# APPLICATION OF IBGE DATABASES FOR OPERATIONAL PLAN IN COMPLEX SAMPLE STUDY SEROPREVALENCE

Raynner Augusto Moreira Parente
(Superior Health Science School, Brasilia, Brazil, raynnerparente@hotmail.com)

Roberto de Melo Dusi*
(University of Brasília (UnB), Brasilia, Brazil, robertodusi@hotmail.com)

Mariana Fehr Nicácio
(UnB, Brasilia, Brazil, mari.fehr@hotmail.com)

## Abstract

**Hantaviruses is an acute viral anthropozoonosis with a high mortality rate, ranging around 20-70%. Given its high lethality and the lack of studies using probabilistic methods, the serum prevalence studies using statistical resources enable better knowledge person, time and place disease distribution. For rural areas this kind of research is not available. Hantaviruses human population-based seroprevalence study for Brazilian Federal District (DF) was designed with cartographic and population of census tracts, which sample size was 495 people. Considering the blood dropping refusal, it was increased by 20%, and a number of 595 households was recommended. The aim is to present the use of random sampling techniques in hantaviruses Seroprevalence study, at DF, with public data. Sampling plan of 595 addresses randomly select, using complex sampling techniques, combining systematic and simple sampling were applied to prepare household list. Census tracts were selected if they had georeferenced points of probable transmission locations (PTL). DF census tract geographical mesh and public data, including the list address, were obtained from the Brazilian Institute of Geography and Statistics (IBGE) domain. There are no list of each dwellers neither family chief name. Random address were listed for each census tract. The source of the probable transmission locations (PTL) data was DF environmental health surveillance reports. Data sheet and its calculations, laptops, geographic positioning device, IBGE and Google Earth™ images were the tools used. Non-probabilistic sampling was obtained during the field visit. Distribution of 595 household for each one of 81 selected census tracts, using 'selection interval', regarding weighed systematic sampling distribution, produced decimal results for several census tracts. Then the rounding was always done to the upper integer number value, resulting in 631 addresses to visit. The household randomly ordered list had 319 (50.6%) addresses classified as rural and 312 (49.4%) as urban. Virtual maps and 81 lists with specific household addresses were prepared. There were 133 (21.1%) addresses with geographic coordinates, and 453 (71.8%) only with directions. Forty five (7.1%) addresses were not useful. When all addresses of one census tract list were useless others techniques were used to provide substitute addresses. For while, 478 (96.6%) blood samples of randomly sampling were dropped. Devices and data were prepared to support field logistic plan to achieve the study sampling. The free image software and public databases allowed a feasible field logistic plan.**

*Key words*: probabilistic sampling; population-based; complex sampling; hantavirus

## 1. Introduction

Hantaviruses is an acute viral anthropozoonosis which etiological agent is Hantavirus, genus of Bunyaviridae family (Schmaljohn *et al.,* 1983). Hantaviruses are most commonly transmitted by

aerosol inhalation or contact with secretions and feces of wild rodents of the Muridae family. The fatality rate of American form has been high, with an average of 43.8%, and at Brazilian 'Savanna' area, it ranged from 20% to 70% (Nunes *et al.*, 2011).

Given its high fatality rate, it is necessary to study more deeply this disease. Infection Seroprevalence studies in the rural population have been conducted, but few used statistical resources to enable greater accuracy. There are few papers with randomly design regarding hantavirus infection worldwide. In 2003 Campos *et al.* reported a well done study, with probabilistic sampling, at Jardinopolis, Brazil, with 14,3 % positive blood samples for hantavirus infection.

A random and representative sampling was designed for DF rural areas. The aim is to present the use of random sampling techniques in hantavirus infection Seroprevalence study, at DF, with public data, from Brazilian Institute of Geography and Statistics (IBGE) available at the internet domain.

## 2. Intermediate Section

With the sample of 595 household, calculated by probabilistic statistical methods, it was necessary to create a routine for systematic random selection (Bolfarine & Bussab, 2005). The addresses to be visited were located in rural or periurban area, which demand prior knowledge of the addresses lists in the regions. The public database available from IBGE on its website was used (IBGE, 2013). The availability of addresses in the IBGE electronic site allowed the preparation of random list of households to be visited. This public database used, with addresses listing, made this process cheaper, easier and faster than visiting, listing and counting all households of 81 census tracts, as a geographic recognition at the field. After this household counting it was also necessary to draw 'n' households for each census tract.

Eighty one (1,9%) census tracts were selected from a total of 4,358 DF census tracts, distributed in 14 districts. These census tracts were selected when the probable transmission locations (PTL) georeferenced points were included in their polygons areas. These points were obtained from hantaviruses autochthonous cases investigation from 2004 to 2012, by health environmental surveillance (Bredt *et al.*, 2004). The study population was all people with 10 years old or more from these 81 polygons, totaling 34 838 inhabitants, regardless any territorial contiguity of these sectors (Fig.1).

Considering that the IBGE census was done four years before this study, a preliminary geographic field recognition was made and it was found that some houses previously labeled as under construction or unoccupied, were now inhabited and the ones labeled as occupied continued to be so. Therefore, the researchers decided to include addresses listed in the 2010 census as under construction or unoccupied. By the other side, few households labeled as occupied, during the visit were without dwellers or even demolished.

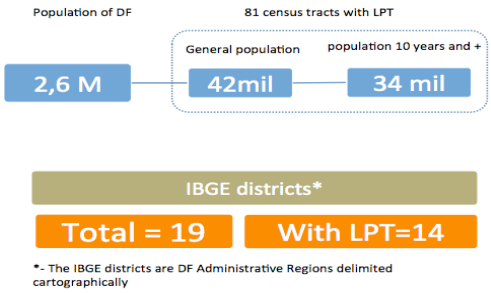


Figure 1. General population and target population used in the study and its distribution into districts.

The addresses selected were listed by census tract. Then all the addresses of each list were sorted by a spreadsheet random function of the software LibreOffice™ and the first 'n' addresses were chosen according to the sample number weighed for each tract. The sample size in each census tract (ni), obtained by procedure with weighted distribution was calculated according to the size of its population aged 10 or older, corresponding as selection interval:

Census tract sample size (ni) = sample size of the study (n) X (sector population 10 years and over / study population (P) (with 10 years or more).

This formula represents the selection interval (Si) calculation, relevant 'step' in systematic sampling (Si = study population / sample size), and was applied in the final stage of addresses drawing. The distribution of 595 households for each 81 select census tracts, using 'selection interval', regarding weighed systematic sampling distribution, produced decimal results for several census tract. Then the rounding was always done to the upper integer number value, resulting in an extra household number increase totaling 631 addresses to visit.

During a test field visit, the researchers tried to use the ordered list of 2010 census. The draw made during this visit produced a list of addresses which house were near among themselves, suggesting that the census recording technique contained components that could compromise the randomness.

Then an alternative way found was to use a random function in a spreadsheet to order households addresses randomly, shown in the last column at figure 2. For each tract, the firsts addresses of the random ordered list, that matched the amount (ni), were included in a separate list for each census tract, resulting in a total of 81 different sheets.

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Code of geographical unit | Current situation | Address | Complement | Locality | Species address | Type of establishment | Block number | Face Number | CEP | Random function |
| 2 | 530010805060447 | Urbano | EDF SOFN Q 1 | LOTE 1 | RA GUARA | Domicilio particular | | 1 | 6 | 70634100 | 0,007285576 |
| 3 | 530010805060447 | Urbano | EDF SOFN Q 2 CL A | BLOCO A ENTRADA 45 SALA 3 | RA GUARA | Domicilio particular | | 1 | 12 | 70634200 | 0,007867304 |
| 4 | 530010805060447 | Urbano | EDF SOFN Q 5 Cj A | LOJA SN | RA GUARA | Estabelecimento de outras finalidades | OFICINA MECANICA | 1 | 33 | 70634510 | 0,009964505 |
| 5 | 530010805060447 | Urbano | EDF SOFN Q 2 CL B | BLOCO B LOJA 37 | RA GUARA | Estabelecimento de outras finalidades | MECANICAOFICIN | 1 | 13 | 70634205 | 0,010805998 |
| 6 | 530010805060447 | Urbano | EDF SOFN Q 2 CL B | BLOCO B LOJA 69 ANDAR 2 AP | RA GUARA | Domicilio particular | | 1 | 13 | 70634205 | 0,013019391 |

Figure 2. The random function (K column) applied for IBGE census tract list.

The household randomly ordered list had 319 (50.6%) addresses classified as rural and 312 (49.4%) as urban. Virtual maps and 81 lists with specific household addresses were prepared. There were 133 (21.1%) addresses with geographic coordinates, and 453 (71.8%) only with directions, sufficient to guide the search at the field. Fortunately 120 (90.0%) of addresses with geographic coordinates were rural. Forty five (7.1%) addresses registered in census tract list obtained from IBGE database did not have any specific direction and were not useful (figure 3). When all addresses of one census tract list were useless others techniques were used to provide substitute addresses.

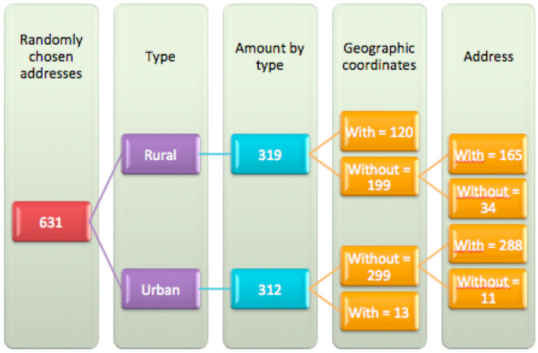


Figure 3. Selected households aspects.

The addresses with geographic coordinates were "plotted" on the map in Google Earth™, as well as the 301 of the 312 urban addresses with appropriate direction. The remaining rural addresses without geographic coordinates were searched in the polygon area, identified by a GPS device (over polygon image on laptop screen), looking for their numbers at fences or asking neighboring dwellers. It was installed on laptop a "software" with the census tracts meshes, provided by IBGE (IBGE, 2013) in .kmz format, facilitating the location of the destinations (Fig. 4). The 34 rural and 11 urban addresses, totaling 45 households, with no possibility of location were replaced by subsequent houses, still following the order of randomization, preserving the randomization of the study. In few cases, all census tract list was useless, and direct substitution was not possible. Then others approaches were necessary, such as the Technical Assistance and Rural Extension Enterprise (EMATER-DF), local schools and health unit maps, or Quantum-Gis™ software. If in these institutions there was a list of households well organized, the addresses were drawn again.

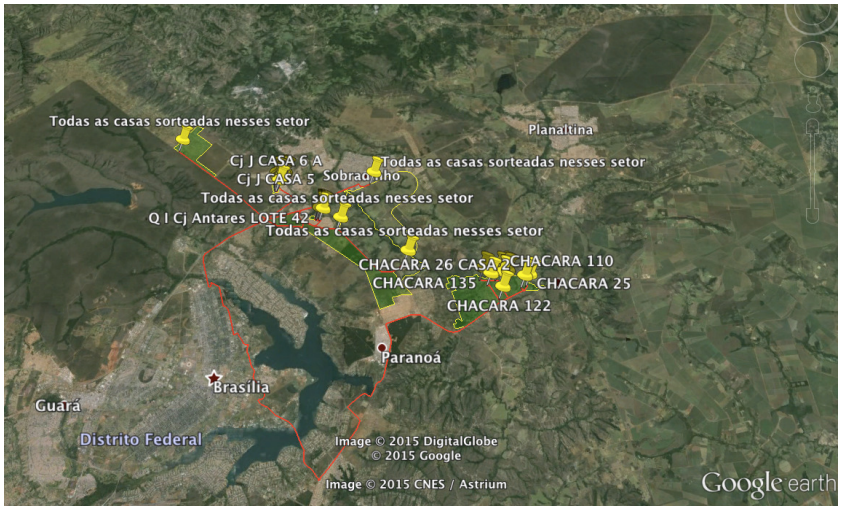


Figure 4. Construction of paths to find the addresses.

A geographic positioning device (GPS) was used to find the field location of the addresses with coordinates, guiding the team according to a previous script. Many satellite images were obtained as seen at figure 4 map, and facilitated to find these locations with an accuracy of 5 meters. For the addresses that had directions recognized by Google™ maps, the route was also set and they were easily found using GPS. Addresses without coordinates and with direction not recognized by Google™ maps, but with possible location, received prior visiting before data collection. The location position was marked on the map and a visit was planned later. If the address had a house number at the list but dwellers did not recognize it (neither printed on the wall or fences), an IBGE protocol for house counting was used. The counting started at the houses on the left and thus following concentrically.

An ethical approach was made for households participants and a written informed consent was obtained. After locating the address and confirming that it was part of the polygon, a list of all eligible dwellers was made and one of them was drawn and invited to participate of the study. For these drawings an android device software or a coin (face/crown) were used. This was a simple sampling step.

Electronic and paper forms were designed in order to proceed the data collection. These forms were filled with (1) the location of the address, (2) the name of the person drawn and the other residents, (3) the participant phone number, (4) the geographic coordinates, (5) the situation of the data collection, (6) exposure situations and (7) clinical data, which was classified into: complete, pending or loss. There are two types of loss: the first, which occurred 15 times (preliminary count), were cases

that resident - eligible person - did not accept to perform the blood test. In such cases, at least three visits were made to confirm the refuse; and there were not substitution. The second type of loss, which summed 92 addresses (preliminary count), was due to nonconformity in the address or less than six month of living in the census tract. Others reasons for these losses were demolished houses, lack of housing in the building or the drawn address and street addresses not found. In the second case a new random household was included. For while, 478 (96.6%) blood samples of randomly sampling were dropped.

## 3. Conclusion

Despite the problems with the address list, IBGE database seems to be a good quality source for rural directions and geographic coordinates. For studies concentrated only in urban sites, this source has a great quality, with small cases of addresses described incorrectly. The data collection is still running.
Finally, this study established a method for the determination of hantavirus seroprevalence in DF population. This study can serve as a basis for future epidemiological studies in the health field, showing the possibility of carrying out scientific products using the probabilistic approach, bringing the results found closer to reality.

## Acknowledgment

We would like to knowledge Professor Lucio Vivaldi, Professor Claudete Ruas, Professor Pedro Sadi Monteiro, Professor Pedro Luiz Tauil, Professor Helenda Costa Gurgel, Renato José Furigo Lélis e Renata Garcia Dusi.

## References

Bolfarine, H., & Bussab, W. (2005). Elementos de Amostragem. São Paulo, Brazil, Ed. Blucher.

Bredt A., Massunaga P.N.T., Maia E.S., Santos D.E., Silva J.A.M. (2004). Análise ambiental dos locais prováveis de infecção para hantavirose no Distrito Federal. Anais do III Simpósio Internacional sobre Arbovírus dos Trópicos e Febres Hemorrágicas.

Campos G.M., Souza R.L.M., Badra, S.J., Pane, C., Gomes, U.A. & Figueiredo L.T.M. (2003). Serogical Survey of Hantavirus in Jardinopolis county, Brazil. J. Med. Virol. 71:417-422.

Google Earth. DF images (2013), Brasília - DF. Available in: <http://www.google earth.com.br/>. Access in 2013, December 12th.

Google Maps. DF images (2014), Brasília - DF. Available in: <http://www.google earth.com.br/>. Access in 2014, Jannuary 10th.

IBGE :: Instituto Brasileiro de Geografia e Estatística. (2013). avaliable: < http://www.ibge.gov.br/home/download/estatistica.shtm >, access in 2013 dec 15th.

Nunes, M.L., Maia-Elkhoury A.N.S., Pelissari D.M., & Elkhoury M.R. (2011). Caracterização clínica e epidemiológica dos casos confirmados de hantavirose com local provável de infecção no bioma Cerrado Brasileiro, 1996 a 2008. Epidemiol Serv Saude 20:537-545.

Schmaljohn, C.S. & Dalrymple, J.M. (1983). Analysis of Hantaan virus RNA: evidence for a new genus of Bunyaviridae. Virology. 131:482–491.