CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

AmostraBrasil: um pacote R para amostragem domiciliar em municípios brasileiros

AmostraBrasil: an R package for household sampling in Brazilian municipalities

AmostraBrasil: un paquete R para muestras domiciliarias en municipios brasileiros

Ricardo Cordeiro [1]
Celso Stephan [1]
Maria Rita Donalísio [1]

doi: 10.1590/0102-311X00069516

Resumo

Frente à relevância dos inquéritos epidemiológicos e às dificuldades em se estabelecer um adequado plano amostral para a sua realização, este artigo apresenta o pacote AmostraBrasil, integrante do software R, de livre acesso, que automatiza a obtenção de amostras aleatórias – simples, sistemáticas e estratificadas – de domicílios de quaisquer municípios do Brasil. Além disso, o pacote possibilita a obtenção automática das coordenadas geográficas dos domicílios amostrados, bem como shapefiles *com o perímetro do município e a distribuição espacial da amostra. São descritos os passos para a sua instalação e utilização no sistema operacional Windows. São apresentados exemplos de aplicações do pacote: amostragem e distribuição espacial de 2.500 domicílios residenciais da cidade do Rio de Janeiro e geração de controles na estimativa da distribuição espacial do risco de homicídios em Campinas, São Paulo. São também apresentadas as exigências e limitações da utilização do pacote AmostraBrasil.*

Amostragem; Inquéritos Epidemiológicos; Software

[1] *Faculdade de Ciências Médicas, Universidade Estadual de Campinas, Campinas, Brasil.*

Correspondência
M. R. Donalísio
Departamento de Saúde Coletiva, Faculdade de Ciências Médicas, Universidade Estadual de Campinas.
Rua Tessalia Vieira de Camargo 126, Campinas, SP 13083-887, Brasil.
rita.donalisio@gmail.com

Introdução

Os inquéritos domiciliares amostrais constituem um método comum e importante na obtenção de dados em estudos não apenas epidemiológicos, mas também de diversos outros campos do conhecimento. Em tais estudos, grande parte das informações analisadas são obtidas por meio de entrevistas com moradores de domicílios previamente alocados em processos de aleatorização de complexidade variada. A existência de um cadastro universal de domicílios, de fácil acesso, cobrindo a área do estudo, viabiliza o planejamento de amostras aleatórias nos inquéritos populacionais. A utilização do cadastro garante, ao menos no plano teórico, que cada domicílio da região estudada tenha a mesma chance de ser incluído na amostra, o que pode ser estendido, mediante certos procedimentos, para cada morador da região de interesse, exceção feita aos moradores de rua.

Até recentemente, para grandes áreas urbanas, era praticamente impossível obter um cadastro rigorosamente universal de domicílios no Brasil. As alternativas comumente utilizadas apresentam limitações, em menor ou maior grau. Felizmente, em 2011, o Instituto Brasileiro de Geografia e Estatística (IBGE) disponibilizou listas com os endereços de todos os imóveis dos 316.574 setores censitários do Brasil visitados por seus recenseadores durante o *Censo Demográfico* de 2010 (http://www.censo2010.ibge.gov.br/cnefe/). Nestas listas, em forma de arquivos texto, além dos respectivos endereços, encontra-se discriminado o tipo do imóvel endereçado: domicílio particular, domicílio coletivo, estabelecimento agropecuário, estabelecimento de ensino, estabelecimento de saúde, outros. Desse modo, foi superada uma das grandes dificuldades do planejamento amostral de inquéritos domiciliares. Entretanto, a obtenção e o manuseio destas listas não é tarefa trivial.

O objetivo desta comunicação é apresentar um pacote integrante do software R (The R Foundation for Statistical Computing, Viena, Áustria; http://www.r-project.org), de livre acesso, que automatiza a obtenção de amostras aleatórias de domicílios de quaisquer municípios do Brasil, bem como de suas coordenadas geográficas e distribuição espacial.

Funcionamento do pacote

O pacote se chama AmostraBrasil, oferecido sob a licença GPL (Licença Pública Geral). Foi desenvolvido pelos autores no Laboratório de Análise Espacial de Dados Epidemiológicos (epiGeo) da Universidade Estadual de Campinas (Unicamp) para ser executado sob o software R, a partir da versão 3.1.3. Abaixo, descrevem-se os passos para a sua instalação e utilização.

Para instalar o software R

1. Visite https://www.r-project.org/ e siga as instruções de instalação.

Para instalar o pacote AmostraBrasil

2. Inicialize o software R.
3. No *prompt* do R, digite install.packages("AmostraBrasil") [enter] e siga as instruções.

Os passos 2 e 3 precisam ser dados apenas uma vez em cada computador onde se deseja utilizar o pacote AmostraBrasil.

Para carregar o pacote AmostraBrasil em uma seção de trabalho no R

4. Com o R aberto, digite library(AmostraBrasil) [enter].

Isso faz com que o pacote AmostraBrasil esteja disponível para uso. Quando se pretende usar o pacote, esse passo precisa ser dado apenas uma vez após cada inicialização do R.

Para utilizar o pacote AmostraBrasil

Para utilizar o pacote, pressupõe-se que os passos 1 a 4 foram dados.

Defina os seguintes parâmetros obrigatórios:

* nome do município (grafia oficial) OU código IBGE do município com sete dígitos (ambos podem ser obtidos em http://www.ibge.gov.br/home/geociencias/areaterritorial/area.shtm).

* tamanho da amostra (número inteiro positivo).

Digite amostraBrasil(mun="nome do município", N=tamanho da amostra) [enter]

Observação: (a) o pacote se chama AmostraBrasil, com "A" maiúsculo. Após instalado, a função que o opera se chama amostraBrasil, com "a" minúsculo; (b) o parâmetro que define o tamanho da amostra é "N", maiúsculo; e (c) ao invés de mun, pode-se usar o parâmetro codmun informando-se o código IBGE do município com 7 dígitos, entre aspas.

A partir do comando acima, AmostraBrasil faz uma amostra aleatória simples sem reposição dos domicílios particulares do município indicado, com o tamanho especificado. Como *output* na tela, o programa mostra uma tabela com o número de identificação do domicílio, o setor censitário a que pertence, indicação se o domicílio está em zona urbana (1) ou rural (2), confirmação

de que se trata de um domicílio residencial (1) e endereço completo do domicílio. No diretório de trabalho é gravado um arquivo dbf contendo o endereço de todos os domicílios residenciais (urbanos e rurais) do município, que serviu de base para a amostragem.

Para obter a localização espacial (latitude e longitude) dos domicílios amostrados, deve-se acrescentar o parâmetro geocod=T ao comando. Nesse caso, o AmostraBrasil geocodifica os endereços obtidos nas listagens do IBGE, utilizando o serviço Google Maps Geocoding API, e mostra o resultado na tela.

Caso o usuário deseje obter arquivos *shapefile* contendo o perímetro do município escolhido e a localização espacial dos domicílios amostrados, deve acrescentar o parâmetro shape=T, conforme se segue: digite amostraBrasil(mun="nome do município", N=tamanho da amostra, geocod=T, shape=T) [enter].

Os arquivos *shapefile* aparecem como *output* na tela e são gravados no diretório de trabalho em uso no R, podendo então ser acessados por qualquer Sistema de Informações Geográficas (SIG).

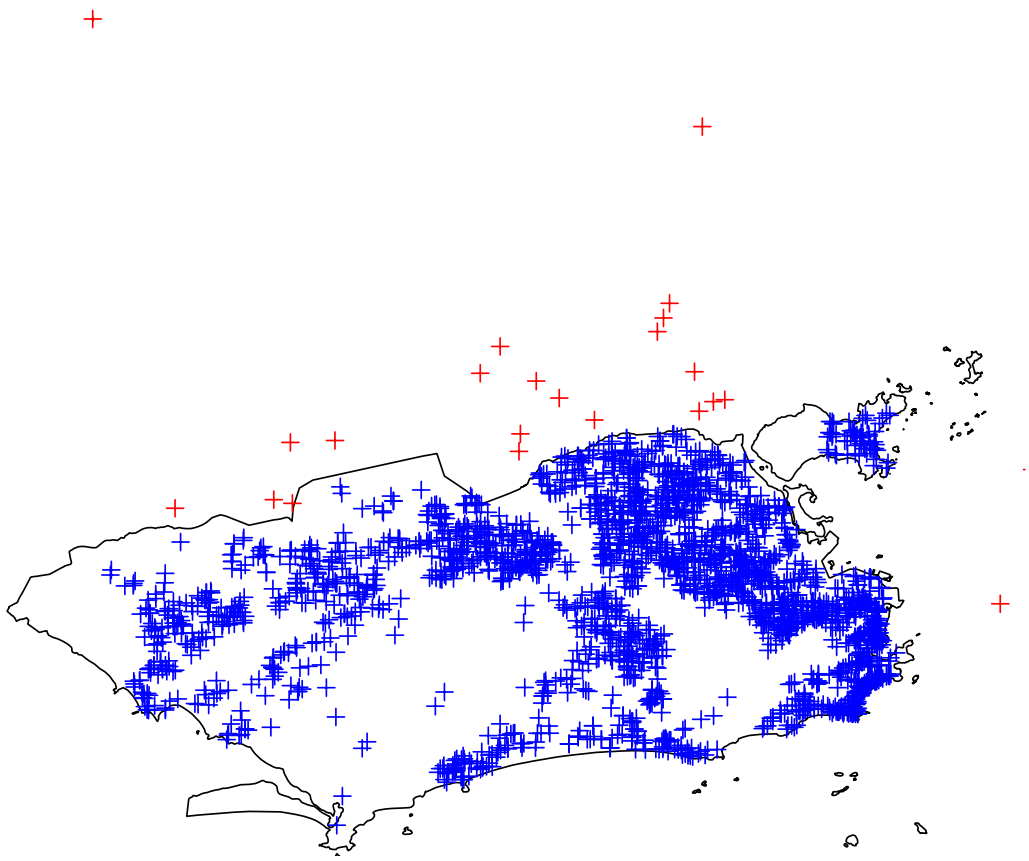
A Figura 1 ilustra os *shapefiles* de área e pontos gerados com uma realização do comando "amostraBrasil(mun="Rio de Janeiro", N=2500, geocod=T, shape=T).

Na Figura 1 são observados que dos 2.500 domicílios amostrados, alguns poucos encontram-se fora do perímetro municipal do Rio de Janeiro. Essa limitação será discutida posteriormente.

O AmostraBrasil, por *default*, realiza amostra aleatória simples de domicílios residenciais do município escolhido. É possível restringir o espa-

Figura 1

Output de tela do pacote AmostraBrasil apresentando a distribuição espacial da realização de uma amostra de 2.500 domicílios residenciais na cidade do Rio de Janeiro, Brasil.



ço amostral dentro do município a setores censitários previamente definidos, com o acréscimo de um parâmetro ao comando, como ilustrado a seguir.

Digite amostraBrasil(mun="nome do muni cípio", N=tamanho da amostra, setor=c(código do IBGE com 15 dígitos de cada setor escolhido, entre aspas e separados por vírgula)) [enter].

Observação: o código IBGE dos setores censitários pode ser obtido abrindo-se a tabela de atributos do *shapefile* do município escolhido, encontrado em ftp://geoftp.ibge.gov.br/malhas_digitais/censo_2010/setores_censitarios.

Assim fazendo, o pacote realiza amostra aleatória simples de domicílios no município escolhido, com o tamanho escolhido, restrita aos setores censitários escolhidos. Uma vez que uma amostra aleatória estratificada pode ser entendida como um conjunto de amostras aleatórias simples realizadas em subpopulações não sobrepostas convenientemente definidas no espaço amostral, isso abre a possibilidade da utilização do AmostraBrasil para realização de amostragem estratificada, com n fixo ou proporcional em cada estrato. Por exemplo, o usuário pode definir estratos de acordo com uma variável indicadora de nível de renda à sua escolha. Para realizar a amostra estratificada, no primeiro estrato definido, deve-se informar ao AmostraBrasil quais setores censitários compõem este estrato, o tamanho amostral e realizar a amostragem; repetindo-se esse procedimento para cada um dos demais estratos definidos.

O AmostraBrasil também possibilita a realização de amostras aleatórias sistemáticas, utilizando-se a listagem produzida de todos os endereços particulares no espaço amostral definido, ordenada de acordo com algum critério conveniente, escolhendo-se um passo conveniente e fazendo-se a seleção em uma planilha eletrônica a partir do arquivo dbf gerado pelo pacote.

Aplicação

São inúmeras as possibilidades de utilização do pacote AmostraBrasil. Além das aplicações comentadas anteriormente, ilustramos a utilização do pacote em um estudo visando estimar a distribuição espacial do risco de homicídio em Campinas, Estado de São Paulo. Para tanto, foram obtidos os locais de ocorrência (latitude/longitude) dos 141 homicídios incidentes na cidade entre moradores de Campinas, no ano 2015. Assumimos que a distribuição espacial da população fonte de homicídios atingindo residentes de Campinas pode ser estimada pela distribuição espacial de domicílios residenciais da cidade. Esta foi obtida por meio de uma amostra aleatória simples, de tamanho 200, utilizando-se o pacote AmostraBrasil. A Figura 2 mostra a distribuição espacial dos locais de ocorrência de homicídios e dos domicílios amostrados nesse estudo.

Aos pontos obtidos, ajustou-se um modelo aditivo generalizado [1], utilizando-se uma função bivariada suave que foi estimada não parametricamente por meio de regressão *spline* penalizada [2]. Detalhes do método utilizado podem ser encontrados em Bailey et al. [3]. A Figura 3 mostra a distribuição do risco relativo espacial de homicídio em Campinas (sob os pressupostos antes referidos), isto é, a razão entre o risco de homicídio em cada ponto da superfície de Campinas e o risco médio no município. A Figura 4 mostra a significância das estimativas de risco relativo espacial obtidas. Observam-se áreas significativas ($p < 0,05$) de risco aumentado ao sul e de risco diminuído ao centro e leste da cidade. O modelo utilizado permite incorporar covariáveis não espaciais ecológicas, herdadas dos setores censitários onde se localizam os homicídios e domicílios analisados.

Limitações

A utilização do AmostraBrasil apresenta algumas limitações, dentre as quais destacamos:

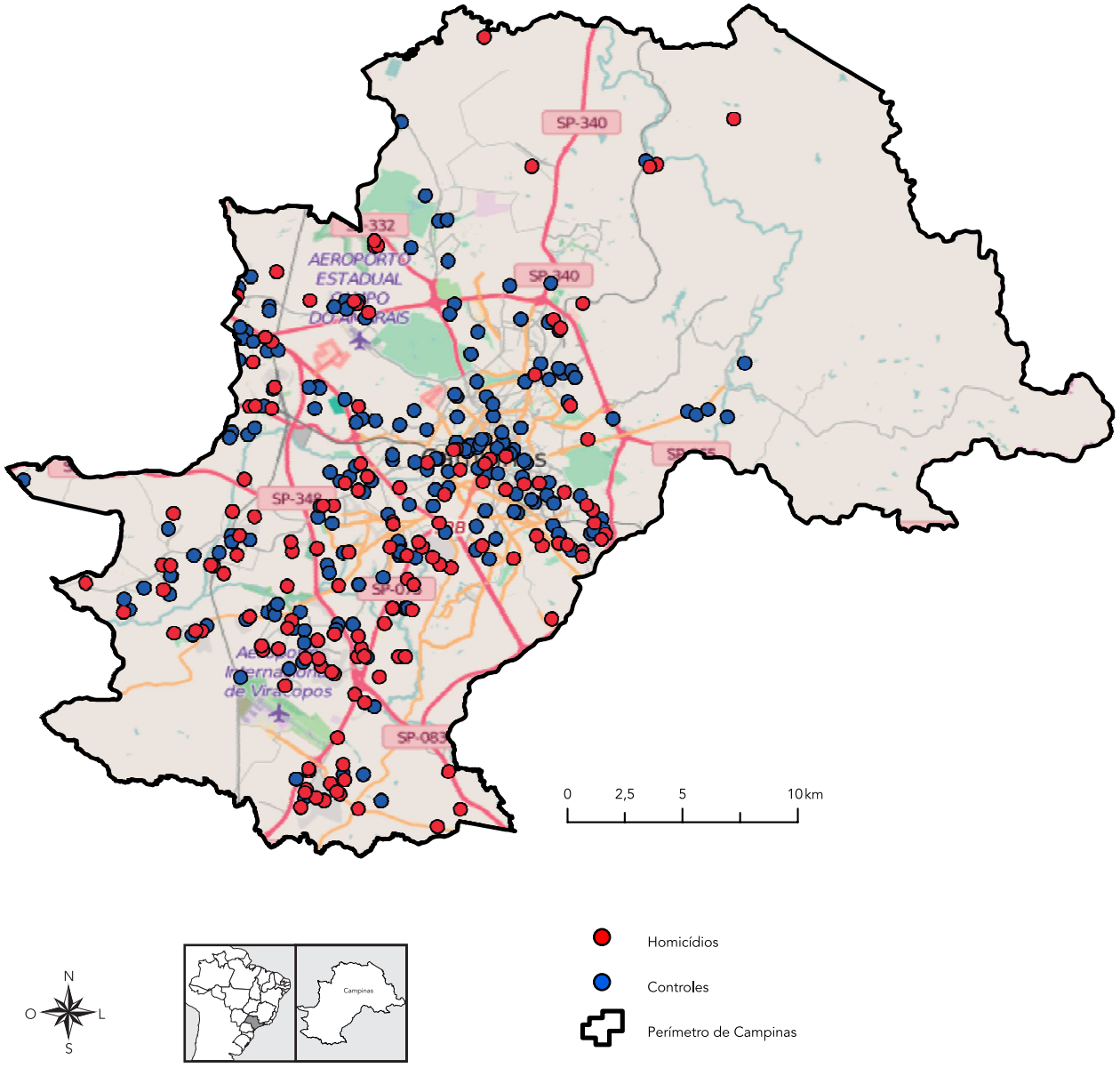
(a) O uso do pacote requer uma conexão veloz com a Internet e boa capacidade de memória quando se deseja realizar amostras em municípios populosos. A base de dados do IBGE para a cidade de São Paulo, por exemplo, conta com cerca de três milhões de domicílios residenciais. Manipular esse volume de informações não é tarefa simples;

(b) O processo de geocodificação apresenta imprecisões próprias, inerentes ao modo como os domicílios são numerados nas ruas de seus respectivos municípios. Além disso, na fase de testes do pacote, os autores observaram que, em média, 1% dos domicílios amostrados apresentam problemas de geocodificação, possivelmente decorrentes de imprecisões no armazenamento de endereços na base de dados do IBGE. Isto é o que explica os pontos fora do perímetro urbano da Figura 1. Sugerimos que ao utilizar o AmostraBrasil se faça uma verificação, utilizando-se a opção geocod=T shape=T, e retirando-se da amostra os domicílios que eventualmente se encontrem fora do perímetro municipal amostrado. Eles são marcados com o valor "0" no campo "dentro", encontrado no arquivo "'nome_do_município'_pts.dbf".

(c) A utilização do AmostraBrasil é dependente da manutenção das bases de dados do IBGE uti-

Figura 2

Distribuição espacial dos locais de ocorrência de homicídios e controles amostrados em Campinas, São Paulo, Brasil, 2015.



lizadas, da preservação inalterada dos endereços dessas bases, bem como da manutenção do livre acesso remoto a elas. O mesmo se aplica ao serviço do Google Maps acessado pelo pacote;

(d) O Google Maps atualmente impõe aos usuários um limite de geocodificação grátis de 2.500 endereços por IP por dia. Para geocodificar mais do que isso com o AmostraBrasil o usuário deve fracionar seu trabalho em vários computadores, ou em vários dias, ou pagar pela geocodificação extra ao Google.

Figura 3

Distribuição do risco relativo espacial de homicídios em Campinas, São Paulo, Brasil, 2015.

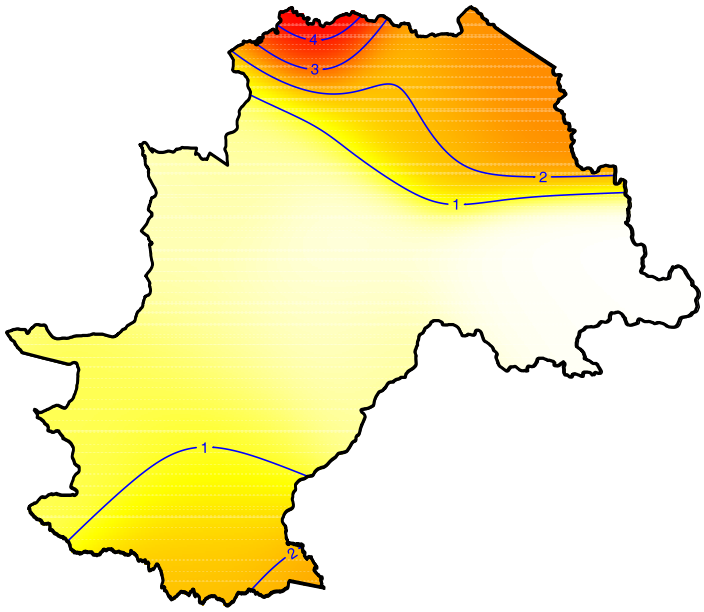


Figura 4

Significância estatística das estimativas de do risco relativo espacial de homicídios em Campinas, São Paulo, Brasil, 2015.

Colaboradores

Todos os autores deram contribuições substanciais para a concepção e desenho do artigo, interpretação dos dados, escrita e revisão do manuscrito.

Referências

1. Hastie TJ, Tibshirani RJ. Generalized additive models. Boca Raton: Chapman & Hall/CRC Press; 1990. (Monographs on Statistics and Applied Probability, 43).
2. Ruppert D, Wand MP, Carroll RJ. Semiparametric regression. Cambridge: Cambridge University Press; 2003.
3. Bailey TC, Cordeiro R, Lourenço RW. Semi-parametric modeling of the spatial distribution of occupational accident risk in the casual labor market, Piracicaba, Southeast Brazil. Risk Anal 2007; 27:421-31.

Abstract

Given the relevance of epidemiological surveys and the difficulties in establishing an adequate sampling plan to conduct them, this article present the AmostraBrasil package, part of the open-access R software, which automatizes the taking of random samples – simple, systematic, and stratified – from households in any Brazilian municipalities (counties). The package also allows automatically obtaining the sampled households' geographic coordinates, was well as shapefiles of the municipality's perimeter and the sample's spatial distribution. The article describes the steps for installing and using the package in the Windows OS. Examples are provided of the package's applications: sampling and spatial distribution of 2,500 residential households in the city of Rio de Janeiro and generation of controls in estimating risk spatial distribution.

Sampling Studies; Health Surveys; Software

Resumen

Frente a la relevancia de las encuestas epidemiológicas y las dificultades de establecer un plan adecuado de muestras para su realización, este artículo presenta el paquete AmostraBrasil, integrante del software R, de libre acceso, que automatiza la obtención de muestras aleatorias -simples, sistemáticas y estratificadas- de domicilios de cualquier municipio de Brasil. Asimismo, el paquete posibilita la obtención automática de las coordenadas geográficas de los domicilios de la muestra, así como como shapefiles con el perímetro del municipio y la distribución espacial de la muestra. Se describen los pasos para su instalación y utilización en el sistema operacional Windows. Se presentan ejemplos de aplicaciones del paquete: muestra y distribución espacial de 2.500 domicilios residenciales de la ciudad de Río de Janeiro y generación de controles en la estimativa de la distribución espacial del riesgo.

Muestreo; Encuestas Epidemiológicas; Programas Informáticos

Recebido em 24/Abr/2016
Versão final reapresentada em 24/Ago/2016
Aprovado em 08/Set/2016