PROBABILITY SAMPLING BY CONNECTING SPACE WITH HOUSEHOLDS USING GIS/GPS TECHNOLOGIES

XINGUANG CHEN*
HUI HU
XIAOHUI XU
JIE GONG
YAQIONG YAN
FANG LI

Sampling methods for survey studies are challenged by the replacement of landline telephones with mobile phones, the lack of timely census data, and the growing need for studies to address new health challenges. GIS/GPS-assisted methods provide a promising alternative, but these methods need further improvement. We established a stratified 3-stage GIS/GPS-assisted sampling method in which residential areas of a target population are divided into mutually exclusive cells – geographic units (geounits) as the primary sampling frame (PSF). Geounits with residential households were randomly selected from the PSF with a semi-automatic algorithm implemented in R. Novel methods were used to sample households and participants. Simulations and application studies indicated adequate feasibility, efficiency and validity of the method in sampling rural-to-urban migrants from a large city with complex residential arrangements. With this method, researchers can determine sample size and number of geounits, households and participants to be sampled; optimally allocate geounits; determine area size of sampled geounits and estimate sample weights; and complete sampling for field data collection in a short period. Our method adds an integrative approach for GIS/

XINGUANG CHEN, Professor, and HUI HU, Research Assistant Scientist, are with the Department of Epidemiology, University of Florida, Gainesville, FL 32610, USA.

XIAOHUI XU, Associate Professor, is with the Department of Epidemiology & Biostatistics, Texas A&M Health Science Center, College Station, TX 77843, USA.

JIE GONG, Professor, YAQIONG YAN, Associate Professor, and FANG LI, Associate Professor, are with the Wuhan Center for Disease Prevention and Control, Wuhan 430022, China.

*Address correspondence to Xinguang Chen, Department of Epidemiology, University of Florida, Gainesville, FL 32610, USA; E-mail: jimax.chen@ufl.edu.

doi: 10.1093/jssam/smx032 Advance access publication 23 January 2018
© The Author 2018. Published by Oxford University Press on behalf of the American Association for Public Opinion Research. All rights reserved. For permissions, please email: journals.permissions@oup.com
GPS-assisted random sampling with a de facto population assumption. Additional evaluation studies are needed to assess the utility of this method in different settings.

KEYWORDS: Epidemiology; GIS/GPS-assisted sampling; Probability sampling; Survey studies.

1. INTRODUCTION

Modern empirical sciences in public health and medicine have been established largely with survey data collected from random or probability samples. Design-based statistical inferences require data collected from a probability sample, and results from a sample of participants can be generalized to a large population only when individual participants are selected with a known probability (Kish 1965; Cochran 1977; Chen, Yin, and Peng 1999; Levy and Lemeshow 1999; Groves Fowler, Couper, Lepkowski, and Singer 2009). Despite its importance, a review of the literature indicates that probability sampling is infrequent in studies published in even very prestigious peer-reviewed journals. For example, a review of articles published in *Journal of Acquired Immune Deficiency Syndrome* and *AIDS and Behavior* during June–December of 2014, we found only two out of seventy-two (3%) and six out of 109 (6%) of the population-based survey studies, respectively, used a probability sample for data collection.

1.1 Methodology Barrier to Probabilistic Sampling

One primary barrier preventing researchers from using a probabilistic sample design could be the lack of appropriate methods (Landry and Shen 2005; Shannon, Hutson, Kolbe, Stringer, and Haines 2012; Wampler, Rediske, and Molla 2013; Escamilla, Emch, Dandalo, Miller, Martinson et al. 2014; Chen, Yu, Zhou, Zhou, Gong et al. 2015). One approach is landline telephone number-based random digit dialing (Kish 1965; Cochran 1977). Although this method is very efficient for sampling, incomplete population coverage has been an issue (Groves et al. 2009). Approximately 3–5% of the households in the United States do not have a landline telephone and thus are not included in the sampling frame (Groves et al. 2009). People living in these households are more likely to be low in socioeconomic status, drug users, sex workers, and/or undocumented migrants (Groves et al. 2009; Singh and Clark 2012), all of whom are at increased risk for poor health. More threatening than incomplete coverage is the replacement of landline telephones by wireless communication technologies that makes it impossible to implement the random digital dialing method.

Most methods attempt to achieve randomness by using census data with detailed demographic information at the household level to construct the
sampling frame (Kish 1965; Cochran 1977; Groves et al. 2009). However, such data are often not available in a timely manner in all developed countries and unavailable in many resource-limited lower- and middle-income countries. In developed countries, census data are collected only for selected years (usually five or ten years apart), and many resource-limited countries do not collect population census data on a regular basis. Even if census data are available, they may fail to count people who are at high risk for poor health, such as temporary and undocumented immigrants (Groves et al. 2009; Singh et al. 2012).

In survey research, sometimes a study population can be operationally defined: for example, non-institutional residents; gender; racial/ethnic minority groups in a country; high school students in a state; or hospitalized patients in a region. In this case, methods are readily available to draw probability samples, such as the multi-stage random sampling methods for national household surveys supported by census data (Cochran 1977), telephone surveys supported by random digit dialing with published telephone numbers (Groves et al. 2009), and school surveys using system sampling methods with complete lists of schools and classes (Eaton, Kann, Kinchen, Shanklin, Flint et al. 2012). However, sometimes the sampling frame for a study population may be clear conceptually but hard to define operationally. Challenging examples for drawing probability samples include mobile migrants, sex workers, drug users, and persons living with HIV (Groves et al. 2009; Singh et al. 2012).

Timing can be another challenge for random sampling when survey studies are needed to address an urgent public health and medical issue (Heeringa and O’Muircheartaigh 2010a; Heeringa and Zinniel 2012). Typical examples include studies of outbreaks and vaccination of infectious diseases, such as HIV/AIDS, severe acute respiratory syndrome (SARS) (Tong 2005; He, Zhuang, Zhao, Dong, Peng et al. 2007), Ebola (Weyer, Grobbelaar, and Blumberg 2015; Boisen, Hartnett, Goba, Vandi, Grant et al. 2016), and Zika (Boeuf, Drummer, Richards, Scoullar, and Beeson 2016; Deseda 2017). Innovative methods have been attempted for timely sampling without using a sampling frame. Well-known examples include venue-day-time sampling, where participants are selected from locations within time ranges when participants are often present (Mansergh, Naorat, Jommaroeng, Jenkins, Jeeyapant et al. 2006); the capture-recapture method derived from agriculture and wild life studies (Tilling 2001); and respondent-driving sampling (RDS), in which study participants are selected by working with a few seed persons to nominate others within their network connections (Heckathorn 1997, 2002). Although these methods allow for timely sampling of study participants, their validity in ensuring probability and representative samples is unclear.
1.2 GIS/GPS-Assisted Methods as an Alternative

Technological advances in geographic information systems (GIS) and global positioning systems (GPS) have encouraged numerous researchers to develop speedy probabilistic sampling methods with adequate geographic and population coverage, with minimal data requirements (Murray, O’Green, and McDaniel 2003; Landry et al. 2005; Galway, Bell, Sae, Hagopian, Burnham et al. 2012; Shannon et al. 2012; Chen et al. 2015). A number of GIS/GPS-assisted probability sampling methods have been developed to deal with specific settings, such as sampling in remote rural areas (Wampler et al. 2013; Escamilla et al. 2014; Kondo, Bream, Barg, and Branas 2014; Haenssgen 2015; Pearson, Rzotkiewicz, and Zwicky 2015), mobile populations (Landry et al. 2005; Singh and Clark 2013; Chen et al. 2015), and other special conditions (Murray et al. 2003; Galway et al. 2012).

A review of the published studies reveals that most GIS/GPS-assisted sampling methods can be characterized as geographically stratified multi-stage sampling. These methods can be summarized in seven steps:

1. Define targeted study population and geographic area,
2. Construct primary sampling frame (PSF) and define residential area to determine the primary sampling units (PSUs),
3. Randomly select PSU with a probabilistic procedure (simple random, proportion to or stratified by population density),
4. Select households from each sampled PSU through random routes or other methods, and enumerate households to construct the secondary sampling frame (SSF),
5. Randomly select a pre-determined number of participants from SSF,
6. Compute sample weights across all sampling stages,
7. Estimate descriptive statistics for the study population, taking into account the sample design and sampling weights.

1.3 Challenges to Implementing a GIS/GPS-Assisted Sampling Method

Despite much progress, additional research is needed on GIS/GPS-assisted sampling methods. First, it is challenging to pre-determine the sample size for several reasons (Landry et al. 2005; Singh et al. 2012; Valliant, Dever, and Kreuter 2013b). The method consists of two steps: sample geographic areas, then sample participants in selected areas. Sample size is easy to determine if a study only needs to draw geographic samples (Balch, Drapeau, Bowler, Booth, Goes et al. 2004; Chen, Zhao, Gao, Henkelmann, and Schramm 2006; Conway 2006; Daly, Lei, Teixeira, Muir, Castillo et al. 2007; Huang, Zhao, Shi, Yu, Zhao et al. 2007; Valliant et al. 2013b: Pearson et al. 2015). However, it is not possible to know exactly how many persons are present in a randomly
sampled geographic area before the area is selected (Landry et al. 2005; Shannon et al. 2012; Valliant et al. 2013b; Chen et al. 2015). One solution is to enumerate all households in a randomly sampled area after a geographic area is selected. This method has often proved infeasible (Landry et al. 2005; Escamilla et al. 2014; Chen et al. 2015) because of high and variable population density, complex residential structure, and presence of high-rise and multi-function buildings in selected geographic areas.

Second, GIS/GPS-assisted sampling method needs to distinguish residential from non-residential housing. Methods for distinguishing between the two have proven very time-consuming to implement (Chen et al. 2009; Singh et al. 2012; Escamilla et al. 2014; Kondo et al. 2014; Pearson et al. 2015). Recent methods have been developed to recognize visually or digitally residential areas/housing with widely available aerial images (Chang et al. 2009; Wampler et al. 2013; Escamilla et al. 2014; Haenssgen 2015; Pearson et al. 2015). These methods are fast, inexpensive, and highly feasible. However, correctly recognizing residential houses remains a big problem even with the assistance of people from local communities. For example, Pearson et al. (2015) conducted a study to determine residential households using aerial images. With assistance of local experts after computerized sampling, five out of 175 determined residential household structures were verified in the field to be non-residential. Although the error rate is not high, this study was conducted in a semi-nomadic pastoral area, a setting much simpler for random sampling than that of a modern urban area. More research is needed to improve this method for use in sampling complex residential areas (Escamilla et al. 2014; Chen et al. 2015; Haenssgen 2015).

Third, stratification has been used in GIS/GPS-assisted sampling to deal with heterogeneities in population density (Kumar 2007; Galway et al. 2012; Valliant et al. 2013b; Kondo et al. 2014). Galway and colleagues have used this approach in their studies and generated promising results (Galway et al. 2012). However, for stratification to be effective, detailed demographic data at the population level by individual grid cells across a jurisdiction are needed (Galway et al. 2012; Valliant et al. 2013b), and often such data are not available in resource-limited, low- and middle-income countries. Night-time satellite images provide information regarding population density, but this approach does not work for rural areas and resource-limited countries, and places with no electricity (Sutton 1998; Schneider, Friedl, and Potere 2009).

Last but not least, geographic sampling weights are difficult to determine because of the lack of clear boundaries between residential and nonresidential areas and lack of information on the number of persons living in sampled geographic units at a specific date and time (Landry et al. 2005; Kumar 2007; Shannon et al. 2012; Valliant et al. 2013b; Kondo et al. 2014).
1.4 Purpose of This Study

In this study, we report on our attempts to overcome the challenges described above. Our goal is to promote the use of GIS/GPS-assisted sampling method in survey studies with probability samples to better address medical and public health questions.

2 METHODS AND MATERIALS

2.1 Spatial Sampling

2.1.1 Principle and geographic data. Geographic data for locations where the study population resides (often by country or jurisdictions within a country) can be obtained from different sources, mostly free of charge (e.g., Google Maps and OpenStreetMap) (Haklay and Weber 2008). The area will be divided into mutually exclusive cells (geographic units, or geounits for short) for further sampling. This spatial sampling process is often realized by creating and laying a grid over the target area (figure 1) and then randomly drawing geounits.

2.1.2 Determination of the area size of a geounit. Determination of the area $A$ of a geounit is critical to create the grid network described in the previous and next sections. In traditional spatial sampling, $A$ is simply calculated using the sampling ratio (Stehman and Overton 1995; Maguire, Batty, and Goodchild 2005; Valliant et al. 2013b). For example, if a researcher plans to sample 8 geounits to cover 0.01% ($10^{-4}$) of a geographic area with the total area size of 1,000,000 ($10^6$) km$^2$, the area for individual geounits is 12.5 km$^2$ ($=10^6 / 10^{-4} / 8$).

A more complex method is needed to draw geographic samples for population-based survey studies, because $A$ is determined not by sampling ratio but by the likelihood of covering an appropriate number of households and eligible persons. Using a larger $A$ increases the chance of sampling adequate numbers of households and eligible subjects, but also increases the workload for household enumeration. The appropriate size $A$ can be determined through pilot studies, considering population density and the number of subjects to be recruited per geounit. For example, when conducting a survey targeting the rural-to-urban migrants temporarily living in Wuhan, China, $A = 100 \times 100$ m was determined through intensive pilot tests in the field. This number was estimated by counting all households located in a geographic area with different sizes being measured manually using tape rulers and/or laser scales. This value of $A$ has approximately 80% probability of covering an adequate number of households, ensuring at least 20 subjects per geounit in a city like Wuhan (Chen et al. 2015).
2.1.3 Creation of grid network and selection of geounits. After $A$ is determined, a grid network is then created and overlaid on the target geographic area, dividing the area into mutually exclusive cells (with cell size of $A$). These cells are the primary sampling frame (PSF) for further sampling (figure 1). Different methods are available for grid network creation; it is often simpler to use geographic coordinate systems rather than side length, and the differences between the two approaches are often small for geographic areas on the scale of a city or a state. For studies involving very large geographic areas like a country (e.g., Russia, Canada, China, or the United States), continents, or the globe, distance defined through appropriate projection systems should be employed.

After the PSF is created, a pre-determined number of geounits (to be discussed next) are randomly selected from the PSF. Unlike Pearson’s method, which uses a set of randomly scattered points as place marker (Pearson et al. 2015), we randomly sample a set of geounits with size $A$ in the geographic area where the study population resides. Given large variations in population density across the geographic area, a stratified strategy is used to sample geounits with more geounits being allocated to areas with higher population density. This step is conducted following an optimum allocation approach to enhance work efficiency (Cochran 1977).

Another issue confronted in practice is that a randomly selected geounit has a sizable likelihood of being located in nonresidential areas such as lakes, bridges, highways, and commercial buildings. To overcome this problem and to enhance feasibility while maintaining a probabilistic sampling process, we devised a semi-automatic, computer-assisted, stepwise algorithm with no replacement procedure for implementing the geounit sampling protocol (figure 2). More details and the R codes for implementation are provided in Appendix 1 of the online supplementary material.
2.2 Number of Geounits to Be Sampled

Before sampling, the number of geounits to be sampled $G$ has to be determined. Obviously $G$ depends on the sample size $N$ and the average number of participants to be sampled per geounit $M$. There is a lack of efficient methods to determine $G$ in reported studies (Landry et al. 2005; Kondo et al. 2014). To overcome this limitation, we used the same $M$ for all geounits. With $M$ and sample size $N$ determined through conventional statistical power analysis, $G$ can simply be calculated as:

$$G = \frac{N}{M}. \tag{1}$$

For example, assuming a researcher plans to draw a sample of $N = 1,200$. If twenty subjects are to be sampled from each geounit, based on equation 1, the
2.3 Household and Subject Sampling

2.3.1 Locating the sampled geounits. After completing spatial sampling steps as described in the previous two sections, detailed geographic data for the selected geounits are available and can be directly uploaded to GPS receivers. In this study, we tested our method using the Oregon 450 GPS from Garmin, but any GPS receiver can be used if it can upload sampled geographic units with coordinates and areal images and is able to track specified areas manually.

2.3.2 Sequential household access and random participant selection. After a sampled geounit is located in the field with the assistance of a GPS receiver, data collectors go to all households within the sampled geounit to prepare for subject sampling and recruitment. This procedure is completed in three steps. Step one: Approach individual households sequentially following natural order with the first household being selected randomly from the main entrance of a street in urban areas or the beginning of a village in rural areas. We used this random route approach to ensure each household has a known probability of being selected (de Rada and Martin 2014; Bauer 2016). Step two: List all eligible participants for each selected household. The list for a household constitutes the secondary sampling frame (SSF). Step three: Select participants randomly from the SSF. If only one person in a household is eligible, this person will be included. If more than one is eligible, only one will be randomly selected using the Kish Table or other random digit method (Kish 1949). Households not available at the time of sampling are revisited to reduce missingness.

An innovation of our method is that data collectors can determine the number of households to be enumerated for a geounit. This is because data collectors have already been told the number of households $H$ and the number of subjects per household $S$ to be sampled. Assuming $M = 20$ and only one subject per household is to be sampled ($S = 1$), approximately twenty households are enumerated ($H = 20$). In addition to minimizing work load, this approach moves the sampling probability of a geounit closer to being proportionate to the population density. This is because the ratio of $H$ to the total households in a same-sized geounit will be smaller in a more populated area and larger in a less populated area.

2.3.3 Complementary data collection. After household enumeration, participant recruitment, and data collection for a given geounit, the following complementary data must be collected. 1) Actual geounit area size $A_g$ for the $g^{th}$
sampled geounit \((g = 1, 2, \ldots, G)\) where households are accessed and participants are sampled. Areas where households are not enumerated must be excluded. The actual area size \(A_g\) is determined by GPS receiver recorded tracking data. 2) Total number of households \(T_g\) in the accessed area \(A_g\) and number of households from which participants are sampled \(H_g\). 3) Households and individuals who refused to participate.

2.4 Determination of Residential Areas

To estimate the overall sample weights, the true area size \(R\) where the target population resides must be determined to estimate geographic sample weights. Although the concept of residential area has no ambiguity, it is often difficult to determine \(R\) (Shannon et al. 2012; Singh et al. 2012; Valliant et al. 2013b; Kondo et al. 2014). We have devised two alternative methods for use in different settings.

2.4.1. Method I: estimate residential area with population and geographic sample data. If \(A_g\) = actual geounit area size for the \(g^{th}\) sampled geounit \((g = 1, 2, \ldots, G)\) where households are accessed and participants are sampled (described in Section 2.3.3), then the total area of \(G\) geounits \(B = \sum_{g} A_g\). Let \(R\) = the total residential area, \(P\) = the total population known to reside in the target district, and \(Q\) = the total population covered by all \(G\) sampled geounits (the total population \(Q_g\) for \(g^{th}\) geounit is estimated using total households \(T_g\) and demographic data from the enumerated households \(H_g\)). Both \(B\) and \(Q\) are calculated based on area size and population data obtained from the randomly selected geounits. If \(G\) is adequately large (e.g., twenty or more), the ratio of the two provides an unbiased and reliable estimate of the ratio of \(R\) over \(P\). That is,

\[
\frac{R}{P} \approx \frac{B}{Q},
\]

So we set

\[
R = P \times \frac{B}{Q}.
\]

This approximating method relies on the expectation that households in a sampled geounit are associated only with that geounit. In practice, this can be achieved by carefully determining the appropriate grid size \(A\) through pilot studies.

2.4.2 Method II: estimate residential area with Monte Carlo method. If data for total population \(P\) is not available, the residential area \(R\) can be estimated using a Monte Carlo method (Metropolis and Ulam 1949; Mathews 1972).
The size of a target area $D$ can be obtained through many GIS packages as described in section 2.1. With the Monte Carlo method, a target area is uploaded to computer. A total number of $n$ points (i.e., several hundred) are randomly selected within the whole area. If $n_r$ points fall on residential areas, and $n_{nr}$ non-residential areas, then $n = n_r + n_{nr}$. Since all points are randomly selected, $n_r/n$ provides an unbiased estimate of $R/D$. Thus,

$$ R = \frac{n_r}{n_r + n_{nr}} D $$

(4)

2.5 Sample Weights

The following equation is used to compute sample weights following the principles for stratified, multistage, and disproportionate probability sampling (Kish 1965; Cochran 1977; Groves et al. 2009; Valliant et al. 2013b).

$$ W_i = W_g \times W_{gh} \times W_{ghi} $$

(5)

Where $W_g$ represents the sample weight for $g$th geounit and equals $R/A_g$, where $R$ is the size of total residential area and $A_g$ is the size of $g$th sampled geounit; $W_{gh}$ represents the sample weight for household $h$ in geounit $g$ and equals $T_g/H_g$, where $T_g = \text{total households in geounit } g$ and $H_g = \text{number of households sampled in geounit } g$; and lastly, $W_{ghi}$ = sample weight for individual subject $i$ from household $h$ and equals $N_{gh}/n_{jh}$, where $N_{gh} = \text{total number of eligible persons in household } h$ within geounit $g$, and $n_{gh} = \text{number of persons sampled in the household } h, h = 1, 2, \ldots, H_g$. If only one person per household is sampled, $n_{gh} = 1$ and $W_{ghi} = N_{gh}$. Variance estimation methods are needed to correctly account for the variance inflation due to weighting, (Kish 1965) and the design effect attributable to the clustering of observational units within the sampled area PSUs also needs to be considered (Kish 1965; Heeringa, West, and Berglund 2010b; Valliant, Dever, and Kreuter 2013a). The supplementary Appendix 2 online describes a simulation study conducted to validate this method using both the jackknife replication and the bootstrap for variance estimation (Valliant et al. 2013a).

2.6 Practical Testing

We tested the integrative GIS/GPS-assisted sampling method in Wuhan, China, when conducting an NIH-funded project (R01 MH086322, PI: Chen X) to investigate the relationship between social capital and HIV risk behaviors among rural-to-urban migrants. Wuhan is the capital of Hubei Province with a total population of approximately ten million, per capita GDP of $12,708, and a large number of rural-to-urban migrants (Statistical Bureau of Wuhan 2012).
The field work for sampling and data collection was completed during 2012–2014. Many migrants do not have a permanent urban residence, and all of them are scattered all over the city. In this case, it is not possible to construct a sampling frame using conventional methods.

3. RESULTS

3.1 Geographic Sampling Frame and Geounits

Following the procedure described in this study, a district boundary file of Wuhan was obtained using the ArcGIS. Based on pilot studies for field work efficiency, a grid-system with 100m × 100m cells was created and imposed on the map to divide the geographic area of Wuhan into small and mutually exclusive cells as geounits (see figure 1). These mutual exclusive cells consist of the PSF for further sampling.

A total of sixty geounits with residential housing were randomly drawn from the PSF and stratified by population density following the steps described in section 2.1. A sample size of sixty geounits was chosen to achieve a total sample of 1,200 with approximately twenty participants per geounit. Allocation of the sixty geounits to districts was optimized, considering traveling cost and cost for field data collection. Figure 3 shows the geographic distribution of the sampled geounits. Households within each of these sampled geounits were then accessed and participants recruited following the steps described in section 2.3. Actual geounit area size $A_g$ was determined with data collected during sampling (see Appendix 3 in the online supplementary material for more details).

3.2 Samples of Households and Participants

Overall, sixty sampled geounits covered 12,016 households, of which approximately 10–25% were occupied by rural migrants. Households were selected following their natural order on a street with the beginning of a main entrance street as the start point, and the first household was determined using random numbers. Of the migrant-occupied households, 1,251 were available and agreed to participate at the time of data collection. A total of 1,310 participants were recruited from these households with one participant per gender per household. The total number of households per geounit varied from thirty in least populated areas to 1,600 in the most populated areas with a median [quartile 1, quartile 3] of one hundred [50, 300] and mean (SD) = 200 (242). The number of households agreeing to participate per geounit varied from twelve to forty with median [quartile 1, quartile 3] of twenty [18, 24] and mean (SD) of twenty-one (6). Table 1 shows the detailed results from the sampling. Applying this method to another study conducted in 2012–2014, we estimated...
that approximately fifty-eight thousand [95% CI: 47,000, 68,000] rural-to-urban migrants in Wuhan were MSM with 3,650 [95% CI: 2,960, 4,282] being tested HIV positive (Chen et al. 2015). Official surveillance data from Wuhan indicated that a total of 3,408 (primarily MSM) persons were living with HIV in 2015 (Wuhan Center for Disease Prevention and Control (CDC) 2016). The observed result is within the estimated 95% CI, and the relatively small difference provides some evidence supporting the validity of our method.

4. DISCUSSION AND RECOMMENDATIONS

In this study, we reported a geographically stratified 3-stage (geographic unit, household, and participants) GIS/GPS-assisted sampling method. This method is developed by integrating various reported GIS/GPS-assisted sampling methods (Chang et al. 2009; Wampler et al. 2013; Escamilla et al. 2014; Haenssgen 2015; Pearson et al. 2015), particularly the methods with a stratified cluster sampling approach (Cochran 1977; Groves et al. 2009). Innovations include methods to determine residential area and methods for sample weight calculation. Our method enhances existing approaches to drawing probability samples for local, national, cross-national, and global survey studies (Heeringa et al. 2010a; Heeringa et al. 2012).

4.1. Strengths of Our Method

Our method is based on sound theories for population and geographic sampling, and has minimal data requirements. Conventional stratified sampling
Table 1. Results of GIS/GPS-Assisted Sampling of Rural-to-Urban Migrants, Wuhan, China

| Geounit Id | Actual area ($m^2$) | Weight $W_g$ | Total households | Accessed households | Weight $W_{gh}$ | Participants recruited |
|------------|----------------------|--------------|------------------|--------------------|-----------------|-----------------------|
| Total      | 1,812,600            | 3453.99      | 12,016           | 1,251              | 9.61            | 1,310                 |
| M017       | 10,000               | 108.99       | 400              | 25                 | 16.00           | 25                    |
| M023       | 20,000               | 54.55        | 50               | 21                 | 2.38            | 21                    |
| M024       | 28,750               | 37.89        | 80               | 14                 | 5.71            | 20                    |
| M025       | 10,000               | 108.99       | 528              | 25                 | 21.12           | 25                    |
| M026       | 15,000               | 72.66        | 200              | 24                 | 8.33            | 24                    |
| M033       | 45,000               | 11.22        | 87               | 28                 | 3.11            | 28                    |
| M037       | 15,000               | 33.77        | 312              | 21                 | 14.86           | 22                    |
| M038       | 77,500               | 6.56         | 300              | 22                 | 13.64           | 23                    |
| M039       | 135,000              | 3.78         | 110              | 23                 | 4.78            | 26                    |
| M045       | 10,000               | 50.66        | 60               | 12                 | 5.00            | 12                    |
| M047       | 20,000               | 25.33        | 50               | 19                 | 2.63            | 20                    |
| M049       | 10,000               | 50.66        | 274              | 13                 | 21.08           | 19                    |
| M050       | 27,500               | 18.45        | 70               | 23                 | 3.04            | 23                    |
| M056       | 21,250               | 23.89        | 100              | 29                 | 3.45            | 31                    |
| M057       | 40,000               | 24.44        | 500              | 24                 | 20.83           | 25                    |
| M058       | 15,000               | 65.22        | 600              | 19                 | 31.58           | 21                    |
| M070       | 17,500               | 55.88        | 260              | 25                 | 10.40           | 25                    |
| M072       | 32,500               | 30.11        | 180              | 45                 | 4.00            | 27                    |
| M076       | 10,000               | 97.77        | 140              | 17                 | 8.24            | 18                    |
| M078       | 23,300               | 42.00        | 50               | 20                 | 2.50            | 20                    |
| M079       | 25,000               | 40.66        | 30               | 18                 | 1.67            | 18                    |
| M082       | 36,250               | 28.11        | 100              | 18                 | 5.56            | 19                    |
| M083       | 36,250               | 28.11        | 400              | 40                 | 10.00           | 40                    |
| M086       | 28,750               | 35.44        | 55               | 12                 | 4.58            | 15                    |
| M087       | 22,500               | 45.22        | 100              | 20                 | 5.00            | 20                    |
| M094       | 80,000               | 12.78        | 50               | 26                 | 1.92            | 26                    |
| M095       | 12,500               | 81.44        | 40               | 18                 | 2.22            | 20                    |
| M100       | 11,250               | 90.44        | 500              | 26                 | 19.23           | 26                    |
| M102       | 18,250               | 55.77        | 45               | 21                 | 2.14            | 21                    |
| M106       | 13,500               | 42.55        | 1,600            | 18                 | 88.89           | 21                    |
| M111       | 53,300               | 10.78        | 348              | 20                 | 17.40           | 21                    |
| M114       | 15,000               | 38.33        | 400              | 26                 | 15.38           | 29                    |
| M117       | 35,000               | 16.44        | 300              | 25                 | 12.00           | 25                    |
| M126       | 15,500               | 37.11        | 400              | 20                 | 20.00           | 21                    |
| M128       | 47,500               | 12.11        | 50               | 21                 | 2.38            | 21                    |
| M134       | 75,000               | 7.67         | 500              | 21                 | 23.81           | 23                    |
| M138       | 46,200               | 12.44        | 40               | 11                 | 3.64            | 18                    |
| M144       | 65,000               | 8.89         | 60               | 19                 | 3.16            | 19                    |
| M148       | 49,000               | 11.78        | 50               | 19                 | 2.63            | 21                    |

Continued
strategies can be used in optimizing geounit allocation to deal with large variations in population density and to increase field-work efficiency (Cochran 1977). The size of geounits can be determined through pilot testing to ensure adequate household/participant coverage, while taking work efficiency into account (Chen et al. 2015). The random route method (Bauer 2016) can be used to ensure an equal probability household sample. Data collected using our method can be analyzed with design-based survey methods (Kish 1965; Cochran 1977; Lohr 1999; Groves et al. 2009; Heeringa et al. 2010b; Valliant et al. 2013b). These methods are available in many software packages, including SUDAAN, SAS, STATA (survey module), SPSS, and “survey” package in R.

Many of the sampling tasks of our method can be implemented on computer with open-source software R and free Google imagery data. A more detailed discussion of the application of our methods is provided in Appendix 3 of the

| Geounit Id | Actual area (m²) | Weight W₉ | Total households | Accessed households | Weight W₉gh | Participants recruited |
|------------|------------------|------------|------------------|--------------------|-------------|-----------------------|
| M152       | 27,500           | 29.33      | 70               | 18                 | 3.89        | 19                    |
| M153       | 20,000           | 40.22      | 40               | 19                 | 2.11        | 21                    |
| M156       | 20,000           | 40.22      | 92               | 15                 | 6.13        | 20                    |
| M159       | 37,500           | 21.44      | 50               | 19                 | 2.63        | 18                    |
| M160       | 25,000           | 32.22      | 40               | 22                 | 1.82        | 25                    |
| M166       | 21,250           | 37.89      | 60               | 18                 | 3.33        | 21                    |
| M171       | 16,250           | 49.55      | 50               | 20                 | 2.50        | 21                    |
| M172       | 52,500           | 15.33      | 421              | 19                 | 22.16       | 19                    |
| M189       | 25,000           | 32.22      | 300              | 27                 | 11.11       | 27                    |
| M194       | 46,250           | 17.44      | 147              | 25                 | 5.88        | 27                    |
| M198       | 26,250           | 997.23     | 256              | 18                 | 14.22       | 20                    |
| M199       | 37,500           | 32.11      | 80               | 20                 | 4.00        | 20                    |
| M204       | 25,000           | 48.22      | 250              | 20                 | 12.50       | 21                    |
| M207       | 15,000           | 80.33      | 40               | 19                 | 2.11        | 20                    |
| M211       | 20,000           | 60.22      | 212              | 20                 | 10.60       | 21                    |
| M221       | 14,500           | 83.10      | 100              | 16                 | 6.25        | 16                    |
| M227       | 14,500           | 83.10      | 60               | 14                 | 4.29        | 14                    |
| M229       | 15,000           | 80.33      | 50               | 17                 | 2.94        | 19                    |
| M252       | 23,750           | 50.77      | 60               | 24                 | 2.50        | 24                    |
| M258       | 37,500           | 32.11      | 69               | 14                 | 4.93        | 19                    |
| M260       | 23,300           | 51.77      | 150              | 19                 | 7.89        | 19                    |
| Median     | 23,525           | 37.89      | 100              | 20                 | 5           | 21                    |
| IQR        | 15,000-37,500    | 22.05-55.47| 50-300           | 18-24              | 2.71-13.36  | 19-25                 |
| Mean       | 30,210           | 57.57      | 200.27           | 20.85              | 9.63        | 21.83                 |
| SD         | 21,900           | 126.13     | 241.81           | 5.75               | 12.58       | 4.34                  |
4.2 Recommendation for Application

GIS/GPS-assisted sampling methods are becoming increasingly available. If a target study population is located in sparsely populated and less developed rural areas, methods with satellite images to identify households for random sampling are a better choice. Typical examples include methods reported by Haenssgen (Haenssgen 2015), Wampler (Wampler et al. 2013), and Escamilla and colleagues (Escamilla et al. 2014). However, if a researcher wants to conduct studies in highly developed urban settings with more complicated residential arrangements, our method would be a better choice than many other methods to ensure probability samples (Landry et al. 2005; Galway et al. 2012; Kondo et al. 2014). To ensure successful application of our method in drawing a probability sample to represent a study population, researchers must pay additional attention to the following three aspects.

The first aspect is related to variations in population density. The fundamental mechanism of our method is to link geographic area with varying population density to households using numerous small geounits for further sampling. Therefore, one natural approach to deal with varying population density is application of the classic stratified sampling strategy to optimize geounit allocation (Cochran 1977), as have been commonly used in this and other studies (Galway et al. 2012; Chen et al. 2015). Our method also offers other possibilities to deal with varying population density issues. For example, instead of using a fixed geounit size and sampling grid, with our method researchers can determine the geounit size disproportionate to population density after randomly selecting the pre-determined number of geounits to be selected. Although determination of population density could remain be a challenge resource-limited areas, we may be able to deal with it with satellite imagery that is widely available.

The second aspect is the determination of area size of a geounit. Larger sizes have greater probability of covering an adequate number of households for sampling. However, if a large-sized geounit is randomly selected in a highly populous area, it will prevent researchers from completing the sampling due to high costs of time and money (Landry et al. 2005). We recommend that researchers conduct adequate pilot studies to determine geounit size, considering variations in population density, time, and resources available for sampling.

The third aspect is household selection within a sampled geounit. Although each selected geounit is not large in area size with a relatively fewer number of households, household arrangement can still be complex. In this study, we
used the random route approach (Bauer 2016), by randomly selecting one household as starting point and then following natural order to select other households until the pre-determined number of households was reached. However, our method may lead to biased estimates of parameters that are related to physical distance. This can happen even with carefully planned and well-tested instructions (Bauer 2016). If conditions permit, an ideal approach would be to list all households in a sampled geounit first and then randomly select the pre-determined number of household for further sampling.

4.3 Limitations and Further Research

In this study, we only demonstrate our method in sampling rural migrants in urban China. A full assessment of the value of our approach requires its application to different populations in diverse geographic and residential settings. Like any multistage sampling method, it is a challenge to ensure an equal probability sample of households. The random route provides a good option, but attention must be paid to instructions to the data collectors and random selection of the starting household (Bauer 2016). Data on the size of a geographic unit is often not directly available, and can be obtained only through repeated pilot tests. Given large variations in household and population density in urban settings, large variations in estimated sample weights are anticipated. Such variations may reduce the precision of sample estimates.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

References

Balch, W. M., D. T. Drapeau, B. C. Bowler, E. S. Booth, J. I. Goes, A. Ashe, J. M. Frye (2004), “A Multi-Year Record of Hydrographic and Bio-Optical Properties in the Gulf of Maine: I. Spatial and Temporal Variability,” Progress in Oceanography, 63, 57–98.
Bauer, J. (2016), “Biases in Random Route Survey,” Journal of survey statistics and methodology, 4, 263–287.
Boeuf, P., H. E. Drummer, J. S. Richards, M. J. Scoullar, and J. G. Beeson (2016), “The Global Threat of Zika Virus to Pregnancy: Epidemiology, Clinical Perspectives, Mechanisms, and Impact.” BMC Medicine, 14, 112.
Boisen, M. L., J. N. Hartnett, A. Goba, M. A. Vandi, D. S. Grant, J. S. Schieffelin, R. F. Garry, and L. M. Branco (2016), “Epidemiology and Management of the 2013-16 West African Ebola Outbreak,” Annual Review of Virology, 3, 147–171.
Chang A. Y., M. E. Parrales, J. Jimenez, M. E. Sobieszczyk, S. M. Hammer, D. J. Copenhaver, and R. P. Kulkarni (2009), “Combining Google Earth and Gis Mapping Technologies in a Dengue Surveillance System for Developing Countries,” International Journal of Health Geographics, 8, 49.
Chen, X. (2009), “A Comparison of Health-Risk Behaviors of Rural Migrants with Rural Residents and Urban Residents in China,” *American Journal of Health Behaviour*, 33, 15–25.

Chen, X., P. Yin, and J. Peng (1999). *Medical Research Design and Data Analysis [in Chinese]*, Wuhan: Wuhan University Press.

Chen, X., P. Yu, D. Zhou, W. Zhou, J. Gong, S. Li, and B. Stanton (2015), “A Comparison of the Number of Men Who Have Sex with Men among Rural-to-Urban Migrants and Non-Migrant Rural and Urban Residents in China: A Gis/Gps-Assisted Random Sample Survey,” *PLoS One*, 10, e0134712.

Chen, J. W., H. M. Zhao, L. Gao, B. Henkelmann, and K. W. Schramm (2006), “Atmospheric Pcd/F and Pcb Levels Implicated by Pine (Cedrus Deodara) Needles at Dalian, China,” *Environmental Pollution*, 144, 510–515.

Cochran, W. G. (1977), *Sampling Techniques*, Wiley Series in Probability and Mathematical Statistics—Applied (3rd ed.), New York: John Wiley & Sons.

Conway, E. M. (2006), “Drowning in Data: Satellite Oceanography and Information Overload in the Earth Sciences,” *Historical Studies in the Physical and Biological Sciences*, 37, 127–151.

Daly, G. L., Y. D. Lei, C. Teixeira, D. C. G. Muir, L. E. Castillo, L. M. M. Jantunen, and F. Wania (2007), “Organochlorine Pesticides in the Soils and Atmosphere of Costa Rica,” *Environmental Science and Technology*, 41, 1124–1130.

de Rada, V. D., and V. M. Martin (2014), “Random Route and Quota Sampling: Do They Offer Any Advantage over Probably Sampling Methods?” *Open Journal of Statistics*, 4, 391–401.

Deseda, C. C. (2017), “Epidemiology of Zika,” *Current Opinion in Pediatrics*, 29, 97–101.

Eaton, D. K., L. Kann, S. Kinchen, S. Shanklin, K. H. Flint, J. Hawkins, and W. A. Harris, et al. (2012), “Youth Risk Behavior Surveillance—United States, 2011,” *Morbidity and Mortality Weekly Report*, 61, 1–62.

Escamilla, V., M. Emch, L. Dandalo, W. C. Miller, F. Martinson, and I. Hoffman (2014), “Sampling at Community Level by Using Satellite Imagery and Geographical Analysis,” *Bulletin of the World Health Organization*, 92, 690–694.

Galway, L.P., N. Bell, A. Sae, A. Hagopian, G. Burnham, A. Flaxman, W. M. Weiss, J. Rajaratnam, and T. K. Takaro (2012), “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,” *International Journal of Health Geographics*, 11, 12.

Groves, R. M., F.W. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2009), *Survey Methodology: Wiley Series in Methodology* (2nd ed.), New York: John Wiley & Sons.

Haenssgen, M. J. (2015), “Satellite-Aided Survey Sampling and Implementation in Low- and Middle-Income Contexts: A Low-Cost/Low-Tech Alternative,” *Emerging Themes in Epidemiology*, 12, 20.

Haklay, M., and P. Weber (2008), “Openstreetmap: User-Generated Street Maps,” *IEEE Pervasive Computing*, 7, 12–18.

He, Z., H. Zhuang, C. Zhao, Q. Dong, G. Peng, and D. E. Dwyer (2007), “Using Patient-Collected Clinical Samples and Sera to Detect and Quantify the Severe Acute Respiratory Syndrome Coronavirus (Sars-Cov),” *Virology Journal*, 4, 32.

Heckathorn, D. D. (1997), “Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations,” *Social Problems*, 44, 174–199.

——— (2002), “Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations,” *Social Problems*, 49, 11.

Heeringa, S., and C. O’Muircheartaigh (2010a), “Sampling Designs for Cross-Cultural and Cross-National Survey Programs,” in *Survey Methods in Multinational, Multiregional and Multicultural Contexts*, eds. J. A. Harkness, M. Baun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Mohler, B. Pennell, and T. W. Smith, pp. 251–268, New York: John Wiley and Sons.

Heeringa, S. G., B. T. West, and P. A. Berglund (2010b), *Applied Survey Data Analysis*, Boca Raton, FL: CRC Press.

Heeringa, S., and S. Ziniel (2012), *Sample Design and Procedures for Hepatitis B Immunization Surveys: A Companion to the Who Cluster Survey Manual (Who/Ivb/11.12)*, Geneva: Immunization, Vaccines and Biologicals, World Health Organization.
Huang, B., Y. Zhao, X. Shi, D. Yu, Y. Zhao, W. Sun, H. Wang, and I. Oborn (2007), “Source Identification and Spatial Variability of Nitrogen, Phosphorus, and Selected Heavy Metals in Surface Water and Sediment in the Riverine Systems of a Peri-Urban Interface,” *Journal of Environmental Science and Health Part A-Toxic/Hazardous Substances and Environmental Engineering*, 42, 371–380.

Kish, L. (1949), “A Procedure for Objective Respondent Selection within Household,” *Journal of American Statistical Association*, 44, 380–387.

——— (1965), *Survey Sampling*, New York: John Wiley & Sons.

Kondo, M. C., K. D. Bream, F. K. Barg, and C. C. Branas (2014), “A Random Spatial Sampling Method in a Rural Developing Nation,” *BMC Public Health*, 14, 338.

Kumar, N. (2007), “Spatial Sampling Design for a Demographic and Health Survey,” *Population Research and Policy Review*, 26, 581–599.

Landry, P. F., and M. Shen (2005), “Reaching Migrants in Survey Research: The Use of the Global Positioning System to Reduce Coverage Bias in China,” *Political Analysis*, 13, 1–22.

Levy, P. S., and S. Lemeshow (1999), *Sampling of Populations: Methods and Applications* (3rd ed.), New York: John Willey & Sons, Inc.

Lohr, S. L. (1999), *Sampling: Design and Analysis*, Pacific Grove, CA: Duxbury Press.

Maguire, D. J., M. Batty, and M. F. Goodchild (2005), *Gis, Spatial Analysis and Modeling*, Redlands, CA: ESRI Press.

Mansergh, G., S. Naorat, R. Jommaroeng, R. A. Jenkins, S. Jeeyapant, K. Kanggamrua, P. Phanuphak, J. W. Tapper, and F. van Griensven (2006), “Adaptation of Venue-Day-Time Sampling in Southeast Asia to Access Men Who Have Sex with Men for HIV Assessment in Bangkok,” *Field Methods*, 18, 135–152.

Mathews, J. H. (1972), “Monte Carlo Estimate for Pi,” *Pi Mu Epsilon Journal*, 5, 281–282.

Metropolis, N., and S., Ulam (1949), “The Monte Carlo Method,” *Journal of American Statistical Association*, 44, 335–341.

Murray, J., A. T. O’Green, and P. A. McDaniel, (2003), “Development of a Gis Database for Ground-Water Recharge Assessment of the Palouse Basin,” *Soil Science*, 168, 759–768.

Pearson, A. L., A. Rzotkiewicz, and A. Zwickle (2015), “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,” *International Journal of Health Geographics*, 14, 33.

Schneider, A., M. A. Friedl, and D. Potere (2009), “A New Map of Global Urban Extent from MODIS Satellite Data,” *Environmental Research Letters*, 4.

Shannon, H. S., R. Hutson, A. Kolbe, B. Stringer, and T. Haines (2012), “Choosing a Survey Sample When Data on the Population Are Limited: A Method Using Global Positioning Systems and Aerial and Satellite Photographs,” *Emerging Themes in Epidemiology*, 9, 5.

Singh, G., and B. D. Clark (2012), “Creating a Frame: A Spatial Approach to Random Sampling of Immigrant Households in Inner City Johannesburg,” *Journal of Refugee Studies*, 26, 126–144.

——— (2013), “Creating a Frame: A Spatial Approach to Random Sampling of Immigrant Households in Inner City Johannesburg,” *Journal of Refugee Studies*, 26, 126–144.

Statistical Bureau of Wuhan (2012), *Wuhan Statistical Yearbook-2012*, Beijing: China Statistics Press.

Stehman, S. V., and W. S. Overton (1995), “Spatial Sampling,” in *Practical Handbook of Spatial Statistics*, ed. S. L. Arlinghaus, pp. 31–63, Boca Raton/New York/London: CRC Press

Sutton, P. (1998), “Modeling Population Density with Night-Time Satellite Imagery and Gis,” *Computers, Environment and Urban Systems*, 31, 227–244.

Tilling, K. (2001), “Capture-Recapture Methods—Useful or Misleading?” *International Journal of Epidemiology*, 30, 12–14.

Tong, T. R. (2005), “Sars-Cov Sampling from 3 Portals,” *Emerging Infectious Disease*, 11, 167.

Valliant, R., J. A. Dever, and F. Kreuter (2013a), *Practical Tools for Designing and Weighting Survey Samples*, Springer. New York.

Valliant, R., J. G. Dever, and F. Kreuter (2013b), *Practical Tools for Designing and Weighting Survey Samples*, New York, NY: Springer.
Wampler, P. J., R. R. Rediske, and A. R., Molla (2013), “Using Arcmap, Google Earth, and Global Positioning Systems to Select and Locate Random Households in Rural Haiti,” International Journal of Health Geographics, 12, 3.

Weyer, J., A. Grobbelaar, and L. Blumberg (2015), “Ebola Virus Disease: History, Epidemiology and Outbreaks,” Current Infectious Disease Reports, 17, 480.

Wuhan Center for Disease Prevention and Control (CDC). (2016), “Report of HIV/Aids Epidemic in Wuhan, China,” Technical, Wuhan CDC.