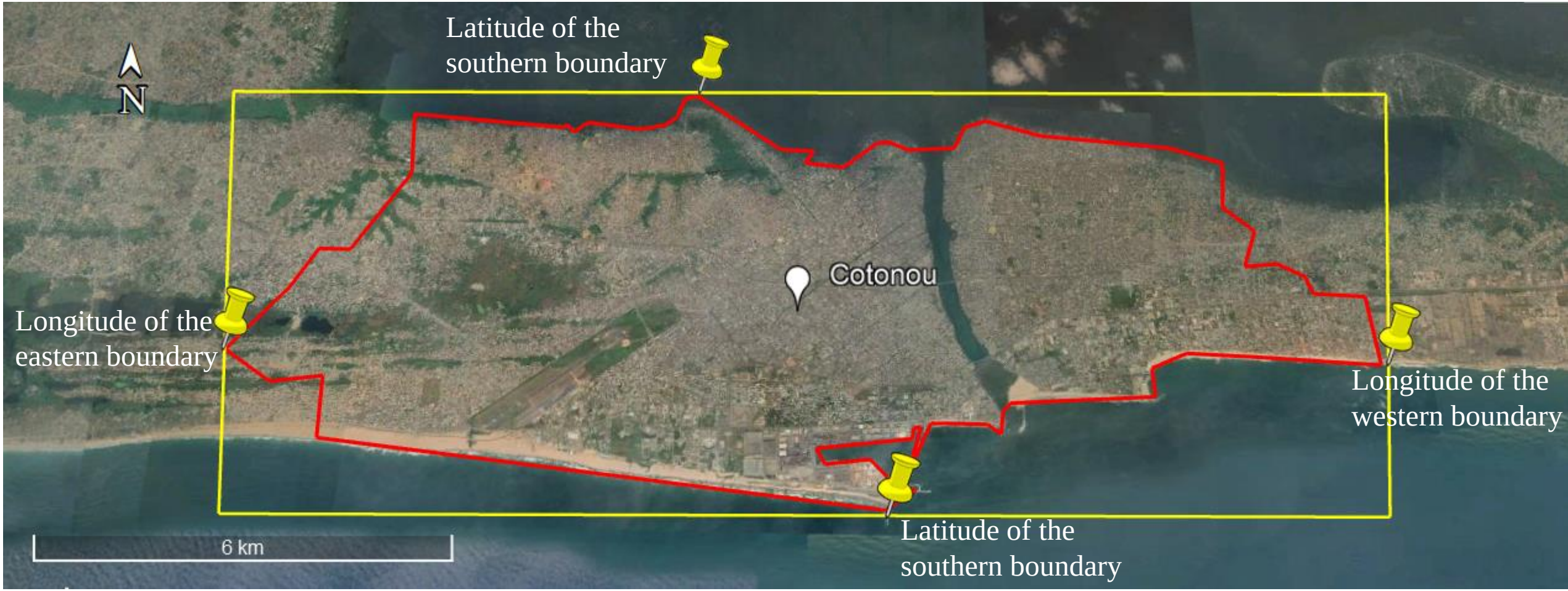
Figure 3 : Illustration of the replacement strategy.

This image is taken from the Google Earth Pro™ satellite view. Image 1 shows the replacement strategy used. The building identified by the geocoded point is designated by "i". The buildings numbered from one to five are, successively and in order of numbering, those included in the replacement strategy. Image 2 presents the proposed new replacement strategy.

Figure 4 : Inventory of generated geocoded points.

Number of interviews conducted at each step of the replacement strategy. One thousand three hundred (1300) geocoded points were randomized with the website Geomidpoint.Com, while 876 geocoded points were randomized within the study area, including the 748 buildings identified on Google Earth Pro™.

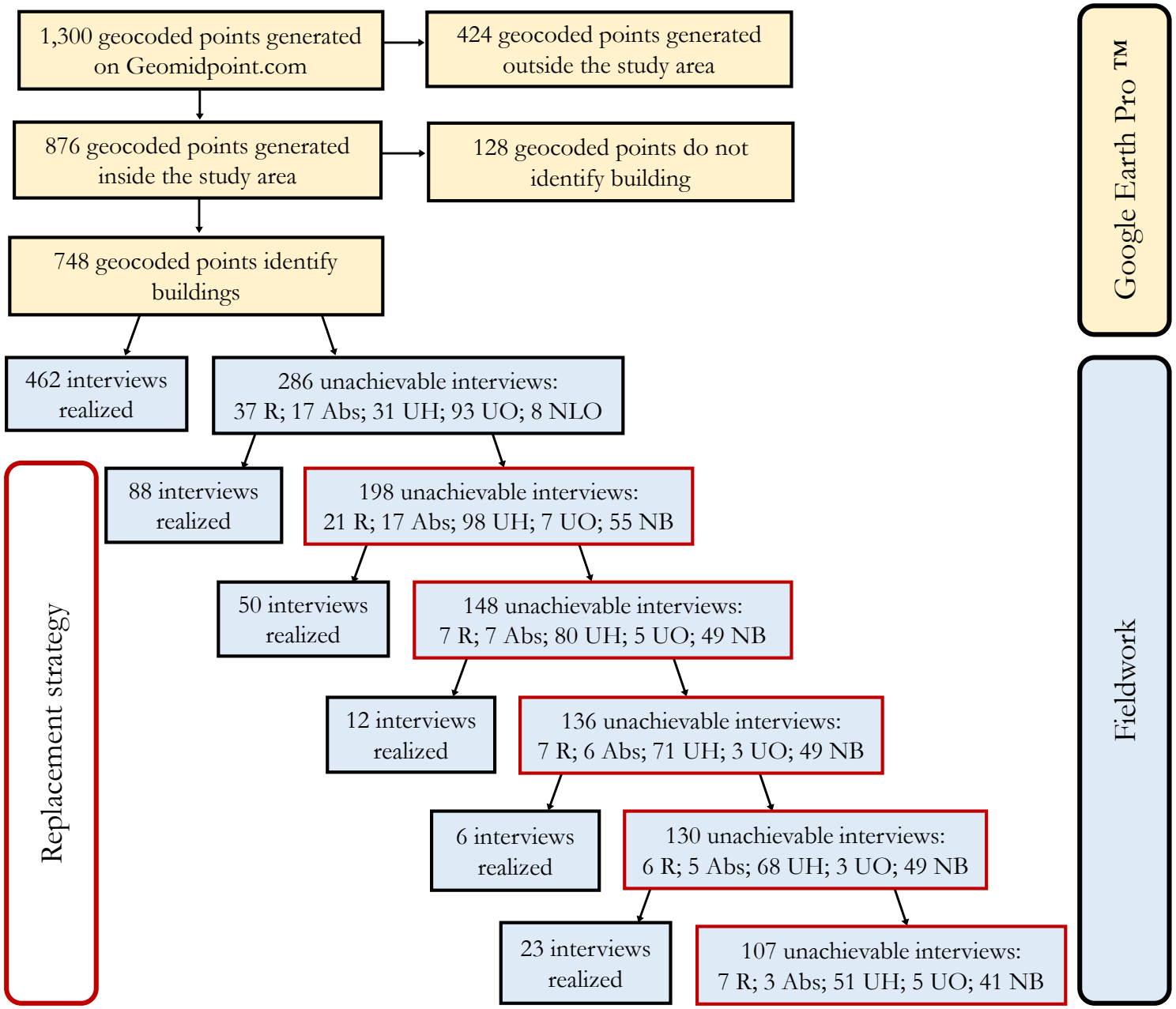
R: Refusal; Abs: Absence; UH: Uninhabited buildings; UO: Unoccupied buildings; NLO: buildings that no longer exist.


Latitude of the
southern boundary
N
Cotonou
Longitude of the
eastern boundary
Longitude of the
western boundary
6 km
Latitude of the
southern boundary


N
Repositioned 39
39
Repositioned 324
1236
324
92
25
100 m
Image © 2017 DigitalGlobe

1

2


1,300 geocoded points generated on Geomidpoint.com
424 geocoded points generated outside the study area
876 geocoded points generated inside the study area
128 geocoded points do not identify building
748 geocoded points identify buildings
462 interviews realized
286 unachievable interviews: 37 R; 17 Abs; 31 UH; 93 UO; 8 NLO
88 interviews realized
198 unachievable interviews: 21 R; 17 Abs; 98 UH; 7 UO; 55 NB
50 interviews realized
148 unachievable interviews: 7 R; 7 Abs; 80 UH; 5 UO; 49 NB
12 interviews realized
136 unachievable interviews: 7 R; 6 Abs; 71 UH; 3 UO; 49 NB
6 interviews realized
130 unachievable interviews: 6 R; 5 Abs; 68 UH; 3 UO; 49 NB
23 interviews realized
107 unachievable interviews: 7 R; 3 Abs; 51 UH; 5 UO; 41 NB
Replacement strategy
Google Earth Pro ™
Fieldwork