Monasterio, Leonardo Monteiro

Working Paper

Sobrenomes e ancestralidade no Brasil

Texto para Discussão, No. 2229

Provided in Cooperation with:
Institute of Applied Economic Research (ipea), Brasília

Suggested Citation: Monasterio, Leonardo Monteiro (2016) : Sobrenomes e ancestralidade no Brasil, Texto para Discussão, No. 2229, Instituto de Pesquisa Econômica Aplicada (IPEA), Brasília

This Version is available at:
https://hdl.handle.net/10419/146665

2229
TEXTO PARA DISCUSSÃO

SOBRENOMES E ANCESTRALIDADE NO BRASIL

Leonardo Monasterio

2229
TEXTO PARA DISCUSSÃO

Rio de Janeiro, setembro de 2016

SOBRENOMES E ANCESTRALIDADE NO BRASIL[1]

Leonardo Monasterio[2]

1. Agradeço os comentários e as sugestões de Claudio Shikida, Pedro Herculano Souza, Rafael Osório, Daniel Franken e dos participantes do Seminário de História Econômica da Universidade da Califórnia em Los Angeles (Ucla). Quaisquer erros remanescentes são de minha inteira responsabilidade. A pesquisa contou com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) – Proc. no. BEX 2549/15-8.
2. Técnico de planejamento e pesquisa da Diretoria de Estudos e Políticas Regionais, Urbanas e Ambientais (Dirur) do Ipea; e professor do Programa de Pós-Graduação em Economia da Universidade Católica de Brasília. *E-mail*: <leonardo.monasterio@ipea.gov.br>.

Texto para Discussão

Publicação cujo objetivo é divulgar resultados de estudos direta ou indiretamente desenvolvidos pelo Ipea, os quais, por sua relevância, levam informações para profissionais especializados e estabelecem um espaço para sugestões.

Texto para discussão / Instituto de Pesquisa Econômica Aplicada.- Brasília : Rio de Janeiro : Ipea , 1990-

ISSN 1415-4765

1.Brasil. 2.Aspectos Econômicos. 3.Aspectos Sociais. I. Instituto de Pesquisa Econômica Aplicada.

CDD 330.908

JEL: J15; N36.

SUMÁRIO

SINOPSE

Este trabalho apresenta um método de classificação da ancestralidade dos sobrenomes dos brasileiros nas seguintes classes: ibérica, italiana, japonesa, alemã e leste europeia. A partir de fontes históricas diversas, montou-se uma base de dados da ancestralidade dos sobrenomes. Essas informações formam a base para a aplicação de algoritmos de classificação de *fuzzy matching* e de *machine learning* nos mais de 46 milhões de trabalhadores da Relação Anual de Informações Sociais (Rais) Migra de 2013. A imensa maioria (96,4%) dos sobrenomes únicos da Rais foi identificada com o processo de *fuzzy matching* e os demais com o método proposto por Cavnar e Trenkle (1994). A comparação dos resultados do procedimento com dados sobre estrangeiros no Censo Demográfico de 1920 e a distribuição geográfica dos sobrenomes não ibéricos reforçam a acurácia do procedimento.

Palavras-chave: imigração; ancestralidade; sobrenomes.

ABSTRACT

This paper presents a method for classifying the ancestry of Brazilian surnames based on historical sources. The information obtained forms the basis for applying fuzzy matching and machine learning classification algorithms to more than 46 million workers in five categories: Iberian, Italian, Japanese, German and East European. The vast majority (96.4%) of the single surnames were identified using a fuzzy matching and the rest using a method proposed by Cavnar and Trenkle (1994). A comparison of the results of the procedures with data on foreigners in the 1920 Census and with the geographic distribution of non-Iberian surnames underscores the accuracy of the procedure.

Keywords: immigration; ancestry; surnames.

1 INTRODUÇÃO

As pesquisas censitárias oficiais no Brasil não cobrem informações sobre a ancestralidade da população. Tradicionalmente, são utilizadas apenas cinco categorias de cor/raça: preto, branco, pardo, amarelo e índio. Apesar de essas categorias possuírem sentido social, muitas vezes são por demais amplas para aplicações específicas, como estudos epidemiológicos ou socioeconômicos.

A contribuição deste *Texto para Discussão* reside em classificar a ancestralidade dos sobrenomes dos brasileiros. O estudo também inova ao utilizar bancos de dados históricos para associar os sobrenomes à ancestralidade e ao aplicar algoritmos de *machine learning* à classificação. Para analisar a distribuição contemporânea dos sobrenomes, o estudo recorreu à Relação Anual de Informações Sociais (Rais) Migra de 2013, em sua versão identificada. A base de dados contém 46,8 milhões de observações sobre brasileiros no mercado de trabalho formal com informações demográficas completas e nomes dos indivíduos.

Há toda uma literatura sobre classificação de sobrenomes e aplicações. Em relação aos algoritmos de classificação, destacam-se os de Cavnar e Trenkle (1994), Mateos (2007), Florou e Konstantopoulos (2011) e Komahan e Reidpath (2014). Já no tocante às aplicações da classificação de sobrenomes, há trabalhos importantes na área de epidemiologia (Lakha, Gorman e Mateos, 2011; Petersen *et al.*, 2011), mas também na ciência política (Dancygier, 2014) e na antropologia (Susewind, 2015). Há estudos mais recentes que utilizam os nomes para questões sociais de longo prazo: Rodríguez-Díaz, Manni e Blanco-Villegas (2015) examinam a distribuição espacial de sobrenomes; Güell, Mora e Telmer (2014), a mobilidade social; enquanto Carneiro, Lee e Reis (2015) usam os primeiros nomes para estudar a assimilação dos imigrantes nos Estados Unidos. Não há registros de estudos brasileiros voltados à classificação de sobrenomes.

Optou-se por classificar apenas cinco ancestralidades dos grupos imigrantes para o Brasil: ibéricos (ou seja, portugueses e espanhóis); italianos; germânicos; europeus orientais; e japoneses.[1] Esses foram os principais países de origem dos imigrantes que chegaram ao Brasil a partir de 1872 (Levy, 1974).

1. O Brasil também recebeu números consideráveis de imigrantes sírios e libaneses, mas, neste *Texto para Discussão,* o foco está nas categorias de imigrantes mais tradicionalmente analisadas.

No caso brasileiro, os métodos de sobrenome não são apropriados para identificar a ancestralidade indígena ou africana. Esses grupos adotaram, ou melhor, foram forçados a adotar sobrenomes ibéricos. Isso faz com que o termo “ancestralidade” se refira, neste trabalho, à ancestralidade do sobrenome e não do indivíduo. Contudo, aqueles dois grupos correspondem, *grosso modo,* às classificações de índio, preto e pardo utilizadas pelo Instituto Brasileiro de Geografia e Estatística (IBGE).[2] Portanto, o método aqui proposto deve ser visto como complemento e não como substituto da classificação padrão de cor/raça. Conforme será visto adiante, existem diferenças marcantes de salário e escolaridade entre as classes de ancestralidade que teriam ficado ocultas caso a análise se limitasse às categorias oficiais de cor/raça.

2 MATERIAL

As principais fontes históricas de ancestralidade dos migrantes foram os registros do Museu da Imigração do Estado de São Paulo (obtidos por *web scrapping*) e os microdados dos censos históricos norte-americanos. Além disso, foram consideradas outras fontes históricas e bancos de dados com listas de sobrenomes. A relação completa de fontes consta da tabela 1. Como se pode ver, construiu-se uma base com quase 5 milhões de registros.

TABELA 1
Descrição das fontes de dados sobre nomes e nacionalidade

Fonte	Nacionalidade	Número de observações
North Atlantic Population Project (microdados dos censos norte-americanos – imigrantes nos Estados Unidos de 1880 e 1910). Disponível em: <https://www.nappdata.org/napp/>. Acesso em: 27 abr. 2016.	Diversas	4.715.496
Museu da Imigração do Estado de São Paulo. Disponível em: <http://museudaimigracao.org.br/acervodigital/livros.php>. Acesso em: 27 abr. 2016.	Diversas	186.193
Sobrenomes japoneses populares.[1] Disponível em: <http://www.ipc.shizuoka.ac.jp/~jjksiro/sei.csv>.	Japonesa	4.000
Heraldica de Apellidos Españoles. Disponível em: <http://www.surnames.org/apellidos/lista.htm>. Acesso em: 27 abr. 2016.	Ibérica	5.138
Banco de sobrenomes de imigrantes italianos para o processo de cidadania italiana. Disponível em: <http://www.santanacidadania.com.br/banco-de-nomes/>.	Italiana	6.573
Emigrazione Veneta. Disponível em: <http://www.emigrazioneveneta.com/>. Acesso em: 27 abr. 2016.	Italiana	60.889
Total	**Diversas**	**4.978.289**

Elaboração do autor.
Nota: [1] Os sobrenomes estavam em caracteres Kanji e foram convertidos para o alfabeto romano pelo Kanji Converter. Disponível em: <http://nihongo.j-talk.com/>. Acesso em: 27 abr. 2016.

2. Ver Piza e Rosemberg (1999) e Osório (2004) para a questão da cor e raça nos censos brasileiros.

Essa base, conforme esperado, contém – especialmente no caso dos nomes dos imigrantes dos censos norte-americanos – erros de transcrição, escrita ou digitação que prejudicariam a precisão dos algoritmos. Graças ao volume inicial de informação, foi possível seguir critérios rigorosos que excluíram sobrenomes que apareciam apenas muito raramente nos microdados do censo.[3] A remoção das repetições, das nacionalidades não analisadas e dos erros resultou em 71.404 pares de sobrenomes-ancestralidade. Essa base será o ponto de partida do processo de *fuzzy matching* e *machine learning* das seções que se seguem. A tabela 2 apresenta a distribuição dos sobrenomes de acordo com a ancestralidade.

TABELA 2
Distribuição do banco de dados de referência sobrenomes-ancestralidade

Ancestralidade	Sigla	Número de observações
Ibérica	IBR	9.782
Italiana	ITA	26.238
Germânica	GER	22.507
Japonesa	JPN	5.376
Leste europeia	EAS	7.501
Total		**71.404**

Elaboração do autor.

Já em relação aos nomes contemporâneos, foram obtidos os dois sobrenomes de cada indivíduo a partir da base de dados da Rais. Como se sabe, a maior parte das pessoas no Brasil possui dois sobrenomes herdados da mãe e do pai, nessa ordem. Além disso, nubentes podem adotar (ou não) os sobrenomes da outra parte após o casamento, mas, tradicionalmente, as mulheres substituem o sobrenome da mãe pelo do marido. Para não se perder informação, optou-se por considerar os dois sobrenomes de cada um dos indivíduos (quando existentes).[4] Isso resultou em cerca de 530 mil sobrenomes únicos.

3. Quando, nas fontes históricas, um mesmo sobrenome foi atribuído a mais de uma categoria, adotou-se o seguinte critério: *i*) no mesmo banco de dados, preservar o nome que tenha mais que 90% de registros associados a uma mesma ancestralidade; e *ii*) dar prioridade à informação do Museu de Imigração de São Paulo.

4. A utilização dos dois sobrenomes diferencia este trabalho dos demais. Foram excluídas as partículas não significativas (de, das, Filho, entre outras) e os primeiros nomes. No apêndice são fornecidos mais detalhes sobre o procedimento de identificação dos primeiros nomes (inclusive os compostos).

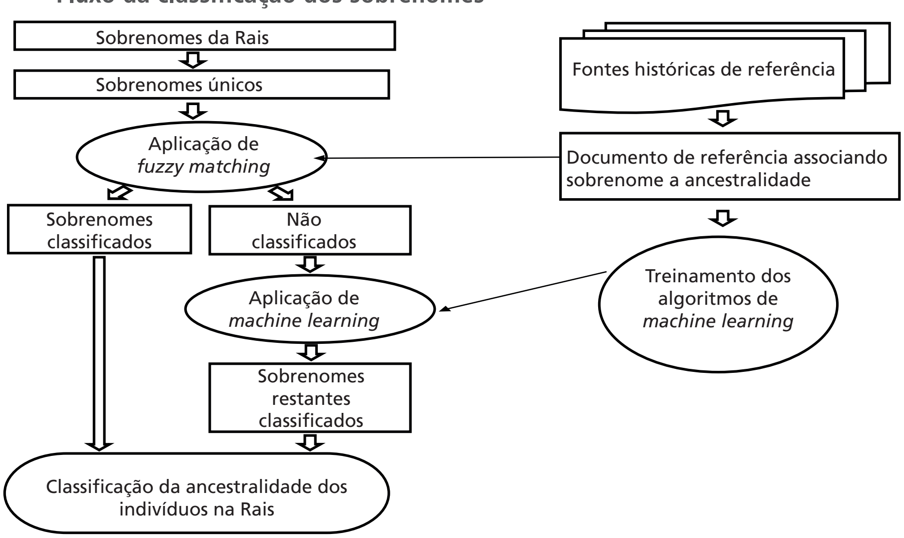
Os cinco sobrenomes mais frequentes no Brasil (Silva, Santos, Oliveira, Souza e Pereira) somam mais de 21,0 milhões dos 46,8 milhões de registros da Rais, ou seja, 45% dessa base. Já aqueles que aparecem apenas uma vez totalizam mais de 204 mil registros. Apesar da qualidade geral dos dados da Rais, especialmente em anos recentes, a simples inspeção visual indica que boa parte desses casos é erro de digitação, quer na Rais, quer em seus documentos de identidade. As técnicas apresentadas na próxima seção buscam superar esse problema.

3 MÉTODOS

3.1 Visão geral

O esquema geral está apresentado na figura 1. A partir dos documentos históricos e atuais, cria-se um documento de referência que associa sobrenomes à ancestralidade. Este arquivo tem função dupla: servir de base para o *fuzzy matching* e para a aplicação dos dois algoritmos de *machine learning*.

FIGURA 1
Fluxo da classificação dos sobrenomes



Elaboração do autor.

Conforme indica a figura, faz-se primeiro a correspondência entre o arquivo de referência e os sobrenomes únicos mediante *fuzzy matching*. Nem todos os nomes serão classificados nesta fase, pois muitos não têm similaridade suficiente com os do arquivo de referência. Assim, apenas os sobrenomes sem correspondência serão alvo da aplicação dos algoritmos de *machine learning,* os quais foram treinados com base no mesmo arquivo de referência já citado. Uma vez que todos os sobrenomes únicos estiverem associados a uma ancestralidade, volta-se à Rais para classificar os sobrenomes de todos os indivíduos. As subseções seguintes detalham os passos.[5]

3.2 *Fuzzy matching*

Os métodos de *fuzzy matching* permitem que duas sequências de caracteres – *strings* – sejam associadas mesmo quando não são idênticas. Nesta aplicação, foram escolhidas técnicas bastante conservadoras para impedir correspondências equivocadas entre os sobrenomes da base de referência e os que constam da Rais. Utilizou-se o critério de *optimal string alignment* (OSA) quando a distância entre duas *strings* era igual ou menor que um valor preestabelecido. Por distância entende-se o número de mudanças (inserções, exclusões ou trocas de posição) necessárias para que as *strings* sejam idênticas. Por exemplo, os sobrenomes Mueller e Miller são *fuzzy matched* pelo critério OSA com d igual a 2. Já com o valor de d igual a 1, como se escolheu neste trabalho, não haveria correspondência. Vale ressaltar que o algoritmo primeiro busca as correspondências perfeitas – com distância igual a zero –, para só em seguida buscar a equivalência quando as distâncias são maiores. Assim, Sousa e Souza, se constarem na base de referência, serão associados aos sobrenomes de indivíduos que tenham esses nomes.

3.3 Algoritmos de *machine learning*

3.3.1 Cavnar e Trenkle

O método de categorização textual proposto por Cavnar e Trenkle (1994) é simples, exige pouco esforço computacional e tem sido utilizado em diversas aplicações de classificação de línguas e sobrenomes. O algoritmo pode ser sintetizado nos passos descritos a seguir.

5. Todas as análises foram feitas com o *software* R 3.2.2 (R Core Team, 2015), com o auxílio dos seguintes pacotes: ngramrr (Chan, 2016); textcat (Hornik *et al.*, 2016); RTextTools (Jurka *et al.*, 2014); e1071 (Meyer *et al.*, 2015); e caret (Kuhn *et al.*, 2016). Agradeço a Chung-Hong Chan, um dos criadores do *software* ngramrr, por atender ao meu pedido de alterar o pacote.

1) Criar *n-grams* do conjunto de palavras com categorização já definida (o chamado *training set*). *N-grams* são sequências com comprimento *n* formadas a partir das palavras. Por exemplo, o sobrenome Lima teria os seguintes 3-*grams*: _ _ L; _LI; LIM; IMA; MA_; e A_ _.[6] No presente caso, isso significa formar *n-grams* para todos os sobrenomes associados a cada uma das ancestralidades.

2) Montar uma tabela com a frequência em ordem decrescente de *n-grams* para cada uma das categorias. Assim, os *n-grams* com *ranking* mais elevado são os mais frequentes em cada uma das categorias. Essa tabela de perfis serve de base para a classificação de novos nomes em categorias.

3) Criar *n-grams* dos novos sobrenomes que são comparados com cada um dos *rankings* das categorias criadas no passo 2.

4) Calcular a "distância total" ("*out of place measure*", nas palavras dos autores) entre a palavra e o *ranking* de *n-grams* de cada categoria. A distância total é o somatório da posição que os *n-grams* dos nomes fora do *training set* ocupam no *ranking* de perfis obtidos no passo 2.

5) Atribuir como categoria do novo sobrenome aquela em que haja menor distância total em relação aos perfis. Na implementação do algoritmo de Cavnar e Trenkle (1994), é preciso definir o valor de *n* dos *n-grams*, bem como quantos desses elementos comporão o *ranking*. Seguindo a literatura, optou-se por considerar o valor de *n* igual a 3 na construção dos *n-grams* e foram considerados os 1.250 *n-grams* mais frequentes. Esses valores foram escolhidos por fornecerem a maior acurácia na classificação.

3.3.2 Naive Bayes

O algoritmo de classificação Naive Bayes também se baseia no padrão de distribuição dos *n-grams* de acordo com as categorias. É o método mais usado para classificação porque, a despeito de sua simplicidade, tende a mostrar resultados surpreendentemente acurados. A intuição se baseia na aplicação da regra de Bayes para se obter a probabilidade de um sobrenome pertencer a uma classe condicionada pelos seus *n-grams*. Em termos formais, considerando o conjunto de sobrenomes *s* associados às classes *y*, a classificação ocorre da seguinte forma:

$$c = argmax\ P(y|s)\ P(s).$$

6. O acréscimo dos espaços em branco no começo e no final das palavras serve para preservar a informação sobre a posição das letras nessas localizações quando são formados os *n-grams*.

Ou seja, cada sobrenome é classificado *a posteriori* de acordo com a maior probabilidade. A distribuição *a priori* é normalmente baseada na participação das classes nos dados utilizados para o treinamento.

O caráter "inocente" do método Naive Bayes refere-se ao pressuposto de que as probabilidades dos *n-grams* são independentes entre si. Por exemplo, as probabilidades de ocorrência dos 3-*grams* SIL e LVA seriam independentes. Isso é obviamente falso. Mesmo assim, o método Naive Bayes apresenta um desempenho notável em aplicações de classificação de texto, mesmo quando comparado com outros métodos que dispensam a hipótese de independência (Eyheramendy, Lewis e Madigan, 2003).

4 RESULTADOS

4.1 Resultados de *fuzzy matching*

O processo de *fuzzy matching* permitiu que a imensa maioria dos indivíduos fosse classificada. Apesar de apenas 293.634 dos 531.009 sobrenomes únicos da Rais terem sido identificados, isso corresponde a 96,4% do total dos indivíduos da Rais. Isso se explica pois os nomes identificados são bem mais populares que os não identificados. Note-se que esse resultado foi obtido mesmo com a prática conservadora de se utilizar a distância máxima de 1 no algoritmo de OSA.

Os sobrenomes dos indivíduos da Rais que não foram classificados por *fuzzy matching* o serão pelos algoritmos de *machine learning*. Mesmo sendo uma porcentagem relativamente pequena de não classificados (3,6%), vale a pena aplicar tais métodos quando o objetivo for a identificação dos imigrantes não ibéricos, pois estes podem estar sobrerrepresentados naquele grupo.

4.2 Acurácia dos procedimentos de *machine learning*

O primeiro passo consiste em dividir os dados no conjunto de treinamento e no de teste. Neste estudo, a partir do arquivo de referência com 71.404 pares de sobrenome e ancestralidade, separou-se um arquivo de teste com 20% dos dados. Os algoritmos calibrados serão aplicados nesse conjunto para que a acurácia possa ser estimada.

4.2.1 Acurácia do procedimento de Cavnar e Trenkle

A partir do conjunto de dados de treinamento, a rotina baseada em Cavnar e Trenkle (1994) criou o perfil de cada uma das nacionalidades com base nos 1.250 *n-grams* (*n* igual a 3) mais frequentes. A tabela 3 mostra a matriz de erros resultante da aplicação do algoritmo ao conjunto de dados de teste.

TABELA 3
Matriz de erros – valores observados e previstos pelo procedimento de Cavnar e Trenkle

		Previsto				
		EAS	GER	IBR	ITA	JPN
Observado	EAS	0,69	0,13	0,11	0,04	0,03
	GER	0,07	0,85	0,04	0,02	0,01
	IBR	0,04	0,05	0,70	0,18	0,04
	ITA	0,01	0,02	0,15	0,79	0,02
	JPN	0,01	0,00	0,02	0,00	0,96

Elaboração do autor.

Como se vê, apesar da simplicidade, o algoritmo Cavnar e Trenkle (1994) se mostrou razoavelmente preciso. No caso dos sobrenomes japoneses (JPN), 96% dos dados de teste foram corretamente classificados. Já o pior resultado foi obtido com os sobrenomes do Leste Europeu (EAS): apenas 69% dos sobrenomes dessa categoria foram classificados como tal. A acurácia estimada foi igual a 80,6%.[7] Já o valor kappa, ou seja, a porcentagem de classificações corretas levando em conta a possibilidade de acerto aleatório, foi igual a 0,73.[8]

4.2.2 Acurácia do procedimento de Naive Bayes

Com certa surpresa, os resultados da aplicação do algoritmo Naive Bayes foram piores que os obtidos com Cavnar e Trenkle (1994). Apesar de, no caso dos germânicos, aquele algoritmo ter sido mais preciso que este, na maior parte dos casos o procedimento bayesiano foi bem pouco acurado nas demais nacionalidades. Os erros mais graves ocorreram na classificação dos sobrenomes do Leste Europeu, dos quais a metade foi classificada como se fosse germânica. Igualmente, o algoritmo de Naive Bayes classificou erroneamente 45% dos sobrenomes ibéricos como italianos.

7. A acurácia é definida como o total de previsões corretas dividido pelo número de observações.
8. Em termos formais, *kappa* = (*acurácia observada* - *acurácia esperada*)/(1 - *acurácia esperada*).

TABELA 4
Matriz de erros – valores observados e previstos pelo procedimento de Naive Bayes

		Previsto				
		EAS	GER	IBR	ITA	JPN
Observado	EAS	0,35	0,42	0,05	0,16	0,02
	GER	0,01	0,86	0,02	0,11	0,00
	IBR	0,00	0,25	0,29	0,45	0,00
	ITA	0,00	0,14	0,07	0,78	0,00
	JPN	0,00	0,14	0,03	0,08	0,74

Elaboração do autor.

A acurácia foi de 69% e o valor kappa foi de 56%, ambos inferiores ao obtido com o método de Cavnar e Trenkle (1994). Outros artigos que aplicaram o Naive Bayes para a classificação de sobrenomes em outros países conseguiram acurácia bem mais elevada que a obtida neste estudo (Komahan e Reidpath, 2014). Talvez as peculiaridades dos sobrenomes brasileiros ou a semelhança linguística entre os sobrenomes das classes analisadas tenham dificultado a classificação. Vale lembrar que, a despeito de os algoritmos de classificação terem tido um desempenho aquém do usual, seus erros de classificação só se aplicam a 3,6% dos indivíduos na Rais.

4.3 Resultados da classificação

Dada a superioridade do algoritmo de Cavnar e Trenkle (1994) em relação ao Naive Bayes, considerou-se apenas o primeiro método na representação dos resultados. Em termos de classificação dos sobrenomes únicos presentes na Rais, o resultado obtido consta da coluna *porcentagem dos sobrenomes únicos* da tabela 5. Essas porcentagens não refletem a ancestralidade da população, pois a concentração de sobrenomes varia dentro de cada ancestralidade. Mais interessante é examinar as colunas *número de indivíduos e porcentagem dos indivíduos*, que apresentam a distribuição da população por ancestralidade.

TABELA 5
Ancestralidade do sobrenome estimada pelo último ou pelo único sobrenome

Ancestralidade do sobrenome	Sigla	Porcentagem dos sobrenomes únicos	Número de indivíduos	Porcentagem dos indivíduos
Ibérica	IBR	27,6	40.969.034	87,5
Italiana	ITA	31,7	3.594.043	7,7
Germânica	GER	21,7	1.525.890	3,3
Leste europeia	EAS	12,6	396.880	0,8
Japonesa	JPN	6,5	315.925	0,7
Total		**100,0**	**46.801.772**	**100,0**

Elaboração do autor.

4.3.1 Análise de cor *versus* previsão

Uma forma imediata de avaliar a validade do procedimento de classificação seria, a princípio, examinar a consistência da classificação de ancestralidade estimada com o dado de cor/raça individual na Rais. Por exemplo, um bom sinal seria se nenhum indivíduo classificado como de ancestralidade oriunda do Leste Europeu fosse identificado na Rais como nativo brasileiro, ou aqueles com sobrenome japonês estivessem identificados como amarelos. Infelizmente, a possibilidade de fazer tal comparação é bastante limitada.

A qualidade da informação da Rais sobre cor é muito baixa. Osório (2004, p. 47) afirma que os dados são providos pelo departamento de recursos humanos das empresas e que, especialmente naquelas com filiais, isso se dá de forma centralizada. Ainda segundo o autor, esse fato leva à baixa precisão da informação sobre cor/raça. De fato, em 23% dos registros da Rais houve a omissão dessa informação.

Uma forma simples de avaliar a qualidade da informação de cor na Rais consiste em verificar se os indivíduos com nomes tipicamente orientais foram classificados como amarelos no banco de dados. Tome-se Tanaka, o mais frequente sobrenome oriental na Rais. Há 1.225 homens com esse sobrenome, mas apenas 323 (ou seja, 26%) constam como amarelos; 46% dos Tanakas constam como brancos. Mesmo considerando somente aqueles que têm um sobrenome Tanaka e outro classificado como japonês pelo algoritmo, 42% constam como brancos na Rais. Talvez estes indivíduos realmente não se considerem descendentes de orientais, ou talvez seja uma imprecisão dos responsáveis na empresa pela entrada de dados. De qualquer forma, esse fato sugere que a informação sobre raça na Rais não é um bom indicador de ancestralidade.[9]

4.3.2 Análise da distribuição geográfica dos sobrenomes

Foi criado um índice de ancestralidade imigrante para fins de síntese. Caso o indivíduo tenha exclusivamente sobrenomes ibéricos, o índice recebe o valor 0. Se tiver exclusivamente sobrenomes não ibéricos, o valor 1 é atribuído ao índice. Se um dos sobrenomes for ibérico e o outro não (sem importar a ordem), o valor é igual a 0,5.

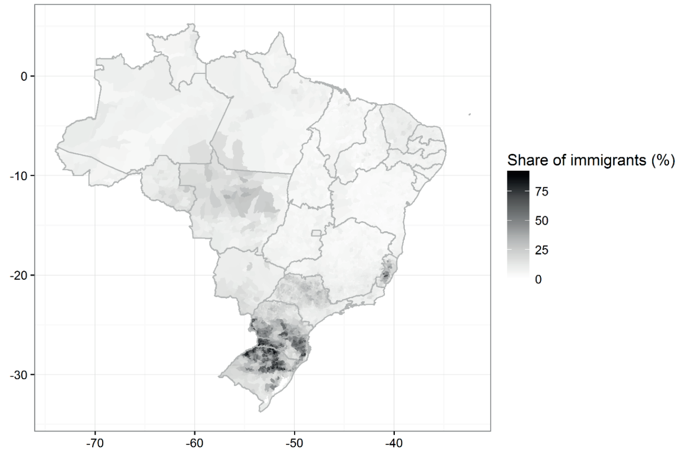
A figura 2 representa a distribuição do índice pelo território brasileiro. A classificação está de acordo com a literatura sobre imigração estrangeira para o Brasil (Levy, 1974). Os municípios com maior parcela de não ibéricos estão concentrados: *i*) nas áreas de destino de colonização da região Sul, do Espírito Santo e do oeste paulista até 1920; e *ii*) nas áreas de expansão da fronteira agrícola de Rondônia, Mato Grosso e Mato Grosso do Sul, que receberam migrantes oriundos daquela região durante as três últimas décadas. Isso sugere que o processo de classificação baseado em *fuzzy matching*, complementado pelo algoritmo de Cavnar e Trenkle (1994), identificou a ancestralidade dos trabalhadores na Rais.

Outra forma interessante de averiguar a validade externa do algoritmo é compará-la com os dados sobre estrangeiros no Censo Demográfico de 1920, quando já havia terminado a grande onda de imigração para o Brasil. Os gráficos 1A, 1B e 1C mostram a relação entre o número de estrangeiros em 1920 e o total dos que foram identificados como descendentes de germânicos, japoneses e italianos em 2013 por estado.[10] A correlação mostra que – apesar das fortes migrações internas ocorridas no período, inclusive a expansão da fronteira agrícola – os descendentes tendem a residir no mesmo estado que seus ancestrais. Esse resultado também dá mais confiança, ao menos no nível agregado, ao procedimento de classificação de sobrenomes.

9. De forma complementar, a pouca correspondência entre a ancestralidade de sobrenomes e a raça pode ter explicações sociológicas. A literatura da área aponta que no Brasil, ao contrário dos Estados Unidos, a cor é definida por aparência e não por ancestralidade (Muniz, 2012). Também já foi percebido que a identificação de cor dos filhos é influenciada pela condição socioeconômica dos pais (Schwartzman, 2007).

10. Utilizaram-se os limites estaduais existentes em 1920. Não foram apresentadas as correlações referentes aos ibéricos, pois, dado seu expressivo contingente, os gráficos apenas refletiriam a mudança na população estadual entre 1920 e 2013.

FIGURA 2
Estimativa da porcentagem de ancestralidade não ibérica



Elaboração do autor.

GRÁFICO 1
Correlação entre o número de estrangeiros residentes em 1920 e o número estimado de descendentes de germânicos, japoneses e italianos em 2013, por estado
1A – Imigrantes germânicos

1B – Imigrantes japoneses

1C – Imigrantes italianos

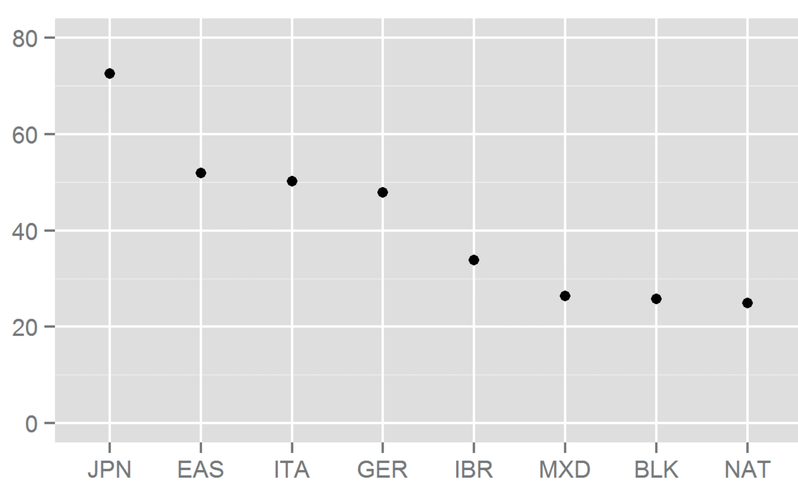
Elaboração do autor.
Obs.: Todos os gráficos estão em escala bilogarítmica.

4.3.3 Diferenciais de salário e escolaridade por ancestralidade

Para que se tenha uma visão abrangente dos trabalhadores brasileiros, fez-se a seguinte reclassificação combinando as informações oficiais de cor/raça e a estimada: utilizou-se a classe obtida pelos algoritmos de classificação de sobrenomes, exceto quando os indivíduos constavam na Rais como possuindo a cor/raça preta, parda e indígena. Nestes casos, foi respeitada a classificação original. Assim, foram obtidas oito categorias de ancestralidade/cor: ibérica (IBR), japonesa (JPN), italiana (ITA), germânica (GER), leste europeia (EAS), preta (BLK), parda (MXD) e indígena (NAT).

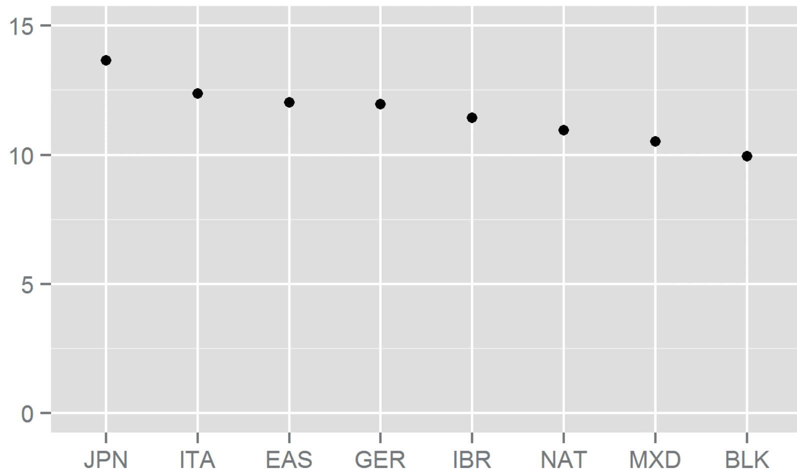
Os gráficos 2 e 3 têm caráter apenas descritivo e não causal. Foram selecionados apenas os brasileiros, entre 23 e 60 anos, que trabalhavam no setor privado com jornada de ao menos 40 horas por semana. Os menores níveis de salário e escolaridade dos trabalhadores brasileiros com cor/raça preta, parda e indígena são bem conhecidos. Mas os gráficos mostram também que indivíduos com sobrenomes não ibéricos têm salários substancialmente maiores que os brancos com sobrenomes ibéricos. Enquanto os com ancestralidade japonesa e leste europeia têm salários em média de R$ 73 e R$ 52 por hora, esse valor não chega a R$ 34 para os ibéricos. No caso da escolaridade, as diferenças são substanciais, mas com amplitude menor. Mais uma vez, os com ancestralidade japonesa e italiana alcançaram a média de 13,6 e 12,4 anos em relação a 11,4 anos dos ibéricos.

GRÁFICO 2
Salários médios por ancestralidade/cor
(Em R$ por hora)



Elaboração do autor.

GRÁFICO 3
Média de escolaridade por ancestralidade/cor
(Em anos)



Elaboração do autor.

Vários fatores econômicos, sociológicos e mesmo geográficos explicam as diferenças de salário e escolaridade, porém essa discussão foge do alcance deste trabalho. De qualquer forma, a classificação da ancestralidade dos sobrenomes revela diferenças que teriam ficado ocultas caso fossem utilizadas apenas as classificações de cor/raça usuais.

5 DISCUSSÃO E CONSIDERAÇÕES FINAIS

Este *Texto para Discussão* mostrou como fontes históricas de sobrenomes combinadas com técnicas contemporâneas de *fuzzy matching* e *machine learning* auxiliam a classificação da ancestralidade de indivíduos. O sistema se mostrou capaz de classificar os sobrenomes por ancestralidade. Apesar da acurácia dos algoritmos de Cavnar e Trenkle (1994) e, especialmente, de o Naive Bayes ter sido inferior ao esperado, isso não parece ter comprometido os resultados gerais. Afinal, a imensa maioria dos indivíduos foi classificada com base em *fuzzy matching,* e apenas 3,6% com o procedimento de Cavnar e Trenkle (1994). Mais ainda, a distribuição geográfica, quer no nível municipal, quer no estadual, reflete o conhecimento histórico e contemporâneo sobre a localização dos imigrantes e seu destino.

Mostrou-se que, no Brasil como um todo, apenas 18% dos indivíduos têm ao menos um sobrenome germânico, italiano, leste europeu ou japonês. Espacialmente, contudo, existem estados com alta concentração de sobrenomes não ibéricos, o que, em geral, coincide com os dados sobre estrangeiros no Censo Demográfico de 1920. Ficou claro também que a ancestralidade de sobrenome está associada a diferenças substantivas de salário e escolaridade.

Obviamente, não há garantias de que a classificação reflete precisamente a ancestralidade cultural ou genômica. Mesmo utilizando ambos os sobrenomes do indivíduo (quando houver), há perda de linhagem matrilinear e adoções, mudanças de nome no casamento, entre outros eventos, que podem fazer reduzir a precisão de tal indicador. Já para dados mais agregados, é provável que muitas dessas idiossincrasias se cancelem e a tendência geral seja mais precisa.

Existem estudos que analisam a autoidentificação de cor/raça e ancestralidade genética no Brasil (Lima-Costa *et al.*, 2015; Pena *et al.*, 2011; Santos *et al.*, 2009; Travassos e Williams, 2004). No futuro, a combinação desta literatura e de tais informações com a análise da ancestralidade dos sobrenomes dos brasileiros tem o potencial de iluminar questões sociais, econômicas e de saúde pública.

REFERÊNCIAS

CARNEIRO, P.; LEE, S.; REIS, H. **Please call me John**: name choice and the assimilation of immigrants in the United States, 1900-1930. London: Centre for Microdata Methods and Practice, 2015. (Cemmap Working Paper, n. 28/15). Disponível em: <http://www.ifs.org.uk/uploads/cemmap/wps/cwp281515.pdf>. Acesso em: 12 abr. 2016.

CAVNAR, W. B.; TRENKLE, J. M. N-gram-based text categorization. **Ann Arbor MI**, v. 48113, n. 2, p. 161-175, 1994.

CHAN, C-H. **ngramrr**: a simple general purpose n-gram tokenizer. [s.l.]: [s.n.], 2016. Disponível em: <https://cran.r-project.org/web/packages/ngramrr/index.html>.

DANCYGIER, R. M. Electoral rules or electoral leverage? Explaining Muslim representation in England. **World Politics**, v. 66, n. 2, p. 229-263, Apr. 2014.

EYHERAMENDY, S.; LEWIS, D.; MADIGAN, D. On the Naive Bayes model for text classification. *In*: INTERNATIONAL WORKSHOP ON ARTIFICIAL INTELLIGENCE AND STATISTICS, 9., 2003, Florida. **Anais**… New Jersey: Society for Artificial Intelligence and Statistics, 2003.

FLOROU, E.; KONSTANTOPOULOS, S. A quantitative and qualitative analysis of Nordic surnames. *In*: NORDIC CONFERENCE OF COMPUTATIONAL LINGUISTICS, 18., 2011, Riga. **Anais**... Stroudsburg: ACL, 2011. Disponível em: <http://users.iit.demokritos.gr/~konstant/dload/Pubs/nodalida11.pdf>. Acesso em: 2 mar. 2016.

GÜELL, M.; MORA, J. V. R.; TELMER, C. I. The informational content of surnames, the evolution of intergenerational mobility and assortative mating. **The Review of Economic Studies**, p. rdu041, 10 dez. 2014.

HORNIK, K. *et al.* **textcat**: n-gram based text categorization. [s.l.]: [s.n.], 2016. Disponível em: <https://cran.r-project.org/web/packages/textcat/index.html>.

JURKA, T. P. *et al.* **RTextTools**: automatic text classification via supervised learning. [s.l.]: [s.n.], 2014. Disponível em: <https://cran.r-project.org/web/packages/RTextTools/index.html>.

KOMAHAN, K.; REIDPATH, D. D. A. "Roziah" by any other name: a simple Bayesian method for determining ethnicity from names. **American Journal of Epidemiology**, p. kwu129, 2014.

KUHN, M. *et al.* **caret**: classification and regression training. [s.l.]: [s.n.], 2016. Disponível em: <https://cran.r-project.org/web/packages/caret/index.html>.

LAKHA, F.; GORMAN, D. R.; MATEOS, P. Name analysis to classify populations by ethnicity in public health: validation of Onomap in Scotland. **Public Health**, v. 125, n. 10, p. 688-696, 2011.

LEVY, M. S. O papel da migração internacional na evolução da população brasileira (1872 a 1972). **Revista de Saúde Pública**, São Paulo, v. 8, número suplementar, p. 49-90, 1974.

LIMA-COSTA, M. F. *et al.* Genomic ancestry, self-rated health and its association with mortality in an admixed population: 10 year follow-up of the Bambui-Epigen (Brazil) cohort study of ageing. **Plos One**, v. 10, n. 12, p. e0144456, 2015.

MATEOS, P. Classifying ethnicity using people's names. *In*: INTERNATIONAL CONFERENCE ON SOCIAL STATISTICS AND ETHNIC DIVERSITY, 2007, Montreal. **Anais**... Québec: QICSS; INED, 2007. Disponível em: <http://ciqss.umontreal.ca/Docs/SSDE/pdf/Mateos.pdf>. Acesso em: 6 mar. 2016.

MEYER, D. *et al.* **e1071**: misc functions of the Department of Statistics, probability theory group (formerly: E1071), TU Wien. [s.l.]: [s.n.], 2015. Disponível em: <https://cran.r-project.org/web/packages/e1071/index.html>.

MUNIZ, J. O. Preto no branco? Mensuração, relevância e concordância classificatória no país da incerteza racial. **Dados**, v. 55, n. 1, p. 251-282, 2012.

OSÓRIO, R. G. O sistema classificatório de cor ou raça do IBGE. *In*: BERNARDINO, J.; GALDINO, D. (Ed.). **Levando a raça a sério**: ação afirmativa e a universidade. Rio de Janeiro: Dp&A, 2004. p. 85-135.

PENA, S. D. *et al.* The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. **Plos One**, v. 6, n. 2, p. e17063, 2011.

PETERSEN, J. *et al.* Names-based classification of accident and emergency department users. **Health & Place**, v. 17, n. 5, p. 1162-1169, 2011.

PIZA, E.; ROSEMBERG, F. Cor nos censos brasileiros. **Revista USP**, n. 40, p. 122-137, 1999.

R CORE TEAM. **R**: a language and environment for statistical computing. [s.l.]: R Foundation for Statistical Computing, 2015. Disponível em: <https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf>.

RODRÍGUEZ-DÍAZ, R.; MANNI, F.; BLANCO-VILLEGAS, M. J. Footprints of middle ages kingdoms are still visible in the contemporary surname structure of Spain. **Plos One**, v. 10, n. 4, p. e0121472, 7 abr. 2015.

SANTOS, R. V. *et al.* Color, race, and genomic ancestry in Brazil: dialogues between anthropology and genetics. **Current Anthropology**, v. 50, n. 6, p. 787-819, 2009.

SCHWARTZMAN, L. F. Does money whiten? Intergenerational changes in racial classification in Brazil. **American Sociological Review**, v. 72, n. 6, p. 940-963, 1º dez. 2007.

SUSEWIND, R. What's in a name? Probabilistic inference of religious community from South Asian names. **Field Methods**, v. 27, n. 4, p. 319-332, 2015.

TRAVASSOS, C.; WILLIAMS, D. R. The concept and measurement of race and their relationship to public health: a review focused on Brazil and the United States. **Cadernos de Saúde Pública**, v. 20, n. 3, p. 660-678, 2004.

APÊNDICE

A frequência de primeiros nomes compostos por duas ou mesmo três palavras principais demandou um procedimento para separá-los dos sobrenomes. Este procedimento está enumerado a seguir.

1) A partir dos dados da própria Relação Anual de Informações Sociais (Rais), foram calculados os quinhentos primeiros nomes mais frequentes no Brasil.
2) Excluiu-se a primeira palavra de todos os nomes.
3) Como, geralmente, os primeiros nomes compostos são formados com base nos primeiros nomes populares (Maria Fernanda, João Paulo, entre tantos outros), excluíram-se os nomes obtidos no primeiro passo quando localizados no começo do que restou da *string* obtida no segundo passo.
4) Foram consideradas como sobrenome do indivíduo a última e a penúltima (se houver) palavra da *string* resultante do terceiro passo.

O nome José Eduardo Silva, por exemplo, perde a primeira palavra no segundo passo e a segunda no terceiro passo (pois Eduardo está entre os nomes mais populares do Brasil). Seu sobrenome será considerado, portanto, apenas Silva. O mesmo procedimento aplicado a José Santos Silva gera dois sobrenomes, pois Santos não é um dos quinhentos primeiros nomes mais populares.

Além desses procedimentos, foram consideradas como primeiros nomes as palavras Socorro, Auxiliadora, Dores, Rosário e Ribamar, que são segundos nomes comuns, mas não estão entre os quinhentos nomes mais frequentes.

Ipea – Instituto de Pesquisa Econômica Aplicada

EDITORIAL

Coordenação
Cláudio Passos de Oliveira

Supervisão
Andrea Bossle de Abreu

Revisão
Camilla de Miranda Mariath Gomes
Carlos Eduardo Gonçalves de Melo
Elaine Oliveira Couto
Laura Vianna Vasconcellos
Luciana Nogueira Duarte
Bianca Ramos Fonseca de Sousa (estagiária)
Thais da Conceição Santos Alves (estagiária)

Editoração
Aeromilson Mesquita
Aline Cristine Torres da Silva Martins
Carlos Henrique Santos Vianna
Glaucia Soares Nascimento (estagiária)
Vânia Guimarães Maciel (estagiária)

Capa
Luís Cláudio Cardoso da Silva

Projeto Gráfico
Renato Rodrigues Bueno

The manuscripts in languages other than Portuguese published herein have not been proofread.

Livraria Ipea

SBS – Quadra 1 - Bloco J - Ed. BNDES, Térreo.
70076-900 – Brasília – DF
Fone: (61) 2026-5336

Correio eletrônico: livraria@ipea.gov.br

Composto em adobe garamond pro 12/16 (texto)
Frutiger 67 bold condensed (títulos, gráficos e tabelas)
Rio de Janeiro-RJ

Missão do Ipea
Aprimorar as políticas públicas essenciais ao desenvolvimento brasileiro por meio da produção e disseminação de conhecimentos e da assessoria ao Estado nas suas decisões estratégicas.

ISSN 1415-4765
9 771415 476001

ipea Instituto de Pesquisa Econômica Aplicada

MINISTÉRIO DO PLANEJAMENTO, DESENVOLVIMENTO E GESTÃO

BRASIL
GOVERNO FEDERAL