ARTICLE TEMPLATE

Adjusting the prevalence estimates of crack cocaine use among homeless persons in the city of Rio de Janeiro for nonignorable nonresponse

Marcus L. Nascimento[a]

[a]Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

ARTICLE HISTORY
Compiled November 17, 2022

ABSTRACT
This

KEYWORDS
Sections; lists; figures; tables; mathematics; fonts; references; appendices

1. Introduction

2. Data set

From 26 to 29 October 2020, the city of Rio de Janeiro conducted a census focused on people experiencing homelessness under the coordination of Pereira Passos Institute (IPP), Municipal Secretariat for Social Services and Human Rights (SMASDH) and Municipal Secretariat of Health (SMS). The main purpose of the census was to assess the number and the profile of people experiencing homelessness in all regions of the city given a period of time.

A fundamental challenge in achieving people experiencing homelessness is to avoid double counting since this population presents a high degree of mobility. The elicitation about this dynamics in the methodology design was constructed from data sets available at SMASDH and the expertise of specialized groups of professionals in the secretariat. This previous knowledge enabled mapping potential target groups and identifying regions where homeless persons are concentrated. Additionally, the SMS guided the definition of scenes of drug use.

In view of all these information, the city was divided into four census districts, namely the Western, Central, Southern and Northern zones, and data from these area was collected on 26, 27, 28, 29 October 2020, respectively. Figure 1 illustrates the geographical division of the territory by planning area (AP) in such a way that Western zone corresponds to AP 4 and AP 5, and Central, Southern and Northern zones correspond to AP 1, AP 2 and AP 3, respectively.

Census takers were organized on three different shifts considering that the presence

CONTACT Marcus L. Nascimento. Email: marcus@dme.ufrj.br

on the streets varies according to the period of the day. The census data collection consisted on the application of different questionnaires in line with three basic scenarios: street, scene of drug use, and shelter. Furthermore, a concise questionnaire was applied for kids under 10 years old, and some observations were made for individuals who did not answer the questionnaire (hereafter, non respondents).

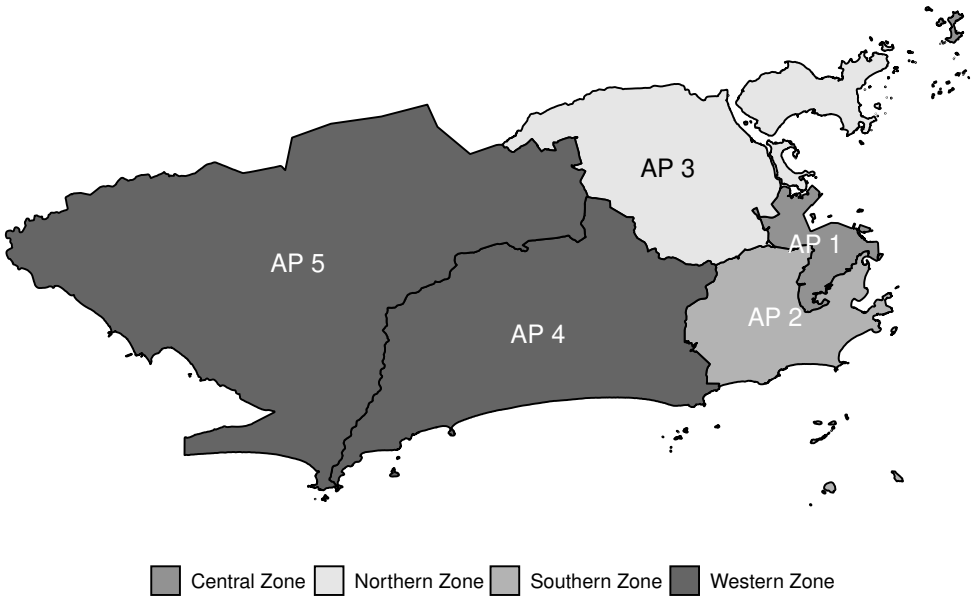

Figure 1. Geographical division of the territory by planning area (AP).

The census identified a total of 7272 people experiencing homelessness in the city of Rio de Janeiro, obtaining valuable information for the development of public policies focused on homeless population. Our analysis is concentrated on the variable about drug use, more specifically, about crack cocaine use, we consider people over 18 years old situated on streets and scenes of drug use, which corresponds to a final data set with 5162 observations.

Crack is a highly addictive and powerful stimulant that is derived from powdered cocaine by dissolving it in a mixture of water and ammonia or sodium bicarbonate (baking soda). In addition to the usual risks associated with cocaine use (constricted blood vessels; increased temperature, heart rate, and blood pressure; and risk of cardiac arrest and seizure), crack users may experience acute respiratory problems, including coughing, shortness of breath, and lung trauma and bleeding. Crack cocaine smoking also can cause aggressive and paranoid behavior.

We study the variable related to crack use jointly with four auxiliary variables: gender, race, age, and local. Gender is a dichotomous variable indicating male and female. Race is a categorical variable separated into black, white and others, which includes asian and indigenous people, for example. Age is also a categorical variable, but separated into adult and elder. Finally, local is a categorical variable referring the place where the individual was identified, street or scene of drug use.

Table 1 introduces some descriptive statistics about our data set. We observe that men correspond to 82.18% of the homeless population in Rio de Janeiro, while black people to around 80%. According to the 2010 Population Census issued by the Brazil-

ian Institute of Geography and Statistics (IBGE), when we look to the whole population in the municipality, these quantities correspond to 46.38% and 47.89% respectively. Moreover, the total of people experiencing homelessness under analysis were distributed around streets (77.24%) and scenes of drug use (22.76%).

Analyzing the prevalence of crack cocaine use among the respondents, the general percentage is 26.64%. We note a slight difference between men and women (26.07% and 29.67%, respectively), and black people (25.92%) and others (25.53) in comparison with white people (29.46). The most relevant difference emerges when we consider the locations, the prevalence of crack cocaine use among people located the streets (22.70%) is substantially lower in comparison with those located on scenes of drug use (52.29%), and when we consider age, since we have a prevalence of 28.57% among adults and a prevalence of 2.91% among elders.

Table 1. Mean, standard deviation and quartiles of dependent, exogenous and endogenous variables in the logarithm scale for all years under analysis.

			Respondents		
Variable	Total	Nonrespondents	User	Non User	Prevalence
	5162	2429 (47.06%)	728 (14.10%)	2005 (38.84%)	26.64%
Gender					
Men	4242 (82.18%)	1937 (45.66%)	601 (14.17%)	1704 (40.17%)	26.07%
Women	920 (17.82%)	492 (53.48%)	127 (13.80%)	301 (32.72%)	29.67%
Race					
Black	4134 (80.09%)	2055 (49.71%)	539 (13.04%)	1540 (37.25%)	25.92%
White	924 (17.90%)	364 (39.39%)	165 (17.86%)	395 (42.75%)	29.46%
Others	104 (2.01%)	10 (9.61%)	24 (23.08%)	70 (67.31%)	25.53%
Age					
Adult	4811 (93.20%)	2284 (47.47%)	722 (15.01%)	1805 (37.52%)	28.57%
Elder	351 (6.80%)	145 (41.31%)	6 (1.71%)	200 (56.98%)	2.91%
Local					
Street	4028 (78.03%)	1817 (45.11%)	464 (11.52%)	1747 (43.37%)	20.98%
Scenes of drug use	1134 (21.97%)	612 (53.97%)	264 (23.28%)	258 (22.75%)	50.77%

Nonresponse occurs when the questionnaire is applied but the individual refuses to answer about crack cocaine use or when the questionnaire application is not possible. Among the major causes of unfeasibility of questionnaire application, we can underline the dispersion due to the approach of interviewer, somnolence, apathy, aggressiveness and torpor induced by drug use. As some of these are possibly related to crack cocaine use, there is an indicative about the association between this variable and the propensity to respond which corroborates with the utilization of nonignorable nonresponse models. In economics and statistics, nonresponse mechanisms are broadly classified as ignorable and nonignorable (Little and Rubin, 2002). The latter arises when the response indicators depend on the missing values, hence nonignorable nonresponse models are applied when the missing data are missing not at random (MNAR).

Besides nonresponse, another aspect from our data set that we intend to explorer in our analysis is the individual location. Although we have mentioned before the high degree of mobility that people experiencing homelessness have, when we examine data about crack cocaine users, it is important to account for a singular characteristic: the existence of cracklands in Brazilian metropolis. Cracklands are marginal areas of public crack cocaine consumption, and the most prominent of these areas is located in the center of São Paulo (Ribeiro et al., 2016). This particularity implies the concentration of substantial numbers of crack users on specific regions, suggesting that location bring some information about our variable of interest.

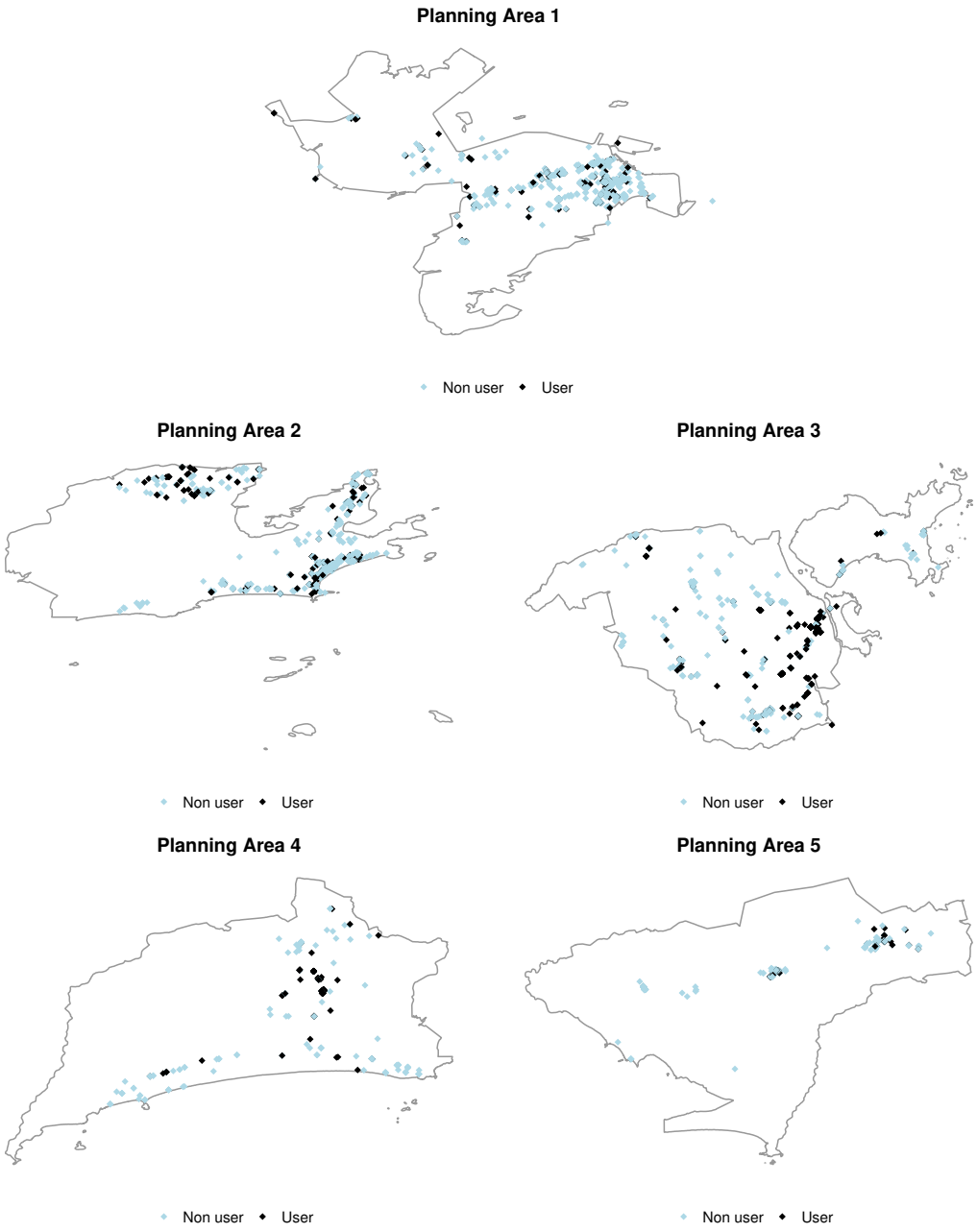

Figure 2. TENTAR MELHORAR A VISUALIZAÇÃO DOS DADOS.

Figure 2 displays the locations among the planning areas. Including non respondents, the AP 1, AP 2 and AP 3 combine more than 80% of people experiencing homelessness in the city, assembling 31.87%, 26.25 % and 27.62% of the population, respectively. These areas also assemble the majority of the residents, and the first two areas are the most relevant from the economic perspective. Examining the maps, it is possible to visualize some well known regions of crack cocaine use. The northern border of AP 2, the southern border and the southeast of AP 3, for example, correspond to an area along Avenida Brasil, an important road connecting downtown area to periphery zones, and Jacare/Jacarezinho and its surroundings. In AP 4, we also see

a cluster of crack users in a neighborhood called Cidade de Deus.

3. Methodology

In this section, we describe a Bayesian model that incorporates the main characteristics outlined on the previous section: dichotomous responses, nonignorable nonresponse and spatial dependence. To that end, we follow Gao, He, and Sun (2014) and construct a pattern mixture model which incorporates a spatial structure in the context of dichotomous responses. Gao, He, and Sun (2014) proposed a Bayesian hierarchical linear mixed model where a conditional autoregressive (Besag, 1974, CAR) prior was set to the random effects. The authors assumed a probit link function to facilitate the application of the Gibbs sampler, and used the data augmentation representation introduced by Albert and Chib (1993).

Our methodology, however, presents two major differences. First, instead of applying the probit link function, we exploit the Pólya-Gamma data-augmented sampler of Polson, Scott, and Windle (2013). Second, we employ a hierarchical Nearest-Neighbor Gaussian Process (Datta et al., 2016, NNGP). Since we are working with a large geostatistical data set, a bottleneck arises from the size of the spatial covariance matrix. As Bayesian inference commonly relies on computational methods for sampling from the joint posterior density, for example, Markov chain Monte Carlo methods (Gamerman and Lopes, 2006, MCMC), each iteration of the algorithm involves decomposing or factorizing this matrix, mostly requiring $\mathcal{O}(n^3)$ floating point operations and memory of the order $\mathcal{O}(n^2)$. The NNGP model induces an exploitable structure in the spatial covariance matrix and makes inference feasible.

3.1. Model

Define $\mathcal{R} = \{s_1, \ldots, s_n\}$, $n = 5162$, as the set of individuals locations. Let $Y(s_i)$ and $Z(s_i)$ denote the indicator of crack cocaine use and response, respectively. Then,

$$Y(s_i) = \begin{cases} 1, & \text{if the } i\text{th individual is a crack user} \\ 0, & \text{if the } i\text{th individual is not a crack user} \end{cases},$$

$$Z(s_i) = \begin{cases} 1, & \text{if the } i\text{th individual is a respondent} \\ 0, & \text{if the } i\text{th individual is not a respondent} \end{cases}.$$

Furthermore, let $x_1(s_i) = 1$ for female and $x_1(s_i) = 0$ for male; $x_2(s_i) = 1$ for white individuals and $x_2(s_i) = 0$ for non-white; $x_3(s_i)$ be an indicator variable to other races; $x_4(s_i)$ be an indicator variable to elder persons; $x_5(s_i)$ be an indicator variable to scenes of drug use; $x_6(s_i)$ be an indicator variable to street.

We assume that

$$Y(s_i) \overset{ind}{\sim} \text{Binomial}\left(n(s_i), \frac{1}{1+\exp\left(-\psi_1(s_i)\right)}\right) \text{ and} \tag{1}$$

$$Z(s_i) \overset{ind}{\sim} \text{Binomial}\left(n(s_i), \frac{1}{1+\exp\left(-\psi_2(s_i)\right)}\right), \tag{2}$$

where $\psi_1(s_i) = \sum_{j=1}^{5} x_j(s_i)\beta_{1j} + w_1(s_i) + \alpha_{sd}x_5(s_i)(1-z(s_i)) + \alpha_{st}x_6(s_i)(1-z(s_i))$ and

$\psi_2(s_i) = \sum_{j=1}^{5} x_j(s_i)\beta_{2j} + w_2(s_i)$. The neighbor sets $n(s_i)$ are defined as $n(s_1) = \{\}$ and $n(s_i) = \min(m, i-1)$ nearest neighbors of s_i in $s_1, \ldots, s_{i-1}$, for $i = 2, \ldots, n$. Then, considering a Nearest-Neighbor Gaussian Process to the vector of random spatial effects $\boldsymbol{w}_k = (w_k(s_1), \ldots, w_k(s_n))$, we have

$$p(\boldsymbol{w}_k) = \prod_{i=1}^{n} p(w_k(s_i)|\boldsymbol{w}_k(n(s_i))), \text{ for } k = 1, 2,$$

in which $\boldsymbol{w}_k \sim N(\boldsymbol{0}, \widetilde{\boldsymbol{C}}(\boldsymbol{\theta}_k))$ and the inverse of the NNGP covariance matrix, $\widetilde{\boldsymbol{C}}(\boldsymbol{\theta}_k)^{-1}$, is sparse. We write $\widetilde{\boldsymbol{C}}(\boldsymbol{\theta}_k) = \sigma_k^2 \widetilde{R}(\boldsymbol{\phi}_k)$ where $\widetilde{R}(\boldsymbol{\phi}_k)$ is the NNGP approximation of the Gaussian Process correlation matrix.

Note that we have three patterns in Equation (1). The first is $\psi_1(s_i) = \sum_{j=1}^{5} x_j(s_i)\beta_{1j} + w_1(s_i)$ for questionnaire respondents ($z(s_i) = 1$). The second is $\psi_1(s_i) = \sum_{j=1}^{5} x_j(s_i)\beta_{1j} + w_1(s_i) + \alpha_{sd}$ for non respondents located on scenes of drug use ($z(s_i) = 0$, $x_4(s_i) = 1$), and the third is $\psi_1(s_i) = \sum_{j=1}^{5} x_j(s_i)\beta_{1j} + w_1(s_i) + \alpha_{st}$ for non respondents located on streets ($z(s_i) = 0$, $x_5(s_i) = 1$). This is the so-called pattern mixture approach (Little and Rubin, 2002).

To avoid introducing $2 \times n$ Metropolis-Hastings steps in every iteration to update the spatial random effects, and, consequently, slow convergence issues, we employ the Pólya-Gamma data-augmented sampler proposed by Polson, Scott, and Windle (2013). Letting $\kappa_1(s_i) = (y(s_i) - n(s_i))/2$, $\kappa_2(s_i) = (z(s_i) - n(s_i))/2$, and introducing the augmented data $\boldsymbol{\omega}_k = (\omega_k(s_1), \ldots, \omega_k(s_n))$ for $k = 1, 2$, we have

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\omega}_1, \boldsymbol{\omega}_2|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{w}_1, \boldsymbol{w}_2) = \prod_{i=1}^{n} \exp\left\{-\frac{\omega_1(s_i)}{2}\left(\frac{\psi_1(s_i) - \kappa_1(s_i)}{\omega_1(s_i)}\right)^2\right\} \\ \times \exp\left\{-\frac{\omega_2(s_i)}{2}\left(\frac{\psi_2(s_i) - \kappa_2(s_i)}{\omega_2(s_i)}\right)^2\right\}. \end{aligned} \tag{3}$$

The likelihood in Equation (3) is similar to the usual likelihood from a spatial mixed linear model with normal distributed errors, but using responses $y^*(s_i) = \kappa_1(s_i)/\omega_1(s_i)$, $z^*(s_i) = \kappa_2(s_i)/\omega_2(s_i)$ and heteroskedastic variances $\tau_1^2(s_i) = 1/\omega_1(s_i)$, $\tau_2^2(s_i) = 1/\omega_2(s_i)$. The additional step we need to execute in this case is updating the $\omega_1(s_i)$ and $\omega_2(s_i)$ for $i = 1, \ldots, n$ as follows

$$\omega_k(s_i) \sim PG(n(s_i), \psi_k(s_i)), \ k = 1, 2,$$

where $PG(a, b)$ denotes the Pólya-Gamma distribution with shape parameter a and tilting parameter b.

We then assume the priors distributions $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, $IG(a_1, b_1)$ and $IG(a_2, b_2)$ for $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{15})$, $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{25})$, σ_1^2 and σ_2^2 respectively.

Following Gao, He, and Sun (2014), we assume the conditional prior distribution for α_{sd} and α_{st} given σ_1^2 as a Cauchy distribution centered at $\boldsymbol{0}$ and and scale of $\sqrt{n\left(\sum_{i=1}^{n}\left(x_4(s_i)(1 - z(s_i))\right)^2\right)^{-1}\sigma_1^2}$ and $\sqrt{n\left(\sum_{i=1}^{n}\left(x_5(s_i)(1 - z(s_i))\right)^2\right)^{-1}\sigma_1^2}$ respectively. REVER ESSAS PRIORIS PARA CABER NO CÓDIGO IMPLEMENTADO

FALAR DAS OPÇÕES PARA $\widetilde{R}(\boldsymbol{\phi}_k)$ ("exponential", "matern", "spherical", and "gaussian") E DAS PRIORIS DE $\boldsymbol{\phi}_1$ E $\boldsymbol{\phi}_2$.

References

Albert, James H., and Siddhartha Chib. 1993. "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association* 88 (422): 669–679.

Besag, Julian. 1974. "Spatial interaction and the statistical analysis of lattice systems." *ournal of the Royal Statistical Society, Series B (Methodological)* 36 (2): 192–236.

Datta, Abhirup, Sudipto Banerjee, Andrew O. Finley, and Alan E. Gelfand. 2016. "Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets." *Journal of the American Statistical Association* 111 (514): 800–812.

Gamerman, Dani, and Hedibert F. Lopes. 2006. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Boca Raton, Florida: Chapman and Hall/CRC.

Gao, Xiaoming, Chong He, and Dongchu Sun. 2014. "A Bayesian spatial model with auxiliary covariates to assess and adjust nonignorable nonresponse." *Spatial Statistics* 8: 122–144.

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. 3rd ed. Hoboken, New Jersey: Wiley.

Polson, Nicholas G., James G. Scott, and Jesse Windle. 2013. "Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables." *Journal of the American Statistical Association* 108 (504): 1339–1349.

Ribeiro, Marcelo, Sérgio Duailibi, Rosana Frajzinger, Ana Leonor S. Alonso, Lucas Marchetti, Anna V. Williams, John Strang, and Ronaldo Laranjeira. 2016. "The Brazilian 'Cracolândia' open drug scene and the challenge of implementing a comprehensive and effective drug policy." *Addiction* 111 (4): 571–573.