



Munich Personal RePEc Archive

## **Brazil without immigrants: microdata long run estimates**

Monasterio, Leonardo and Lopes, Daniel

Ipea, UCB

23 July 2018

Online at <https://mpra.ub.uni-muenchen.de/88170/>  
MPRA Paper No. 88170, posted 26 Jul 2018 17:01 UTC

# Brasil sem imigrantes: estimativas de longo prazo baseadas em microdados\*

Leonardo Monasterio<sup>†</sup>

Daniel Lopes. <sup>‡</sup>

Junho, 2018

Versão preliminar não revista

Comentários são bem-vindos

## Resumo

Este trabalho estima o efeito da imigração não-ibérica para o Brasil com base em microdados históricos e contemporâneos. A base histórica engloba mais 1,7 milhões de registros de imigrantes; já a contemporânea parte de um banco de dados com mais de 165 milhões registros administrativos. O cálculo do enumeramento dos imigrantes permite afirmar que [Stolz, Baten e Botelho \(2013\)](#) subestimaram essa habilidade dos estrangeiros e, portanto, seu impacto no Brasil. Um algoritmo de classificação de sobrenomes categorizou a população brasileira contemporânea em grupos ancestrais. Em seguida, constroem-se duas estimativas contrafatuais do que seria a renda per capita se nunca tivesse havido imigração não-ibérica. Um contrafactual decorre da regressão dessa renda pelos percentuais de cada grupo ancestral nos municípios. O outro, resulta da regressão de salários individuais sobre a ancestralidade de sobrenome de cada trabalhador. Os coeficientes obtidos em ambas são usados para estimar a renda per capita em um Brasil sem descendentes de imigrantes. Nesse caso, estimou-se que a renda per capita brasileira seria entre 12,6% e 17% menor do que a hoje observada.

**Palavras-chaves:** Imigração; Capital Humano ; Enumeramento.

## Abstract

This paper estimates the effect of non-Iberian immigration to Brazil based on historical and contemporary microdata. The historical database encompasses over 1.7 million immigrant records; the contemporary has more than 165 million records. The estimation of immigrant numeracy suggests that [Stolz, Baten e Botelho \(2013\)](#) underestimated their skills and, therefore, their impact on Brazil. An algorithm classified the surnames of contemporary Brazilians according to their ancestral origins. Two counterfactual estimates are constructed in order to estimate the income per capita if there had never been any non-Iberian immigration. The first counterfactual is built upon the regression of income on the percentages of each ancestral group in municipalities. The second, results from the regression of individual wages on the surname ancestry of workers. The coefficients of these regressions are used to estimate the income of a counterfactual Brazil with no descendants of immigrants. It was estimated that in the absence of non-Iberian immigrants today's income would be from 12.6 % to 17 % lower.

**Key-words:** Immigration; Human Capital; Numeracy.

---

\*Agradecemos a Pedro Souza e Claudio Shikida pelos comentários. Somos gratos também a Rodrigo Soares e Fábio Vaz pelo apoio na obtenção e processamento de dados. Detalhes no Apêndice B.

<sup>†</sup>Instituto de Pesquisa Econômica Aplicada (IPEA) e Universidade Católica de Brasília (UCB). email: leonardo.monasterio at gmail.com

<sup>‡</sup>Instituto de Pesquisa Econômica Aplicada (IPEA). email: lopes.daniel at t-online.de

## Introdução

Em 1872, o Brasil possuía cerca de 10 milhões de habitantes. Nas seis décadas que se seguiram, mais de 4 milhões de estrangeiros entraram no Brasil (LEVY, 1974). Se, por um lado, é do senso comum que essa imigração teve efeitos relevantes, por outro, faltam estimativas do real impacto econômico desse choque positivo na oferta de mão-de-obra. Assim, o objetivo deste artigo é estimar os efeitos de longo prazo da imigração não-ibérica.

Para construir o contrafactual que intitula este artigo, parte-se de fontes históricas de registros individuais desses imigrantes, bem como bases contemporâneas de microdados de cidadãos brasileiros. Até onde se sabe, as fontes primárias com dados de imigrantes ainda não haviam sido examinadas de forma agregada. Analisa-se aqui os registros de imigrantes que reúnem mais de 1,7 milhão de observações, de mais de 67 nacionalidades que chegaram ao Brasil entre 1840 e 1958. Igualmente, fez-se uma análise inédita de registros administrativos contemporâneos (Ministério do Trabalho e Emprego (2016), Ministério do Desenvolvimento Social (2015) e Receita Federal (2017)), que somam mais de 165 milhões de observações. O conjunto histórico e o contemporâneo de bases individuais permite visões e análises inéditas sobre o impacto da imigração no Brasil.

O paper de Stolz, Baten e Botelho (2013) se relaciona intimamente com este trabalho. Os autores usaram uma amostra dos dados de imigrantes do Arquivo Nacional do Rio de Janeiro para, obtendo o enumeramento dos brasileiros e estrangeiros, estimar o impacto dos imigração no longo prazo. Eles o fazem com base em uma estimação econométrica por países que relaciona o enumeramento em 1900 com o PIB per capita em anos recentes. Como será visto mais adiante, a utilização de bancos de dados mais amplos dos imigrantes e bases de microdados contemporâneas de grande escala nos levou a resultados bem distintos dos obtidos em Stolz, Baten e Botelho (2013).

Como se sabe, a imigração não-ibérica ao Brasil foi promovida quando se tornou evidente que a escravidão como força de trabalho se tornara insustentável (REIS; REIS, 1988; LAGO, 2014). Houve experimentos com a imigração subsidiada ao longo do século XIX, mas o fluxo mais importante concentrou-se após a emancipação dos escravos em 1888. Entre 1872 e 1920 mais de 3,2 milhões de estrangeiros desembarcaram no Brasil (LEVY, 1974).

Em 1920, 5,1% da população era de origem estrangeira ou naturalizada. Ao longo da década de trinta, especialmente durante o Estado Novo, o país foi se fechando à imigração. E, desde então, não houve mais grandes ondas migratórias para o país. Em 2010, apenas 0,3% da população nasceram no exterior, um dos menores percentuais do mundo.

A imigração subsidiada foi um choque positivo de capital humano. O nível educacional no Brasil era substancialmente baixo, mesmo quando comparado com outros países pobres (CHAUDHARY et al., 2012). Em 1920, apenas 23% dos brasileiros, de todas as idades, conseguiam ler ou escrever. Em geral, os imigrantes não eram altamente qualificados em seus países de origem, porém suas dotações de capital humano eram provavelmente mais elevadas que as dos brasileiros, sendo que 52% dos estrangeiros eram alfabetizados naquele ano (DGE, 1929).

A importância da mão-de-obra estrangeira para a modernização brasileira está bem assentada na literatura. Bresser Pereira (1964) apontou que, na época, 84,3% dos empresários paulistas declararam que a sua origem étnica era estrangeira (em até três gerações). Estrangeiros compunham 40% e 44% da mão-de-obra industrial em Rio de Janeiro e São Paulo, respectivamente, em 1920 (VERSIANI, 1993). Monasterio e Reis (2008) também mostraram que, no Brasil como um todo, a presença de estrangeiros nos municípios esteve associada à manufatura entre 1872 e 1920, mesmo com diversos controles estatísticos.<sup>1</sup>

<sup>1</sup> Os autores também apontam que a presença de imigrantes, junto com a expansão das ferrovias, induziram forças aglomerativas que explicam a concentração espacial da industrialização brasileira

Indiretamente, este trabalho se relaciona com a crescente literatura sobre transmissão cultural, normas e o impacto de fatores culturais sob as instituições e desenvolvimento econômico, em particular [Bisin e Verdier \(2000\)](#), [Alesina, Giuliano e Nunn \(2013\)](#) e [Fernández e Fogli \(2009\)](#).

[Tabellini \(2008\)](#) atribui a persistência das instituições a indicadores relacionados a crenças e valores individuais como confiança e respeito aos outros. Especificamente, buscamos relacionar a importância da cultura para a formação de capital humano (por exemplo, [Guiso, Sapienza e Zingales \(2006\)](#), [Figlio et al. \(2016\)](#), [Sørensen et al. \(2016\)](#)).

Nossa análise dialoga com a vasta literatura sobre determinantes históricos das instituições locais e performance econômica no Brasil. [Naritomi, Soares e Assunção \(2012\)](#) estuda o legado colonial sob instituições, [Musacchio, Martinez e Viarengo \(2014\)](#) foca na economia política da educação após a proclamação da república. Já [Carvalho Filho e Colistete \(2010\)](#) relaciona o ciclo da expansão cafeeira, imigração europeia de trabalhadores rurais e o estabelecimento de instituições de instrução educacional no estado de São Paulo.

O estudo de [Rocha, Ferraz e Soares \(2017\)](#) analisa o caso dos núcleos coloniais em São Paulo e sugere evidências do impacto do capital humano dos imigrantes no desenvolvimento de longo prazo dessas regiões. [Carvalho Filho e Monasterio \(2012\)](#), por sua vez, explora o estabelecimento de colônias patrocinadas pelo governo no Rio Grande do Sul do país e encontra efeitos persistentes sobre variáveis econômicas atuais. Finalmente, [Souza \(2017\)](#) mostra que o impacto das escolas fundadas por imigrantes alemães em São Paulo se dissipou ao longo do tempo, não deixando sinais nas *proxies* locais de capital humano.

Indiretamente, este trabalho também é relacionado com uma ampla literatura sobre os resultados educacionais dos imigrantes, em especial a transmissão intergeracional de escolaridade; incluindo [Algan e Cahuc \(2010\)](#), [Card, DiNardo e Estes \(1998\)](#), [Borjas, Bronars e Trejo \(1992\)](#) e [Borjas \(1995\)](#)

De acordo com as nossas estimações, caso a imigração não-ibérica não tivesse ocorrido no Brasil, a nossa renda per capita seria entre 12,6% a 17% mais baixa do que a observada. Essa estimativa é obtida mediante modelos econométricos elaborados a partir de regressões com dados individuais e municipais que simulam a ausência de descendentes de tais imigrantes no Brasil.

Obviamente, toda estimativa contrafactual tem que ser entendida dentro de seus próprios limites. Aqui, apesar de se tentar mapear alguns dos mecanismos envolvidos, não se consideram efeitos de equilíbrio geral da imigração ou alterações institucionais que possam ter efeitos no longo prazo. Além disso, está implícito que a produtividade do trabalhador é resultado de suas características individuais (com os controles locais) e que efeitos diretamente setoriais não são relevantes.

## 1 Fontes históricas

O trabalho se baseia em três fontes primárias ainda pouco exploradas:

- Arquivo da Hospedaria dos Imigrantes de São Paulo: essa base engloba 1.574.107 registros de imigrantes que passaram pelo local entre 1882 e 1958. Há registros não só de estrangeiros, mas também de brasileiros vindos de outros estados. Livros de Registro das Antigas Hospedarias de Imigrantes (1882 a 1958) Conjunto de mais de 150 livros de Matrícula das Hospedarias do Bom Retiro (1882 a 1887) e Brás (1887 a 1958), contém informações sobre a passagem de estrangeiros por essas instituições. Dentre os possíveis dados encontrados, destacam-se: nome, idade, nacionalidade, data de entrada na Hospedaria, profissão,

parentesco e estado civil.<sup>2</sup>

- Arquivo Público do Rio de Janeiro: 208.824 observações. Contém os registros de Entrada de Estrangeiros no Brasil – Porto do Rio de Janeiro, com informações abrangendo o período de 1875 a 1910, alimentada com o apoio do BNDES. Nela, é possível encontrar o nome do imigrante, nome do navio, data de entrada, nacionalidade, nível de instrução, religião, profissão, estado civil, local de procedência e local de destino do imigrante no Brasil. A escolha das Relações de Passageiros recair sobre o Porto do Rio de Janeiro deveu-se àquele porto ser, na primeira fase da grande leva de imigração (de 1875 a 1910), a principal porta de entrada de imigrantes.
- Arquivo Público do Espírito Santo: 43.293 observações. Projeto Imigrantes Espírito Santo fornece listas contendo os nomes dos imigrantes que entraram no Espírito Santo durante os séculos XIX e XX. Até 10 de março de 2014, foram indexados 54.155 imigrantes que entraram no Espírito Santo (sendo 46.885 no século XIX e 7.146 no século XX) provenientes dos cinco continentes. Apesar da base cobrir mais de 73 nacionalidades dos imigrantes, neste trabalho tem-se apenas os dados referentes aos italianos e alemães, que constituem 75% de todos os imigrantes. Atualmente, as demais nacionalidades acham-se em processamento.

## 1.1 Descrição da base

A figura 1 mostra a chegada de imigrantes não-ibéricos por ano.<sup>3</sup> A comparação das ondas de imigração da nossa base com os dados de entrada apresentados por Levy (1974) são aparentemente semelhantes.<sup>4</sup>

Como se vê, a maior parte dos registros de entrada das nossas bases são anteriores ao início da I Guerra Mundial e os italianos são predominantes. Após o fim desse conflito, a nova onda é marcadamente formada por imigrantes japoneses. A figura 2 mostra a distribuição dos registros da base por grupos nacionais, incluindo os ibéricos e a migração interna de brasileiros que passaram pela Hospedaria de São Paulo.

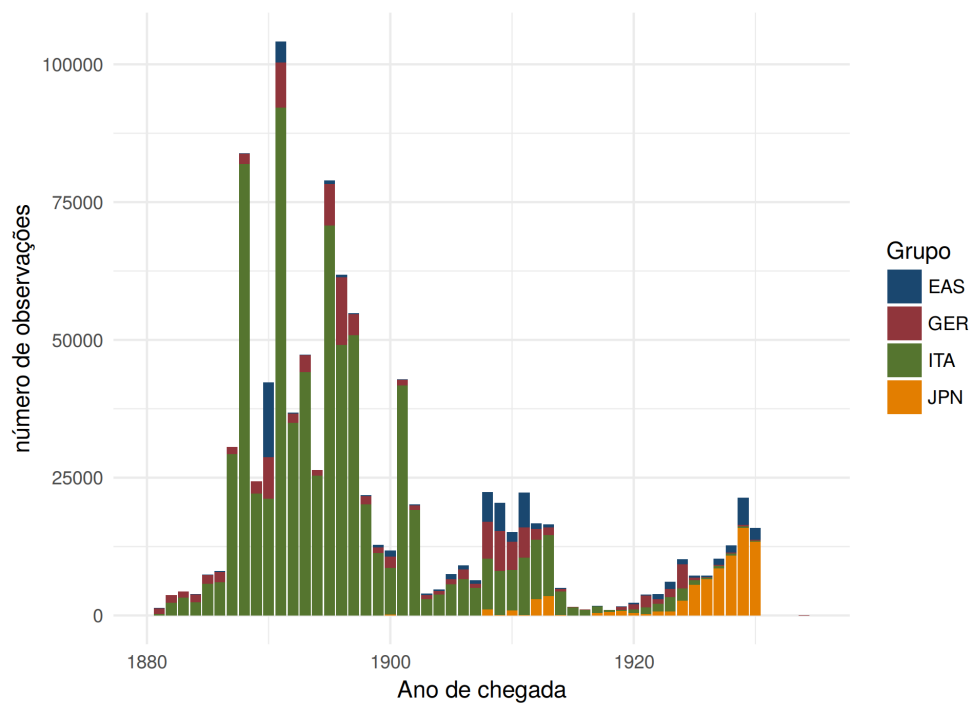
---

<sup>2</sup> Cabe observar que, como esses registros estão transcritos literalmente como aparecem nos Livros, é comum que a grafia de nomes e sobrenomes possa não condizer com a forma correta. Também cabe notar que os registros eram feitos em nome dos “chefes de família” (via de regra o parente do sexo masculino, em idade produtiva, mais velho)

<sup>3</sup> Há um problema no banco de dados do Arquivo Nacional. Algumas datas de chegada referentes ao século XIX foram convertidas para "01/01/1900". Provavelmente, trata-se de um problema causado pelo uso do Microsoft Excel. Esse *software* é incapaz de lidar com datas anteriores ao século XX. Nossos resultados neste trabalho, felizmente, independem da precisão da variável de data de chegada.

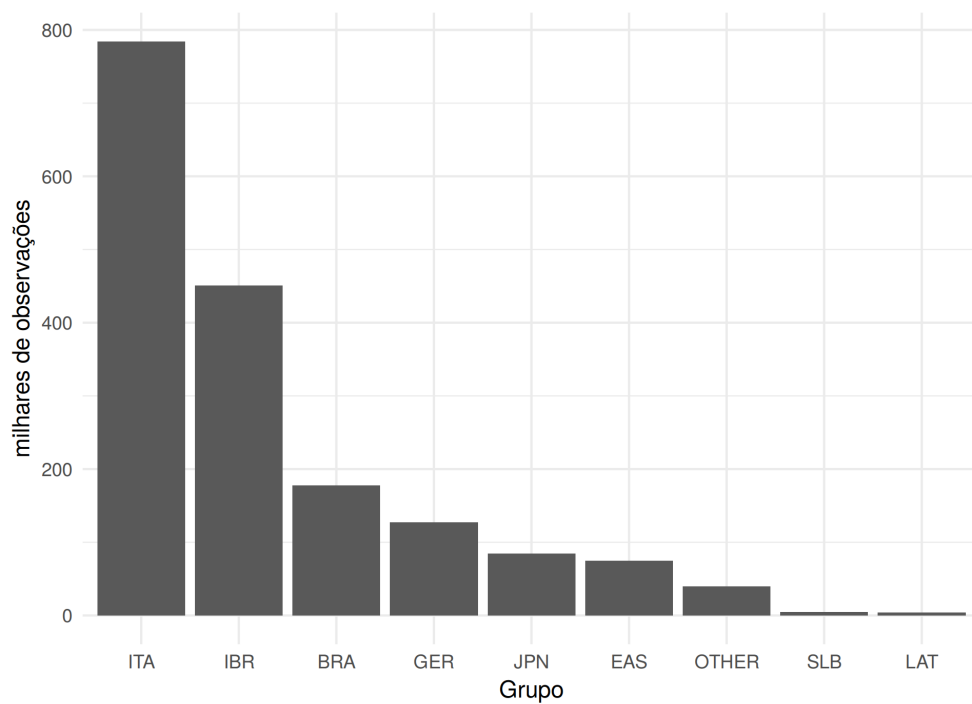
<sup>4</sup> Em versões posteriores deste *paper* será feita uma comparação mais rigorosa das bases com outras fontes de entradas de imigrantes.

Figura 1: Ano de entradas dos imigrantes por grupos ancestrais selecionados



Notes: Cálculos dos autores com base nos registros da Hospedaria de Imigrantes- SP, Arquivo Público-ES e Arquivo Nacional- RJ.

Figura 2: Principais grupos nacionais nos registros dos imigrantes



*Notes:* Cálculos dos autores com base nos registros da Hospedaria de Imigrantes- SP, Arquivo Público-ES e Arquivo Nacional- RJ. BRA: brasileiros migrantes; EAS: Europa Oriental; IBR: espanhóis e ibéricos; GER: germânicos; ITA: italianos; JPN: japoneses; LAT: países latino-americanos; SLB: sírios e libaneses; OTHER: todos os demais países.

## 2 Enumeramento dos imigrantes

### 2.1 Medida de enumeramento

Como medir o capital humano por grupos de imigrantes? Utilizou-se o método proposto por Mokyry (1983) e desenvolvido por A'Hearn, Baten e Crayen (2009). Ele se baseia na tendência dos indivíduos com baixa capacidade quantitativa a arredondar o último dígito de valores em 5 ou 0. O Índice de Whipple mede essa tendência e é definido pelo quíntuplo da razão da soma dos indivíduos, entre 23 e 62 anos, que informam ter idade terminada entre 0 e 5 dividido pela total de indivíduo entre aquelas idades. Assim, o índice varia entre 100 (não haveria preferência por arredondamento) e 500.<sup>5</sup>

$$Whipple = 5 \times \frac{\sum(n_{25} + n_{30} + \dots + n_{60})}{\sum_{i=23}^{62} (n_i)}$$

Onde  $n_i$  = número de pessoas com a idade  $i$ .

Para facilitar a interpretação do índice de Whipple, Crayen e Baten (2010) propuseram o índice ABCC, que é uma transformação linear do índice de Whipple. O ABCC varia entre 0 e 100 e quanto maior o valor, maior o enumeramento do grupo selecionado.

$$Whipple \geq 100 \implies ABCC = \left(1 - \frac{Whipple - 100}{400}\right) \cdot 100$$

$$Whipple < 100 \implies ABCC = 100$$

### 2.2 Resultados do ABCC para a amostra

Stolz, Baten e Botelho (2013) estimaram o ABCC dos brasileiros e dos imigrantes a partir de uma amostra de dados obtida no Arquivo Nacional do Rio de Janeiro. Nós seguimos o mesmo procedimento a partir do nosso banco de dados. Este foi limitado apenas aos indivíduos maiores de 23 e menores de 62 anos, bem como os que chegaram ao Brasil entre 1840 e 1920. Com isso, a amostra passou a ter 774.625 observações individuais.

A distribuição de idades está no gráfico de densidade na Figura 3. A faixa de idade considerada no índice ABCC é representada pela faixa sombreada. Como se vê, mesmo com as restrições na amostra, cobre-se uma parte considerável da distribuição. Vale notar que o fenômeno do arredondamento em dígitos já pode ser percebido na figura, nos pequenos picos nas idades terminadas em 0 e 5.

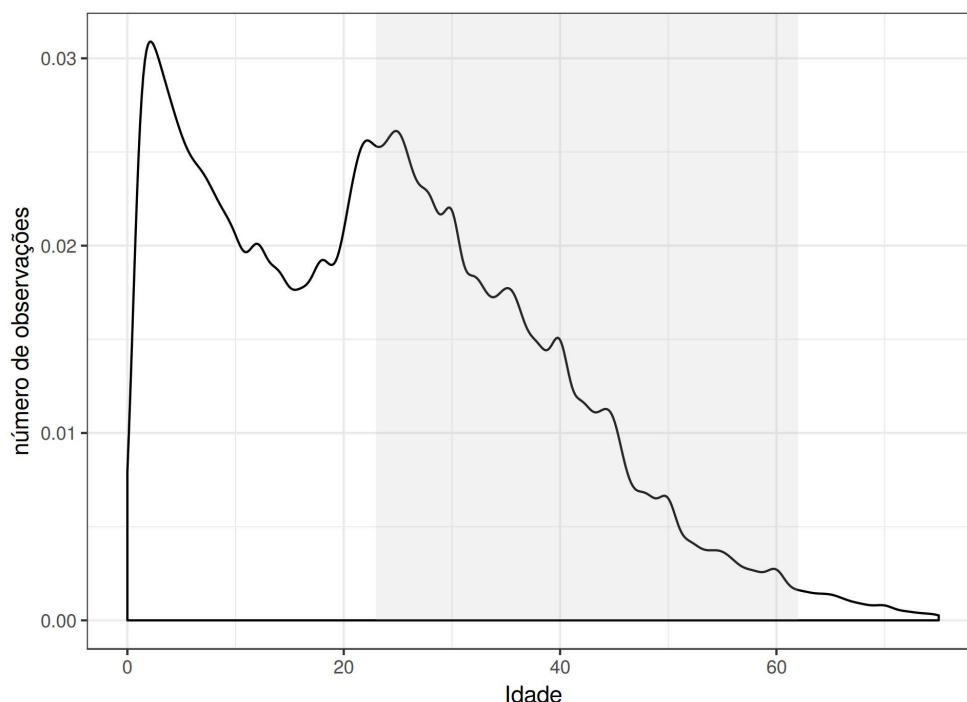
Os resultados do cálculo ABCC por coorte para os principais grupos da nossa amostra estão representados na tabela abaixo. Limitou-se a apresentação dos resultados aos pares de grupos ancestrais-coortes que tivessem ao menos 100 observações. As lacunas na tabela são assim explicadas. É evidente a tendência de incremento do ABCC. A exceção é a coorte entre 1910-1920, onde há uma pequena queda. Os brasileiros, migrantes internos, têm valores baixos, conforme esperado. Pode causar surpresa o valor do ABCC para os oriundos da Europa Oriental, eslavos em sua maioria russos e poloneses. Em todas as coortes, exceto a de 1850, o índice destes imigrantes foi menor do que a dos ibéricos.

Os japoneses tiveram o maior valor do índice ABCC em todas as coortes. Os italianos ficam, em geral, na segunda posição. O enumeramento dos germânicos se mostra maior do que a dos migrantes brasileiros, mas ficam atrás dos ibéricos na coorte de 1870, 1890, e 1910. Os

<sup>5</sup> Ver A'Hearn et al. (2016) para críticas quanto ao uso do arredondamento de idades como *proxy* de capital humano.



Figura 3: Distribuição das idades do imigrantes



Notes: Cálculos dos autores com base nos registros da Hospedaria de Imigrantes- SP, Arquivo Público-ES e Arquivo Nacional- RJ. A área sombreada indica a faixa etária que é utilizada para o cálculo do índice de Whipple e ABCC.

sírio-libaneses possuíam níveis muito baixos do índice ABCC nas coortes em que o número de observações é superior ao mínimo.

Tabela 1: Índice ABCC para toda a amostra por grupos nacionais e coorte

	CNTRY	1840	1850	1860	1870	1880	1890	1900	1910
1	BRA	69.5	65.5	70.2	78.3	83.6	87.4	94.7	
2	EAS	73.8	87.9	96.1	93.8	94.5	95.2	100.0	99.0
3	GER	90.9	95.8	96.6	94.2	98.9	99.1	100.0	92.6
4	IBR	88.9	93.8	95.8	98.0	97.8	98.3	100.0	99.4
5	ITA	95.3	97.4	97.3	97.9	98.3	100.0	100.0	98.7
6	JPN			98.7	100.0	99.6	100.0	100.0	100.0
7	LAT	81.9	80.8	85.7	94.7	100.0			
8	SLB	36.5	36.5	49.6	70.1				

Nota: BRA: brasileiros migrantes; EAS: Europa Oriental; IBR: espanhóis e ibéricos; GER: germânicos; ITA: italianos; JPN: japoneses; LAT: países latino-americanos; SLB: sírios e libaneses; OTHER: todos os demais países.

### 2.3 ABCC, bases de dados e a seleção de imigrantes

Os resultados de enumeramento dos imigrantes em nossos dados são, em geral, maiores do que os observados na amostra do RJ estudada por [Stolz, Baten e Botelho \(2013\)](#). Os autores encontram um valor médio de 94,1 para a coorte de 1900 dos imigrantes. Já na nossa amostra, chega-se a um ABCC de 98,3. Essa diferença de pouco mais de 4,2 pontos pode parecer desprezível, mas é a mesma que separava, por exemplo, o enumeramento dos argentinos ou

norte-americanos da dos hondurenhos na mesma coorte de 1900. A principal diferença da nossa base em relação a amostra de [Stolz, Baten e Botelho \(2013\)](#) é que na nossa há um maior peso dos que passaram pela Hospedaria dos Imigrantes de São Paulo. Logo, podemos concluir que estes tinham, em média, maior enumeramento do que os da base do Arquivo Nacional do Rio de Janeiro.

Vale a pena comparar os imigrantes com as populações de seus países de origem para identificar em que medida houve seleção - positiva ou negativa- dos que para cá vieram. O sítio [Clio-Infra \(2017\)](#) reúne estimativas do índice ABCC para muitos países do mundo, desde 1500, com periodicidade irregular, por coorte. Na Tabela 2 comparou-se o enumeramento dos imigrantes e da população de origem. Optou-se por analisar as coortes de 1870 a 1900. Como se vê, não há um padrão claro. No caso do húngaros e dos romenos, parece ter havido seleção negativa, ou seja, vieram aqueles com menor enumeramento. Já no tocante aos poloneses e dos portugueses, os dados sugerem seleção positiva. No caso dos demais países, as diferenças são menos pronunciadas.

Tabela 2: Índice ABCC para imigrantes e para toda a amostra por grupos nacionais e coorte

COUNTRY	MIGRANTS	POPULATION
GERMAN	98.3	99.9
HUNGARIAN	96.7	99.7
ITALIAN	99.0	99.8
JAPANESE	99.9	99.2
POLISH	94.4	93.7
PORTUGUESE	98.5	93.6
ROMANIAN	95.2	99.8
RUSSIAN	97.0	97.0
SPANISH	98.1	99,4

Média do índice ABCC para as coortes de 1870 a 1900 para a amostra e para os países de origem. Fonte: ver texto e [Clio-Infra \(2017\)](#). A lista de países é incompleta pois foram omitidos os países da amostra em que não houvesse ao menos 100 observações por coorte.

### 3 Novas estimativas para o impacto da imigração

Para [Stolz, Baten e Botelho \(2013\)](#) o impacto da imigração para o Brasil foi de USD 75 por ano. Essa estimativa foi baseada em uma regressão do PIB per capita de 2000 sobre o índice ABCC de 1900 por um conjunto de controles geográficos e institucionais por país. Segundo os cálculos dos autores, os imigrantes teriam um ABCC de 94,1 e os nascidos no Brasil, 89,7 para a coorte de 1900. Este valor contrafactual, ou seja, qual seria o ABCC sem a chegada dos imigrantes, é inserido na regressão estimada por país para chegar ao valor citado. Em termos percentuais, o efeito da imigração seria de cerca de 1% do PIB per capita.

Para rever essa estimativa, nesta seção, criar-se-á outros contrafatuais. Na sessão anterior, mostrou-se que nossa estimativa do enumeramento dos imigrantes é sensivelmente maior do que a obtida por [Stolz, Baten e Botelho \(2013\)](#). Agora, em vez de estimar o impacto da imigração com base em uma regressão por país, será feita uma regressão contrafactual a partir de dados contemporâneos individuais e por município.

Em ambas estratégias de construção do contrafactual, utilizou-se o algoritmo baseado em [Monasterio \(2017\)](#) para classificar sobrenomes dos indivíduos contemporâneos de acordo com a ancestralidade. O procedimento tem início com a construção de uma base de dados históricos que

associa sobrenomes a ancestralidades únicas. No artigo citado anteriormente, foram utilizadas diversas fontes: uma amostra obtida por *web scraping* no [Museu da Imigração \(2015\)](#), informações sobre o tráfico inter provincial de escravos ([FRAGOSO; FERREIRA, 2002](#)), microdados dos censos históricos norte americanos ([North Atlantic Population Project, 2016](#)), entre muitas outras. As bases foram agregadas em pares únicos de sobrenome-ancestralidade. Essa base tem um duplo papel: a) servir para o *fuzzy matching* com a base de sobrenomes observados; b) calibrar o algoritmo de *machine learning*, baseado em [Cavnar e Trenkle \(1994\)](#), que identifica as nacionalidades com base em um padrão linguístico. Existem 464.613 sobrenomes distintos entre os trabalhadores na base da RAIS 2016, mas alguns concentram um grande número de observações. Os SILVA, SANTOS, OLIVEIRA, SOUZA e PEREIRA somam 45% das pessoas. Assim, a maior parte dos sobrenomes únicos são, na verdade, erros de digitação e observados apenas uma ou duas vezes. Ocorreu *match* exato em 40.496 sobrenomes, o procedimento de *fuzzy matching* permitiu encontrar 336.258 sobrenomes; os demais (87.859) foram classificados com o auxílio do algoritmo de [Cavnar e Trenkle \(1994\)](#).

A classificação de sobrenomes tem limitações. Além dos problemas habituais de acurácia do algoritmo de classificação, o problema principal é a perda da informação matrilinear dos sobrenomes. No padrão de escolha de nomes no Brasil, o sobrenome de uma avó estrangeira, por exemplo, estará perdido já no seu neto. Contudo, caso haja endogamia entre os grupos de ancestralidade, mesmo que seu sobrenome específico esteja perdido, há chances maiores de que seu neto tenha ancestrais do mesmo grupo.<sup>6</sup> Na aplicação presente, de qualquer forma, o problema persiste pois, por simplificação, será usado apenas o último sobrenome do indivíduo.

Para a classificação do perfil de grupos ancestrais de cada município utilizou-se as seguintes bases de dados contemporâneas: RAIS ([Ministério do Trabalho e Emprego \(2016\)](#); 67.144.598 observações), o CadÚnico ([Ministério do Desenvolvimento Social \(2015\)](#); 80.957.412 observações) e a base de sócios da Receita ([Receita Federal \(2017\)](#), 17.780.860 observações). A combinação desses dados cobre um amplo espectro social do Brasil. O CadÚnico contém informações sobre os mais pobres, a maioria no setor informal da economia. A RAIS, como foi dito, as faixas intermediárias, os empregados formais. A lista de sócios, por fim, cobre os representantes dos estratos mais altos, mas também os microempreendedores individuais (MEI)<sup>7</sup>.

Os dados dessas três bases foram classificados conforme a ancestralidade de acordo com o procedimento apresentado na sessão anterior e agrupados por municípios. Utilizou-se a informação de cor/raça disponível na RAIS para o caso de indivíduos registrados como "índio"(NAT), "pardo"(MXD) e "preto"(BLK). Já para os "brancos" e "amarelos" foi utilizado o algoritmo de classificação de sobrenomes.

### 3.1 Regressão municipal

Regrediu-se a renda per capita domiciliar mensal observada em 2010 sobre um conjunto amplo de variáveis ao nível municipal.<sup>8</sup> A inovação é a inclusão de variáveis do lado direito das regressões que representem o impacto de longo prazo da imigração na população. Esse efeito será medido como as participações dos grupos ancestrais de sobrenome na base composta (CadÚnico, RAIS, e Sócios) por município.

Além dessas variáveis relacionadas com a ancestralidade, foram utilizadas *proxies* para o nível educacional do município, suas instituições, população, *dummies* estaduais e um grupo

<sup>6</sup> Segundo [DGE \(1916\)](#), 77% dos filhos de estrangeiros nascidos entre 1908 e 1912 em capitais selecionadas do Brasil tinham o cônjuge igualmente estrangeiro.

<sup>7</sup> A base da Receita Federal não informa o município de residência, apenas o CNPJ da empresa da qual o indivíduo é sócio. Foi feito um cruzamento dessa base com a RAIS-Estabelecimentos e atribuiu-se ao sócio o mesmo município da sede da sua firma

<sup>8</sup> As variáveis utilizadas e suas fontes constam do Anexo A.

exaustivo de variáveis geográficas e climáticas (altura, distância da capital do estado, pluviosidade, temperatura e aptidão dos solos).

A Tabela 3 apresenta os resultados. A participação do grupo com ancestralidade ibérica é a variável de referência. Tomando-se a especificação mais completa e o grupo IBR como referência, tem-se que as participações dos grupos NAT, MXD, BLK, SLB estão associadas a reduções da renda per capita municipal; já os grupos GER, ITA e JPN, a aumentos desse valor. Podem causar surpresa inicial os valores negativos associados à participação de indivíduos com sobrenomes da Europa Oriental. Vale lembrar contudo que o enumeramento desses imigrantes foi menor do que o dos ibéricos na maior parte das coortes. Já o valor absoluto dos estimadores associados à participação dos japoneses é realmente impressionante.

Qual o significado preciso dos coeficientes associados aos percentuais de cada grupo ancestral de sobrenomes? Quais os canais exatos que levam a essa associação? Como os coeficientes continuam significativos mesmo após a inserção das variáveis educacionais e institucionais, existem duas possibilidades: ou as *proxies* não captam totalmente o efeitos dessas dimensões; ou existem canais não contemplados na análise. Será evitada uma resposta definitiva neste artigo. A próxima seção também busca avançar um pouco nessa questão, mas, no estado atual da pesquisa e frente ao presente objetivo, seria imprudente defender um mecanismo causal de forma categórica.

### 3.1.1 Análise contrafactual - dados municipais

Para estimar o impacto da imigração, criou-se um contrafactual em que foram zeradas as parcelas da população com ancestralidade não-ibérica por município. Recalculadas as participações para os grupos étnicos restantes, estes valores foram inseridos na especificação mais completa da tabela 3 e então recalculada a renda per capita brasileira. A comparação entre a renda per capita contrafactual (sem imigrantes) e a observada representam o impacto local da imigração não-ibérica.

A renda per capita observada em 2010 foi igual a R\$ 781,75. Já a contrafactual a partir da equação 4 foi de R\$ 683,18. Ou seja, a renda per capita brasileira seria cerca de 12,6% menor da observada caso não houvesse imigrantes. Ou seja, trata-se de um efeito bem maior do que o 1% estimado por [Stolz, Baten e Botelho \(2013\)](#). Note-se que esse resultado não é o efeito trivial da retirada de indivíduos com salários mais altos da amostra. Afinal, a especificação preferida foi feita com base nas parcelas de cada grupo e mantidos todos os controles municipais: variáveis institucionais, geográficas, população municipal e *dummies* por UF.

## 3.2 Regressão individual

Segue-se aqui a mesma estratégia empírica da seção anterior, mas com dados individuais dos trabalhadores. O problema dessa escolha é que apenas os ocupados no setor formal são examinados. Por outro lado, a vantagem do método consiste em que podem ser utilizadas *dummies* municipais que capturam os possíveis efeitos geográficos, econômicos ou institucionais de caráter local.

Mesmo contendo mais de 67 milhões de observações, a base da RAIS cobre apenas os trabalhadores formais brasileiros. Limitou-se a amostra aos trabalhadores no setor privado, brasileiros, entre 23 e 60 anos e que trabalhassem ao menos 40 horas por semana. Com isso, restaram 23.899.675 registros de trabalhadores com CPF único. Apenas com base nesses dados, os grupos mais presentes são IBR (48%), MXD (35%) e BLK (5,4%). As participações dos não-ibéricos são as seguintes; EAS:0,8 %; GER: 2,0%; ITA: 7,2%; JPN: 0,6 %.

Devido a restrições computacionais, fez-se uma amostra aleatória de um milhão de observações. E, no modelo mais completo, com mais de 5500 *dummies* municipais, a amostra foi

Tabela 3: Regressões com dados municipais

<i>Dependent variable:</i>				
Income per capita by municipality				
	Model 1	Model 2	Model 3	Model 4
% NAT	-510,8*** (47,9)	-179,2*** (40,5)	-182,0*** (43,7)	-219,1*** (41,6)
% MXD	-418,6*** (11,2)	-82,4*** (13,1)	-88,4*** (13,2)	-164,5*** (13,6)
% BLK	-54,0 (42,3)	-155,7*** (37,6)	-120,9*** (37,8)	-125,5*** (37,3)
% GER	201,9*** (29,4)	220,6*** (26,2)	231,6*** (27,0)	277,7*** (26,1)
% EAS	-1.034,5*** (89,3)	-657,2*** (70,5)	-591,2*** (71,3)	-461,6*** (67,9)
% ITA	609,4*** (30,8)	550,3*** (26,8)	552,8*** (27,8)	632,6*** (28,3)
% SLB	3.671,4*** (603,8)	418,4 (475,9)	44,0 (495,3)	-202,3 (464,1)
% JPN	8.898,1*** (604,6)	4.692,5*** (523,6)	3.379,4*** (509,2)	2.375,5*** (480,5)
Education		-12,2*** (0,3)	-10,0*** (0,3)	-7,8*** (0,3)
Governance			23,8*** (2,4)	12,6*** (2,3)
Land Gini			-146,2*** (22,1)	-151,9*** (20,9)
Access to justice			44,5*** (2,1)	12,3*** (2,4)
Political persistence			3,8 (3,2)	3,7 (3,0)
Population (ln)				48,9*** (2,1)
State dummies		Y	Y	Y
Geographical controls				Y
Observations	5.505	5.505	4.900	4.898
Adjusted R <sup>2</sup>	0,6	0,8	0,8	0,8

*Note:*

\*p&lt;0,1; \*\*p&lt;0,05; \*\*\*p&lt;0,01

reduzida para 100 mil observações para reduzir o tempo de processamento.

Em seguida, seguindo a lógica da seção anterior, os parâmetros dessa regressão servem para estimar os salários contrafatuais em uma base sem os trabalhadores com ancestralidade não-ibérica. Mais uma vez, a diferença entre os valores observados e simulados dos salários por trabalhador representa o impacto da imigração não-ibérica.

A Tabela 4 apresenta os resultados individuais. As variáveis de interesse são as *dummies* de ancestralidade de sobrenomes, mas os controles tiveram sinal e significância esperadas. Omitiu-se dos resultados os coeficientes associados ao sexo, grau de instrução, idade, idade ao quadrado, unidade federativa e *dummies* municipais (conforme o caso).

Na especificação mais completa, com controles individuais e *dummies* municipais, os coeficientes associados a variável de ancestralidade medida pelo sobrenome são estatística e substantivamente significativos. O grupo de referência são os indivíduos com sobrenome IBR. Conforme esperado, os grupos BLK, MXD e NAT estão associados a salários mais baixos. Já os indivíduos com ancestralidade de sobrenomes JPN auferem salários 17,8% mais elevados na especificação preferida, i.e. a mais completa. Os coeficientes dos possuidores dos demais sobrenomes não-ibéricos também são positivos, mas com valores mais baixos. Surpreende que os trabalhadores com sobrenome sírio-libaneses tenham salários 10% maiores do que os ibéricos e até mesmo maiores do que os GER, EAS e ITA. Não é clara a razão para tal fenômeno.

### 3.3 Análise contrafactual - dados individuais

A criação do contrafactual foi feita pela substituição da variável de ancestralidade de sobrenomes dos descendentes de imigrantes não-ibéricos (EAS, GER, ITA e JPN) por ibéricos (IBR). O salário médio por hora da amostra observada é de R\$58,04 reais. Já os salários contrafatuais foram de R\$48,29. Isso sugere que sem a imigração não-ibérica os salários seriam cerca de 17% mais baixos.

Isso pode ser considerado o limite inferior da estimativa, pois em vez de criar um contrafactual que reproduzisse as participações observadas com outros grupos ancestrais (MXD, BLK e NAT), optou-se apenas pela simples substituição dos portadores de sobrenomes não-ibéricos por ibéricos.

## 4 Discussão

Supondo que as estimativas estejam corretas, quais seriam os canais que fizeram com que a imigração tenha tal impacto no crescimento econômico? O primeiro candidato é, claro, o capital humano, em seu sentido mais amplo. Lopes (2017) mostrou que crianças de 8 anos com ancestralidade de sobrenome não-ibérica obtêm desempenho sensivelmente superior às brancas com sobrenome ibérico medido por testes padronizados, mesmo com controles para *background* socioeconômico dos pais e efeitos fixos por sala de aula. Em termos substantivos, no caso daqueles estudantes com ancestrais japoneses, o efeito equivale a um ano a mais de escolaridade na disciplina de matemática. Ao que parece, há uma transmissão intergeracional do capital humano que não é capturada adequadamente pelas *proxies* de escolaridade usadas nas regressões.

Ehrl e Monasterio (2017), por sua vez, analisaram o impacto de longo prazo imigração na produtividade via aumento da diversidade local de habilidades dos trabalhadores do Rio Grande do Sul. Os autores encontraram efeitos bastante robustos. Porém, sua abordagem não é capaz de explicar integralmente os resultados aqui encontrados. Se a diversidade fosse o canal, os controles locais nas regressões individuais deveriam capturar todo o efeito das variáveis de sobrenome. A mesma lógica indica que explicações baseadas em capital social ou instituições locais também são insuficientes.

Tabela 4: Regressão individual

	<i>Dependent variable:</i>		
	log wage per hour		
	Model 1	Model 2	Model 3
BLK	-0,095*** (0,002)	-0,067*** (0,002)	-0,094*** (0,007)
EAS	0,078*** (0,006)	0,072*** (0,006)	0,053*** (0,018)
GER	0,084*** (0,003)	0,077*** (0,003)	0,087*** (0,010)
ITA	0,090*** (0,002)	0,065*** (0,002)	0,072*** (0,006)
JPN	0,223*** (0,006)	0,179*** (0,006)	0,164*** (0,020)
MXD	-0,137*** (0,001)	-0,050*** (0,001)	-0,081*** (0,004)
NAT	-0,103*** (0,011)	-0,043*** (0,011)	-0,067* (0,036)
SLB	0,086*** (0,007)	0,085*** (0,006)	0,100*** (0,020)
State dummies		Y	
Municipal dummies			Y
Observations	997.682	997.682	99.768
R <sup>2</sup>	0,374	0,407	0,454
Adjusted R <sup>2</sup>	0,374	0,407	0,432
Residual Std. Error	0,503	0,490	0,483

*Note:* \*p<0,1; \*\*p<0,05; \*\*\*p<0,01

Quais outros canais tornariam as estimativas enganosas? Se os salários mais altos forem explicados integralmente por discriminação ao nível da firma em favor dos possuidores de sobrenomes não-ibéricos, os mesmos resultados econométricos seriam obtidos nas regressões ao nível individual. Porém, nesse caso, seria necessário explicar o porquê dos municípios com maiores participações de não-ibéricos também estão associados a rendas per capita mais altas.

## Considerações finais

Este trabalho estimou o impacto da imigração não-ibérica para o Brasil com base em microdados históricos e contemporâneos. Enquanto estudo anterior (STOLZ; BATEN; BOTELHO, 2013) calculou um impacto de cerca de 1%, aqui estimou-se que renda per capita do brasileiro seria entre 12,6% e 17% mais baixa sem a imigração.

A partir dos registros de entrada de mais de 1,7 milhão de imigrantes, foi calculado o enumeramento por país de origem e mostrou-se que, em geral, este era bem maior do que o dos brasileiros e também maior do que indicavam as estimativas anteriores.

A elaboração do contrafactual- isto é, o Brasil sem descendentes de imigrantes não-ibéricos- foi feita a partir de regressões de salários e de renda per capita ao nível individual e municipal sobre variáveis de ancestralidade de sobrenomes. Por mais críticas que esse o método possa sofrer, muitas delas pertinentes, ele permite que se tenha um entendimento mais preciso dos efeitos de longo prazo da imigração para o Brasil

Os próximos passos da pesquisa em curso se dirigem em três direções: a) aumentar ainda mais as bases de dados históricas pela inclusão de novas fontes; b) incrementar o algoritmo para classificar sobrenomes com base em origens subnacionais (ex. oriundos do Norte/Sul da Itália, Okinawa, entre outras); c) buscar identificar os nexos causais que fizeram com que a imigração não-ibérica tivesse os efeitos aqui identificados.



## Referências

- A'HEARN, B.; BATEN, J.; CRAYEN, D. Quantifying quantitative literacy: age heaping and the history of human capital. *The Journal of Economic History*, Cambridge University Press, v. 69, n. 3, p. 783–808, 2009.
- A'HEARN, B. et al. Rethinking age-heaping. a cautionary tale from nineteenth century italy. *Laboratory of Economics and Management Working Paper Series*, v. 35, p. 2016–35, 2016.
- ALESINA, A.; GIULIANO, P.; NUNN, N. On the origins of gender roles: Women and the plough. *The Quarterly Journal of Economics*, v. 128, n. 2, p. 469–530, 2013.
- ALGAN, Y.; CAHUC, P. Inherited trust and growth. *American Economic Review*, v. 100, n. 5, p. 2060–92, December 2010.
- BISIN, A.; VERDIER, T. “beyond the melting pot”: Cultural transmission, marriage, and the evolution of ethnic and religious traits. *The Quarterly Journal of Economics*, v. 115, n. 3, p. 955–988, 2000.
- BORJAS, G. J. Ethnicity, neighborhoods, and human-capital externalities. *The American Economic Review*, American Economic Association, v. 85, n. 3, p. 365–390, 1995.
- BORJAS, G. J.; BRONARS, S. G.; TREJO, S. J. Self-selection and internal migration in the united states. *Journal of Urban Economics*, v. 32, n. 2, p. 159 – 185, 1992.
- CARD, D.; DINARDO, J.; ESTES, E. *The More Things Change: Immigrants and the Children of Immigrants in the 1940s, the 1970s, and the 1990s*. [S.l.], 1998. (NBER, Working Paper Series, 6519).
- CARVALHO FILHO, I. de; COLISTETE, R. P. MPRA Paper, *Education Performance: Was It All Determined 100 Years Ago? Evidence From São Paulo, Brazil*. 2010.
- CARVALHO FILHO, I. de; MONASTERIO, L. Immigration and the origins of regional inequality: Government-sponsored European migration to southern Brazil before World War I. *Regional Science and Urban Economics*, v. 42, n. 5, p. 794–807, 2012.
- CAVNAR, W. B.; TRENKLE, J. M. N-gram-based text categorization. *Ann Arbor MI*, v. 48113, n. 2, p. 161–175, 1994.
- CHAUDHARY, L. et al. Big BRICs, weak foundations: The beginning of public elementary education in Brazil, Russia, India, and China. *Explorations in Economic History*, Elsevier, v. 49, n. 2, p. 221–240, 2012.
- Clio-Infra. *Clio Infra projec*. 2017. Disponível em: <<https://www.clio-infra.eu/>>.
- CRAYEN, D.; BATEN, J. Global trends in numeracy 1820–1949 and its implications for long-term growth. *Explorations in Economic History*, Elsevier, v. 47, n. 1, p. 82–99, 2010.
- DGE. *Anuario Estatístico do Brazil, 1908-1912*. Rio de Janeiro: Typographia Estatística, 1916.
- DGE. *Recenseamento do Brazil*. [S.l.]: Typ. da Estatística, 1929.
- EHRL, P.; MONASTERIO, L. *Inherited cultural diversity and wages in Brazil*. Clermont-Ferrand, France: 10th AFD-World Bank Migration and Development Conference, 2017.

- FERNÁNDEZ, R.; FOGLI, A. Culture: An empirical investigation of beliefs, work, and fertility. *American Economic Journal: Macroeconomics*, v. 1, n. 1, p. 146–77, January 2009.
- FIGLIO, D. et al. *Long-Term Orientation and Educational Performance*. [S.l.], 2016. (NBER, Working Paper Series, 22541).
- FRAGOSO, J. L. R.; FERREIRA, R. G. *Final Report and database of the IPEA/LIPHIS project : “Tráfico de Escravos e Relações Comerciais no Sudeste do Brasil: Primeira Metade do século XIX”*. [S.l.]: Ipea, 2002.
- GUISO, L.; SAPIENZA, P.; ZINGALES, L. Does culture affect economic outcomes? *Journal of Economic Perspectives*, v. 20, n. 2, p. 23–48, June 2006.
- IPEA. *IPEADATA*. 2018. Disponível em: <<http://www.ipeadata.gov.br/>>.
- LAGO, L. A. C. do. *Da escravidão ao trabalho livre: Brasil, 1550-1900*. [S.l.]: Editora Companhia das Letras, 2014.
- LEVY, M. S. F. O papel da migração internacional na evolução da população brasileira (1872 a 1972). *Revista de Saúde Pública, SciELO Brasil*, v. 8, n. supl., p. 49–90, 1974.
- LOPES, D. A. F. *Culture, institutions and school achievement in Brazil*. Dissertação (Mestrado) — Universidade Católica de Brasília, Distrito Federal, 2017.
- Ministério do Desenvolvimento Social. *Cadastro único*. 2015.
- Ministério do Trabalho e Emprego. *RAIS identificada*. 2016.
- MOKYR, J. *Why Ireland starved: a quantitative and analytical history of the Irish economy, 1800-1850*. [S.l.]: Routledge, 1983.
- MONASTERIO, L. Surnames and ancestry in brazil. *PloS ONE*, Public Library of Science, v. 12, n. 5, p. e0176890, 2017.
- MONASTERIO, L.; REIS, E. Mudanças na concentração espacial das ocupações nas atividades manufatureiras no brasil: 1872-1920. Instituto de Pesquisa Econômica Aplicada (Ipea), 2008.
- MUSACCHIO, A.; MARTINEZ, A.; VIARENGO, M. *Colonial Institutions, Commodity Booms, and the Diffusion of Elementary Education in Brazil, 1889-1930*. [S.l.], 2014. (NBER, Working Paper Series, 20029).
- Museu da Imigração. *Museu da Imigração*. 2015. Disponível em: <<http://museudaimigracao.org.br/acervodigital/livros.php>>.
- NARITOMI, J.; SOARES, R. R.; ASSUNÇÃO, J. J. Institutional Development and Colonial Heritage within Brazil. *The Journal of Economic History*, v. 72, n. 02, p. 393–422, 2012.
- North Atlantic Population Project. *North Atlantic Population Project*. 2016. Disponível em: <<https://www.nappdata.org/napp/>>.
- PEREIRA, L. C. B. Origens étnicas e sociais do empresário paulista. *Revista de Administração de Empresas, SciELO Brasil*, v. 4, n. 11, p. 83–106, 1964.
- PNUD. *Atlas do desenvolvimento humano*. 2013. Disponível em: <<http://www.atlasbrasil.org.br/2013/>>.
- Receita Federal. *Dados Abertos do CNPJ*. 2017. Disponível em: <<http://idg.receita.fazenda.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoas-juridicas-cnpj/dados-abertos-do-cnpj>>.

- REIS, E. J.; REIS, E. P. As elites agrárias e a abolição da escravidão no Brasil. *Dados: Revista de Ciências Sociais*, v. 31, n. 3, p. 309–341, 1988.
- ROCHA, R.; FERRAZ, C.; SOARES, R. R. Human capital persistence and development. *American Economic Journal: Applied Economics*, v. 9, n. 4, p. 105–36, 2017.
- SØRENSEN, R. J. et al. *Culture and school performance: Evidence from second generation immigrants to Norway*. [S.l.], 2016. (BI, Working Paper).
- SOUZA, B. G. W. d. Immigration and the path dependence of education: the case of German-speakers in São Paulo, Brazil (1840–1920). *The Economic History Review*, Wiley Online Library, 2017.
- STOLZ, Y.; BATEN, J.; BOTELHO, T. Growth effects of nineteenth-century mass migrations: "Fome Zero" for Brazil? *European Review of Economic History*, Oxford University Press, v. 17, n. 1, p. 95–121, 2013.
- TABELLINI, G. Presidential address institutions and culture. *Journal of the European Economic Association*, Blackwell Publishing Ltd, v. 6, n. 2-3, p. 255–294, 2008.
- VERSIANI, F. R. Imigrantes, trabalho qualificado e industrialização: Rio e São Paulo no início do século. *Revista de Economia Política*, v. 13, n. 4, p. 52, 1993.

## APÊNDICE A – Fontes de dados

**Dados de imigrantes** : Dados obtidos em arquivos digitais, *web scraping* e OCR:

- Hospedaria de São Paulo;
- Arquivo Nacional do Rio de Janeiro;
- Arquivo Público do Espírito Santo.

**Variáveis institucionais** : Dados fornecidos por Rodrigo Soares e utilizados em [Naritomi, Soares e Assunção \(2012\)](#)

- *land Gini*: coeficiente de Gini para a distribuição de terra construído com bases no Censo Agrícola de 1996;
- *political persistence*: número de repetições de ao menos um sobrenome de família no nome dos prefeitos nas eleições de 1996, 2000 e 2004;
- *governance*: média simples de indicadores qualitativos sobre práticas da administração municipal, calculados com dados entre 1997 e 2000;
- *access to justice*: média de variáveis binárias que indicam existência de Tribunal de Pequenas Causas, Conselho Tutelar e Comissão de Defesa do Consumidor em 2001.

**Variáveis geográficas** : dados obtidos no [IPEA \(2018\)](#)

- Chuvas: estimativas da pluviosidade em cada uma das estações do ano;
- Temperatura: estimativas das médias trimestrais de temperatura em centígrados em cada uma das estações do ano ;
- Distância à capital estadual em km;
- Altitude: altura da sede do município.

**Variáveis sociais e demográficas** dados obtidos no [PNUD \(2013\)](#)

- Educação: taxa de alfabetização dos maiores de 25 anos ;
- Renda per capita: renda domiciliar per capita.

**Bases cadastrais contemporâneas** : Arquivos digitais de larga escala

- [Ministério do Trabalho e Emprego \(2016\)](#): RAIS identificada com os dados dos vínculos dos empregados, arquivo digital de acesso restrito;
- [Ministério do Desenvolvimento Social \(2015\)](#): Cadastro Único. arquivos digital de acesso restrito com dados individual dos beneficiados. Fábio Vaz preparou o banco de dados;
- [Receita Federal \(2017\)](#) Dados abertos do CNPJ. Álvaro Justen criou um *script* em Python para download e limpeza dos dados da Receita Federal (<https://github.com/turicas/socios-brasil>).

## APÊNDICE B – Tabulações complementares

Tabela 5: Distribuição das fontes por grupos de ancestralidade

	CNTRY	Arq. Público- ES	Arq. Nacional - RJ	Hosp. dos Imigrantes - SP
1	BRA	0	0	178065
2	EAS	0	37997	37032
3	GER	4596	66844	56070
4	IBR	0	40128	410769
5	ITA	36569	12698	735048
6	JPN	0	167	84238
7	LAT	0	3536	475
8	OTHER	0	18072	21747
9	SLB	0	4874	95

Nota: BRA: brasileiros migrantes; EAS: Europa Oriental; IBR: espanhóis e ibéricos; GER: germânicos; ITA: italianos; JPN: japoneses; LAT: países latino-americanos; SLB: sírios e libaneses; OTHER: todos os demais países.

Tabela 6: Distribuição das fontes por países

	COUNTRY	es	rj	sp
1	ALBANESE	0	0	91
2	ARAB	0	4084	27
3	ARGENTINIAN	0	3088	379
4	ARMENIAN	0	137	85
5	AUSTRALIAN	0	0	3
6	AUSTRIAN	0	31940	29378
7	BELGIAN	0	2177	1605
8	BOLIVIAN	0	91	29
9	BRAZILIAN	0	0	178065
10	BRITISH	0	0	1510
11	BULGARIAN	0	72	195
12	CANADIAN	0	0	905
13	CHILEAN	0	280	31
14	CHINESE	0	34	272
15	COLOMBIAN	0	23	29
16	CROACIA	0	0	15
17	CUBAN	0	0	13
18	CZECH	0	1163	1122
19	DANISH	0	0	857
20	DUTCH	0	0	904
21	EGYPTIAN	0	0	39
22	EQUATORIAN	0	16	9
23	ESTONIAN	0	0	346
24	FINNISH	0	0	35
25	FRENCH	0	1332	4233
26	GERMAN	0	34904	26692
27	GREEK	0	0	4031
28	HUNGARIAN	0	576	3209
29	INDIAN	0	0	300
30	IRAKI	0	0	1
31	IRANIAN	0	13	25
32	IRISH	0	0	3
33	ISRAELI	0	46	261
34	ITALIAN	0	12698	735048
35	JAPANESE	0	167	84238
36	KOREAN	0	0	4
37	LEBANESE	0	142	64
38	LETHONIAN	0	0	3163
39	LITHUANIAN	0	0	4740
40	MALTESE	0	0	324
41	MEXICAN	0	92	6
42	MONTENEGRIN	0	28	24
43	MORROCAN	0	0	31
44	NORWEGIAN	0	0	21
45	OTHER	0	7342	343
46	PALESTINIAN	0	0	9
47	PARAGUAYAN	0	97	2
48	PERUVIAN	0	108	12
49	POLISH	0	4684	10896

50	PORTUGUESE	0	13145	155643
51	ROMANIAN	0	388	1924
52	RUSSIAN	0	30651	7736
53	SERBIAN	0	233	4
54	SLOVAKIAN	0	0	5
55	SLOVENIAN	0	0	1
56	SPANISH	0	26983	255126
57	STATELESS	0	0	1521
58	SWEDISH	0	0	1414
59	SWISS	0	0	2035
60	SYRIAN	0	4732	31
61	TUNISIAN	0	2244	581
62	TURKISH	0	0	1
63	UCRANIAN	0	230	1531
64	URUGUAYAN	0	355	78
65	USA	0	0	143
66	VENEZUELAN	0	21	9
67	YUGOSLAVIAN	0	0	2132

Tabela 7: ABCC por país e coorte

	COUNTRY	1840	1850	1860	1870	1880	1890	1900	1910
1	ARGENTINIAN		84.7	87.0	93.8	99.2			
2	AUSTRIAN	90.3	96.1	96.8	93.0	97.5	99.5	95.9	
3	BRAZILIAN	69.5	65.5	70.2	78.3	83.6	87.4	94.7	
4	CZECH		94.7	95.3	100.0		97.8	100.0	95.7
5	GERMAN	91.7	95.3	96.2	95.0	99.5	98.8	100.0	98.4
6	HUNGARIAN		97.4	96.4	95.6	96.0	95.1	100.0	94.1
7	ITALIAN	95.4	97.4	97.3	97.8	98.4	99.9	100.0	98.8
8	JAPANESE			98.7	100.0	99.6	100.0	100.0	100.0
9	LETHONIAN			90.9	94.6	93.2	93.0	100.0	97.9
10	LITHUANIAN				92.8	96.2	93.8	99.6	100.0
11	POLISH	78.6	89.5	94.6	94.2	87.3	96.2	100.0	100.0
12	PORTUGUESE	93.1	96.0	98.1	98.5	98.2	97.5	100.0	99.2
13	ROMANIAN				91.9	94.8	94.3	100.0	
14	RUSSIAN	72.4	86.6	97.3	92.9	97.1	98.0	100.0	98.3
15	SPANISH	85.5	92.4	94.6	97.8	97.5	100.0	97.2	99.8
16	SYRIAN	37.5	36.8	49.3	70.8				
17	UCRANIAN							100.0	99.9
18	YUGOSLAVIAN					99.1	95.3	100.0	98.7