

Apêndice técnico

Contratos, fórmulas e validações

Patrick Andrade

2026-09-23

Contratos

`distribution_bins` possui uma linha por ano, geografia, ranking e grupo. `income_components` acrescenta o tipo de rendimento, dedução, imposto, ativo ou passivo. Todos os registros conservam a fonte e a aba original.

Centis hierárquicos

A visão disjunta contém os grupos 1–99, 101–109 e 111–120. O grupo 100 deve reconciliar com 101–110, e 110 deve reconciliar com 111–120. As tolerâncias são zero para contagens, admitindo diferença máxima de uma unidade por arredondamento, e um centavo para somas monetárias.

As contagens reconciliam em todas as 2.406 distribuições. A fonte apresenta divergências monetárias internas em 140 distribuições RTB subnacionais/exterior de 2017–2021. Esses casos permanecem como `warn`, são exportados integralmente e nunca têm seus valores substituídos.

Fórmulas

Todos os indicadores são estimados a partir de dados agrupados: cada grupo contribui com sua média, sua contagem e seus limites monetários, sem informação sobre a variação interna [1]. O Gini agrupado é obtido pela integral trapezoidal da curva de Lorenz e corresponde ao limite inferior do Gini compatível com os dados divulgados [2]. O Theil T utiliza as médias dos grupos [3]. O Atkinson principal usa $\epsilon = 0,5$ [4]; no caso-limite $\epsilon = 1$, o índice usa a média geométrica ponderada:

$$A_1 = 1 - \frac{\exp\left(\sum_i w_i \log y_i / \sum_i w_i\right)}{\mu}.$$

Wolfson segue a forma do índice Foster–Wolfson apresentada no manual da Statistics Canada [5]. Com m como mediana aproximada e $L(0,5)$ como a participação acumulada da metade inferior:

$$P_W = 2 \frac{\mu}{m} \left[2 \left(\frac{1}{2} - L\left(\frac{1}{2}\right) \right) - G \right].$$

Esteban–Ray é suplementar e usa $\alpha \in \{1, 0; 1, 3; 1, 6\}$ [6]. Nenhuma dessas medidas recupera a variação dentro dos grupos divulgados.

Limites do Gini com dados agrupados

Seguindo J. L. Gastwirth [2], além do limite inferior (trapézio sobre as médias), calcula-se um limite superior somando a máxima desigualdade intragrupo compatível com a média m_i e os limites monetários $[a_i, b_i]$ de cada grupo — a distribuição de dois pontos que preserva a média. Para grupos internos, o máximo intragrupo é $\frac{(m_i - a_i)(b_i - m_i)}{m_i(b_i - a_i)}$; para o grupo superior, aberto, usa-se o supremo $1 - a_i/m_i$, o que torna o limite conservador. As colunas `gini_lower_bound` e `gini_upper_bound` acompanham todas as distribuições em `distribution-metrics.parquet`.

Alíquota efetiva

A alíquota efetiva média de cada grupo é $\tau_i = \text{imposto devido}_i / \text{renda}_i$, com a renda no conceito do próprio ranking. Sob RB4, que inclui rendimentos isentos e de tributação exclusiva, a curva mede a carga sobre a renda ampla declarada [7]. Grupos com renda nula ou não divulgada recebem NA. Como toda medida deste projeto, trata-se de médias de grupos.

Interpolação de Pareto generalizada (trabalho futuro)

As participações de topo somam grupos disjuntos divulgados, sem modelo intragrupo. A literatura de tabulações fiscais recomenda interpolação de Pareto generalizada [8], [9] para recuperar a curva completa e reduzir o viés de agrupamento; a adoção dessa técnica permanece na agenda de pesquisa do projeto. As estimativas brasileiras de topo com dados tributários seguem M. Medeiros, P. H. G. F. Souza, e F. A. Castro [10], M. Morgan [11] e P. H. G. F. Souza [12] como referências de comparação.

Qualidade

- `unique_distribution_key`: **pass** — Chaves únicas
- `raw_bin_count`: **pass** — 0 distribuições sem os códigos 1 a 120 exatamente uma vez
- `series_coverage`: **warn** — 1 combinação(ões) ausente(s) na fonte: 2023-SP-RB9
- `hierarchical_contributors`: **pass** — 0 distribuições com divergência de contagem
- `hierarchical_amounts`: **warn** — 140 distribuições com divergência monetária preservada da fonte; ver `hierarchy-reconciliation.csv`
- `leaf_bin_count`: **pass** — A visão analítica deve conter 118 grupos disjuntos
- `effective_rate_range`: **pass** — Alíquotas efetivas dentro de [0, 1]
- `effective_rate_coverage`: **warn** — 60003 de 283908 grupo(s) sem alíquota interpretável (imposto acima da renda do conceito ou renda nula); concentrados nos conceitos estreitos RB5, RB9 e RB10
- `bounded_index_range`: **pass** — Gini e Atkinson dentro de [0, 1]
- `unstable_index_values`: **warn** — 118 distribuição(ões) com Wolfson > 1 (mediana baixa) ou Palma > 50 (base com participação baixa)
- `gini_year_over_year`: **warn** — 16 salto(s) nacional(is) de Gini acima de 0.05: RB1/2018, RB1/2019, RB2/2018, RB2/2019, RB3/2018, RB3/2019, RB4/2018, RB4/2019, RB6/2018, RB6/2019, RB7/2018, RB7/2019, RB8/2018, RB8/2019, RTB/2018, RTB/2019
- `debt_income_ratio_range`: **pass** — 26562 de 27376 grupos com razão dívida/renda definida e não negativa
- `asset_total_identity`: **pass** — 10260 grupos com soma das famílias igual ao total divulgado
- `top_counts_nesting`: **pass** — 928 contagens de topo aninhadas corretamente

Proveniência

O manifesto registra SHA-256, tamanho, data de recuperação, URL direta e página de origem. Uma mudança de conteúdo sob uma URL conhecida exige revisão humana.

Comparabilidade

O dicionário `config/schema/rankings.csv` registra a definição dos conceitos RTB e RB1–RB10. O layout de 2017 é monolítico; o de 2024 distribui rankings nacionais e estaduais em arquivos distintos. O parser converte ambos para o mesmo contrato sem depender exclusivamente do nome do arquivo.

O crosswalk anual registra uma lacuna da própria fonte: SP–RB9 não está presente em 2023. O gate `series_coverage` mantém esse caso como aviso e não imputa a distribuição ausente.

As somas e componentes de 2017–2021, divulgados em milhões de reais, são multiplicados por 10^6 no estágio de harmonização. Limites e médias, já publicados em reais, não são reescalados. Os arquivos atuais permanecem em reais.

Bibliografia

- [1] F. A. Cowell, *Measuring Inequality*, 3º ed. Oxford: Oxford University Press, 2011.
- [2] J. L. Gastwirth, “The Estimation of the Lorenz Curve and Gini Index”, *The Review of Economics and Statistics*, v. 54, n. 3, p. 306–316, 1972, doi: 10.2307/1937992.
- [3] H. Theil, *Economics and Information Theory*. Amsterdam: North-Holland, 1967.

- [4] A. B. Atkinson, "On the Measurement of Inequality", *Journal of Economic Theory*, v. 2, n. 3, p. 244–263, 1970, doi: 10.1016/0022-0531(70)90039-6.
- [5] M. C. Wolfson, "When Inequalities Diverge", *The American Economic Review*, v. 84, n. 2, p. 353–358, 1994.
- [6] J.-M. Esteban e D. Ray, "On the Measurement of Polarization", *Econometrica*, v. 62, n. 4, p. 819–851, 1994, doi: 10.2307/2951734.
- [7] S. W. Gobetti e R. O. Orair, "Taxation and Distribution of Income in Brazil: New Evidence from Personal Income Tax Data", *Brazilian Journal of Political Economy*, v. 37, n. 2, p. 267–286, 2017, doi: 10.1590/0101-31572017v37n02a01.
- [8] T. Blanchet, J. Fournier, e T. Piketty, "Generalized Pareto Curves: Theory and Applications", *Review of Income and Wealth*, v. 68, n. 1, p. 263–288, 2022, doi: 10.1111/roiw.12510.
- [9] T. Blanchet, L. Chancel, I. Flores, e M. Morgan, "Distributional National Accounts Guidelines: Methods and Concepts Used in the World Inequality Database", technical report, 2024. [Online]. Disponível em: <https://wid.world/methodology/>
- [10] M. Medeiros, P. H. G. F. Souza, e F. A. Castro, "O Topo da Distribuição de Renda no Brasil: Primeiras Estimativas com Dados Tributários e Comparação com Pesquisas Domiciliares (2006–2012)", *Dados*, v. 58, n. 1, p. 7–36, 2015, doi: 10.1590/00115258201537.
- [11] M. Morgan, "Extreme and Persistent Inequality: New Evidence for Brazil Combining National Accounts, Surveys and Fiscal Data, 2001–2015", *WID.world Working Paper*, n. 2017/12, 2017, [Online]. Disponível em: <https://wid.world/document/extreme-persistent-inequality-new-evidence-brazil-combining-national-accounts-surveys-fiscal-data-2001-2015-wid-world-working-paper-201712/>
- [12] P. H. G. F. Souza, *Uma História de Desigualdade: A Concentração de Renda entre os Ricos no Brasil, 1926–2013*. São Paulo: Hucitec, 2018.