The Journal of Open Source Software

genderBR: Predicting gender from Brazilian first names

Fernando Meireles[1]

1 Instituto de Estudos Sociais e Politicos, Universidade do Estado do Rio de Janeiro, Brazil

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

Summary

Quantitative social science research frequently relies on large-scale administrative or scraped datasets that lack demographic indicators – including gender. While inferring gender from first names is often used to bridge this gap, existing tools mostly target Anglophone populations and use non-public, commercial data and methods. `genderBR` is an R package for inferring binary gender probabilities from Brazilian first names using official data from the 2010 and 2022 Brazilian censuses. The package offers a fast offline mode via an internal dataset and an online mode that queries IBGE's API and supports state-level filters, allowing calibration to temporal and regional variations in name–gender associations. By using census microdata, `genderBR` thus enables academics, journalists, and others to obtain binary gender probabilities to study aggregate demographic groups in a reproducible manner.

Statement of need

A common challenge in quantitative social science is the lack of demographic information in large-scale datasets. For instance, researchers frequently utilize administrative records and other data sources that do not contain gender information to study topics as diverse as elections and descriptive representation (Colner 2025; Lucas et al. 2021; Warshaw, Benedictis-Kessner, and Velez 2022); labor market inequalities (Banerjee et al. 2025; Wu 2020); academic publishing patterns and career trajectories (Dion, Sumner, and Mitchell 2018; Gaule and Piacentini 2018; Hagan et al. 2020; Liu et al. 2023; Mulders, Hofstra, and Tolsma 2024); the effects of gender composition of jury trials (Flanagan 2018); among many others. As these datasets often include first names, scholars often resort to imputing gender based on first names, exploiting naming conventions that correlate with a binary gender classification.

Despite these limitations, obtaining gender labels from first names remains a popular method in the absence of direct measures. For example, solutions such as Gender-API (n.d.), Namsor (n.d.), and Genderize (n.d.) offer commercial APIs that allow users to query first names and obtain gender predictions based on proprietary datasets. In the open-source domain, packages such as `gender` (Mullen 2016; see also Shi and Tong 2025; Tzioumis 2018) rely on historical datasets from the United States or Europe, such as name counts by gender from the US Social Security Administration, to predict gender. While useful for Anglophone contexts, these tools are not well-suited for other populations due to differences in naming conventions. In a comparative assessment, VanHelene et al. (2024) show that these methods are highly accurate for Western populations, with both Gender-API (n.d.) and Genderize (n.d.) achieving accuracies exceeding 98% in classifying a global dataset of mostly Western names. However, performance dropped below 82% for names from South Korea, China, Singapore, and Taiwan. The open-source alternative,

the `gender` package (Mullen 2016), achieved an even lower overall accuracy, of only 79.8% using the IPUMS method and 85.7% using US Social Security Administration data. In a different study, Santamaría and Mihaljević (2018) corroborates these findings for Asian populations, showing also that, even among Western names, commercial solutions can leave as many as 20% of names unclassified due to limited name coverage.

`genderBR` addresses these gaps by offering an open-source solution tailored to a particular non-Anglophone country – Brazil, the world's largest Portuguese-speaking country with over 200 million people. First released in 2017, when few similar tools existed, the package is grounded in Brazilian Institute of Geography and Statistics (IBGE) census microdata (Instituto Brasileiro de Geografia e Estatística 2023), which provides standardized counts of first-name occurrences by gender as recorded in the census – coded as a binary classification based on self-reports. The package supports both national and state-level predictions for 2010 and 2022 via IBGE's public API. Unlike API-based solutions that rely on private or non-representative scraped data, `genderBR` leverages official census data to provide broad coverage of Brazilian names[1], allowing accurate and reproducible gender predictions for Brazilian populations.

Software Design

`get_gender` is the package's core function: it cleans names (converts to lowercase, strips accents, and keeps only the first token that in Portuguese corresponds to the first name), aggregates duplicates to reduce computation time, and returns either binary labels or female-use probabilities. Users can switch between census years (`year = 2010` or `year = 2022`), request probabilities (`prob = TRUE`), tune decision threshold (`threshold`, default is 0.9 to maximize accuracy), and choose the data source. When `internal = TRUE` (the default) and no state is provided, results come from a bundled probability table (`nomes`), enabling fully offline analyses with the R package `data.table` (Barrett et al. 2025) as backend for fast joining operations. This internal data is stored as compressed `.Rda` to minimize package size while ensuring availability. Conversely, when a state is supplied or `internal = FALSE`, the function queries IBGE's API (with rate-limit handling), which obtains state-level data needed for specific regional predictions. This design choice avoids inflating the package size with state-level microdata for every census year. `internal`, thus, triggers different internal modules, non-exposed to users, depending on the requested functionality.

To make it easier to use `genderBR` on large datasets, the internal table with all female-by-name probabilities is available as a standalone `data.frame` named `nomes`, which users can join to their own data via first names. The distribution of unique names per census year can be seen in Table 1.

Table 1: Unique first names covered in the package's dataset per census year.

Year	Unique names
2010	125,294
2022	123,733
Unique (2010 & 2022)	141,742

[1] 1

Classification algorithm

After installing and loading the package, users can call the main function as follows. To obtain national predictions with binary labels for a vector of names, user can pass a vector of Brazilian first or full names, with both accented and unaccented characters and mixed case, to the `get_gender` function:

```
# Return labels
name_vec <- c("João", "Ana Maria da Silva", "alex")
get_gender(name_vec)
```

```
## [1] "Male"   "Female" "Male"
```

```
# Return probabilities of given names being female
get_gender(name_vec, prob = TRUE)
```

```
## [1] 0.005753 0.995264 0.007937
```

To obtain the probabilities calibrated to specific period and Brazilian states, users can provide a `state` vector with the corresponding two-letter state abbreviations (e.g., "RJ" for Rio de Janeiro), in which case the function will query IBGE's API directly, instead of using the internal data, to calculate name-by-gender-state probabilities.

Internally, `genderBR` computes $p_{\text{female}} = \frac{f}{f+m}$ for each name (and state, when provided), where f and m are IBGE female and male counts, respectively. The returned label applies the rule: female if $p_{\text{female}} > \tau$, male if $p_{\text{female}} < 1 - \tau$, and unknown otherwise, with the default $\tau = 0.9$. Names not found in the census data are returned as missing values (`NA`). This decision threshold is deterministic, but users can tune τ to trade off precision and coverage in their applications.

Validation

As a validation exercise, I test the package's accuracy on a dataset of 463,681 candidates from Brazil's 2024 municipal elections. Their self-reported binary gender data were obtained from official electoral records using the `electionsBR` R package (Meireles, Silva, and Costa 2016). This data includes the full names and reported gender for candidates running for mayor or city councilor positions across all Brazilian municipalities. In particular, I use `genderBR` to obtain the probabilities of each candidate being female. To classify names into gender labels, I adopt varying thresholds from 0.5 (most inclusive) to 0.9 (most strict).

As summarized in Table 2, `genderBR` achieves high accuracy across all thresholds, never falling below 99%, for covered names (considering only names that were classified as male or female). Additionally, the method obtains almost complete coverage, leaving as few as 3% of the names unclassified (when they were not found in the census data) and less than 2.5% as unknown (when the predicted probability fell within the uncertain range), even at the strictest threshold of 0.9.

Table 2: Validation results using 2024 municipal election candidates' dataset

Threshold	Accuracy (2010)	Accuracy (2022)	Unknown (2010)	Unknown (2022)	Unclassified (2010)	Unclassified (2022)
0.5	99.0%	99.0%	0.0%	0.0%	2.9%	2.9%
0.6	99.2%	99.2%	0.5%	0.5%	2.9%	2.9%
0.7	99.4%	99.4%	1.0%	0.9%	2.9%	2.9%
0.8	99.6%	99.5%	1.6%	1.6%	2.9%	2.9%
0.9	99.7%	99.7%	2.6%	2.5%	2.9%	2.9%

Research Impact

The `genderBR` package was released on CRAN in September 2017 and has since been downloaded over 41,000 times (as of January 2025).[2] In academia, the package has collected over 21 citations on Google Scholar in the author's profile[3], although the actual number of papers using the package is likely higher as it did not have a formal publication until now – and many researchers may have used it without proper citation. A search for the term "genderBR" on Google Scholar, for example, returns 81 entries, including working papers, preprints, and published articles[4]. Substantively, the package has been used in a variety of research areas, including economics, business, social sciences, scientometrics, and more (Borenstein, Perlin, and Imasato 2022; Corradini, Lagos, and Sharma 2025; Fanton et al. 2024; Mastella et al. 2021; Minasi, Mayer, and Santos 2022).

Ethical Considerations

While `genderBR` provides a fast and reproducible method for imputing gender from first names, it is crucial to acknowledge the implications of using this approach. By relying on a binary gender classification derived from naming conventions recorded at birth, the provided method is unable to differentiate between non-binary gender identities or changes in gender identity over time. In line with recommendations from similar packages (Mullen 2016), users should avoid using `genderBR` to impose binary classifications to individuals or in contexts where misclassification may lead to harm or discrimination against groups. Instead, the package should be regarded as an estimator for aggregate, large populations – to approximate the proportion of female partisan affiliates in the whole country, for example. In sum, `genderBR` should be viewed as a last resort tool when self-identified gender data is lacking and inferring it from first names does not pose risks to groups under study.

AI Usage Disclosure

From the package version 1.2.0 onwards, AI tools (GitHub Copilot with GPT-5.2 Codex Max) were used to assist in extending tests, updating documentation, and revising this paper. All code and text generated by AI tools were reviewed, edited, and validated by the author.

[2] 2
[3] 3
[4] 4

Acknowledgements

I thank the Brazilian Institute of Geography and Statistics for providing the public census name data that make this work possible, and the R and academic communities that contributed to the development of the package with feedback and suggestions over the years.

References

Banerjee, Rakesh, Tushar Bharati, Adnan MS Fakir, Yiwei Qian, and Naveen Sunder. 2025. "Gender Differences in Preferences for Flexible Work Hours: Experimental Evidence from an Online Freelancing Platform." *Labour Economics*, 102813.

Barrett, Tyson, Matt Dowle, Arun Srinivasan, Jan Gorecki, Michael Chirico, Toby Hocking, Benjamin Schwendinger, and Ivan Krylov. 2025. *Data.table: Extension of 'Data.frame'*. https://doi.org/10.32614/CRAN.package.data.table.

Borenstein, Denis, Marcelo S Perlin, and Takeyoshi Imasato. 2022. "The Academic Inbreeding Controversy: Analysis and Evidence from Brazil." *Journal of Informetrics* 16 (2): 101287.

Colner, Jonathan. 2025. "Running Toward Rankings: Ranked Choice Voting's Impact on Candidate Entry and Descriptive Representation." *American Journal of Political Science* 69 (3): 1010–28.

Corradini, Viola, Lorenzo Lagos, and Garima Sharma. 2025. "Collective Bargaining for Women: How Unions Can Create Female-Friendly Jobs." *The Quarterly Journal of Economics*, qjaf024.

Dion, Michelle L, Jane Lawrence Sumner, and Sara McLaughlin Mitchell. 2018. "Gendered Citation Patterns Across Political Science and Social Science Methodology Fields." *Political Analysis* 26 (3): 312–27.

Fanton, Marcos, Hugo Ribeiro Mota, Carolina de Melo Bomfim Araújo, Mitieli Seixas da Silva, and Raquel Canuto. 2024. "Philosophical Research in Brazil: A Structural Topic Modeling Approach with a Focus on Temporal and Gender Trends." *Metaphilosophy* 55 (3): 457–501.

Flanagan, Francis X. 2018. "Race, Gender, and Juries: Evidence from North Carolina." *The Journal of Law and Economics* 61 (2): 189–214.

Gaule, Patrick, and Mario Piacentini. 2018. "An Advisor Like Me? Advisor Gender and Post-Graduate Careers in Science." *Research Policy* 47 (4): 805–13.

Gender-API. n.d. "Gender-API: AI-Powered Gender Prediction from Names." https://gender-api.com/.

Genderize. n.d. "Genderize API Reference." https://genderize.io/.

Hagan, Ada K, Begüm D Topçuoğlu, Mia E Gregory, Hazel A Barton, and Patrick D Schloss. 2020. "Women Are Underrepresented and Receive Differential Outcomes at ASM Journals: A Six-Year Retrospective Analysis." *MBio* 11 (6): 10–1128.

Instituto Brasileiro de Geografia e Estatística. 2023. "Censo Demográfico 2022: API de Nomes." https://censo2022.ibge.gov.br/nomes/.

Liu, Fengyuan, Petter Holme, Matteo Chiesa, Bedoor AlShebli, and Talal Rahwan. 2023. "Gender Inequality and Self-Publication Are Common Among Academic Editors." *Nature Human Behaviour* 7 (3): 353–64.

Lucas, Jack, Reed Merrill, Kelly Blidook, Sandra Breux, Laura Conrad, Gabriel Eidelman, Royce Koop, Daniella Marciano, Zack Taylor, and Salomé Vallette. 2021. "Women's Municipal Electoral Performance: An Introduction to the Canadian Municipal Elections Database." *Canadian Journal of Political Science/Revue Canadienne de Science Politique* 54 (1): 125–33.

Mastella, Mauro, Daniel Vancin, Marcelo Perlin, and Guilherme Kirch. 2021. "Board Gender Diversity: Performance and Risk of Brazilian Firms." *Gender in Management:*

An International Journal 36 (4): 498–518.

Meireles, Fernando, Denisson Silva, and Beatriz Costa. 2016. *electionsBR: R Functions to Download and Clean Brazilian Electoral Data*. http://electionsbr.com/novo/.

Minasi, Sarah Marroni, Verônica Feder Mayer, and Glauber Eduardo de Oliveira Santos. 2022. "Desigualdade de gênero No Turismo: A Mulher No Ambiente Profissional No Brasil." *Revista Brasileira de Pesquisa Em Turismo* 16: e–2494.

Mulders, Anne Maaike, Bas Hofstra, and Jochem Tolsma. 2024. "A Matter of Time? Gender and Ethnic Inequality in the Academic Publishing Careers of Dutch Ph. Ds." *Quantitative Science Studies* 5 (3): 487–515.

Mullen, Lincoln. 2016. *Gender: Predict Gender from Names Using Historical Data*. https://cran.r-project.org/package=gender.

Namsor. n.d. "Namsor: The #1 AI for Name Origin, Ethnicity & Gender Detection." https://namsor.app/.

Santamaría, Lucía, and Helena Mihaljević. 2018. "Comparison and Benchmark of Name-to-Gender Inference Services." *PeerJ Computer Science* 4: e156.

Shi, Dongbo, and Sherry T Tong. 2025. "An Open Dataset of Chinese Name-to-Gender Associations for Gender Prediction in Broad Scientific Research." *Scientific Data* 12 (1): 1962.

Tzioumis, Konstantinos. 2018. "Demographic Aspects of First Names." *Scientific Data* 5 (1): 1–9.

VanHelene, Alexander D, Ishaani Khatri, C Beau Hilton, Sanjay Mishra, Ece D Gamsiz Uzun, and Jeremy L Warner. 2024. "Inferring Gender from First Names: Comparing the Accuracy of Genderize, Gender API, and the Gender r Package on Authors of Diverse Nationality." *PLOS Digital Health* 3 (10): e0000456.

Warshaw, Christopher, Justin de Benedictis-Kessner, and Yamil Velez. 2022. "Local Representation in the United States: A New Comprehensive Dataset of Elections." In *Local Representation in the United States: A New Comprehensive Dataset of Elections: Warshaw, Christopher/ Ude Benedictis-Kessner, Justin/ uVelez, Yamil*. [Sl]: SSRN.

Wu, Alice H. 2020. "Gender Bias Among Professionals: An Identity-Based Interpretation." *Review of Economics and Statistics* 102 (5): 867–80.