PONDERAÇÃO DA BASE DE DADOS DE MEDIÇÕES SIMET EM ESCOLAS PÚBLICAS

Marcelo Pitta[1], Thiago Meireles[1] & Pedro Luis do Nascimento Silva[2]

marcelopitta@nic.br

1 *Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação (Cetic.br) - São Paulo, Brasil.*

2 *Escola Nacional de Ciências Estatísticas e Science - Rio de Janeiro, Brasil*

1 Introdução

O Ajuste dos Escores de Propensão tem sido amplamente utilizado para redução de vieses relacionados a erros de não cobertura e não resposta, bem como para amostras não probabilísticas (Elliot, 2009; Elliott & Valliant, 2017). Este artigo apresenta uma aplicação em amostras voluntárias para populações conhecidas em combinação com uma técnica de seleção de variáveis baseada em árvores aleatórias (*Random Forest*). A análise parte de uma amostra numerosa de escolas básicas que aderiram voluntariamente ao Simet (Sistema de Medição de Tráfego Internet), seja via *firmware* instalado no roteador ou *software* no computador para coleta automática de medidas de qualidade da Internet. Sendo a população de escolas básicas conhecida e suas características medidas pelo Censo Escolar, a metodologia de ponderação apresentada tem como objetivo permitir a estimação de medidas de qualidade da Internet para o conjunto de todas as escolas a partir das medições disponíveis somente para as escolas da amostra de escolas participantes no Simet.

No entanto, uma vez que a instalação dos medidores nas escolas não se dá por um processo de amostragem probabilística, não são conhecidas a probabilidade de uma escola receber o monitor bem como a população representada pelas escolas que instalaram o medidor. Para alcançar o objetivo de estimar as medidas de qualidade da Internet a partir da ponderação baseada nas informações coletas em escolas que possuam o medidor, é necessário:

(a) Determinar que população alvo pode ser representada pelas escolas que possuem

medidores instalados;

(b) Construir pseudo-pesos amostrais para as escolas que possuem medidores para expandir os resultados de qualidade obtidos para a população estabelecida em (a).

2 Dados

As bases de dados utilizadas para a análise das medidas de qualidade da Internet nas escolas são (a) o Censo Escolar mais atualizado[1], doravante Censo, e (b) a base de dados da amostra das escolas com medidores Simet[2]. O Censo possui informações sobre os estabelecimentos de ensino, alunos, gestores e profissionais nos estabelecimentos de ensino básico. Já os dados do Simet possuem informações sobre a qualidade da Internet medida por um *firmware* instalado no roteador ou por *software* no computador. Eles registram medições periódicas enquanto os dispositivos estiverem em funcionamento de forma independente. Para a análise, foi consolidada uma base agregando a informação da existência ou não de medidores Simet em cada escola às demais informações coletadas no Censo para todas as escolas.

2.1 Determinação da população representada pelas escolas com medidores

A coleta de dados sobre a qualidade da conexão das escolas é realizada por um *firmware* instalado no roteador ou por um *software* em um computador conectado à Internet. Ambos realizam medições da qualidade da Internet de forma automática em espaços de tempo previamente estabelecidos, sem a necessidade de intervenção humana. Para que elas sejam realizadas, apenas é necessário que o dispositivo esteja ligado e conectado.
A partir deste enquadramento foram definidos os pré-requisitos para a medição de qualidade da Internet nas escolas, quais sejam, (a) existência de acesso à Internet na escola e (b) existência de computador ligado à Internet na escola. Quando observados os pré-requisitos e as estatísticas de instalação dos medidores por tipo de administração escolar[3], ver Tabela 1, verificou-se que a população de escolas que poderiam ser re-

[1]Realizado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)

[2]Disponibilizado pelo Centro de Estudos e Pesquisas em Tecnologia de Redes e Operações (Ceptro.br) vinculado ao Núcleo de Informação e Coordenação do Ponto BR (NIC.br)

[3]Estatísticas relativas a extração de base Simet de abril de 2023 e dados do Censo de 2022. Essa análise é realizada a cada atualização de bases. Não há mudança nas premissas necessárias à instalação

presentadas pelas escolas com medidores é formada pelo conjunto de escolas públicas municipais e estaduais da educação básica com acesso à Internet e com computador.

Tabela 1: Escolas do Censo: existência de medidores por dependência administrativa

Tipo de administração	Não	Sim	Total
Pública Federal	735	-	735
Pública Estadual	18.049	15.411	33.460
Pública Municipal	98.570	34.059	132.629
Privada	57.803	22	57.825
Total	175.157	49.492	224.649

Fonte: Bases Censo Escolar 2022 e Simet (Abril/2023).

Dessa forma, a base final para ajuste dos pseudo-pesos é composta pelas escolas públicas estaduais e municipais que possuem computador e acesso à Internet declarado no Censo. Adicionalmente, foi tratada para a exclusão de observações atípicas (*outliers*) e de escolas que possuem o Simet instalado a despeito de declararem que não possuíam computadores e/ou acesso à Internet no Censo. A exclusão deste grupo de escolas que possuem o Simet instalado no processo de estimação dos pseudo-pesos se justifica por não estarem no recorte do universo, mas serão reincluídas na base de medições de qualidade com peso igual a 1 (um), representando apenas elas mesmas. Da mesma forma, escolas classificadas como *outliers* seguirão a mesma regra de inclusão na base de medidas de qualidade caso possuam Simet instalado[4].

3 Representatividade das escolas com medidores e estimação de pseudo-pesos

Uma vez que a instalação dos medidores Simet não ocorre de forma aleatória, não é possível, a priori, considerar o conjunto das escolas com medidores como representativo do conjunto de escolas estabelecido para a análise. Dessa forma é necessário avaliar a "representatividade"das escolas desta amostra de escolas com medidores para expandir os resultados para as escolas municipais e estaduais de educação básica com computador e acesso à Internet. Uma das possibilidades encontradas na literatura para correção de vícios de autosseleção (*self-selection*) é a construção de pseudo-pesos. Seguindo esta

e funcionamento do Simet nas escolas.

[4]Não foram excluídas escolas *outliers* que não tivessem o Simet instalado até a atualização da base do Censo de 2022.

abordagem, se buscaria estimar a probabilidade de uma escola ter um medidor ativo instalado a partir de um cadastro completo da população, que aqui seria o Censo. Estas probabilidades, chamadas de escores de propensão (*propensity scores*), seriam transformadas em pseudo-pesos correspondentes ao inverso da propensão de ter um medidor ativo instalado. Dessa forma, seria possível generalizar os resultados da amostra para todo o conjunto de escolas elegíveis.

A estimação dos pseudo-pesos se dá a partir do ajuste de um modelo de regressão logística no qual a variável resposta é um indicador binário (Y_i) que identifica a presença de um medidor na escola i. Já as variáveis explicativas seriam as disponíveis no Censo, sendo 190 variáveis categóricas e 113 numéricas. Por se tratar de ajustes de modelos logísticos para um banco de mais de 100.000 registros, tanto a seleção de variáveis relevantes quanto as estatísticas de bondade de ajuste do modelo não são eficientes.

Dessa forma, optou-se por uma metodologia em duas etapas. Na primeira foram ajustados dois modelos de árvores aleatórias (*Random Forest*, RF) tendo Y_i como variável resposta, mas separando as variáveis categóricas e as numéricas como variáveis explicativas. Um dos resultados obtidos é o grau de importância de cada variável explicativa na predição da existência de medidor. Assim, em uma segunda etapa, foram ajustados modelos logísticos para a estimação dos pseudo-pesos incluindo gradualmente variáveis categóricas e/ou numéricas a partir da ordem do grau de importância determinado pela análise de RF. O modelo logístico é dado pela fórmula:

$$log\left(\frac{P(Y_i=1)}{1-P(Y_i=1}\right) = \alpha + \beta_1 X_i^c + \beta_2 X_i^n \tag{1}$$

Onde: Y_i é a variável resposta, assumindo o valor 1 se a escola i possui medidor Simet instalado, e valor 0, caso contrário; X_i^c é um vetor com os valores de variáveis explicativas categóricas da escola i selecionadas a partir da importância estabelecida pelo modelo de RF; X_i^n é um vetor com os valores de variáveis explicativas numéricas da escola i selecionadas a partir da importância estabelecida pelo modelo RF, $P(Y_i = 1)$ representa a probabilidade da escola i ter medidor instalado, e α e β_1, β_2 são parâmetros do modelo, a serem estimados.

Já as estimativas para $P(Y_i = 1)$ fornecidas pela expressão

$$\hat{P}(Y_i=1) = \left(\frac{exp(\hat{\alpha} + \hat{\beta}_1 X_i^c + \hat{\beta}_2 X_i^n)}{1 + exp(\hat{\alpha} + \hat{\beta}_1 X_i^c) + \hat{\beta}_2 X_i^n}\right) \tag{2}$$

são os chamados escores de propensão considerados na metodologia, sendo que $\hat{\alpha}$, $\hat{\beta}_1$ e $\hat{\beta}_2$ são as estimativas dos parâmetros obtidas com base no modelo de regressão logística

ajustado. Foram ajustados diversos modelos logísticos, a cada passo incrementando o número de variáveis explicativas consideradas (tanto as categóricas como as numéricas). Para cada modelo foram construídos os pseudo-pesos dados por $w_i = \frac{1}{\hat{P}(Y_i = 1)}$.
Estes pseudo-pesos foram calibrados para o total de escolas considerando distribuições marginais por Unidades Federativas (UFs), tipo de administração e localização (urbano/rural)[5]. A partir dos pseudo-pesos calibrados foram calculadas seis estatísticas resumo[6]. Estas estatísticas, que sumarizam desvios entre estimativas obtidas com emprego dos pseudo-pesos calculados e os correspondentes valores populacionais calculados com dados de toda a população, são apresentadas a seguir:

- Soma dos desvios absolutos para variáveis categóricas:
$$SDA_n = \sum_{b=1}^{B} |\hat{p}_j^b - p_j^b| \quad (3)$$

- Soma dos desvios absolutos relativos para variáveis categóricas:
$$SDAR_c = \sum_{a=1}^{A} \sum_{j=1}^{K_a} \left(\frac{|\hat{p}_j^a - p_j^a|}{p_j^a} \right) \quad (4)$$

Onde: a é uma variável categórica existente na base, A é o total de variáveis categóricas da base, j é uma categoria da variável categórica a existente na base, K_a é o total de categorias da variável categórica a da base, e p é a proporção.

- Soma dos desvios absolutos para variáveis numéricas:
$$SDA_n = \sum_{b=1}^{B} |\hat{m}_j^b - m_j^b| \quad (5)$$

- Soma dos desvios absolutos relativos para variáveis numéricas
$$SDAR_n = \sum_{b=1}^{B} \left(\frac{|\hat{m}_j^b - m_j^b|}{m_j^b} \right) \quad (6)$$

Onde: b variável numérica existente na base, B última variável numérica da base, e m é a média; Var_c é o número de variáveis categóricas em que ao menos um intervalo de confiança de 95% para uma proporção de uma categoria não contém a proporção observada no Censo; e Var_n é o número de variáveis numéricas em que o intervalo de confiança de 95% para a média não contém a média observada no Censo.
As variáveis de interesse para divulgação do resultados (UF, tipo de administração e localização) não foram consideradas nas estatísticas das variáveis categóricas por

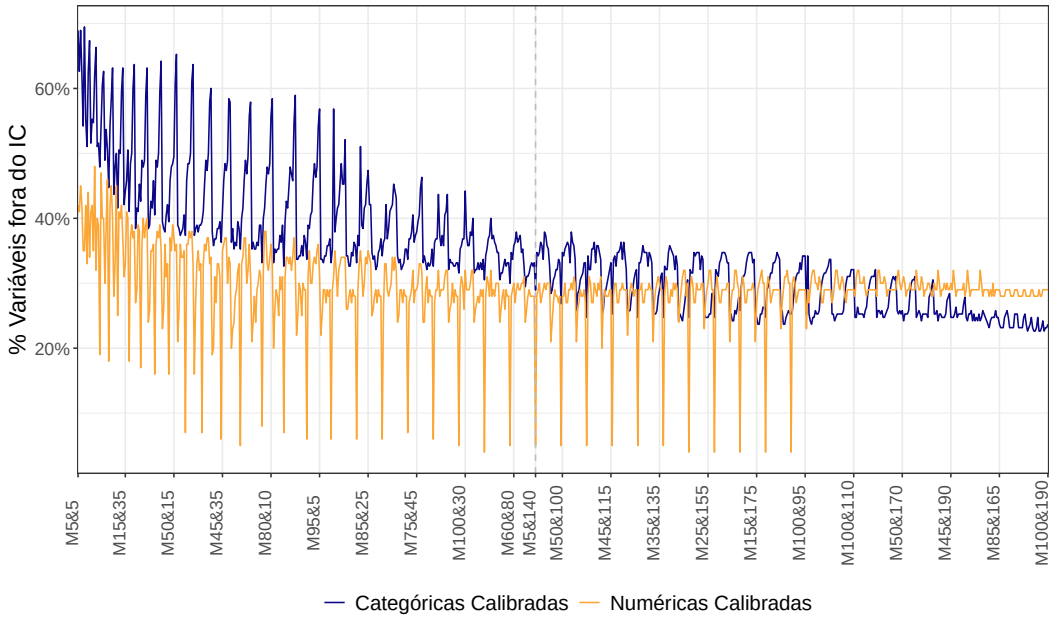
[5]São as variáveis de interesse para divulgação dos resultados, de forma que se buscou uma solução que permita a divulgação dos resultados como se a amostra fosse representativa para o universo em questão.

[6]Os pseudo-pesos e planos amostrais derivados foram declarados e as estatísticas calculadas por meio do Pacote *survey* da linguagem R.

passarem pelo processo de calibração dos totais conhecidos do universo da pesquisa por pós-estratificação[7].

4 Resultados

Figura 1: Porcentagem de variáveis alvo fora do Intervalo de Confiança de 95%*

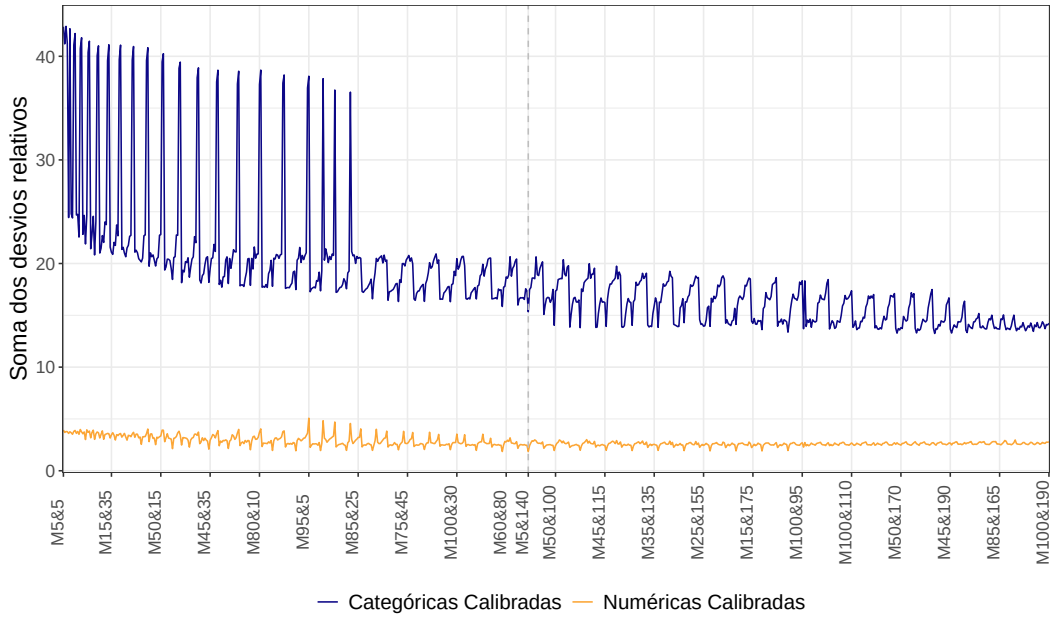


Fonte: Bases Censo Escolar, Simet e modelos.
*Os modelos foram ordenados seguindo (1) o menor número total de variáveis e (2) o menor número de variáveis numéricas

A seleção do modelo seguiu a observação dos gráficos com os resultados para o conjunto completo de modelos ajustados (combinações possíveis de variáveis numéricas e categóricas). Optou-se por um modelo mais parcimonioso, com um menor número de variáveis, que mantivesse um vício total próximo aos valores mais baixos. Assim, foi selecionado o modelo com 5 variáveis numéricas e 140 categóricas – M5&140 destacado nos gráficos.

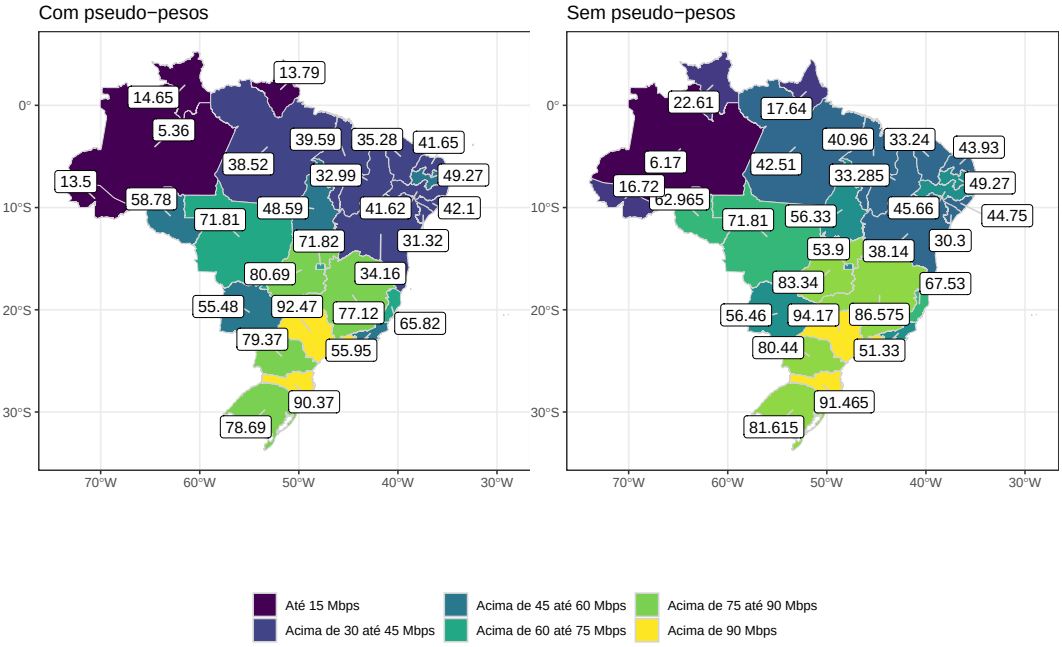
[7]O método só é passível de aplicação quando todos os estratos possuírem ao menos uma observação. Caso algum dos estratos não possuir observação será aplicado o método *rake*.

Figura 2: Soma dos desvios relativos por modelo ajustado*



Fonte: Bases Censo Escolar, Simet e modelos.
*Os modelos foram ordenados seguindo (1) o menor número total de variáveis e (2) o menor número de variáveis numéricas

Figura 3: Comparação da mediana por unidade fedarativa com e sem pseudo-pesos*



Fonte: Bases Censo Escolar, Simet e modelos.
*Os modelos foram ordenados seguindo (1) o menor número total de variáveis e (2) o menor número de variáveis numéricas

Como exemplo, é possível ver que as estimativas da mediana por unidade federativa para a velocidade de download apresentam diferenças significativas. Além de ser representativa apenas para as escolas que possuem o medidor (e não para as escolas com computador e Internet), elas mostram as diferenças geradas pelo viés de autosseleção.

5 Conclusão

A aplicação de modelos para obtenção de pseudo-pesos seguida de calibração destes para algumas distribuições marginais de interesse resulta em uma base de dados que permite a estimação das medições de qualidade da Internet para as escolas públicas estaduais e municipais com computador e acesso à Internet. A atualização das estimativas utilizando o processo de análise de dados e estimação de pseudo-pesos é necessária caso sejam incluídas novas escolas, tanto por alterações na base de dados da população alvo de escolas no Censo quanto pela inclusão de novas escolas com medidor Simet.

Referências

Bache, S. M., & Wickham, H. (2022). *magrittr: A Forward-Pipe Operator for R* [R package version 2.0.3]. https://CRAN.R-project.org/package=magrittr

Dayanand Ubrangala, R, K., Prasad Kondapalli, R., & Putatunda, S. (2022). *SmartEDA: Summarize and Explore the Data* [R package version 0.3.9]. https://CRAN.R-project.org/package=SmartEDA

Dever, J. A. (2018). Combining probability and nonprobability samples to form efficient hybrid estimates: An evaluation of the common support assumption. *Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference*, 15.

Dowle, M., & Srinivasan, A. (2023). *data.table: Extension of 'data.frame'* [https://r-datatable.com, https://Rdatatable.gitlab.io/data.table, https://github.com/Rdatatable/data.tab

Elliot, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, *2*(6).

Elliott, M. R., & Valliant, R. (2017). Inference for Nonprobability Samples. *Statistical Science*, *32*(2), 249–264. https://doi.org/10.1214/16-STS598

Greifer, N. (2023). *cobalt: Covariate Balance Tables and Plots* [R package version 4.5.1]. https://CRAN.R-project.org/package=cobalt

Hlavac, M. (2022). *stargazer: Well-Formatted Regression and Summary Statistics Tables* [R package version 5.2.3]. Social Policy Institute. Bratislava, Slovakia. https://CRAN.R-project.org/package=stargazer

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of official statistics*, *22*(2), 329.

Li, Y. C. P., & Wu, C. (2020). Doubly Robust Inference With Nonprobability Survey Samples. *Journal of the American Statistical Association*, *115*(532), 2011–2021. https://doi.org/10.1080/01621459.2019.1677241

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22. https://CRAN.R-project.org/doc/Rnews/

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley Series in Probability and Statistics.

Lumley, T. (2023). *survey: analysis of complex survey samples* [R package version 4.2].

Mc Overton, J., Young, T., & Overton, W. (1993). Using 'found' data to augment a probability sample: Procedure and case study. *Environmental monitoring and assessment*, *26*(1), 65–83. https://doi.org/10.1007/bf00555062

Müller, K. (2020). *here: A Simpler Way to Find Your Files* [R package version 1.0.1]. https://CRAN.R-project.org/package=here

Müller, K., & Wickham, H. (2023). *tibble: Simple Data Frames* [R package version 3.2.1]. https://CRAN.R-project.org/package=tibble

Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference. (2018).

Schauberger, P., & Walker, A. (2022). *openxlsx: Read, Write and Edit xlsx Files* [https://ycphs.github.io/
https://github.com/ycphs/openxlsx].

Smit, V. I. C. (2021). Correcting Selectivity in Datasets with Pseudo-Weights: a Simulation Study.

Smith, T. M. F. (1983). On the Validity of Inferences from Non-Random Samples. *Journal of the Royal Statistical Society: Series A (General)*, *146*(4), 394–403. https://doi.org/https://doi.org/10.2307/2981454

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional Variable Importance for Random Forests. *BMC Bioinformatics*, *9*, 11. https://doi.org/https://doi.org/10.1186/1471-2105-9-307

Valliant, R. (1993). Poststratification and Conditional Variance Estimation. *Journal of the American Statistical Association*, *88*(421), 89–96. https://doi.org/10.1080/01621459.1993.10594298

Valliant, R. (2019). Comparing Alternatives for Estimation from Nonprobability Samples. *Journal of Survey Statistics and Methodology*, *8*(2), 231–263. https://doi.org/10.1093/jssam/smz003

Valliant, R., & Dever, J. A. (2011). Estimating Propensity Adjustments for Volunteer Web Surveys. *Sociological Methods & Research*, *40*(1), 105–137. https://doi.org/10.1177/0049124110392533

Vaughan, D., & Dancho, M. (2022). *furrr: Apply Mapping Functions in Parallel using Futures* [R package version 0.3.1]. https://CRAN.R-project.org/package=furrr

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation* [R package version 1.1.2]. https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2023). *purrr: Functional Programming Tools* [R package version 1.0.1]. https://CRAN.R-project.org/package=purrr

Wickham, H., Miller, E., & Smith, D. (2023). *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files* [https://haven.tidyverse.org, https://github.com/tidyverse/haven, https://github.com/WizardMac/ReadStat].

Yoshimura, O. (2004). Adjusting responses in a non-probability web panel survey by the propensity score weighting. *ASA Proceedings of the Joint Statistical Meetings, Survey Methodology Section*, 4660–4665.