

# PaNasc: Uma solução computacional para extração e análise de dados do SINASC

Luan Augusto Sousa Nascimento<sup>1</sup>, Francisco Gustavo Braga Batista<sup>1</sup>,  
Abel Brasil Ramos Da Silva<sup>2</sup>

<sup>1</sup>Dep. de Estatística e Matemática Aplicada – Universidade Federal do Ceará (UFC)  
Campus do Pici - Centro de Ciências - Bloco 910 - 60440-900 - Fortaleza-CE

<sup>2</sup>Unidade de Sistemas da Informação e Inteligência de Dados – Complexo Hospitalar da UFC  
R. Pastor Samuel Munguba, 1290 - Rodolfo Teófilo - 60430-372 - Fortaleza - CE

{luan.augusto,gustavobraga}@alu.ufc.br, abel.silva@ebserh.gov.br

**Abstract.** *The availability of data on live births is essential for monitoring public health and making data-driven decisions. However, the tools provided by DATASUS, such as Tabwin and Tabnet, have limitations and are difficult to access for users without experience in database systems and management. To address this issue, PaNasc was developed. PaNasc is a project that facilitates access to and analysis of SINASC data by creating a process written in the R programming language that performs ETL on these data and provides an intelligence dashboard with fast and interactive analyses of these indicators.*

**Resumo.** *A disponibilidade de dados sobre os nascidos vivos é fundamental para monitoramento da situação sanitária e tomada de decisão baseada em evidências. No entanto, as ferramentas disponibilizadas pelo DATASUS, como TabWin e TabNet, apresentam limitações e são de difícil acesso para usuários sem experiência em sistemas e gerenciamento de banco de dados. Para suprir essa lacuna foi desenvolvido o PaNasc. O PaNasc é um projeto que facilita o acesso e análise dos dados do SINASC através da criação de um processo escrito em linguagem R que executa o ETL desses dados e disponibiliza um painel de inteligência com análises rápidas e interativas desses indicadores.*

## 1. Introdução

O Departamento de Informática do Sistema Único de Saúde (DATASUS), criado em 1991, é responsável por manter sistemas de informações que cobrem diversos aspectos relacionados à saúde populacional. Dentre os sistemas mantidos, estão os de natureza epidemiológica, como o Sistema de Informações de Nascidos Vivos (SINASC) e o Sistema de Informações de Mortalidade (SIM), assistencial, como o Sistema de Informações Hospitalares do SUS (SIH/SUS), dentre outros [da Saúde 2008].

Atualmente, os dados do DATASUS podem ser acessados por meio de duas interfaces: TabWin e TabNet. No entanto, essas ferramentas apresentam algumas limitações. O TabWin, por exemplo, exporta os dados no formato DBC, que exige um software específico para leitura. O TabNet é uma ferramenta de tabulação que permite a consulta online de dados e a geração de planilhas. Porém, esses sistemas não possibilitam a apresentação dos dados de forma gráfica, assim como uma análise mais aprofundada dos dados através da interatividade dos números.

Para superar essas dificuldades, foram desenvolvidos três produtos: um pacote na linguagem R, seguindo as recomendações de desenvolvimento de pacotes [Leisch 2008], composto por três funções que automatizam a extração e o pré-processamento dos dados do SINASC, um painel de indicadores no Power BI, fornecendo uma visão interativa sobre os nascimentos, além do desenvolvimento de um site capaz de realizar o download dos microdados do SINASC e disponibilizá-los para os usuários em formato de planilha Excel.

Na literatura, existem outros pacotes R utilizados para coleta e tratamento de dados da saúde, que foram utilizados como referência para a elaboração do presente artigo, como o *microdatasus* [Saldanha et al. 2019], que realiza o download dos dados brutos de alguns sistemas de informações de saúde brasileiros, e o *csapAIH* [Nedel 2017], que automatiza a classificação e a descrição das Condições Sensíveis à Atenção Primária (CSAP). Também foi utilizado como referência o Painel de Nascimentos da cidade do Rio de Janeiro [de Inteligência Epidemiológica sd].

Em relação a esses trabalhos, o trabalho apresentado neste artigo se diferencia por realizar todo o fluxo ETL diretamente no R, além de disponibilizar de uma plataforma web e um painel interativos voltados não só à visualização, mas também à customização e exportação dos dados com rótulos legíveis. Ademais, o escopo do projeto é nacional e permite a atualização contínua com base nos dados mais recentes do SINASC, preenchendo lacunas observadas nas soluções existentes.

## 2. Método proposto

Trata-se de uma pesquisa de caráter exploratório, realizada em três etapas: desenvolvimento de um pacote computacional capaz de realizar o download e tratamento dos arquivos referentes ao SINASC, a fim de facilitar o acesso aos dados, criação de um painel de dados em Power BI com os principais indicadores sobre os nascidos vivos, selecionados após pesquisas em referências bibliográficas e desenvolvimento de um site utilizando o pacote do R *Shiny*. O projeto foi realizado de acordo com o seguinte fluxograma, mostrado na Figura 1.

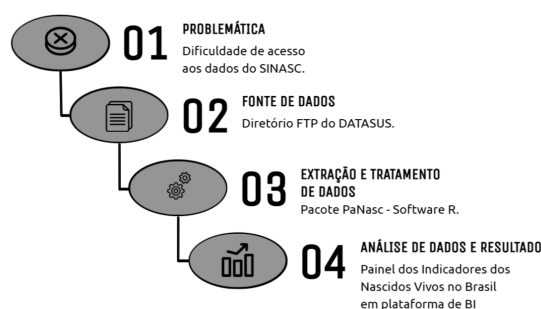


Figura 1. Fluxograma do projeto

### 2.1. Instalação do software R

Para realizar a instalação do R, deve-se acessar sua página eletrônica (<https://www.r-project.org/>) e seguir os passos de instalação. Após isso, deve-se instalar o pacote dev-

tools, utilizado para realizar a instalação de pacotes hospedados no GitHub, e carregá-lo, antes de prosseguir para a próxima etapa.

```
install.package("devtools")
```

## 2.2. Instalação do pacote PaNasc no R

O repositório do pacote está disponível publicamente para download no GitHub (<https://github.com/abelbrasil/PaNasc>). No R, digita-se o seguinte código para realizar a instalação e carregamento do pacote:

```
devtools::install_github("abelbrasil/PaNasc")  
library("PaNasc")
```

O processo de instalação do pacote deve ser realizado apenas uma vez, porém, o carregamento do pacote, realizado através da função `library`, deve ser realizado após cada inicialização do software.

### 2.2.1. Função *download.sinasc*

A função `download.sinasc` realiza a busca e o download dos dados brutos do SINASC, presente no Diretório FTP do DATASUS, além do pré-tratamento dos dados. A importação dos dados é realizada utilizando-se o pacote *read.dbc* [Petruszalek 2016].

A função possui quatro argumentos. Sendo eles:

- *inicio*: O ano em que será iniciada a coleta dos dados.
- *fim*: O último ano a ser coletado os dados.
- *UF*: Os estados do qual se deseja obter os dados. Por padrão, é realizado o download dos dados de todos os estados dos respectivos anos selecionados. Vale ressaltar que os estados o qual deseja-se extrair deve ser escritos em um vetor, representados por suas siglas, ex: `c("CE","PE","PB")`.
- *cod\_estab*: Código CNES do estabelecimento de saúde a qual se deseja extrair os dados.

### 2.2.2. Função *process.sinasc*

Após aplicar a função `download.sinasc` para realizar o download dos dados do SINASC, os dados são carregados para o software R em seu formato bruto, sendo necessário realizar a decodificação das colunas que possuem variáveis categóricas, substituindo os códigos pelos seus respectivos rótulos.

### 2.2.3. Função *panel.sinasc*

Esta função é utilizada para preparar os dados para serem carregados para o painel do Power BI. Ela é responsável por selecionar e tratar as colunas necessárias para o painel.

## 2.3. Painel PaNasc

Além do pacote computacional no R, existe a possibilidade de obter-se um painel de dados em Power BI com o quantitativo de nascidos vivos agregados pelos indicadores selecionados. O painel encontra-se disponível para download no repositório público do GitHub `abelbrasil/PaNasc`, na pasta `inst/extdata`, intitulado de `Painel Nascidos Vivos.pbix`.

## 2.4. Aplicação Web

Caso o usuário não deseje ou não possua habilidades computacionais para utilizar o pacote R ou o painel em Power BI, ele pode optar por baixar os microdados do SINASC através do site desenvolvido utilizando o Shiny.

O Shiny é um framework para o desenvolvimento de aplicações web em R e Python. Ele permite ao desenvolvedor criar uma interface interativa onde o usuário insere os dados, que são processados no servidor para gerar e exibir os resultados [Wickham 2021].

No site PaNasc, o usuário deve informar o período em que deseja realizar o download dos dados, assim como os estados e, se for o caso, os estabelecimentos específicos que deseja-se realizar o download dos microdados.

## 3. Resultados

### 3.1. Exemplo Hipotético

Nesta subseção, apresenta-se uma situação hipotética para o entendimento do fluxo do processo. Suponha que se deseja realizar o download e tratamento dos dados do SINASC referente ao estado do Ceará nos anos de 2022 e 2023.

```
> library(PaNasc)
> data <- download.sinasc(inicio = 2022, fim = 2023,
                          UF = "CE")
> data_processed <- process.sinasc(data)
> fNasc_Vivos <- panel.sinasc(data_processed)
> save("fNasc\_Vivos",
       file = "caminho_do_arquivo/dataset.Rdata")
```

### 3.2. Painel interativo

Para facilitar a análise dos indicadores de nascidos vivos, foi desenvolvido um painel interativo no Microsoft Power BI. Esse painel utiliza os dados gerados pelo pacote em R, a partir do arquivo .RData, para realizar as análises de forma dinâmica. Ele é estruturado em quatro seções, organizadas conforme as agregações pertinentes: Visão Geral, Recém-Nascido, Gestação/Parto e Dados da Mãe.

#### 3.2.1. Visão Geral

Nesta seção são apresentados os primeiros indicadores dos Nascidos Vivos, sendo eles: Número de Nascidos Vivos Global, Número de Nascidos Vivos por ano, mês e estabelecimento de saúde. É possível filtrar os dados através do painel de filtros, podendo os dados serem filtrados por: Ano, Mês-Ano, Região de Nascimento, UF/Município de Nascimento, Vínculo (EBSERH ou outros), Estabelecimento, UF/Município de Residência da Mãe, Ocupação da Mãe, Sexo do bebê e Local de Nascimento.

#### 3.2.2. Recém-Nascido

Na segunda aba é possível observar outros indicadores sobre os Recém-Nascidos, sendo eles: Média do peso ao nascer, Nascidos Vivos por Sexo, Raça/Cor, Faixa de APGAR5,

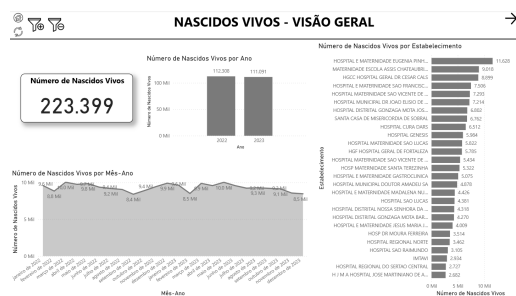
## Faixa de Peso e Nascidos Vivos com Anomalia Congênita.

### 3.2.3. Gestação e Parto

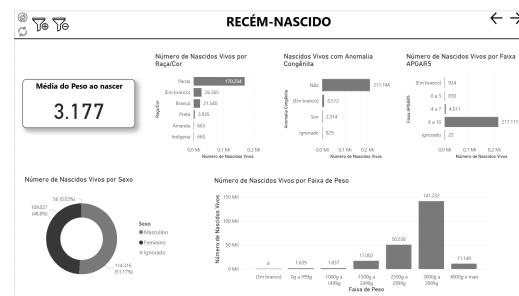
Além de analisar os indicadores sobre os Recém-Nascidos, é importante realizar a análise dos indicadores sobre a gestação e parto. Os indicadores selecionados para o painel foram: Número de Nascidos Vivos por N° de consultas pré-natal, tipo de gravidez, quantidade de semanas de gestação e tipo de parto.

### 3.2.4. Dados da mãe

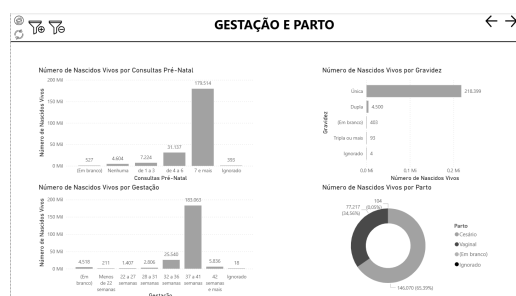
Por fim, o usuário tem a opção de analisar os indicadores da mãe do recém-nascido, em que os indicadores gerados foram: Números de Nascidos Vivos por escolaridade da mãe, raça/cor da mãe, faixa etária da mãe, estado civil da mãe e ocupação da mãe.



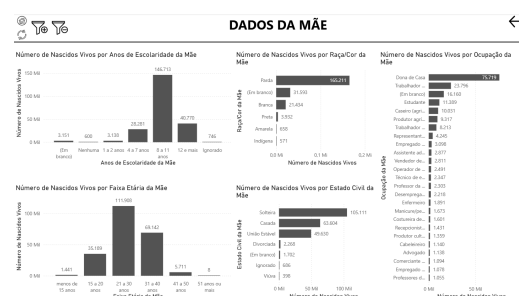
(a) Visão Geral



(b) Recém-Nascido



(c) Gestação e Parto



(d) Dados da mãe

Figura 2. Seções do Painel PaNasc

### 3.3. Site

O site PaNasc oferece uma interface intuitiva, apresentando informações sobre a aplicação e instruções para que o usuário utilize a ferramenta de forma correta. É possível acessá-lo em: <https://w3sy5o-luan0augusto-sousa0nascimento.shinyapps.io/PaNasc-web/>.

**Figura 3. Site PaNasc**

## 4. Conclusão

O presente trabalho teve como objetivo facilitar o acesso e a análise dos dados do SINASC, por meio de um pacote e um painel interativo desenvolvidos em R e Power BI. A solução torna o processo de tratamento dos dados mais ágil, dispensa softwares adicionais e oferece visualizações e filtros prontos para análise exploratória. Também foi disponibilizado um site para download dos dados em Excel e o código-fonte do projeto no GitHub (<https://github.com/abelbrasil/PaNasc/>), permitindo personalizações e contribuições da comunidade.

Como trabalhos futuros, pretende-se realizar uma avaliação sistemática da solução desenvolvida para acompanhar o desempenho e impacto da solução implementada. Além disso, está em desenvolvimento uma solução computacional com objetivo similar à apresentada, porém com dados provenientes do Sistema de Informação de Mortalidade - SIM, com o objetivo de estender a aplicabilidade da proposta para outro importante banco de dados da saúde pública brasileira.

## Referências

- da Saúde, O. P.-A. (2008). Indicadores básicos para a saúde no Brasil: conceitos e aplicações.
- de Inteligência Epidemiológica, C. (s.d.). Painéis Epidemiológicos Nascimentos - EpiRio.
- Leisch, F. (2008). Creating r packages: A tutorial.
- Nedel, F. B. (2017). csapAIH: uma função para a classificação das condições sensíveis à atenção primária no programa estatístico R. *Epidemiologia e Serviços de Saúde*, 26:199–209.
- Petruszalek, D. (2016). READ. DBC: um pacote para importação de dados do datasus na linguagem R. In *Anais do XV Congresso Brasileiro de Informática em Saúde [Internet]*, pages 27–30.
- Saldanha, R. d. F., Bastos, R. R., and Barcellos, C. (2019). Microdatasus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). *Cadernos de Saúde Pública*, 35:e00032419.
- Wickham, H. (2021). *Mastering shiny*. "O'Reilly Media, Inc."