CSP CADERNOS DE SAÚDE PÚBLICA
REPORTS IN PUBLIC HEALTH

Microdatasus: pacote para *download* e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS)

Microdatasus: a package for downloading and preprocessing microdata from Brazilian Health Informatics Department (DATASUS)

Microdatasus: paquete para descarga y preprocesamiento de microdatos del Departamento de Informática del SUS (DATASUS)

Raphael de Freitas Saldanha [1]
Ronaldo Rocha Bastos [2]
Christovam Barcellos [1]

doi: 10.1590/0102-311X00032419

Resumo

O objetivo do estudo foi desenvolver um algoritmo capaz de realizar o download *e o pré-processamento de microdados fornecidos pelo Departamento de Informática do SUS (DATASUS) para diversos sistemas de informações em saúde para a linguagem de programação estatística R. O pacote desenvolvido permite o* download *e o pré-processamento de dados de diversos sistemas de informação em saúde, com a inclusão da rotulagem dos campos categóricos nos arquivos. A função de* download *foi capaz de acessar diretamente e reduzir o volume de trabalho para a seleção de arquivos e variáveis de microdados junto ao DATASUS. Já a função de pré-processamento foi capaz de efetuar a codificação automática de diversos campos categóricos. Dessa forma, a utilização desse pacote possibilita um fluxo de trabalho contínuo no mesmo programa, no qual esse algoritmo permite o* download *e o pré-processamento, e outros pacotes do R permitem a análise de dados dos sistemas de informação em saúde do Sistema Único de Saúde (SUS).*

Software; Processamento Eletrônico de Dados; Sistemas de Informação em Saúde

Correspondência
R. F. Saldanha
Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fundação Oswaldo Cruz.
Av. Brasil 4365, Pavilhão Haity Moussatché, Rio de Janeiro, RJ 21040-900, Brasil.
raphael.saldanha@icict.fiocruz.br

[1] Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Fundação Oswaldo Cruz, Rio de Janeiro, Brasil.
[2] Departamento de Estatística, Universidade Federal de Juiz de Fora, Juiz de Fora, Brasil.

Introdução

A melhoria dos sistemas de saúde e de seus processos de decisão depende fortemente da produção de dados sobre o seu funcionamento [1]. No Brasil, após a *Constituição Federal* de 1988 ter estabelecido o Sistema Único de Saúde (SUS), foi criado o Departamento de Informática do SUS (DATASUS) em 1991 visando à coleta e organização de dados referentes ao SUS [2].

Os sistemas de informação em saúde mantidos pelo DATASUS, ou em colaboração com ele, cobrem diversos aspectos da saúde populacional. Alguns são de natureza epidemiológica, como o Sistema de Informações sobre Mortalidade (SIM) e o Sistema de Informações sobre Nascidos Vivos (SINASC), que utilizam dados dos cartórios. Já outros sistemas têm objetivos administrativos, como o Sistema de Internações Hospitalares (SIH) e o Sistema de Informações Ambulatoriais (SIA), usando dados provenientes diretamente da assistência à saúde. Mesmo os sistemas de informação de origem administrativa e financeira contêm dados relevantes acerca da situação de saúde brasileira [3].

A disseminação desses dados é realizada pelo DATASUS por meio de duas interfaces oferecidas aos usuários: TabNet e TabWin [4]. O TabNet é uma interface de produção de tabelas de dados agregados por meio do acesso a microdados contidos nos seus servidores de dados. Ele também possibilita a consulta de dados e indicadores de diferentes sistemas de informação em saúde agregados em unidades de tempo ou unidades geográficas.

O TabWin é uma interface de acesso do tipo "cliente", que permite a leitura dos arquivos de microdados do DATASUS disponibilizados no formato DBC. A utilização dos microdados anonimizados disponibilizados pelo DATASUS permite a realização de pesquisas com maior flexibilidade e detalhamento por parte do usuário por não estarem agregados em unidades preestabelecidas de tempo ou região.

Contudo, o TabWin apresenta algumas limitações [4,5]. Dentre elas, pode-se mencionar que ele pode ser executado apenas em um sistema operacional (Microsoft Windows) e não oferece a opção de *download* direto dos microdados, que devem ser baixados e organizados previamente pelo usuário. Além disso, a análise de dados no TabWin é limitada e precisa ser realizada em pacotes estatísticos dedicados.

Como alternativa à interface ao TabWin e superando algumas de suas limitações, este trabalho desenvolveu e disponibilizou um pacote para o programa estatístico R (http://www.r-project.org), com funções de *download* e pré-processamento de microdados do DATASUS. Ainda que o R apresente uma curva de aprendizado árdua, considera-se que esta dificuldade inicial é recompensada pela gama de possibilidades de manipulação e análise de dados que o programa permite.

Métodos

O programa estatístico R é uma linguagem de programação versátil, que permite desde a manipulação de dados à sua análise por meio de métodos estatísticos.

A quantidade e finalidade dos comandos ou funções do R podem ser livremente ampliadas pela criação de novos pacotes de funções. O algoritmo desenvolvido foi disponibilizado num pacote denominado *microdatasus*, seguindo as recomendações de desenvolvimento de pacotes [6]. Ele oferece funções para o download e o pré-processamento de microdados disponibilizados pelo DATASUS.

Instalação do pacote

O pacote pode ser instalado por meio de seu repositório no *website* GitHub (https://github.com/). Com o pacote *devtools* previamente instalado no R, o pacote *microdatasus* pode ser instalado da seguinte forma:

```
devtools::install_github("rfsaldanha/microdatasus")
```

Função *fetch_datasus*

Para o acesso, *download* e leitura dos arquivos de microdados no formato DBC foi desenvolvida a função *fetch_datasus*. Após especificar à função qual sistema de informação em saúde e qual a cobertura no tempo desejada, ela realiza o *download* dos respectivos microdados disponíveis no *website* do DATASUS (http://datasus.saude.gov.br/). A importação dos dados é realizada utilizando-se, interna mente, o pacote *read.dbc* (https://CRAN.R-project.org/package=read.dbc).

A função apresenta a seguinte estrutura básica:

```
fetch_datasus(year_start, month_start, year_end, month_end,
              uf = "all", information_system, vars = NULL)
```

Onde *year_start* é o ano do início da cobertura dos dados; *month_start* é o mês do início da cobertura dos dados; *year_end* é o ano do fim da cobertura dos dados; *month_end* é o mês do fim da cobertura dos dados; *uf* é a lista das Unidades Federativas que devem ser baixadas, informadas por intermédio de suas siglas oficiais [p.ex.: uf = c("RJ", "MG", "SP")]; *information_system* é a sigla do sistema de informação de saúde que deve ser acessado; e *vars* é a lista de variáveis que deve ser mantida após o *download*. Por padrão, a função mantém todas as variáveis encontradas nos microdados.

Em sua versão atual, a função *fetch_datasus* permite o *download* e a leitura dos arquivos de microdados dos sistemas SIM, SINASC, SIH, CNES (Cadastro Nacional dos Estabelecimentos de Saúde) e SIA, conforme o Quadro 1.

Cabe destacar que os argumentos relativos à época de cobertura dos dados (*year_start*, *month_start*, *year_end* e *month_end*) são referentes aos anos e meses de processamento dos casos pelo DATASUS. Por exemplo, óbitos ocorridos em dezembro de 2017 podem ser encontrados em arquivos de dezembro de 2017 e janeiro de 2018.

Os argumentos *month_start* e *month_end* são aplicados no *download* de dados dos sistemas de informação cujos arquivos de microdados são mensais, conforme Quadro 1.

A utilização do argumento *vars* pode ser de interesse quando pretende-se aplicar a função em um grande período temporal ou grande abrangência regional de cobertura, o que resulta em um grande número de registros. Para poupar recursos computacionais nesse caso, pode-se limitar o número de variáveis de interesse por meio do argumento *vars*, reduzindo o tamanho final do *data.frame*.

Cabe destacar que o *download* dos arquivos de microdados são realizados para uma pasta temporária do R no computador cliente e eliminados após a execução da função.

Funções de pré-processamento

Foram criadas funções de pré-processamento específicas para cada sistema de informação em saúde. Essas funções realizam o tratamento dos campos de acordo com o sistema de informação em saúde obtido, convertendo cada variável ao seu formato correto (texto, número inteiro, número decimal ou categórico) e imputando rótulos dos campos categóricos.

A versão atual do pacote permite o pré-processamento do SIM (todas as subdivisões), SINASC e SIH-RD, conforme o Quadro 2.

Os argumentos dessas funções são: *data* (objeto criado com a função *fetch_datasus*); *municipality_data* (enriquecimento dos dados referentes ao município de residência); e *information_system* (sistema de informações em saúde específico).

O argumento *municipality_data* acrescenta informações referentes ao município de residência do caso, como o nome do município, nome da Unidade Federativa, latitude, longitude da sede administrativa e área territorial.

O argumento *information_system* existe apenas na função dedicada a dados do SIH. Atualmente, essa função suporta apenas dados do SIH-RD, e toma esse argumento como padrão. Futuramente, a função será expandida para suportar arquivos de outros subsistemas do SIH.

A rotulagem dos campos categóricos é a tarefa principal dessa função. As informações referentes aos códigos dos campos categóricos foram obtidas baseando-se nos arquivos de definição (extensão

Quadro 1

Sistemas de informação, subdivisões, siglas e períodos de cobertura.

SISTEMA	SUBDIVISÃO	SIGLA	COBERTURA	MENSAL
SIH	AIH reduzida	SIH-RD	1992 – atual	Sim
	AIH rejeitada	SIH-RJ	1992 – atual	Sim
	AIH serviços profissionais	SIH-SP	1992 – atual	Sim
	AIH rejeitadas com código de erro	SIH-ER	1992 – atual	Sim
SIM	Declarações de óbitos	SIM-DO	1979 – atual	Não
	Declarações de óbitos fetais	SIM-DOFET	1979 – atual	Não
	Declarações de óbitos por causas externas	SIM-DOEXT	1979 – atual	Não
	Declarações de óbitos infantis	SIM-DOINF	1979 – atual	Não
	Declarações de óbitos maternos	SIM-DOMAT	1996 – atual	Não
SINASC	Declarações de nascidos vivos	SINASC	1994 – atual	Não
CNES	Leitos	CNES-LT	Outubro/2005 – atual	Sim
	Estabelecimentos	CNES-ST	Agosto/2005 – atual	Sim
	Dados complementares	CNES-EQ	Agosto/2005 – atual	Sim
	Equipamentos	CNES-EQ	Agosto/2005 – atual	Sim
	Serviço especializado	CNES-SR	Agosto/2005 – atual	Sim
	Habilitação	CNES-HB	Março/2007 – atual	Sim
	Profissional	CNES-PF	Agosto/2005 – atual	Sim
	Equipes	CNES-EP	Abril/2007 – atual	Sim
	Regra contratual	CNES-RC	Março/2007 – atual	Sim
	Incentivos	CNES-IN	Novembro/ 2007 – atual	Sim
	Estabelecimentos de ensino	CNES-EE	Março/2007 – atual	Sim
	Estabelecimento filantrópico	CNES-EF	Março/2007 – atual	Sim
	Gestão e metas	CNES-GM	Junho/2007 – atual	Sim
SIA	APAC de acompanhamento à cirurgia bariátrica	SIA-AB	Janeiro/2008 – Março/2013	Sim
	APAC de acompanhamento pós-cirurgia bariátrica	SIA-ABO	Abril/2013 – atual	Sim
	APAC de confecção de fístula arteriovenosa	SIA-ACF	Junho/2014 – atual	Sim
	APAC de laudos diversos	SIA-AD	Janeiro/2008 – atual	Sim
	APAC de medicamentos	SIA-AM	Janeiro/2008 – atual	Sim
	APAC de nefrologia	SIA-AN	Janeiro/2008 – Outubro/2014	Sim
	APAC de quimioterapia	SIA-AQ	Janeiro/2008 – atual	Sim
	APAC de radioterapia	SIA-AR	Janeiro/2008 – atual	Sim
	APAC de tratamento dialítico	SIA-ATD	Junho/2014 – atual	Sim
	Produção ambulatorial	SIA-PA	Junho/1994 – atual	Sim
	Psicossocial	SIA-PS	Janeiro/2013 – atual	Sim
	Atenção domiciliar	SIA-SAD	Novembro/2012 – atual	Sim

AIH: Autorização de Internação Hospitalar; APAC: Autorização de Procedimento de Alta Complexidade; CNES: Cadastro Nacional dos Estabelecimentos de Saúde; SIA: Sistema de Informações Ambulatoriais; SIH: Sistema de Internações Hospitalares; SIM: Sistema de Informações sobre Mortalidade; SINASC: Sistema de Informações sobre Nascidos Vivos.

Fonte: elaborado pelos autores com base em informações constantes no website do DATASUS (http://datasus.saude.gov.br/).

Quadro 2

Sistemas de informação e funções de pré-processamento.

SISTEMA	FUNÇÃO DE PRÉ-PROCESSAMENTO
SIM	`process_sim(data, municipality_data = TRUE)`
SINASC	`process_sinasc(data, municipality_data = TRUE)`
SIH	`process_sih(data, information_system = "SIH-RD", municipality_data = TRUE)`

SIH: Sistema de Internações Hospitalares; SIM: Sistema de Informações sobre Mortalidade; SINASC: Sistema de Informações sobre Nascidos Vivos.
Fonte: elaboração própria.

DEF) criados para o aplicativo TabWin. Destaca-se que algumas variáveis de alguns sistemas não são contempladas por esses arquivos de definição do TabWin, não permitindo assim a rotulagem dessas variáveis.

Os esquemas de rotulagem completos estão documentados detalhadamente na página do projeto, acessível em https://github.com/rfsaldanha/microdatasus.

Resultados

A utilização das duas funções contempladas no pacote *microdatasus* permitiu realizar o *download* e o pré-processamento conforme observa-se:

Consulta ao SIM-DO, ano de 2014 para todas as Unidades Federativas

Download dos dados:

```
> dados_brutos <- fetch_datasus(year_start = 2014, year_end = 2014,
                                information_system = "SIM-DO")
```

Pré-processamento:

```
> dados <- process_sim(data = dados_brutos)
```

Resultados:

```
> dim(dados)
[1] 1227039    103

> table(dados$SEXO)

 Feminino Masculino
  532.362   693.922

> prop.table[table(dados$RACACOR, useNA = "ifany")]*100

   Amarela     Branca  Indígena      Parda     Preta      <NA>
 0.5507567 51.2397731 0.2963231 35.4251984 7.5520012 4.9359474
```

Foram encontradas 1.227.039 declarações de óbitos no Brasil para o período, compondo um banco de dados com 103 variáveis. Dessas declarações, 532.362 óbitos foram do sexo feminino e 693.922 do sexo masculino. Sobre raça/cor, 51,24% dos óbitos foram identificados como de pessoas da cor branca e 4,93% não continham informações.

SIM-DO, 2013 a 2014, campos específicos, Estado do Rio de Janeiro

Aquisição dos dados:

```
dados_brutos <- fetch_datasus(year_start = 2013, year_end = 2014,
                    information_system = "SIM-DO",
                    vars = c("CODMUNRES", "DTOBITO", "CAUSABAS"),
                    uf = c("RJ"))
```

Pré-processamento:

```
dados <- process_sim(data = dados_brutos)
```

Resultados:

```
> dim(dados)
[1] 261076     11

> table(dados$SEXO)

 Feminino  Masculino
  559.956    587.510
```

Foram encontradas 261.076 declarações de óbitos para o Estado do Rio de Janeiro no período especificado. Além das três variáveis selecionadas na consulta, outras oito variáveis foram adicionadas sobre o município de residência com a função *process_sim*, em que o argumento *municipality_data* é verdadeiro por padrão.

SINASC, 2013, UFs específicas

Aquisição dos dados:

```
> dados_brutos <- fetch_datasus(year_start = 2013, year_end = 2013,
                              information_system = "SINASC",
                              uf = c("RJ", "SP", "MG", "ES")
```

Pré-processamento:

```
> dados <- process_sinasc(data = dados_brutos)
```

Resultados:

```
> dim(dados)
[1] 1147627      70
> prop.table[table(dados$LOCNASC, useNA = "ifany")]*100

                     Domicílio                           Hospital
                   0.227774355                       99.431696884
Outro estabelecimento de saúde                             Outros
                   0.248774210                        0.088094825
                          <NA>
                   0.003659726
```

Foram encontrados 1.147.627 declarações de nascidos vivos na Região Sudeste para os arquivos processados no ano de 2013. Nota-se que 99,43% desses nascimentos ocorreram em hospitais.

SIH-RD, 1º semestre de 2014, sem campos adicionais

Aquisição dos dados:

```
> dados_brutos <- fetch_datasus(year_start = 2014, month_start = 1,
                              year_end = 2014, month_end = 6,
                              information_system = "SIH-RD")
```

Pré-processamento:

```
> dados <- process_sih(data = dados_brutos,
                      information_system = "SIH-RD",
                      municipality_data = FALSE)
```

Resultados:

```
> dim(dados)
[1] 5722339     113
```

```
> prop.table[table(dados$COMPLEX, useNA = "ifany")]*100
```

```
 Alta complexidade Média complexidade
          6.373635          93.626365
```

Foram encontradas 5.722.339 Autorizações de Internação Hospitalar (AIH) no Brasil para o primeiro semestre de 2014. Dessas, 6,37% foram classificadas como de alta complexidade.

Discussão

A criação deste pacote tornou possível o acesso aos dados dos sistemas de informação em saúde mantidos pelo DATASUS diretamente por meio do programa R. Isso torna possível a adoção de um fluxo de trabalho linear, sem a necessidade de utilização de diferentes programas para aquisição, pré-processamento e análise de dados, dando celeridade e otimizando o processo de trabalho e a organização dos dados pelo usuário.

Políticas sobre dados abertos em entidades governamentais estão sendo amplamente discutidas atualmente e pode-se afirmar que o Ministério da Saúde, por intermédio do DATASUS, detém ampla experiência neste quesito, demonstrando um inegável pioneirismo na disseminação de dados em um sistema de cobertura universal.

Melhorias e modernizações são possíveis nos mecanismos de disseminação desses dados [5], como na documentação dos arquivos, metadados, estratégias de versionamento e melhor distinção entre data de processamento e real data do evento. Isso vem reforçar a necessidade de manutenção e investimento nos sistemas de informação em saúde para o conhecimento amplo e incondicional das condições de saúde da população brasileira.

Colaboradores

R. F. Saldanha contribuiu com a concepção e delineamento do artigo, aplicação do método e redação do manuscrito. R. R. Bastos e C. Barcellos contribuíram com o delineamento do artigo e sua revisão crítica.

Informações adicionais

ORCID: Raphael de Freitas Saldanha (0000-0003-0652-8466); Ronaldo Rocha Bastos (0000-0001-9597-5967); Christovam Barcellos (0000-0002-1161-2753).

Agradecimentos

Este artigo é parte integrante da tese de doutorado vinculada ao Programa de Pós-graduação Stricto Sensu em Informação e Comunicação em Saúde (PPGICS), do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (Icict), da Fundação Oswaldo Cruz (Fiocruz), intitulada *Da Aquisição a Visualização de Dados: Aplicações do Processo KDD em Saúde*. O presente trabalho foi em parte financiado pela Fiocruz.

Referências

1. Handley K, Boerma T, Victora C, Evans TG. An inflection point for country health data. Lancet Glob Health 2015; 3:e437-8.
2. Ministério da Saúde. DATASUS: trajetória 1991-2002. Brasília: Ministério da Saúde; 2002.
3. Bittencourt SA, Camacho LAB, Leal MC. O Sistema de Informação Hospitalar e sua aplicação na saúde coletiva. Cad Saúde Pública 2006; 22:19-30.
4. Silva NP. A utilização dos programas TABWIN e TABNET como ferramentas de apoio à disseminação das informações em saúde [Dissertação de Mestrado]. Rio de Janeiro: Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz; 2009.
5. Jorge MHPM, Laurenti R, Gotlieb SLD. Avaliação dos sistemas de informação em saúde no Brasil. Cad Saúde Colet (Rio J.) 2010; 18:7-18.
6. Wickham H. R packages: organize, test, document, and share your code. Sebastopol: O'Reilly; 2015.

Abstract

This study aimed to develop an algorithm for downloading and preprocessing microdata furnished by the Brazilian Health Informatics Department (DATASUS) for various health information systems, using the R statistical programming language. The package allows downloading and preprocessing data from various health information systems, with the inclusion of labeling categorical fields in the files. The download function was capable of directly accessing and reducing the workload for the selection of microdata files and variables in DATASUS, while the preprocessing function enabled automatic coding of various categorical fields. The package thus enables a continuous workflow in the same program, in which the algorithm allows downloading and preprocessing and other packages in R allow analyzing data from the health information systems in the Brazilian Unified National Health System (SUS).

Software; Electronic Data Processing; Health Information Systems

Resumen

El objetivo del estudio fue desarrollar un algoritmo capaz de realizar la descarga y pre-procesamiento de microdatos, proporcionados por el Departamento de Informática del SUS (DATASUS), para diversos sistemas de información en salud, así como para el lenguaje de programación estadístico R. El paquete desarrollado permite la descarga y preprocesamiento de datos de diversos sistemas de información en salud, con la inclusión del rótulo de los campos categóricos en los archivos. La función de descarga se mostró capaz de acceder directamente y reducir el volumen de trabajo para la selección de archivos y variables de microdatos a través del DATASUS, mientras que la función de pre-procesamiento fue capaz de efectuar la codificación automática de diversos campos categóricos. De esta forma, la utilización de este paquete posibilita un flujo de trabajo continuo en el mismo programa, donde este algoritmo permite la descarga y preprocesamiento y otros paquetes del R permiten el análisis de datos de los sistemas de información en salud del Sistema Único de Salud (SUS).

Programas Informáticos; Procesamiento Automatizado de Datos; Sistemas de Información en Salud

Recebido em 20/Fev/2019
Versão final reapresentada em 01/Mai/2019
Aprovado em 17/Jun/2019