TEXTO PARA DISCUSSÃO

2229

Rio de Janeiro, setembro de 2016

SUMÁRIO EXECUTIVO

SOBRENOMES E ANCESTRALIDADE NO BRASIL

Leonardo Monasterio

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>.

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.

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.

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, o que explica os nomes identificados serem bem mais populares que os não identificados. Já na aplicação dos métodos de *machine learning*, o algoritmo de Cavnar e Trenkle (1994)[1] apresentou um resultado mais acurado que o conhecido Naive Bayes.

Estimou-se que, no Brasil como um todo, apenas 18% dos indivíduos têm ao menos um sobrenome germânico, italiano, leste europeu ou japonês. A concentração espacial obtida desses indivíduos está de acordo com a literatura sobre imigração estrangeira para o Brasil. 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), corretamente identificou a ancestralidade dos trabalhadores na Rais.

A análise exploratória indicou 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. Também 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 comparação a 11,4 anos dos ibéricos.

SUMÁRIO EXECUTIVO

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