Qualidade dos Dados no Programa Dinheiro Direto na Escola (PDDE): Superando o Desafio do Garbage In, Garbage Out (GIGO)

Guilherme Ataíde Dias
Doutor, Universidade de São Paulo (USP), São Paulo, SP, Brasil.
Professor Titular, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
http://lattes.cnpq.br/9553707435669429

Wagner Junqueira de Araújo
Doutor, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
Professor Associado III, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
http://lattes.cnpq.br/6762905361803183

Adriana Valéria Santos Diniz
Doutora, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
Professora Associado II, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
http://lattes.cnpq.br/7196551398849603

Flavio Ribeiro Córdula
Doutor, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
Analista de TI, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
http://lattes.cnpq.br/7466802181232338

Paulo Roberto Santos Costa
Mestre, Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
Cargo ocupado, Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB), João Pessoa, PB, Brasil.
http://lattes.cnpq.br/7257600761427884

Submetido em: 16/10/2023. Aprovado em: 30/04/2024. Publicado em: 18/07/2024.

RESUMO

Introdução: A investigação apresenta e discute a relevância da qualidade dos dados na Ciência de Dados, trazendo o conceito de "*Garbage In, Garbage Out*" (GIGO) ao Programa Dinheiro Direto na Escola (PDDE). **Objetivos:** Teve como objetivo descrever o processo de extração, transformação e carga de dados para a geração de painéis de informação pelo Centro Colaborador de Apoio ao Monitoramento e à Gestão de Programas Educacionais (CECAMPE/NE). **Metodologia:** A metodologia empregada configura-se como quantitativa e utilizou ferramentas típicas da Ciência dos Dados. Os dados relacionados com o PDDE foram coletados de sistemas transacionais e de enquetes com as populações envolvidas. **Resultados:** Os resultados mostraram uma redução satisfatória do GIGO, embora tenham sido identificados desafios como divergências de atualização de dados, falta de documentação adequada das tabelas de dados e problemas com campos de digitação livre nos formulários das enquetes. **Conclusão:** A análise reitera que o conceito de GIGO é um desafio significativo para a utilização eficaz dos dados do PDDE e destaca a necessidade de elaboração de

melhores práticas de gestão de dados no contexto do programa que todos os profissionais que usam dados como a matéria-prima para a realização de suas atividades profissionais devem estar conscientes desses desafios e trabalhar em prol de soluções eficazes.

Palavras-chave: Programa Dinheiro Direto na Escola; *garbage in garbage out*; qualidade de dados; painéis de informação; centro colaborador de apoio ao monitoramento e à gestão de programas educacionais.

> "*Garbage in, garbage out*" provavelmente é a primeira lição que os aspirantes a cientistas de dados aprendem sobre suas futuras empreitadas analíticas. " (Ozminkowski, 2021, p. 1, tradução nossa)[1].

1 Original: "*Garbage in, garbage out* is probably the first lesson budding data scientists learn about their forthcoming analytic endeavors"(Ozminkowski, 2021, p. 1).

INTRODUÇÃO

Esse trabalho aborda sobre uma questão importante que muitos cientistas de dados podem enfrentar nas etapas iniciais de suas investigações, a garantia da qualidade dos dados obtidos. Qualidade esta que poderá ter repercussões positivas ou negativas nas análises a serem realizadas, bem como em todos os entregáveis resultantes dos conjuntos de dados utilizados. A questão do impacto da qualidade dos dados em processos computacionais e de informação são bastante reconhecidos desde meados do século XX, visto que, a partir do desenvolvimento e popularização dos primeiros computadores digitais com uma maior capacidade de processamento (*mainframes*) a quantidade de geração e processamento de dados vem crescendo, explodindo literalmente a partir dos anos 1990 com a massificação dos microcomputadores e uma pletora de outros dispositivos das Tecnologias Digitais da Informação e Comunicação (TDICs).

Associado ao processo de tratamento de dados, mencionamos o conceito de *Garbage In, Garbage Out* (GIGO). Literalmente, essa expressão idiomática em Língua Inglesa significaria em Língua Portuguesa algo como "Lixo entra, Lixo sai". Stenson (2016, *online*, tradução nossa[2]) explica que: "A ideia por trás de GIGO remonta ao próprio amanhecer da computação, no início do século XIX, quando Charles Babbage apresentou o projeto de sua 'máquina diferencial' ao Parlamento da Inglaterra". O mesmo autor explica ainda que possivelmente a expressão foi criada por um funcionário da IBM no ano de 1958 ou 1959 durante um treinamento no computador 305 RAMAC para clientes da empresa em Nova Iorque.

O conceito de GIGO é mais impactante hoje do que em qualquer época, devido à nossa dependência que temos dos dados em todas as instâncias das atividades humanas. Mesmo com os avanços nos produtos de *software* que possibilitam a consistência nos processos de tratamento de dados, os desafios persistem. O impacto da possibilidade de entrada de "lixo" afeta todos os tipos de sistema, desde os sistemas tradicionais de folha de pagamento até os mais sofisticados sistemas de Inteligência Artificial (AI), baseados em modelos de aprendizado profundo, que são absolutamente dependentes de vastos volumes de dados.

Nesse contexto, a pesquisa tem como objetivo descrever o processo de extração, transformação e carga de dados para a geração de painéis de informação pelo Grupo de Monitoramento do Centro Colaborador de Apoio ao Monitoramento e à Gestão de Programas Educacionais (CECAMPE/NE). Os dados obtidos são relativos ao Programa Dinheiro Direto na Escola (PPDE), uma iniciativa do Fundo Nacional de Desenvolvimento da Educação (FNDE). Diniz *et al.,* (2022, p. 7-8) explicam que:

> O PDDE é um programa federal de transferência suplementar de recursos direto às instituições de ensino da educação básica pública. Ao lado de outros programas federais, a exemplo do Programa Nacional de Transporte Escolar (PNATE), Programa Caminho da Escola, Programa Nacional de Alimentação Escolar (PNAE), Programa

2 Original: "The idea behind GIGO actually dates to the very dawn of computation, the early 19th century, when Charles Babbage presented the design for his "difference engine" to England's Parliament" (Stenson, 2016, online).

> Nacional de Livro Didático (PNLD), promove a melhoria física e pedagógica das instituições de ensino, contribuindo com o fortalecimento da gestão escolar, uma vez que abre espaço para a comunidade participar da tomada de decisão, tanto no que se refere aos aspectos administrativo-financeiros e didático-pedagógicos. (Diniz *et al*., 2022, p. 7-8).

Córdula *et al*., (2022, p. 68) explicam o que são os Centros Colaboradores de Apoio ao Monitoramento e à Gestão de Programas Educacionais:

> Os Centros Colaboradores de Apoio ao Monitoramento e à Gestão de Programas Educacionais (CECAMPEs) são universidades vinculadas do Fundo Nacional de Desenvolvimento da Educação (FNDE) que realizam atividades de assistência técnica e monitoramento a estados, municípios e escolas, dando suporte para que estes possam aprimorar a execução e o desempenho do Programa Dinheiro Direto na Escola (PDDE) e suas Ações Integradas, do Programa Caminho da Escola e do Programa Nacional de Apoio ao Transporte do Escolar (PNATE).

No Nordeste, o CECAMPE está representado pela Universidade Federal da Paraíba, doravante referido como CECAMPE/NE. Com relação a esse CECAMPE, Diniz *et al*., (2022, p. 8) esclarecem que:

> O CECAMPE da Região Nordeste é formado por 9 Estados, 1.794 Municípios, organizado em 30 Polos, e 51.169 Entidades do PDDE. Adotamos como ponto de partida para a organização do Projeto, o estudo e análise do Índice de Desempenho da Gestão Descentralizada do PDDE (IdeGES) na região, destacando a preponderância de escolas situadas no IdeGES abaixo de 06.

Os dados gerados no âmbito do PDDE após os devidos tratamentos (reduzindo a incidência de GIGO) foram utilizados para a elaboração de painéis de informações pela equipe de monitoramento do CECAMPE/NE. Esses painéis fornecem subsídios necessários para que o FNDE, o CECAMPE/NE e toda a sociedade possam acompanhar os mais diversos indicadores acerca do PDDE. Destaca-se a possibilidade das equipes de campo do CECAMPE/NE realizarem intervenções quando em atividades nas escolas, mediante a consulta aos painéis, seja em uma escola da capital ou em uma escola situada em uma área rural no interior do Nordeste[3].

METODOLOGIA

Esta pesquisa configura-se como bibliográfica, descritiva, de abordagem mista e aplicada (Richardson, 2017). A pesquisa bibliográfica foi realizada ao longo do ano de 2021 e segundo semestre de 2022, período em que foi realizada a investigação. Foram consultados o Portal de Periódicos da Capes, o *Google Scholar*. Os dados obtidos para a elaboração

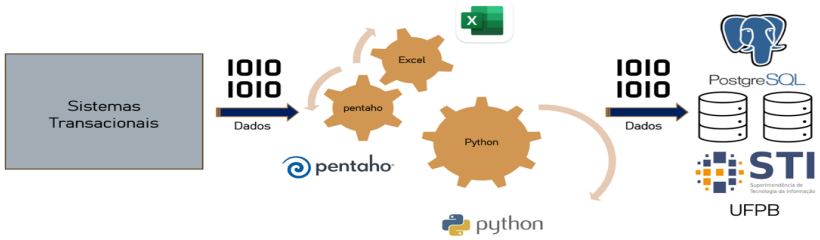
3 Os painéis disponibilizados pelo CECAMPE/NE podem ser acessados através do seguinte URL: https://www.cecampe.ufpb.br/paineisdeinformacoes.

dos painéis vieram de quatro fontes: 1) do FNDE, que aconteceu, primeiramente, a partir da Plataforma Ágil de Serviços de Dados do Banco Central do Brasil, conhecida como Olinda, e, posteriormente, por meio da plataforma de aplicação *Web* da Microsoft conhecida com *Sharepoint*; 2. Dados obtidos por meio de de formulários gerados a partir de enquetes realizadas com o *Google Forms* com as populações atendidas pelo PDDE; 3. Microdados do Censo Escolar da educação básica, extraídos do *site* oficial do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP); e 4. da divisão territorial brasileira, que teve como fonte o *site* do Instituto Brasileiro de Geografia e Estatística (IBGE).

O período de coleta de dados compreendeu os anos de 2021 e 2023.

Quanto aos dados obtidos através da Plataforma Olinda e *Sharepoint*, eles foram submetidos às atividades de ETL (***Extract***, ***Transform*** and ***Load***) (Vide **Figura 1**). Esse processo envolveu a extração de dados com *scripts* específicos da plataforma e sua subsequente transformação na estação de trabalho do cientista de dados, foram utilizados os seguintes produtos de *software*: *pentaho*, *Python, R* e *Microsoft Excel*. O formato final dos dados, após as devidas transformações, foi o *Comma-Separated Values* (CSV). Esses dados no formato CSV foram carregados em uma base de dados do gerenciador de banco de dados *PostgreSQL* da Superintendência de Tecnologia da Informação da UFPB (STI). O acesso a esse banco de dados é concedido apenas aos usuários devidos credenciados através do IP 150.165.130.10, sendo que, atualmente, essa comunidade é constituída majoritariamente pelos desenvolvedores de painéis.

Figura 1 – Captura de dados dos sistemas transacionais



Fonte: Desenvolvimento dos autores (2023)

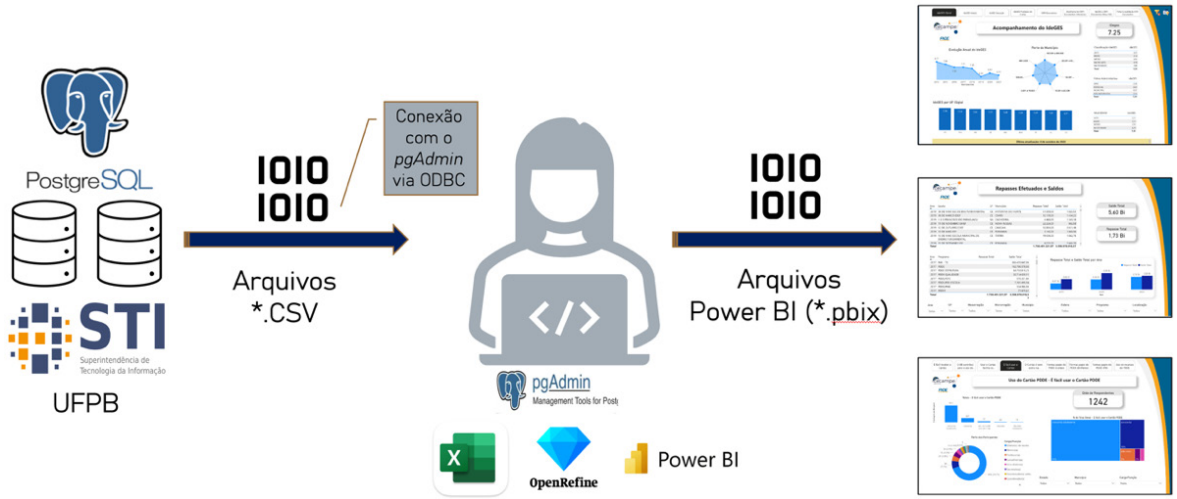
É importante mencionar que os dados extraídos da Plataforma Olinda, durante o período inicial dos desenvolvimentos, apresentavam divergências. Para sanar estas divergências, conforme descrito no relatório apresentado pelo CECAMPE/NE (Brasil, 2022), foi necessária a utilização de cinco bases distintas, providas pelo FNDE, conforme disponibilidade dos dados.

Os dados obtidos por meio de formulários gerados a partir de enquetes realizadas com o *Google Forms* foram exportados no formato CSV para o programa de planilhas eletrônicas *Microsoft Excel*, após passarem por uma etapa de transformações com o *software OpenRefine*. Este é um produto *Open Source* que facilita a limpeza e a transformação de dados, disponibilizando diversos algoritmos para verificar a consistência deles. Os dados

obtidos mediante enquetes realizadas com o *Google Forms* não foram disponibilizados no gerenciador de banco de dados *PostgreSQL* da STI/UFPB, mas armazenados no sistema de arquivos das estações de trabalho dos Cientistas de Dados.

A **Figura 2** ilustra o processo de importação de dados a partir do gerenciador de banco de dados *PostgreSQL* da STI/UFPB pelos cientistas de dados e desenvolvedores de painéis.

Figura 2 – Produção de painéis a partir dos Bancos de SQL do STI/UFPB



Fonte: Desenvolvimento dos autores (2023)

Após a importação dos dados e a realização de eventuais ajustes que sejam necessários com o *Open Refine* ou outra ferramenta, esses dados são transferidos para o programa de *Business Analytics* da Microsoft, conhecido como *Power BI*. A aplicação gerada é então disponibilizada na *Web* para consumo por toda a comunidade.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

Após a apresentação dos procedimentos metodológicos, indicamos que a redução do GIGO no processo de extração, transformação e carga dos dados do PDDE foram consideradas satisfatórias pelos cientistas de dados do CECAMPE/NE. Ficou evidenciado que as anomalias identificadas nos processos iniciais de extração contribuíram para minorar resultados inconsistentes nos painéis de dados.

Dentre os desafios encontrados no tratamento dos dados obtidos a partir da Plataforma Olinda, mencionamos: 1. as divergências no intervalo temporal entre a atualização dos dados disponibilizados pelas bases do FNDE, SIGEF e SAE. Isso pode causar divergências em resultados calculados por aplicações que usem uma ou outra fonte de dados. O FNDE orientou que os dados usados para a elaboração de painéis passassem a ser obtidos através de sua plataforma interna Microsoft *Sharepoint* (Dias, 2022); 2. a ausência de documentação associada às tabelas e a muitos dos seus respectivos atributos. Frequentemente, os identificadores utilizados para nomear as tabelas não eram suficientes para possibilitar

uma compreensão efetiva da semântica desse objeto. Uma situação análoga também foi detectada nos nomes dados a alguns atributos das tabelas (Dias, 2022). Por exemplo, uma tabela que modela os fornecedores de uma empresa devia ser chamada de “fornecedores” e não algo como “abend” ou “xpto”. A mesma lógica se aplica aos atributos da tabela; 3. O uso de valores fora do padrão, como, por exemplo, o símbolo usado como separador de casas decimais ou formatos de datas diferentes do utilizado por padrão no Brasil.

Em relação aos desafios encontrados no tratamento dos dados obtidos por meio dos formulários gerados a partir das enquetes realizadas com o *Google Forms*, consideramos importantes mencionar um problema associado ao preenchimento de dados pelos respondentes dos questionários em campos de livre digitação. Por permitir a inserção de qualquer valor, um campo de livre digitação possibilita a captura de um mesmo conceito, representado por sintaxes diversas, o que demanda uma etapa adicional no tratamento dos dados. O **Quadro 1** ilustra algumas das possibilidades associadas ao cargo de “Diretor(a)”.

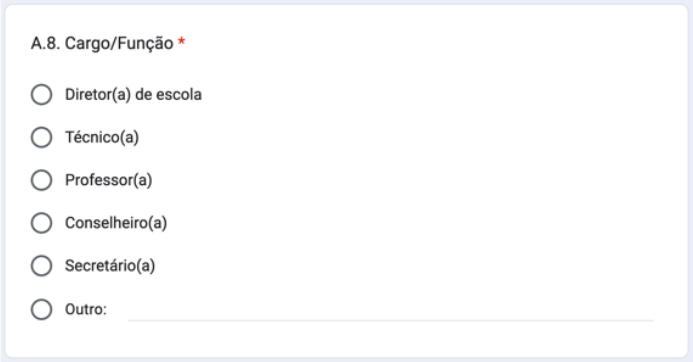
Quadro 1 – Problemas terminológicos em dados provenientes de enquetes

Diretor	Diretor	Diretora Adjunta	Diretor adjunto
Diretor Adjunto	Diretor administrativo	Diretor Administrativo da Educação	Diretor de Cultura e Eventos
Diretor de Departamento de Gestão	Diretor de ensino	Diretor de Ensino	Diretor de escola Adjunto
Diretor de escola(a)	Diretor de projetos especiais	Diretor Escolar	Diretor Escolar Adjunto
Diretor Financeiro (Tesoureiro)	Diretor geral da Secretaria	Diretor na SEDUC	Diretora
Diretora da Direc	Diretora adjunta	Diretora Adjunta	Diretora Adjunta escolar
Diretora administrativa e financeira	Diretora da Semed	Diretora da área de prestação de contas	Diretora de Ensino
Diretora Financeira	Diretora NTE		

Fonte: Dias *et al*., (2022, p. 84).

O problema mencionado ocorreu na realização das primeiras enquetes, posteriormente os campos de livre digitação foram substituídos por opções de resposta com valores fixos (vide **Figura 3**).

Figura 3 – Produção de painéis a partir dos Bancos de SQL do STI/UFPB.



Fonte: Dias *et al*., (2022, p. 84)

O aprendizado alcançado pela equipe de cientistas de dados do CECAMPE/NE na redução do GIGO possibilitou a elaboração de produtos de informação mais fidedignos, com potencial para atender de maneira mais assertiva toda a comunidade de usuários.

CONCLUSÕES

Após o término do período de dois anos de coleta dos dados do PDDE, foi possível constatar a existência de questões importantes relacionadas à qualidade desses dados, conforme apresentado na seção anterior. Essas questões têm o potencial para afetar negativamente as análises e as conclusões que emergem desses dados, materializadas na forma de painéis, caso não tivessem sido efetivamente abordadas.

O conceito de GIGO, popular nos centros de processamento de dados na segunda metade do século XX, volta a ser proeminente em nossa contemporaneidade, caracterizado pela “datificação” de todos os aspectos da atividade humana. A partir desta investigação, podemos concluir que a máxima GIGO representa um desafio real para a utilização efetiva dos dados do PDDE. De modo a perpassar este desafio, sugerimos a implementação de melhores práticas na gestão de dados no contexto do FNDE/PDDE, atentando para a padronização de todas as etapas relacionadas com um ciclo de vida de dados no contexto de atuação dos gestores de bancos de dados, cientistas de dados, desenvolvedores de aplicações e de todos(s) aqueles(as) que usam os dados como a matéria-prima para a realização de suas atividades profissionais.

Com relação às sugestões para direcionarmos futuros desenvolvimentos de aplicações e pesquisas no contexto do PDDE, identificamos a aplicação das técnicas de Inteligência Artificial (IA) como uma perspectiva promissora para enfrentar e solucionar os desafios referentes à qualidade dos dados. Esta possibilidade tem o potencial de aprimorar a confiabilidade e a utilidade dos dados provenientes do PDDE, enriquecendo, assim, as análises e fundamentando decisões de forma mais precisa e embasada em evidências. As técnicas de IA podem ser utilizadas para identificar e corrigir automaticamente

anomalias nos dados, validar informações e preencher lacunas de valores ausentes, além de oferecer abordagens analíticas e interpretativas avançadas, contribuindo de forma efetiva para a qualidade e confiabilidade das informações utilizadas. Diante disso, encorajamos aprofundamentos posteriores na exploração dessas técnicas de IA, identificando-as como uma orientação promissora para pesquisas futuras. Este direcionamento visa maximizar a eficácia e o valor dos dados do PDDE no contexto das análises e na fundamentação de decisões no âmbito educacional.

REFERÊNCIAS

CORDULA, F.; DIAS, G. A.; COSTA, P. R. S.; ARAÚJO, W.; DINIZ, A. V. S. Desenvolvimento da Integração de Dados para o Projeto Cecampe – NE no Contexto do "Programa Dinheiro Direto Na Escola". *In*: SEMINÁRIO DO CECAMPE NORDESTE, 1., 2022, João Pessoa. **Anais** [...]. João Pessoa: Editora do CCTA, 2022. v. 1, p. 67-68.

DIAS, G. A.; COSTA, P. R. S.; ARAÚJO, W. J.; CORDULA, F. R.; DINIZ, A. V. S. Dados e painéis informacionais: insumos e tecnologias habilitadoras para a disseminação do conhecimento no contexto do "Programa Dinheiro Direto na Escola". *In*: DINIZ, A. V. S.; PRESTES, E. M. T.; SANTOS, J. L. B.; FITTIPALDI, I.; PEREIRA, M. A. N.; ARAÚJO, W. J. (org.). **Os novos gerenciamentos de ações para o fortalecimento do programa dinheiro direto na escola**. João Pessoa: Editora do CCTA, 2022. v. 1, p. 79-87.

DINIZ, A. V. S.; PEREIRA, M. A. N.; ARAÚJO, W. J.; FITTIPALDI, I. O fortalecimento do Programa Dinheiro Direto na Escola na Região Nordeste como estratégia para a gestão democrática e para a qualidade da educação. *In*: SEMINÁRIO DO CECAMPE NORDESTE, 1., 2022, João Pessoa. **Anais** [...]. João Pessoa: Editora do CCTA, 2022. v. 1, p. 6-23.

MINISTÉRIO DA EDUCAÇÃO (Brasil). **O fortalecimento do Programa Dinheiro Direto na Escola na região Nordeste como estratégia para a gestão democrática e para a qualidade da educação CECAMPE-região Nordeste**: relatório técnico: análise do IDEGes e sua evolução. João Pessoa-PB, 2022. Disponível em: https://doi.org/10.5281/zenodo.7599128. Acesso em: 21 maio 2023.

OZMINKOWSKY, R. Garbage In, Garbage Out. **Towards Data Science**. [*S. l.*], nov. 2021. Disponível em: https://bit.ly/46bbreK. Acesso em: 18 maio 2023.

RICHARDSON, J. R. **Pesquisa social**: Métodos e técnicas. 4. ed. São Paulo: Atlas, 2017. 424 p.

STENSON, R. Is This the First Time Anyone Printed, 'Garbage In, Garbage Out'?. **Atlas Obscura**. [*S. l.*], mar. 14, 2016. Disponível em: https://www.atlasobscura.com/articles/is-this-the-first-time-anyone-printed-garbage-in-garbage-out. Acesso em: 18 maio 2023.

AGRADECIMENTOS

A realização dessa pesquisa foi possível com resultado do Conselho Nacional de Pesquisas. Vide processo CNPq número 311563/2018-0.