READ.DBC - UM PACOTE PARA IMPORTAÇÃO DE DADOS DO DATASUS NA LINGUAGEM R

Daniela Petruzalek

Resumo: Este artigo apresenta um novo pacote para a linguagem R com a finalidade de ler arquivos de dados do DATASUS em formato DBC. **Métodos:** Foi realizada uma análise do formato do arquivo DBC e desenvolvido um *software* de conversão deste formato para DBF. O código foi encapsulado em um pacote para distribuição. **Resultados:** O pacote *read.dbc* simplificou o processo de importação de dados do DATASUS na linguagem R, permitindo a leitura direta do arquivo DBC sem necessidade de conversão prévia. **Conclusão:** Este trabalho contribui para a democratização do acesso aos dados do DATASUS, flexibilizando as tecnologias empregadas na análise de dados para além das ferramentas disponíveis até o momento.

Palavras-chave: Sistema Único de Saúde (SUS), Análise de Dados, Informática Médica.

Abstract*: This article presents a new package for the R language with the objective of reading the DBC file format used by DATASUS. Methods: an analysis of the DBC file format has been made and a software was written to convert it to the DBF format. The resulting code was encapsulated in a package structure for deployment. Results: The read.dbc package has simplified the data import process from DATASUS in the R language, allowing the direct read of the DBC file without previous conversion. Conclusion: This work contributes to the democratization of the access to DATASUS data, enabling new technologies to be employed in data analysis beyond the tools available at the present date.*

Keywords*: Unified Health System, Data Analysis, Medical Informatics.*

Introdução

A informatização do Sistema Único de Saúde (SUS) é essencial para a descentralização das atividades de saúde, além da viabilização da participação popular e do controle social sobre a utilização dos recursos disponíveis[1]. A disponibilização de dados referentes aos processos de funcionamento do SUS tem a potencialidade de promover a auto regulação, um atributo fundamental conforme especificam as diretrizes do SUS[2].

O Departamento de Informática do Sistema Único de Saúde (DATASUS) é o agente responsável pela manutenção das bases de dados nacionais do SUS e da Agência Nacional de Saúde Suplementar (ANS), disponibilizando informações que podem servir para subsidiar análises objetivas da situação sanitária, tomadas de decisão baseadas em evidências e elaboração de programas de ações de saúde[3].

Para atender as necessidades de obtenção de informações de modo ágil e rápido, por parte dos atores sociais vinculados ao SUS, o DATASUS desenvolveu um *software* próprio de tabulação, hoje em sua terceira versão, denominado Tabwin[4]. Esse tabulador trabalha com arquivos no formato DBF, introduzido pelo software de gerenciamento de banco de dados dBase.

Devido ao grande volume de dados produzidos pelos mais diversos sistemas de saúde pública, o DATASUS desenvolveu uma extensão do formato de arquivo DBF, no qual os dados são armazenados

de modo compactado, o arquivo DBC[5]. Por ser um formato próprio do DATASUS, os arquivos DBC apenas conseguem ser lidos pelas ferramentas disponibilizadas pelo Ministério da Saúde, dentre elas o Tabwin e uma ferramenta de linha de comando chamada DBC2DBF[6].

Ambas as ferramentas disponibilizadas no sitio do DATASUS são ferramentas livres, porém de código fechado e disponíveis exclusivamente para a plataforma Windows, o que impõe severas limitações à implementação de melhorias e customizações no *software* de tabulação, ao desenvolvimento de processos automatizados (*batches*) e na escalabilidade das plataformas de *software*, além de vincular o desenvolvimento de qualquer produto derivado a um único sistema operacional.

O sitio do DATASUS documenta a possibilidade de executar o código desenvolvido para Windows em sistemas operacionais Linux através da biblioteca Wine[7], porém tal abordagem adiciona uma camada indesejável de complexidade à topologia da solução, além de não solucionar todas as limitações expostas acima.

Diante das limitações existentes nas ferramentas atualmente disponibilizadas pelo DATASUS, foi desenvolvido um pacote para leitura de arquivos DBC como uma extensão da linguagem R. O objetivo deste trabalho é demonstrar as capacidades do pacote desenvolvido, o *read.dbc,* quanto à leitura de arquivos DBC sem a necessidade de conversão prévia entre formatos DBC e DBF.

Métodos

A linguagem R possui suporte para leitura e escrita no formato DBF através do pacote *foreign*[8], mantido pelo R Core Team, o grupo de desenvolvedores responsável pela manutenção e evolução da linguagem. O pacote *foreign* disponibiliza este suporte através das funções *read.dbf* e *write.dbf*, capazes de ler e gravar arquivos DBF, respectivamente. Para adicionar suporte ao arquivo DBC, é necessário implementar a descompactação do arquivo DBC para DBF em tempo de execução.

O grande desafio da leitura de arquivos DBC é o fato de que este é um formato desenvolvido internamente pelo DATASUS com pouca informação disponível ao seu respeito. Existem no mercado, alguns tipos de arquivo que compartilham a extensão DBC, como por exemplo, o *database file* do Microsoft FoxPro, porém estes arquivos não são compatíveis com o formato do DATASUS. No entanto, o estudo de um projeto *open source* chamado *blast-dbf* [9], um programa de linha de comando escrito na linguagem C que tem a funcionalidade de converter (descompactar) arquivos DBC para DBF, viabilizou o entendimento do mecanismo de compressão empregado nos arquivos DBC.

Uma análise do arquivo DBC e do código fonte do programa *blast-dbf*, revelaram que o cabeçalho do arquivo DBF está praticamente intacto, segundo a especificação do formato dBase Version 7, mas com algumas modificações: após o campo entitulado *field descriptor terminator* (assinalado pelo código 0x0D), ao invés da estrutura *field properties structure* temos uma sequência de 4 *bytes* indicando o CRC32 do arquivo, seguido pelos dados compactados com o algoritmo *implode* da empresa PKware.

A linguagem R oferece suporte nativo para a importação de funções e bibliotecas escritas em outras linguagens, como o C, C++ e FORTRAN. Aproveitando-se desta capacidade, o código de descompressão de arquivos DBC escrito em C foi transformado em uma *shared library* para permitir o seu acesso pelo interpretador da linguagem R.

Em seguida, foi concebida uma função em R, chamada *dbc2dbf*, servindo de *wrapper* para a função de descompactação exposta pela *shared library.* Através desta função é possível gerar em tempo execução um arquivo DBF a partir de qualquer arquivo DBC que pode então ser lido pela função *read.dbf* do pacote *foreign*.

Para encapsular a complexidade de trabalhar com arquivos DBF temporários outra função foi concebida, denominada *read.dbc*. Esta função é a responsável por orquestrar o acesso aos arquivos e invocar a função *read.dbf,* no arquivo temporário, retornando um *data.frame* com os dados do arquivo

para o chamador. Desta forma, é possível obter uma interface transparente para o usuário abstraindo as complexidades da implementação.

A última etapa do projeto foi organizar os códigos fontes de acordo com a estrutura de pacotes do R[10]. Este processo foi facilitado pelo uso dos pacotes *roxygen*[11] e *devtools*[12]. Adicionalmente, o código-fonte e a documentação foram adequados aos padrões estabelecidos pela *The Comprehensive R Archive Network* (CRAN), a biblioteca central da linguagem R, o que permitiu o seu aceite para publicação na mesma.

Publicar um pacote na CRAN traz inúmeros benefícios. Dentre eles, pode-se citar um ganho de qualidade, em virtude dos processos rigorosos de revisão; garantia de funcionamento em múltiplas plataformas, incluindo Linux, Windows e Solaris; e também a facilidade para o usuário final de instalar o pacote por meio da função *install.packages*, da mesma forma que todos os principais pacotes do R são instalados.

Resultados e Discussão

Para demonstrar as potencialidades do pacote *read.dbc* desenvolvido através desse trabalho, foram realizados testes de funcionamento e aplicabilidade. O primeiro passo é a instalação do pacote, que pode ser realizado conforme mostra a Figura 1.

```
> install.packages("read.dbc")
Installing package into '/home/dani/R/x86_64-pc-linux-gnu-library/3.3'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/src/contrib/read.dbc_1.0.4.tar.gz'
Content type 'unknown' length 20918 bytes (20 KB)
==================================================
downloaded 20 KB

* installing *source* package 'read.dbc' ...
** package 'read.dbc' successfully unpacked and MD5 sums checked
** libs
gcc -std=gnu99 -I/usr/share/R/include -DNDEBUG      -fpic  -g -O2 -fstack
or=format-security -D_FORTIFY_SOURCE=2 -g  -c blast.c -o blast.o
gcc -std=gnu99 -I/usr/share/R/include -DNDEBUG      -fpic  -g -O2 -fstack
or=format-security -D_FORTIFY_SOURCE=2 -g  -c dbc2dbf.c -o dbc2dbf.o
gcc -std=gnu99 -shared -L/usr/lib/R/lib -Wl,-Bsymbolic-functions -Wl,-z,r
lib -lR
installing to /home/dani/R/x86_64-pc-linux-gnu-library/3.3/read.dbc/libs
** R
** inst
** preparing package for lazy loading
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (read.dbc)

The downloaded source packages are in
        '/tmp/RtmpqEOUDw/downloaded_packages'
>
```

Figura 1: Captura de tela demonstrando a execução da função *install.packages* para a instalação do pacote *read.dbc* no R.

Uma vez que o pacote esteja instalado, com poucas linhas de código é possível carregar qualquer arquivo do DATASUS ou da ANS que seja publicado no formato DBC. A Figura 2 contém a listagem

de um código fonte em R demonstrando o uso da ferramenta para carregar o arquivo de Declarações de Óbitos Fetais (DOFET) do ano de 2013.

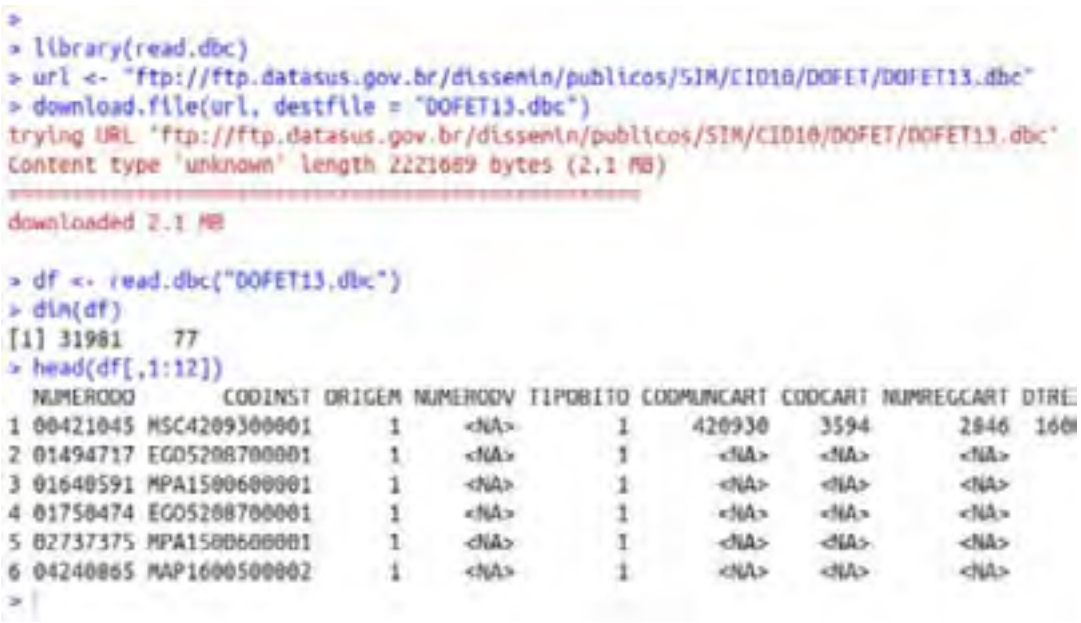

Figura 2: Captura de tela demonstrando a carga do arquivo de Declarações de **Ó**bitos Fetais (DOFET) do ano de 2013 utilizando a função *read.dbc*.

O fato do pacote *read.dbc* ter auferido sucesso junto aos testes da CRAN garante o seu funcionamento em qualquer ambiente R a partir da versão 3.3.0 (versão na qual o pacote foi homologado). Este fator traz inúmeros benefícios tanto para os meios de pesquisa acadêmica quanto para a gestão da saúde como um todo, pois abre a possibilidade para executar análises utilizando qualquer ferramenta que ofereça suporte à linguagem R, e não apenas o Tabwin.

Uma das limitações mais marcantes da linguagem R tradicional é o fato da ferramenta não escalar além do computador local, ou seja, o limite de tamanho de um conjunto de dados a ser analisado é determinado pela quantidade de memória e poder de processamento disponível na máquina que fará a análise. Este problema pode ser resolvido através do uso de tecnologias mais completas de processamento de dados, como por exemplo, o uso de bancos de dados relacionais que possuem suporte à linguagem R ou mesmo de distribuições comerciais da linguagem.

Conclusão

As bases de dados do DATASUS possuem, atualmente, cerca de 31 terabytes de informações a respeito de estabelecimentos de saúde, produção ambulatorial e hospitalar, mortalidade e vigilância epidemiológica, entre outros[1]. A exploração das informações disponibilizadas pelos diferentes sistemas operantes junto ao DATASUS enfrentava como obstáculo a limitação técnica das ferramentas até então disponíveis.

O pacote *read.dbc* contribui para a democratização do acesso aos dados do DATASUS, uma vez que permite flexibilizar as tecnologias empregadas na análise de dados através de uma interface simples e intuitiva. O fato de ser um pacote publicado na CRAN, fornece uma garantia adicional de confiabilidade, estabilidade e compatibilidade com múltiplas plataformas, abrindo a possibilidade para a exploração de novas tecnologias que expandem o universo da linguagem R. Além disso, a partir do seu desenvolvimento torna-se possível integrar os dados disponibilizados pelo DATASUS

a linguagens amplamente utilizadas para fins de mineração de dados. A disseminação do presente pacote favorece a construção de estudos em saúde contribuindo para o fortalecimento dos pilares do SUS.

Finalmente, por se tratar de um *software* livre e de código aberto, a comunidade científica pode, de maneira autônoma, propor melhorias e correções de erros, dando maior poder aos seus usuários. Cabe ainda ressaltar que este projeto só foi possível graças à colaboração entre desenvolvedores de código aberto, demonstrando a importância de tal abordagem.

Agradecimentos

À Thiago Augusto Hernandes Rocha pelo incentivo e apoio na revisão do trabalho e à Pablo Marcondes Fonseca por sua contribuição em código aberto e disponibilidade para discutir a tecnologia.

Referências

[1] Ministério da Saúde. DATASUS Relatório Executivo da Gestão 2011 - 2014. Brasília; 2015. Disponível em: http://www2.datasus.gov.br/DATASUS/download/201501. Acesso em 30 de jun. 2016.

[2] ROLIM LB, CRUZ R de SBLC, SAMPAIO KJA de J. Participação popular e o controle social como diretriz do SUS: uma revisão narrativa. Saúde em Debate. 2013;37(96):139–47.

[3] Informações de Saúde [Internet]. Disponível em: http://datasus.saude.gov.br/informacoes-de-saude. Acesso em 30 de jun. 2016.

[4] Brasil, Ministério da Saúde, Secretaria Executiva, Departamento de Informática do SUS. DATASUS Trajetória 1991-2002. Brasília: Ministério da Saúde; 2002. 62 p. Disponível em: http://bvsms.saude.gov.br/bvs/publicacoes/trajetoria_datasus.pdf. Acesso em 30 de jun. 2016.

[5] Brasil, Ministério da Saúde, Secretaria Executiva, Departamento de Informática do SUS. Tab para Windows Versão 2 [Internet]. [citado 2016 Jun 30]. p. 139. Disponível em: ftp://ftp.datasus.gov.br/tabwin/tabwin/TabWin.pdf

[6] TABWIN [Internet]. [citado 2016 Jun 30]. Disponível em: http://www2.datasus.gov.br/DATASUS/index.php?area=060805&item=6

[7] Novidades no TabWin 3.0 a 3.6 [Internet]. [citado 2016 Jun 30]. Disponível em: http://www2.datasus.gov.br/DATASUS/tabwin/DocTabWin.htm

[8] R Core Team. foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ... [Internet]. 2015. Disponível em: https://cran.r-project.org/package=foreign. Acesso em 30 de jun. 2016.

[9] Fonseca PM. Code to convert from dbc to dbf [Internet]. Disponível em: https://github.com/eaglebh/blast-dbf. Acesso em 30 de jun. 2016.

[10] R Core Team. Writing R Extensions [Internet]. Vol. 0. 2016. p. 176. Disponível em: https://cran.r-project.org/doc/manuals/r-release/R-exts.pdf. Acesso em 30 de jun. 2016.

[11] Wickham H, Danenberg P, Eugster M. roxygen2: In-Source Documentation for R [Internet]. 2016. Disponível em: https://github.com/klutometis/roxygen. Acesso em 30 de jun. 2016.

[12] Wickham H, Chang W. devtools: Tools to Make Developing R Packages Easier [Internet]. 2016. Disponível em: https://cran.r-project.org/package=devtools. Acesso em 30 de jun. 2016.

Contato

e-mail: daniela.petruzalek@gmail.com
Telefone: (42) 99116-1104