RESPONDENT DRIVEN SAMPLING AND SPATIAL AUTOCORRELATION

E Scott Morris[1], Vaishnavi Thakar[2], Daniel A. Griffith[3]

University of Texas at Dallas, 800 W. Campbell Road, Richardson, Texas 75080,
[1]Email: exm123030@utdallas.edu
[2]Email: vxt110130@utdallas.edu
[3]Email: dagriffith@utdallas.edu

Abstract

Respondent driven sampling (RDS) is a type of sampling method used to survey rare and hard to reach populations. RDS was developed to address the issue of bias associated with snowball sampling in qualitative research.Although, RDS has evolved by addressing major issues involved with the snowball sampling method, the issue of how the presence of spatial autocorrelation (SA) affects RDS had not been studied. SA refers to the clustering of similar attribute values in geographic space. Quantitative studies show that the presence of positive SA leads to an underestimation of the appropriate sample size. If RDS is not affected by SA, then the samples are expected to be dispersed in geographic space and not clustered around a sampling seed that initiates a sequence of respondents. This paper presents impacts of SA on RDS when a social network displays a geographic pattern. The geographic distribution of the samples and associated socio-economic and demographic variables are analyzed with respect to sequences of respondents. Social network RDS data for Rio de Janiero, Brazil are analyzed. Preliminary results indicate that in these social network RDS data, samples are clustered around their initial seeds and do not spread out in geographic space as the sequence of respondents progresses. The result is increased sampling variance, which raises a concern about appropriate sample size determination in RDS.

Keywords: respondent driven sampling, snowball sampling, spatial autocorrelation, social network.

1. Introduction

Snowball sampling, introduced by Goodman in 1961, is a survey strategy initialized by selecting a group of participants known as seeds. Once surveyed, each seed recommends potential respondents (nodes) with shared connections (edges) on the basis of a research topic and whom each referrer believes likely to also participate. This process proceeds in a similar fashion over a series of waves, and the nodes and edges define a social network. Respondent driven sampling (RDS), developed by Heckathorn in 1997, is a formalized method, based on the snowball strategy, that compensates for the non-random process of data collection. Previous studies have concluded that little bias exists among RDS results compared to simple random sampling (SRS). However, subsequent research has indicated the presence of a variance inflation factor (VIF) and increased design effect among underlying attributes of the members of the RDS networks versus SRS. The result of this effect is the tendency to underestimate the appropriate sample size among RDS surveys. A prominent contributor to the VIF may be positive spatial autocorrelation (SA) attributed to the geographic configuration of the population being. If a social network displays a geographic pattern, the variance for a RDS is likely to be impacted by SA sampled (Rudolph et al., 2015). To analyze this impact, a simulation can be designed based upon real world data from a RDS survey conducted with heavy drug users in Rio de Janiero, Brazil. This simulation proceeds through the obtained social network based upon empirical probabilities determining the number of nodes for each subsequent wave. The purpose of this paper is to establish a basis for designing this type of simulation experiment, and demonstrate that the VIF attributable to SA is an outcome.

2. Data

2.1 Network

The network analyzed in this research is from an RDS study conducted in Rio de Janiero in 2009. The network consists of 611 heavy drug users defined as having injected illegal narcotics in the last 6 months and/or using illicit drugs, other than marijuana or hashish, at least 25 days in the last 6 months. Respondents are over the age of 18 and meet the protocol of the study. The original study utilizes RDS as a technique for surveying hard to reach populations, specifically, HIV transmission associated with heavy drug users (Toledo, et. al. 2009). Network data are configured in two tables. First, the node data consist of anonymous respondent identification numbers (ID) and their corresponding administrative regions (AR) of residence (Table 1); 140 of the respondents' locations are unknown.

ANCHIETA	10	MADUREIRA	91
BANGU	2	MEIER	3
BARRA DA TUUCA	1	PAQUETA	0
BOTAFOGO	4	PAVUNA	2
CAMPO GRANDE	2	PENHA	1
CENTRO	30	PORTUARIA	35
CIDADE DE DEUS	0	RAMOS	5
COMPLEXO DA MARE	0	REALENGO	4
COMPLEXO DO ALEMAO	0	RIO COMPRIDO	8
COPACABANA	0	ROCINHA	0
GUARATIBA	1	SANTA CRUZ	4
ILHA DO GOVERNADOR	67	SANTA TERESA	4
INHAUMA	2	SAO CRISTOVAO	153
IRAJA	0	TIJUCA	29
JACAREPAGUA	1	VIGARIO GERAL	0
JACAREZINHO	1	VILA ISABEL	9
LAGOA	2	UNKNOWN	140
TOTAL		611	

Table 1. RDS respondent count per region of Rio de Janerio

2.3 Demographics

The study area consists of 33 ARs. Demographic data were obtained from the Instituto Brasileiro de Geografia e Estat ística (IBGE) from the 2010 census online at http://www.ibge.gov.br/home. Four attribute variables were selected assuming that they demonstrate correlation with heavy drug use, and reflect at least a moderate degree of SA indicative of the geographic configuration. The demographic variables are: population density, median income, unemployment percentage, and illiteracy percentage.

2.2 Transformation and Mapping

Thirty three ARs in the Rio de Janeiro municipality from the state of Rio de Janeiro, Brazil were selected for this study (Figure 1. a.). The following socio-economic variables were obtained at the AR level for all 33 spatial units: population density, median income including no income, median income excluding no income, percentage of unemployment, and percentage of illiteracy. These variables were mapped in order to visualize their spatial distributions (Figures 1- b through 2-f).

2.3 Spatial Autocorrelation

Tobler's first law of geography states that "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p. 236). Hubert et al. define SA as "Given a set *S* containing *n* geographical units, the relationship between some variable observed in each of the *n* localities and a measure of geographical proximity defined for all *n(n–1)* pairs chosen from *S*" (Hubert et al. 1981, p. 224). Standard inferential statistics assumes complete randomness (of observations, which is referred to as an independent random process (IRP), or complete spatial randomness (CSR)). But, spatial data violate this assumption due to the presence of SA. Positive SA causes variance inflation. Hence, spatial statistics measures are used to quantify the degree of self-correlatedness of a variable as a function of nearness. A spatial weight matrix is needed in order to measure SA, which gives the information about relative location of pairs of adjacent neighboring locations (binary rook, queen neighbors; first, second order neighbors) or all other locations (distance based – Euclidian, rectilinear, or network).

Two widely used indices of SA are provided by Moran (1948) and Geary (1954). Global Moran's I tests for spatial randomness (null hypothesis) and detects the nature (positive or negative) as well as degree of SA. The Moran's *I* values range from roughly -1 (high negative SA; dissimilar values cluster on a map) to 1 (high positive SA; similar values cluster on a map)—this lower bound can range between -1 and -0.5, whereas this upper bound can range from 0.8 to more than 1.3—Moran's *I* value denoting no SA is $-1/(n-1)$, which is slightly less than zero.

The global Geary's c (null hypothesis of no SA) values range from roughly 0 (extreme positive SA), 1 (no SA) to 2+ (extreme negative SA). The extreme values are functions of eigenvalues associated with the spatial weights matrix**.**

The Moran Coefficient and the Geary Ratio for variable y are given by:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(y_i - \bar{y})\,(y_j - \bar{y})}{\sum_{i=1}^{n}(\,y_i - \bar{y})^2}$$

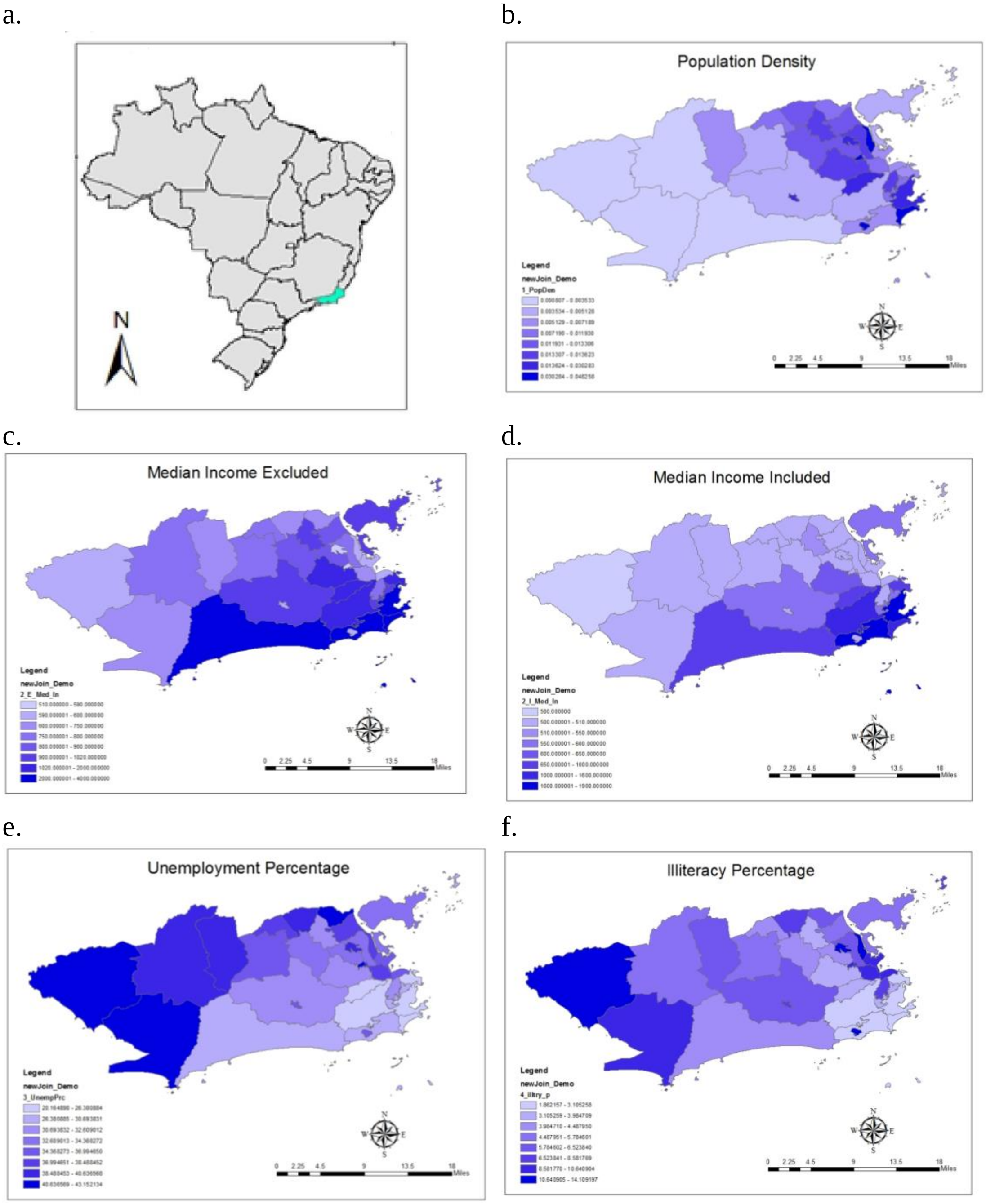


Figure 1. a. Study area representing 33 administrative regions in the Rio de Janeiro municipality in the state of Rio de Janeiro, Brazil, b. through f. socio-economic variables.

$$c = \frac{n-1}{2(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij})} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(y_i - y_j)^2}{\sum_{i=1}^{n}(\,y_i - \bar{y})^2}$$

where $\bar{y}$ = mean of the variable y;

y_i = variable value at a particular location *I*; and,

w_{ij} = weight indexing for location *i* relative to *j* (a spatial weights matrix, **W**, cell entry)

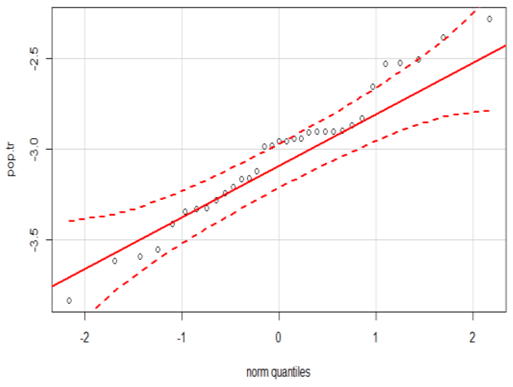
Previous quantitative geographic research documents that human attributes tend to display moderate positive SA. This idea was captured by 2005 Nobel Prize winning economist Thomas Schelling in his model of segregation (1969, 1971). Keeping this observation in mind, the Moran and Geary SA indices were used to measure and map the nature and degree of SA for the selected Rio de Janeiro socio-economic variables.

SA indices need a spatial weight matrix which can define polygons sharing common boundaries as neighbors. The Rio study area polygon shapefile includes two disconnected islands, which were manually edited and connected to the main land before creating the spatial weight matrix. This was done by observing the bridges connecting the islands with the main land, so that each of the spatial units (polygons) share at least one common boundary or edge. A row standardized spatial weight matrix was created using the modified polygon shapefile with connected islands.

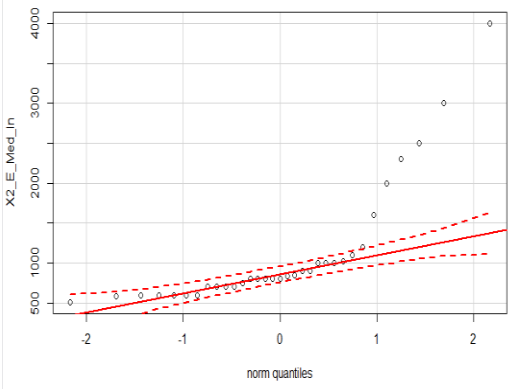
Before proceeding with the Moran Coefficient and Geary Ratio calculations, a Box-Cox power transformation (log) was performed on population density to make its frequency distribution more bell-shaped. Table 2 summarizes the Moran Coefficient and Geary Ratio values of SA for each socio-economic variable; all five variables contain slight to moderate positive SA. Figure 4 displays the normal quantile plots for each variable.

Variable Name	Moran Coefficient	Z-scores (Moran's I)	Geary Ratio	Z-scores (Geary's c)
Population Density	0.23	2.2135	0.56	3.299
Median Income Excluded	0.47	4.5227	0.71	1.7969
Median Income Included	0.56	5.274	0.55	2.899
Percentage of Unemployment	0.37	3.3796	0.65	2.4992
Percentage of Illiteracy	0.33	3.2141	0.68	2.1011

Table 2. Summary of Moran Coefficient and Geary Ratio values for socio-economic variables in the Rio de Janeiro municipality.

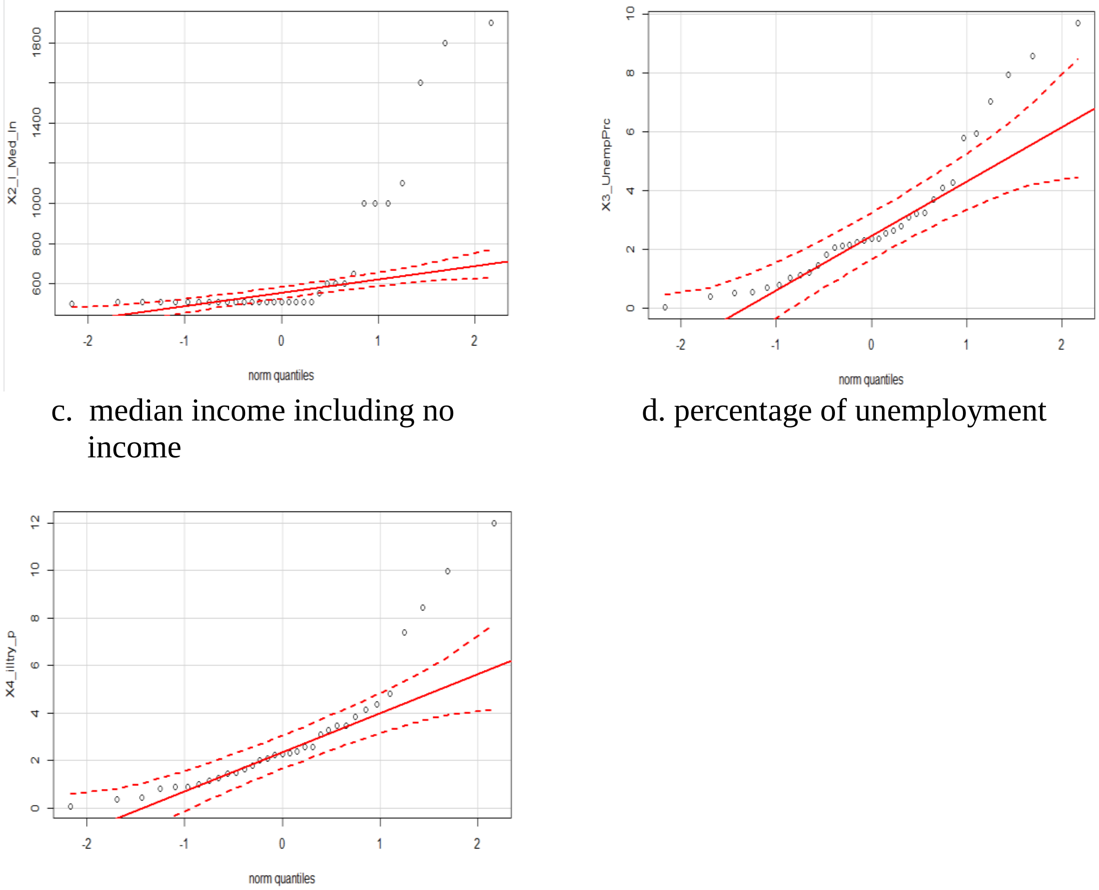


a. population density (transformed)



b. median income excluding no

income



c. median income including no income

d. percentage of unemployment

e. percentage of illiteracy

Figure 4. Normal quantile plots for selected socio-economic variables.

3. Methodology

3.1 Network Chains

A network is developed by respondents recommending subsequent participants. The process is initiated by the selection of initial participants, or seeds, from whom the social network emanates. The network analyzed in this study originates with six seeds, one of which yielded no referrals. Based upon the referral connectivity, the network comprises 377 chains. Each chain is defined by a seed—a respondent whom is never referred—and an end—a respondent with no referrals. Chain lengths range from 2-12 respondents (Table 3).

NODES	2	3	4	5	6	7	8	9	10	11	12	TOTAL
CHAINS	5	13	21	24	42	44	31	80	53	46	18	377
SUB-CHAINS	2656	2279	1907	1548	1210	896	624	396	199	82	18	11815

Table 3. Number of chains and sub-chains by length.

3.2 Simulation Design

Analysis of the network can be conducted by simulation based upon the network connectivity. Starting points and chain lengths could be selected, and sub-chains could be traced based upon underlying empirical probabilities. Therefore, the 377 complete chains are partitioned into sub-chains, providing a total number of 11,815 possible chains. For example, one chain of length 5 consists of two sub-chains of length 4, three of length 3, and four of length 2. The prevalence of missing locational data introduces complications and constraints to such a simulation experiment.

4. Anticipated Results

The anticipated result of this research is that the geographic distribution of a social network displays SA. In turn, this feature inflates the sampling variance. Inflation occurs because of two factors: (1) the sampling probabilities no longer are equal; and, (2) covariance between individuals no longer is zero. This covariation is a function of the structure of a social network coupled with SA in its geographic landscape, which are correlated. One way to capture this latter effect is to couple a social network with its corresponding spatial weights matrix, conceptualizing the social network as being articulated first. Griffith (2005) outlines the VIF attributable to SA. Extending this specification, and considering only the case of positive SA,

$$1 \leq \text{VIF} \leq \text{TR}(\mathbf{V}_s^{-1})\text{TR}(\mathbf{V}_N^{-1})/n^2 \quad ,$$

where n is the number of observations (i.e., individual or areal units), $\mathbf{V}_S$ denotes the spatial autoregressive variance component [e.g., $(\mathbf{I}-\rho_s\mathbf{W})^T(\mathbf{I}-\rho_s\mathbf{W})$, where ρ_s is the spatial autoregressive parameter], and $\mathbf{V}_N$ denotes the network autoregressive variance component. This specification indicates that network autocorrelation inflates variance beyond what spatial autocorrelation does, and vice versa. It also could be modified by including a geographic aggregation matrix, which would smooth the spatial autocorrelation effects.

5. Acknowledgements

This research was conducted with supported by the U.S. National Science Foundation, grant BCS-1262717; any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the National Science Foundation. We thank Dr. Francisco Inacio Bastos of the Oswaldo Cruz Foundation for providing the data facilitating this research, and Dr. Parmanand Sinah for his initial work on retrieving the social network data.

6. References

Authority F, 1973, Stating the obvious: an interdisciplinary approach, *Journal of Entirely Predictable Results*, 63(2):1037-1068.

Fotheringham, A, Stewart and Rogerson P, 2009, *The Sage Handbook of Spatial Analysis*. London: Sage Publications.

Fudgit B, Publish HWP and Writer AB, 1997, *Looming Deadlines and How to Deal with Them*. Partridge & Co, Norwich, UK.

Geary, R, 1954, The contiguity ratio and statistical mapping, *Incorporated Statistician*, 5:115-41.

Goel S, Salganik MJ, 2010, Assessing respondent-driven sampling, *PNAS*, 107(15):6743-6747.

Griffith D, 2005, Effective geographic sample size in the presence of spatial autocorrelation, *Annals*, Association of American Geographers, 95: 740-760.

Heckathorn, D. D., 1997, Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems*, 44(2):174-199.

Hubert LJ, Golledge RG, Costanzo CM, 1981, Generalized procedures for evaluating spatial autocorrelation, *Geographical Analysis*, 13(3):224-233

Learned C and Expert M, 1982, Reworking previous publications for fun and profit. In: Doctor K and Professor B (eds), *Proceedings of the 2nd International Conference on Something You Thought Was Relevant But Isn't Really*, Los Angeles, USA, pp. 120-149.

Moran, P, 1948, The interpretation of statistical maps, *Journal of the Royal statistical Society*, B, 10:243-51.

Rudolph A, Young A, Lewis C, 2015, Assessing the geographic coverage and spatial clustering of illicit drug users recruited through respondent-driven sampling in New York City, *Journal of Urban Health*, 92:352-378.

Schelling T C, 1969, Models of segregation, *American Economic Review*, 59(2):488-493.

Schelling T C, 1971, Dynamic models of segregation, *Journal of Mathematical Sociology*, 1(2):143-186.

Tobler WR, 1970, A computer movie simulating urban growth in the Detroit region, *Economic Geography* 46(2):234-240.

Toledo L, Codeço CT, Bertoni N, Albuquerque E, Malta M, Bastos FI, 2011, Putting respondent_driven sampling on the map: insights from Rio de Janeiro, Brazil, *Acquir Immune Defic Syndr*, 57(3), pp. S136-S143.