

ANÁLISE DE DADOS E ESCRITA DE UM ARTIGO CIENTÍFICO USANDO O CHATGPT

Eduardo Gonçalves Paterson Fox^{1*} (PO - ofox@biof.ufrj.br)

¹Universidade Estadual de Goiás – Câmpus Sudoeste – Sede Quirinópolis. Avenida Brasil, nº 435, Conjunto Hélio Leão, CEP: 75862-196, Quirinópolis, Goiás.

Resumo: Os recursos mais modernos de Inteligência Artificial (IA) trazem potenciais vantagens, bem como dilemas, para o ambiente acadêmico e para a produtividade científica. Uma das ferramentas mais conhecidas é a interface ChatGPT, que responde a perguntas e pedidos feitos pelo usuário em uma entrada de comando (*prompt*) na forma de uma conversa natural. O presente manuscrito relata a experiência pessoal do autor no seu primeiro uso do ChatGPT para ajudar na interpretação estatística e discussão de dados de um projeto dentro de sua área de especialidade. Tecnicamente, a interface foi usada de suporte para gerar códigos de programação (*scripts*) para o software RStudio, no sentido de aplicar diferentes abordagens de análises multivariadas a um conjunto de dados de medidas de duas populações de objetos (formigas obtidas em campo). Como resultado, ao longo de cerca de 36 horas de interação, a plataforma ajudou na produção de 6 gráficos de análises, geradas por cerca de 250 linhas de codificação. A plataforma também deu suporte em dúvidas de interpretação e técnica, no sentido de ajudar na escrita da discussão dos resultados. O manuscrito por detrás destes dados está já em adiantada fase de desenvolvimento. O presente autor não acredita que poderia ter completado as análises sem a ajuda técnica da plataforma de IA, e que recebeu suporte satisfatório para avançar na discussão dos dados. Fica evidente que a IA, se usada com responsabilidade e devido embasamento técnico, pode fornecer ajuda significativa e fundamental para o avanço e melhoria da produtividade científico-acadêmica.

Palavras-chave: Escrita em Espelho; Raciocínio Computacional; Análise Secundária de Dados

Introdução

A Inteligência Artificial (IA) pode acelerar o processo de pesquisa acadêmica ao facilitar o acesso a informações e dados relevantes. Ferramentas como *chatbots* e assistentes virtuais ajudam acadêmicos a encontrarem mais rapidamente as informações necessárias, tornando o processo de pesquisa ágil e democrático (ROSSONI e CHAT, 2022). ChatGPT trata-se de um modelo de *software* do tipo *chatbot* desenvolvido pela empresa OpenAI lançado publicamente em novembro de 2022. A alcunha "ChatGPT" vem justamente sua funcionalidade de conversação (do inglês "Chat"), enquanto que a partícula "GPT" vem de *Generative Pre-trained Transformer* (i.e. Transformador Pré-treinado Generativo), referente ao seu modelo de linguagem. Funciona por *deep learning* utilizando a arquitetura *Transformer*, que foi especialmente projetada para processar e gerar informação na forma de texto imitando a linguagem humana. Ao buscar informações para interagir com o usuário, o modelo ChatGPT acessa um grande banco de dados textuais oriundo da internet,

permitindo que o modelo compreenda e responda de forma compreensível perguntas e solicitações.

Resumidamente, ao receber um comando de entrada (o *prompt*) do usuário, o modelo interpreta o contexto e tenta gerar respostas coerentes simulando uma conversa humana. Demais *chatbots* tradicionais seguem *scripts* pré-definidos, ao passo que o ChatGPT aprende com as interações, ajustando as respostas de acordo com as entradas dos *prompts* (DENG e LYIN, 2002). Além de sustentar diálogos sobre diversos assuntos, o modelo ChatGPT pode responder perguntas gerais e técnicas, criar conteúdo original (textuais, como histórias e poemas), bem como resolver problemas matemáticos. Assim, possui grande aplicabilidade em diversas esferas, como Entretenimento, Organização pessoal (listas de tarefas e agendamento de compromissos), sistemas de suporte ao Cliente, e como ferramenta de Educação (PRADANA, 2023). Nesta última, tem sido bastante usado no aprendizado de novos conceitos e mesmo de idiomas.

Uma das aplicabilidades mais frequentes tem sido escrever sequências de linhas de comando (*scripts* de códigos) na linguagem de programação dinâmica orientada a objetos bem conhecida como R, desenhada para auxiliar na análise estatística de dados. Essa prática tem se tornado cada vez mais comum dentre programadores e analistas de dados de diversas esferas que buscam maior agilidade na criação de códigos em *scripts* e na solução de problemas (ou *bugs*) relacionados à manipulação de dados e visualização de resultados (MORGAN, 2023). O ChatGPT gera códigos sintaticamente corretos para diversas tarefas em R, sendo mais ideal para tarefas relativamente pouco complexas, como simulação, manipulação de dados e criação de gráficos. Os erros de interpretação e de sintaxe se fazem mais presentes na medida em que as tarefas se tornam menos usuais e mais confusas. No âmbito das análises comuns de dados experimentais, então, fica evidente a utilidade do ChatGPT em auxiliar pesquisadores, especialmente aqueles especialistas não-programadores, na padronização e otimização de seus *scripts* de código R, favorecendo a transparência e a reprodutibilidade de resultados publicados de pesquisa científica.

O presente contexto explora a interface desta última esfera com a ciência acadêmica. De forma análoga ao aprendizado de idiomas, hoje existem muitos usuários utilizando o ChatGPT para facilitar a análise de dados, tais como na estatística de resultados de pesquisas. O presente manuscrito descreve a primeira experiência pessoal de um cientista estabelecido no uso do modelo ChatGPT como uma ferramenta central de suporte para a análise de dados experimentais simples (medidas morfológicas) e no desenvolvimento da discussão dos padrões encontrados, no contexto da escrita de um manuscrito científico. O tema do manuscrito trata do uso de proporções corporais (morfometria) para diferenciação de duas espécies muito comuns e semelhantes de formigas, que ocorrem em uma mesma região geográfica. O autor é especialista em formigas deste grupo (Insecta: Hymenoptera: Formicidae: *Solenopsis* spp.), experiente na morfologia das duas espécies-alvo e tendo já autorado diversos scripts de códigos na linguagem R publicados (<https://github.com/eduardofox2>). No entanto, o autor é essencialmente inexperiente em abordagens de análises multivariadas, como aquelas atualmente aplicáveis neste tipo de estudo. Análise multivariada (MVA) é uma técnica estatística para examinar simultaneamente múltiplas variáveis a partir de um mesmo conjunto de dados, buscando identificar padrões e relações entre essas variáveis.

Considerações Metodológicas

Os dados utilizados advêm de um projeto conjunto na discriminação de duas espécies muito semelhantes de formigas do grupo das lava-pés, que neste manuscrito chamaremos de A e B, para simplificar o contexto. Os dados são oriundos de duas fontes: (i) um conjunto de 14 medidas originais de estruturas corporais de um conjunto de 115 operárias selecionadas ao acaso, para cada uma das duas espécies do presente trabalho; (ii) um outro conjunto de 14 medidas equivalentes para as mesmas duas espécies, mas oriundas de outras partes do Brasil, e disponibilizado publicamente por Tschinkel et al. (2013).

O conjunto de medidas foi analisado seguintes métodos de MVA: (i) Análise de Componentes Principais (PCA), para reduzir a dimensionalidade dos dados, identificando aquelas variáveis mais explicativas das variâncias; (ii) Análise de Cluster por *Linear Discriminant Analysis* (LDA), buscando maximizar as diferenças entre

diferentes conjuntos de objetos para tentar agrupar as observações semelhantes em *clusters*.

Os dados foram introduzidos e trabalhados por meio da plataforma RStudio v.2022.02.1, utilizando os pacotes “MASS”, “ggplot2”, “ggfortify”, “ggforce”, “plotly” e “dplyr”. A interface do ChatGPT (<https://chat.openai.com/>) foi utilizada para consultas acerca de como analisar os dados utilizando as abordagens mencionadas acima, utilizando-se de *prompts* pedindo esclarecimentos acerca das diferenças intrínsecas de cada abordagem e scripts de códigos para transformação e análise dos dados por cada abordagem, bem como de scripts para plotagem gráfica dos resultados e padrões encontrados. A interface ChatGPT também foi utilizada para questionar as melhores formas de se comparar os resultados obtidos das diferentes ferramentas, bem como possíveis explicações para os padrões encontrados.

Resultados e Discussão

O ChatGPT forneceu informações precisas e informativas acerca das abordagens de análises multivariadas, bem como *scripts* bastante funcionais (depois de edições necessárias) e suporte satisfatório no caso de eventuais mensagens de aviso ou de erros na execução das operações. Em todos os casos, foi necessária alguma inclusão de contexto nos *prompts*, bem como reinterpretação das respostas sob a luz das particularidades de cada situação. Em todos scripts providos foi necessária alguma adaptação para este funcionar, geralmente sobre os nomes dos objetos e conversão de formato ou classe dos objetos de entrada contendo os dados. O *script* de código gerado contou com aproximadamente 250 linhas, e produz 6 imagens de resultados, exemplificadas na Figura 1, abaixo.

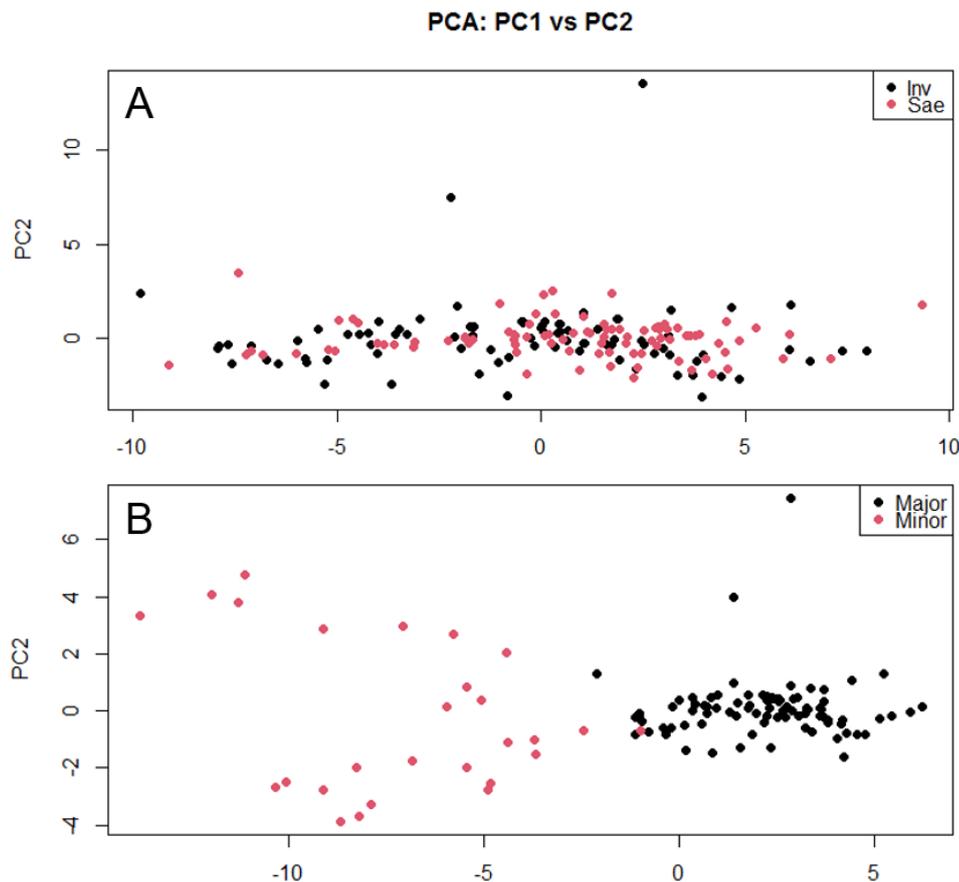


Figura 1. *Scatterplot* de resultados de análise de PCA separando populações distintas de um mesmo objeto (espécies de formigas), obtido com o suporte da interface de IA conhecida como ChatGPT para a geração de scripts de código de R.

A interação com o ChatGPT levou um total de cerca de 36h, distribuídas ao longo de 12 dias, para completar a essência das análises desejadas. Estima-se que o tempo de interação com o RStudio executando os scripts tenha sido de, ao menos cerca de 150h, ou seja, 5 vezes maior. O manuscrito em seu estado atual conta com quase 20 páginas, sem contar com o conteúdo de arquivos de *scripts*, que serão submetidos como material suplementar, no momento da submissão do manuscrito final para publicação. O autor não acredita que teria sido capaz de efetuar estas análises sem a ajuda do ChatGPT, e que teria parado nos testes usando o PCA, tendo por usado por base inicial para sua execução um *script* escrito e disponibilizado por outro colega, da área de ecologia, e que não estava bem ajustado às suas necessidades técnicas atuais. Em suma, a interação com o *prompt* de comandos permitiu ao autor ajustar as análises ao seu contexto específico de pesquisa,

principalmente quando apoiado pela literatura especializada nas estratégias de análises sugeridas aplicadas ao mesmo assunto.

A utilização do ChatGPT para escrever scripts em R é uma tendência crescente entre desenvolvedores e analistas (ZAVERIA, 2023). A ferramenta oferece uma maneira rápida de gerar código (MORGAN, 2023). Hoje em dia existem muitos usuários compartilhando suas experiências com o uso do ChatGPT para escrever códigos de análise de dados, sobretudo com a linguagem R; tais discussões correm normalmente em tom bastante informal, em comunidades de discussão online, como fóruns de programação (p.ex. <https://www.r-bloggers.com/2023/04/how-to-r-code-faster-with-chatgpt/>). É impressão pessoal do autor que a maior parte destas discussões foca na precisão e na necessidade de revisão do código gerado, onde muitos usuários afirmam que ele pode não atender exatamente às necessidades específicas de cada um. No entanto, deve-se notar que o contexto destas discussões é de grupos de programação, denotando que pode haver uma complexidade maior de tarefas desejadas, bem como operações que não sejam tão correntes, como manipulações bem específicas dos dados.

Intrinsicamente, o modelo ChatGPT possui limitações em relação à compreensão profunda do fluxo lógico do código, logo é natural que o usuário precise revisar e ajustar o código gerado para suas necessidades específicas.

Com relação à escrita da discussão dos resultados, ChatGPT oferece diversas vantagens, atuando como uma ferramenta auxiliar que ajuda os pesquisadores a economizar tempo valioso durante o processo de escrita, permitindo que se concentrem em outras áreas da pesquisa. A ferramenta pode ajudar a expandir ideias e fornece sugestões criativas que os autores podem não ter considerado inicialmente. Isso é especialmente útil na fase de busca de conceitos e ideias no desenvolvimento de projetos (ROSSONI e CHAT, 2022).

Por outro lado, há fortes dilemas. Uma das principais preocupações éticas é o potencial de plágio. Pesquisadores podem usar o ChatGPT indevidamente para gerar seções inteiras de texto sem a devida atribuição, levando à desonestidade acadêmica. Também o ChatGPT pode produzir informações enganosas ou incorretas, o que pode ser particularmente problemático em contextos científicos onde a precisão é crucial (DENG e LYIN, 2022). Pesquisadores podem inadvertidamente confiar nessas imprecisões se não avaliarem criticamente os resultados da IA, o que tende a

acontecer quando há total desleixo ou desconhecimento acerca do tema sendo abordado (PARK, 2023).

No presente caso, o pesquisador conhece bastante do modelo de estudo e da pergunta experimental sendo formulada. A interface ajudou principalmente na confecção de respostas analíticas acerca das metodologias de análise sendo aplicadas, que não são do núcleo de domínio técnico do pesquisador. Como exemplo de resposta útil gerada, segue uma pergunta técnica seguida de um trecho da resposta provida.

What does it mean if LDA does separate my two groups well, but I am unsure about their Gaussian distribution nature?



If LDA separates your two groups well but you are unsure about the Gaussian distribution of your data, it can indicate several important things:

1. LDA's Robustness to Some Deviations from Normality

- LDA is fairly robust to mild deviations from normality. While LDA assumes that the data in each group follows a multivariate normal distribution, it may still work well even if this assumption is somewhat violated. This is especially true if:
 - The group means differ enough to allow good separation, even without perfect normality.
 - The covariance structure (relationships between variables) is not too complex or different between the groups.

In such cases, LDA can still produce meaningful group separation even if the data isn't perfectly Gaussian, especially when the differences between group means are large and the within-group variances are not too different.

2. Possibility of Overfitting

- If your groups are well separated by LDA, but the underlying distributions are far from Gaussian, there is a chance that the separation may be overfitted to your specific data. This means that while the LDA performs well on your training data, it might not generalize as well to new, unseen data.
- This is particularly true if you have a small sample size or if the separation seems too "clean." LDA might be finding patterns in the data that are specific to your sample, rather than capturing general differences between the groups.

Figura 2. Exemplo de interação com a plataforma de IA conhecida como ChatGPT, acerca de informar-se dos aspectos técnicos de uma abordagem estatística, como suporte técnico na escrita da discussão dos dados de um artigo científico.

Em suma, o presente autor sentiu a interação com a plataforma como muito valiosa na finalização de uma análise de dados de um projeto importante de sua pesquisa, que não teria sido possível por conta própria. Fica evidente que, corretamente aplicada, a ferramenta de IA pode ajudar em muito na escrita científica.

Considerações Finais

Embora o ChatGPT seja uma ferramenta interessante e de interface amigável para facilitar a análise de dados, é essencial saber dosar o uso com cautela. Os usuários devem permanecer cientes de suas limitações, buscando garantir o manuseio ético dos dados e sempre priorizarem a lógica e supervisão humanas para mitigar o desenvolvimento de uma dependência excessiva pelo tipo de suporte gerado por IA.

Também na escrita, por mais que o modelo ofereça significativos recursos de apoio na geração de ideias, interpretação de padrões e resultados, e mesmo na organização de pensamentos para desenvolvimento de textos, o uso do ChatGPT na escrita científica é ainda permeado de uma aura de dilemas éticos, e preocupações generalizadas com a coerência e com a responsabilização técnica e ética sobre o conteúdo publicado. O presente autor, em sua primeira experiência, achou a ferramenta adequada para conseguir avançar no uso de técnicas de análise estatísticas que pesquisadores devem abordar sua aplicação com cautela, garantindo transparência e adesão aos padrões de integridade acadêmica, mas é fundamental que os usuários verifiquem e ajustem o conteúdo retornado pela interface para garantir que atende às suas necessidades específicas, bem como aos preceitos de seriedade e ética das comunidades científica e acadêmica.

Agradecimentos

O autor foi financiado com verba PDCTR, processo FAPEG/CNPQ no. 317847/2021-0.

Referências

DENG, J.; LIN, Y. The benefits and challenges of ChatGPT: An overview. **Frontiers in Computing and Intelligent Systems**, v. 2, n. 2, p. 81-83, 2022.

ISLAM, I.; ISLAM, M. N. Opportunities and challenges of ChatGPT in academia: A conceptual analysis. **Authorea Preprints**, 2023.

MORGAN, D. L. Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. **International journal of qualitative methods**, v. 22, p. 16094069231211248, 2023.

PARK, J.-Y. Could ChatGPT help you to write your next scientific paper?: concerns on research ethics related to usage of artificial intelligence tools. **Journal of the Korean Association of Oral and Maxillofacial Surgeons**, v. 49, n. 3, p. 105-106, 2023.

PRADANA, M.; ELISA, H. P.; SYARIFUDDIN, S. Discussing ChatGPT in education: A literature review and bibliometric analysis. **Cogent Education**, v. 10, n. 2, p. 2243134, 2023.

ROSSONI, L.; CHAT, G. P. T. A inteligência artificial e eu: escrevendo o editorial juntamente com o ChatGPT. **Revista eletrônica de ciência administrativa**, v. 21, n. 3, p. 399-405, 2022.

TSCHINKEL, W. R. The morphometry of *Solenopsis* fire ants. **PloS one**, v. 8, n. 11, p. e79559, 2013.

ZAVERIA. Top 8 ChatGPT R Programming Tools in 2023. Analytics Insight, ChatGPT [online website]. Disponível em:
<https://www.analyticsinsight.net/chatgpt/top-8-chatgpt-r-programming-tools-in-2023>.
Consultado em: 21/11/2024.