Pesquisa Operacional para o Desenvolvimento
Vol. 11, n. 1, p. 6-25, 2019

ISSN 1984-3534

INCORPORANDO TÉCNICAS DE *MACHINE LEARNING* AO ESTUDO DA MIGRAÇÃO INTERNA NO BRASIL[1]

Pier Francesco De Maria[a*]

[a]Escola Superior de Administração, Marketing e Comunicação (ESAMC), Campinas-SP, Brasil.

Recebido 01/01/2018, aceito 13/03/2019

RESUMO

Embora o estudo da migração interna, de modo geral, seja feito sobre recortes geográficos que avaliam os fluxos de origem e destino em determinada área, a compreensão da inserção dos municípios e dos migrantes na malha migratória nacional fica aquém do possível. A partir dos microdados do Censo de 2010, este artigo se propõe a resumir o conjunto de fluxos migratórios no Brasil nos principais padrões, bem como agrupar os diversos fluxos e avaliar quais são as principais variáveis a serem incluídas nos estudos migratórios. A pesquisa é conduzida combinando uma análise de componentes principais com a obtenção de regras de associação pelo algoritmo *a priori*. Ademais, lança-se mão do algoritmo *random forest* e da análise de clusters para ter um perfil mais completo dos fluxos migratórios intermunicipais brasileiros.

Palavras-chave: Migração intermunicipal, Aprendizado de máquina, Big microdata.

ABSTRACT

Although the study of internal migration is generally done on geographic cuttings that assess the flows of origin and destination in a given area, the understanding of the insertion of municipalities and migrants in the national migratory network is far from the possible. Based on the 2010 Census microdata, this article proposes to summarize the set of migratory flows in Brazil in the main patterns, as well as to group the different flows and to evaluate the main variables to be included in the migratory studies. The research is conducted by combining a principal component analysis with the obtaining of association rules by the *a priori* algorithm. In addition, we uses the *random forest* algorithm and a clusters analysis to have a more complete profile of the intermunicipal migratory flows in Brazil.

Keywords: Intermunicipal migration, Machine learning, Big microdata.

* Autor para correspondência. E-mail: dpierf@gmail.com
DOI: 10.4322/PODes.2019.002

[1]*Todos os autores assumem a responsabilidade pelo conteúdo do artigo.*

1. Introdução

O estudo da dinâmica da migração interna, de modo geral, é realizado a partir de recortes geográficos que avaliam os fluxos de origem e destino em determinada área. Ademais, o uso dos dados sobre migração do Censo Demográfico brasileiro, via de regra, se restringe à análise do volume de migrantes, bem como ao cálculo de taxas brutas de imigração e emigração. Nesta direção, as pesquisas na área de migração têm se concentrado em compreender os fluxos inter-regionais e inter-estaduais, bem como aqueles intra-metropolitanos (Cunha, 2005). Embora estas pesquisas sejam válidas e relevantes para compreender a dinâmica migratória local, duas limitações aparecem. De um lado, os recortes geográficos limitam a compreensão da inserção dos municípios e dos migrantes na malha migratória nacional. De outro lado, o uso de dados gerais sobre fluxos mascara os porquês por trás da escolha de sair de um município para se dirigir até outro.

As reconfigurações pelas quais têm passado as relações entre a migração e as questões socioeconômicas (Baeninger, 2012), ao se estudar este tema a partir dos recortes e das limitações acima, são entendidas menos claramente. Tais reconfigurações se processam de forma dinâmica no tempo e no espaço, afetadas por dimensões individuais-familiares, bem como por questões relativas a toda a sociedade (Vignoli, 2007). Neste sentido, a análise da migração interna precisa levar em consideração as diversas escalas nas quais o fenômeno, necessariamente, se processa (Brandão, 2007; Vainer, 2002), entre as quais aparecem a família, o município e o país. Ademais, é preciso recordar que a dinâmica migratória é permeada por um amplo conjunto de seletividades, já que o migrante é a exceção à regra na população (Campos, 2015). Tais seletividades migratórias se originam de duas formas: (1) dos perfis socioeconômico, demográfico, laboral e educacional do migrante; e (2) das características dos municípios de origem e de destino destes fluxos migratórios. Além disso, é preciso considerar a expressiva extensão do território brasileiro, combinada às heterogeneidades econômica, social e cultural inerentes a este espaço.

Estudar a migração interna demanda usar informações de dois tipos. O primeiro (relativo ao migrante) inclui: (1) os perfis sociodemográfico e econômico dos fluxos; (2) as diferenças, nos municípios, entre migrantes e 'nativos'; e (3) as diferenças entre os que saíram de um município e os que não migraram. O segundo (relacionado aos municípios) abarca: (1) os perfis econômico, demográfico e social dos municípios; e (2) as diferenças existentes entre municípios de origem e de destino. Este conjunto de informações, se adequadamente processado, tem potencial para subsidiar a compreensão da constituição dos espaços migratórios (Baeninger, 1999). A fim de compreender a dinâmica migratória da forma como aqui se sugere, é preciso lançar mão de um conjunto mais amplo de dados, os quais têm um nível de detalhamento muito maior.

Nos últimos anos, a Demografia tem se aberto à revolução de dados que tem ocorrido, com o aparecimento do chamado '*big data*' (Letouzé, 2015). Entretanto, é preciso frisar que, de modo geral, ter grandes volumes de dados não é algo novo para os estudos demográficos (Billari e Zagheni, 2017), uma vez que recenseamentos são realizados desde o século XVIII. Além do mais, desde meados do século XX até hoje, tem ocorrido uma revolução em termos de disponibilização de microdados (McCaa e Ruggles, 2002). Esta revolução nos microdados tem motivado o uso do termo '*big microdata*' (Ruggles, 2014), justificado pelo fato de os microdados censitários terem algumas peculiaridades compatíveis com as que caracterizam o '*big data*'.

Entre as peculiaridades, Ruggles (2014, p. 295) destaca: (1) o alto nível de estruturação dos microdados; (2) os baixos níveis de não-resposta; e (3) o fato de os dados fornecerem informações abrangentes sobre a população. Para tirar o maior proveito possível do '*big microdata*', é preciso utilizar métodos e técnicas que sejam capazes de transformar o conjunto de dados em conhecimento que tenha sentido lógico (González-Bailón, 2013). Dentro das Ciências Sociais, o uso de técnicas de aprendizado de máquina (*machine learning*) tem sido uma das principais abordagens para descoberta de conhecimento a partir de bancos de (micro)dados (Foster et al., 2016).

Entretanto, a literatura sobre aplicações de técnicas de aprendizado de máquina à migração é recente e escassa, com raras implementações para dados censitários. Franco-Arcega et al. (2014) aplicam técnicas de mineração de dados (*data mining*) para identificar padrões nos fluxos migra-

tórios do Estado de Hidalgo a partir de dados do Censo Demográfico do México de 2010. Pande e Rajan (2015), por sua vez, utilizando os Censos da Índia de 1991 e 2001, fazem uma Análise Exploratória de Dados Espaciais (AEDE) e ajustam uma regressão linear para identificar e estimar os padrões migratórios do Estado de Andhra Pradesh. Por fim, em sua dissertação, Lammers (2017) aplica técnicas de aprendizado de máquina para prever fluxos migratórios internacionais com dados da Organização para a Cooperação e o Desenvolvimento Econômico (OCDE).

Por outro lado, há mais estudos que se valem de dados não censitários. Barchiesi et al. (2015) aplicam técnicas de aprendizado de máquina para estabelecer a probabilidade de algum fluxo migratório acontecer (com o objetivo de quantificar e modelar padrões de mobilidade humana), a partir de dados georreferenciados do *site* Flickr. Por sua vez, Lindström (2017) tenta prever o volume de fluxos migratórios na Suécia com registros administrativos dos últimos 50 anos, por meio de regressões logísticas e redes neurais. Estas aplicações (com dados censitários ou fontes alternativas) mostram a dificuldade de obter um modelo generalista para análise e predição da migração (interna e/ou internacional), bem como para identificação de padrões.

Houve, porém, duas tentativas recentes nesta direção. De um lado, Simini et al. (2012), a partir de um aperfeiçoamento da lei da gravitação universal de Newton, propuseram um modelo para estudo da migração e da mobilidade, (o '*radiation model*'). De outro lado, Robinson e Dilkina (2018) apresentam um modelo de aprendizado de máquina que seria aplicável a qualquer contexto, sem restrições em termos de variáveis empregadas. Em seus testes, este modelo tem desempenho melhor do que os modelos tradicionais para estudo da mobilidade humana - incluindo desdobramentos como o de Simini et al. (2012).

Este levantamento bibliográfico mostrou que há poucos trabalhos que aplicam aprendizado de máquina aos dados censitários para compreender a migração interna no Brasil. Diante disto, este artigo tem por objetivos: (1) identificar os principais padrões existentes nos fluxos migratórios brasileiros; (2) analisar comparativamente a composição dos diversos padrões migratórios; (3) identificar as principais variáveis que afetam a migração no Brasil; e (4) agrupar e diferenciar os fluxos migratórios de grande porte (acima de 1.000 migrantes). Tais objetivos dialogam com os anseios de Ruggles (2014, p. 295, tradução minha), para o qual "é preciso de novas estratégias de pesquisa, modelagem e mineração de dados para capitalizar, em escala e escopo, tais fontes".

2. Materiais e Métodos

2.1. Fonte de Dados e Variáveis Utilizadas

Para alcançar os objetivos propostos, são levantadas informações para todos os 5.565 municípios brasileiros, no ano de 2010, coletando variáveis sobre os seguintes temas: (1) Composição sociodemográfica; (2) Perfil educacional; (3) Nupcialidade e família; (4) Pobreza e desigualdade; (5) Mercado de trabalho; e (6) Dinâmica econômica. Ao todo, 55 variáveis quantitativas contínuas (descritas no Apêndice 1) foram selecionadas de forma arbitrária[1], buscando contemplar a maior parte dos temas trabalhados no Censo Demográfico. As variáveis de dinâmica econômica, por não serem levantadas pelo Censo, foram extraídas da pesquisa "Produto Interno Bruto dos Municípios" (realizada, assim como o Censo, pelo Instituto Brasileiro de Geografia e Estatística - IBGE).

2.2. Redução de Dimensionalidade

As variáveis passaram por um processo de redução (extração) de atributos, o qual ajuda a manter apenas as componentes mais relevantes - algo que pode ser feito manualmente, mas também por meio de métodos automáticos (Witten e Frank, 2005). Como vantagens da redução da dimensionalidade, os autores apontam (Witten e Frank, 2005, p. 289):

- O algoritmo de aprendizado tem um ganho de performance;

[1]Há mais variáveis disponíveis no Censo Demográfico do que as selecionadas. Neste estudo, foram escolhidas aquelas que trariam informações mais relevantes em termos dos temas apontados.

- O modelo final é mais simples e fácil de interpretar;
- O foco do pesquisador se direciona para as principais variáveis.

Algumas técnicas comuns para extração de atributos são: a Análise de Componentes Principais (PCA), a Árvore de Decisão, a Regressão (Han et al., 2012) e o Escalonamento Multidimensional. Uma vez que os atributos são numéricos, além de que estamos buscando relações de interdependência entre as variáveis e não temos um atributo-meta envolvido neste primeiro banco de dados, a extração de atributos será feita utilizando PCA. Este método, algebricamente descrito por Jolliffe e Cadima (2016), tem o objetivo de maximizar a variância de uma função linear composta por p coeficientes (os atributos), cada qual tem um coeficiente a atribuído. Ao adicionar a restrição $a^T a = 1$, para que o problema de maximização tenha solução bem definida, a questão pode ser resolvida por meio de multiplicadores de Lagrange, de modo que a solução do PCA seja um autovalor λ para cada autovetor **a**:

$$\sum_{j=1}^{p} a_j \mathbf{x}_j = \mathbf{Xa} \Rightarrow var(\mathbf{Xa}) = \mathbf{a}^T\mathbf{Sa} \qquad var(\mathbf{Xa}) = \mathbf{a}^T\mathbf{Sa} = \lambda a^T a = \lambda \tag{1}$$

O próprio Jolliffe (1972, 1973) demonstrou, para pesquisas envolvendo o uso de PCA, que valores de λ abaixo de 0.70 são desprovidos de utilidade prática, podendo este valor ser utilizado como corte para seleção do número de atributos a reter. Adicionalmente, para evitar que uma variável se sobressaia em relação a outros atributos (devido à excessiva amplitude), todos eles passarão por um processo de normalização pelo método de *scores* Z, algebricamente descrito a seguir. O novo valor da j-ésima variável x para a i-ésima observação é calculado levando-se em consideração a média μ e o desvio-padrão σ da variável x_j:

$$z_{ij} = \frac{x_{ij} - \mu(x_j)}{\sigma(x_j)}, \quad z \in [-\infty; +\infty] \Longrightarrow \begin{cases} \mu(z_j) = 0 \\ \sigma(z_j) = 1 \end{cases} \tag{2}$$

2.3. Construção dos Bancos de Dados

As variáveis selecionadas via PCA foram utilizadas em um novo banco de dados. Este, ao invés de conter informações sobre os municípios, é montado a partir das transações (fluxos migratórios) existentes entre os municípios. Cada observação do banco de dados contém a informação do número total de migrantes e sua distribuição segundo o perfil sociodemográfico, bem como as variáveis relativas ao perfil dos municípios de origem e de destino. Deste modo, os atributos do banco de dados final (o qual é utilizado daqui em diante) são:

- Geocódigos dos municípios de origem (A) e de destino (B);
- Valor da transação (A $\rightarrow$ B), em número de migrantes;
- Proporção de migrantes por sexo e idade média do fluxo;
- Proporção de migrantes por escolaridade e nível de renda;
- Variáveis (selecionadas por PCA) do município A;
- Variáveis (selecionadas por PCA) do município B.

Os atributos relativos aos municípios A e B já foram extraídos por PCA; por sua vez, o fluxo migratório entre os municípios e sua composição sociodemográfica foram obtidos a partir da adaptação de uma rotina já elaborada em SAS® (Maria, 2017) (disponível no Apêndice 2), a qual permite extrair uma tabela de vetores origem-destino, com todos os atributos acima elencados.

Depois de a tabela final de atributos, a qual conta com um total de 298.494 transações ocorridas entre 2005 e 2010[2], o último passo é criar um outro banco com todas as variáveis discretizadas, uma vez que nosso interesse é também trabalhar com algoritmos que gerem um conjunto

[2]Do total de transações efetivamente ocorridas (298.494), foram subtraídos 47 fluxos que envolveram o município de Nazária (PI, geocódigo 2206720), por este ter sido criado depois de 2005.

de "regras" acerca dos fluxos migratórios no Brasil. A discretização foi realizada manualmente, criando categorias relevantes aos interesses desta pesquisa (de modo a identificar se algum segmento populacional específico compõe os fluxos migratórios). Os atributos do banco de dados final, bem como suas categorias, são apresentados no Quadro 1.

Quadro 1: Classes criadas por meio do processo de discretização manual de atributos.

Atributo	**Classes criadas**
Sexo	Abaixo de 50%: mais mulheres – Acima de 50%: mais homens
Raça/cor(1)	Até 25%: muito mais negros – De 25 a 50%: mais negros
	De 50 a 75%: mais brancos – De 75 a 100%: muito mais brancos
Escolaridade	Até 25%: muito mais SE/EF – De 25 a 50%: mais SE/EF
	De 50 a 75%: mais EM/ES – De 75 a 100%: muito mais EM/ES
Idade	Até 25: idade média jovem – De 25 a 39: idade média adulta-jovem
	De 40 a 59: idade média adulta-idosa – Acima de 59: idade média idosa
RDPC (SM)(2)	Até 1.0 SM – De 1.0 a 2.0 SM – De 2.0 a 3.0 SM
	De 3.0 a 5.0 SM – Mais de 5.0 SM
Fluxo	Até 24: irrelevante – De 25 a 49: muito baixo – De 50 a 99: baixo
	De 100 a 499: médio – De 500 a 999: alto – Acima de 1.000: muito alto
PCA(3)	Até -2.0: muito baixo – De -1.9 a -1.0: baixo – De -0.9 a 0.9: médio
	De 1.0 a 1.9: alto – Acima de 2.0: muito alto

(1) 'Brancos' inclui amarelos. 'Negros' incorpora pretos, pardos e indígenas. (2) Salário mínimo de 2010: R$ 510, definido pela Lei 12.255/2010. (3) Cortes de atributos oriundos do PCA feitos com base no número de desvios-padrão.

Fonte: Elaboração do autor. Classes criadas manualmente, a critério do autor.

2.4. Técnicas Empregadas

As técnicas utilizadas para análise do banco de dados final, a fim de alcançar os objetivos apresentados na introdução deste artigo, foram três: (1) regras de associação; (2) análise de clusters; e (3) *random forest* (combinado com a taxa de ganho de informação para seleção de atributos). Com a combinação destas três técnicas, é possível: (1) descobrir padrões relevantes em termos de fluxos migratórios no Brasil; (2) alocar os diversos fluxos municipais em grupos, bem como avaliar as diferenças entre eles; e (3) definir os principais atributos que diferenciam as migrações internas.

As regras de associação são utilizadas para descobrir quais atributos são comumente mais combinados (Hastie et al., 2008). Com isto, nosso interesse é descobrir quais classes de cada atributo (bem como quais atributos) podem estar associados aos fluxos migratórios, por serem as combinações mais frequentes. Para aplicação deste método de aprendizado não supervisionado, recorremos ao algoritmo *a priori*, proposto por Agrawal e Skirant (1994). Para análise de quais sejam as principais regras de associação, recorremos a duas métricas - o suporte (T) e a confiança (C) - definidas abaixo em termos probabilísticos. Como critérios para a pesquisa, foram utilizados T_{min} de 0.10 e C_{min} de 0.60.

$$A \Rightarrow B, \quad \begin{cases} C(A \Rightarrow B) = \dfrac{T(A \Rightarrow B)}{T(A)} = Pr(B|A) \\ T(A \Rightarrow B) = Pr(A \cup B) \end{cases} \tag{3}$$

Na sequência, a aplicação do algoritmo *random forest* (RF) é utilizada para descobrir quais são os principais atributos a explicarem os fluxos migratórios analisados. O algoritmo RF, apresentado por Breiman (2001), é um aprimoramento do *bagging*, fazendo com que as árvores (construídas a partir de amostras de treino obtidas por *bootstrap*) sejam não-correlacionadas entre si, obtendo um modelo de decisão mais confiável e realista (James et al., 2015). Este procedimento é

preferível a outros métodos de comitê (como o *boosting*) por ser mais robusto e menos suscetível a erros e a *outliers*, além de ser resistente a *overfitting* (Han et al., 2012).

Por fim, adotamos a clusterização para verificar como todas as transações analisadas podem ser agrupadas por afinidade em termos dos atributos usados. Como não estamos interessados em relações de dominância-dependência (e os atributos do banco final são categóricos), o agrupamento é feito via particionamento, utilizando o algoritmo *k-medoids*. Este método de aprendizado não-supervisionado elege um candidato (centroide) representativo do agrupamento; na sequência, atribuem-se cada instância ao cluster com o qual tiver maior similaridade com o representante, a partir de um critério de erro absoluto (Han et al., 2012). O objetivo é minimizar esta função para as p instâncias a serem alocadas nos $i = (1, \ldots k)$ agrupamentos.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p,\, o_i) \tag{4}$$

2.5. *Software* Utilizado

A fim de realizar todas as análises propostas, os datasets são processados por meio do *software* Weka®, versão 3.8.1[3]. A aplicação de todos os procedimentos aqui utilizados (PCA, regras de associação, *random forest* e clusterização) no *software* está exaustivamente descrita em Frank et al. (2016), nas seções 2.4 a 2.7.

3. Resultados e Discussão

A primeira etapa dos resultados consiste na execução do PCA. O número de componentes a ser escolhido levou em consideração os diversos critérios comentados na seção 2.2. As primeiras 17 componentes tiveram autovalor superior a 0.70 (limiar estabelecido pelo critério de Jolliffe). Uma vez que o número de componentes extraído foi excessivo, recorreu-se a outros critérios para seleção do número de componentes. Hair et al. (2009) sugerem o critério da variância acumulada e o método do cotovelo como instrumentos auxiliares na definição do número de componentes. Para os autores, em pesquisas na área de Ciências Sociais, é possível extrair um número de componentes que explique em torno a 60% da variância. Deste modo, optou-se por extrair as primeiras 4 componentes principais, responsáveis por 58.1% da variância acumulada.

As componentes estão descritas no Quadro 2. Estas são muito bem definidas (no sentido de as principais variáveis poderem ser compreendidas de maneira conjunta), de modo que nomeá-las é um processo relativamente fácil. Todas as componentes envolvem, majoritariamente, variáveis sociais e populacionais, embora diversas informações a respeito da dinâmica econômica também apareçam. Disto, observa-se que os principais pontos de diferenciação entre os municípios são o social e o econômico - o que vai de encontro com o apontado por Magalhães e Miranda (2009). As componentes extraídas permitem algumas inferências gerais:

- Quando o *score* da componente 1 for elevado, há indícios de maiores níveis de vulnerabilidade e pobreza combinados com menor desenvolvimento e renda *per capita*;
- *Scores* maiores na componente 2 refletem maiores níveis de desocupação e uma proporção mais elevada de famílias do tipo 'monoparental';
- Na componente 3, um *scores* mais elevados se relacionam à maior presença do setor de serviços frente ao agrícola, bem como a uma maior a proporção de idosos;
- Por fim, *scores* maiores na componente 4 estão associados à maior desigualdade e à menor participação industrial e/ou de empregados com carteira no município.

[3]O programa pode ser baixado gratuitamente em `https://www.cs.waikato.ac.nz/ml/weka/`.

Quadro 2: Componentes extraídas da tabela de informações municipais, principais variáveis e nome atribuído a cada componente.

PC	Principais variáveis	Nome
1	1- IDHM (-)	**Vulnerabilidade e pobreza**
	2- % de pobres (+)	
	3- Renda per capita (-)	
	4- IVS (+)	
2	1- Índice de envelhecimento (-)	**Dimensão demográfica**
	2- % de casados (-)	
	3- Taxa de desocupação (+)	
	4- % de famílias "mãe com filhos"(+)	
3	1- % de famílias "mãe com filhos"(+)	**Participação econômica**
	2- Participação dos Serviços no VAB (+)	
	3- Participação da Agricultura no VAB (-)	
	4- Índice de envelhecimento (+)	
4	1- Índices de Gini e Theil-L (+)	**Desigualdade social**
	2- Participação da Indústria no VAB (-)	
	3- População total (+)	
	4- % de empregados com carteira (-)	

Nota. Os sinais apresentados indicam a carga de cada variável para definição dos *scores*.

Fonte: Variáveis do Apêndice 1. Tabulação e elaboração do autor.

Após obter as componentes por PCA e discretizá-las (conforme Quadro 1), foram construídas as regras de associação para o conjunto de fluxos migratórios brasileiros[4]. Todavia, mesmo com o baixo nível de confiança exigido para a criação de uma regra (60%), o procedimento não obteve êxito. Deste modo, optou-se por extrair regras de associação individualmente (isto é, para cada classe de fluxos migratórios), utilizando o nível de renda *per capita* (também discretizado conforme o Quadro 1) como atributo-meta. Com isto, foi possível observar quais as principais características encontradas para cada perfil de renda em cada classe de fluxo.

Os resultados obtidos (que nos auxiliam em alcançar os objetivos 1 e 2 desta pesquisa) se encontram na Figura 1. É possível observar uma série de repetições em todas as classes de fluxo, o que mostra que as migrações são um fenômeno que, via de regra, abrange um grupo socioeconômico específico. Menores níveis de renda, escolaridade baixa e pessoas na fase adulta (geralmente homens pretos e pardos) são as características que tendem a ser as mais relevantes. Tais resultados reforçam as já antigas leis de migração de Ravenstein (1980), bem como os fatores de atração-repulsão e os obstáculos intervenientes de Lee (1980).

Adicionalmente, os municípios de origem e destino são bastante similares entre si; quando isto não ocorre, o município de destino, de modo geral, tende a ter melhores condições de vida. Além disso, o migrante está especialmente de olho nos níveis de pobreza e desigualdade, bem como no nível de desenvolvimento do município que o receberá - resultado que se aproxima das observações de Todaro (1980) sobre migração de mão-de-obra para as áreas urbanas. Por fim, é notável que os fluxos de menor envergadura tenham maiores percentuais de pessoas de mais baixa renda. Neste sentido, fluxos menos intensos (geralmente de cunho local e/ou que não são direcionados a grandes centros urbanos) são, majoritariamente, de pessoas mais pobres.

[4]Na aplicação do algoritmo *a priori*, foram excluídos os fluxos migratórios com menos de 25 pessoas, o que corresponde a 67.4% do total de transações entre 2005 e 2010. Esta exclusão foi realizada por conta da elevada sensibilidade das distribuições de frequência a pequenas amostras.

Figura 1: Características gerais dos fluxos migratórios internos obtidas com base nas regras de associação para cada nível, Brasil (2005-2010).

Fluxo	Perfil de renda	Características dos fluxos
Muito alto (1000+)	1 a 2 SM (36,3%)	1) Jovens-adultos (25-39) e negros, alguns com ensino médio 2) Grandes municípios de destino com desigualdade média-baixa 3) Municípios de origem com altas taxas de desemprego
	2 a 3 SM (20,5%)	
Alto (500-999)	1 a 2 SM (40,4%)	1) Pessoas negras, com pouca escolaridade, nem sempre jovens 2) Municípios de origem com altas taxas de desemprego 3) Municípios de destino semelhantes aos de origem
	0 a 1 SM (25,4%)	
Médio (100-499)	0 a 1 SM (45,1%)	1) Pessoas jovens (0-24), geralmente negros, com baixa escolaridade 2) Origem e destino com níveis elevados de pobreza/vulnerabilidade 3) Equilíbrio entre níveis médios de desigualdade e de desemprego
	1 a 2 SM (34,4%)	
Baixo (50-99)	0 a 1 SM (54,2%)	1) Origem e destino com níveis elevados de pobreza/vulnerabilidade 2) Pessoas de baixa escolaridade, nem sempre homens e/ou jovens 3) Cidades com desigualdade média e equilíbrio de agricultura e serviços
	1 a 2 SM (29,1%)	
Muito baixo (25-49)	0 a 1 SM (57,0%)	1) Origem e destino com níveis elevados de pobreza/vulnerabilidade 2) Jovens (0-24), com pouca escolaridade, nem sempre negros 3) Cidades com desigualdade média e equilíbrio de agricultura e serviços
	1 a 2 SM (26,4%)	

Nota: Excluídos fluxos inter-municipais com menos de 25 migrantes.

Fonte: IBGE - Censo Demográfico de 2010. Tabulação e elaboração do autor.

Na sequência, foi aplicado o algoritmo *random forest* (RF), cujas medidas de avaliação aparecem na Tabela 1 (e nos ajudam a alcançar o objetivo 2 deste artigo). A classe majoritária (fluxos de menor volume) é prevista de maneira 'liberal', admitindo mais falsos positivos (FP), enquanto a classe minoritária (fluxos mais volumosos) é prevista de forma mais rígida, com menos falsos positivos, mas rejeitando muitos verdadeiros positivos (VP). No conjunto, as áreas embaixo da curva ROC (*Receiver Operating Characteristic*) são maiores para fluxos mais volumosos (numericamente infrequentes), enquanto a precisão é maior para as classes majoritárias. O modelo tem uma baixa taxa de verdadeiros positivos (47.8%), enquanto admite também um nível razoável de falsos positivos (33.1%). A partir destes resultados, entendemos que: (1) os fluxos migratórios são expressivamente heterogêneos, o que reduz o poder preditivo do modelo; (2) os atributos utilizados na análise podem ser insuficientes, o que pode indicar a necessidade de incluir mais variáveis ou, ainda, a presença de atributos latentes.

Tabela 1: Medidas de avaliação do algoritmo *random forest* aplicado aos dados sobre fluxos migratórios no Brasil (2005-2010).

Classe de fluxo	Taxa VP	Taxa FP	Precisão	Área ROC
Muito baixo	0,711	0,504	0,576	0,654
Baixo	0,221	0,196	0,305	0,526
Médio	0,319	0,140	0,366	0,665
Alto	0,052	0,007	0,107	0,705
Muito alto	0,145	0,005	0,220	0,824
Média geral	**0,478**	**0,331**	**0,447**	**0,623**

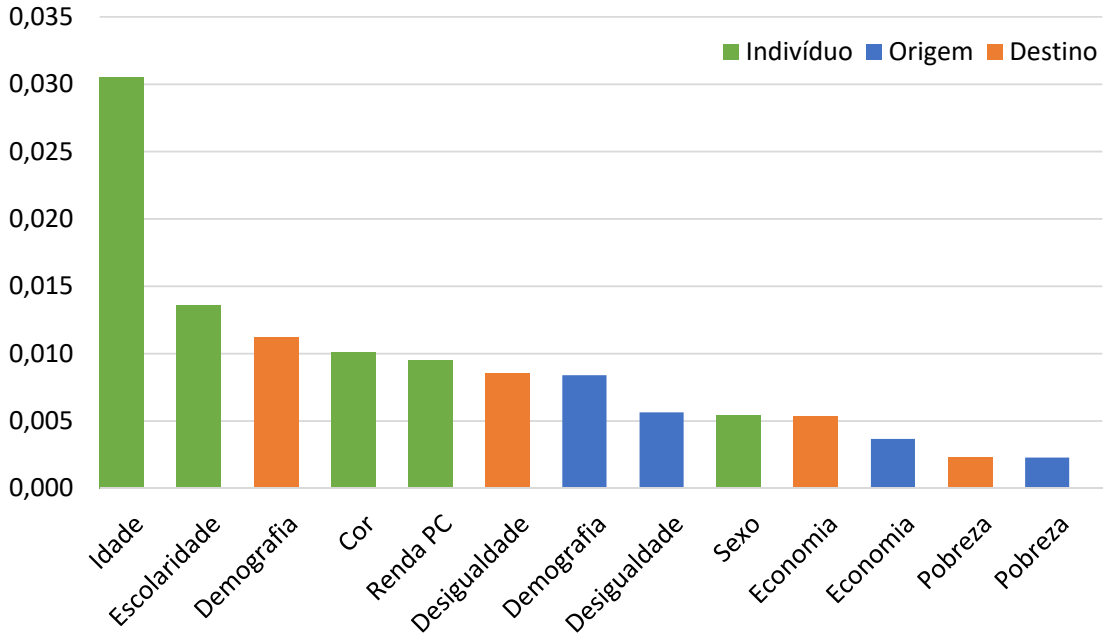
Nota: Excluídos fluxos inter-municipais com menos de 25 migrantes.

Fonte: IBGE - Censo Demográfico de 2010. Tabulação e elaboração do autor.

Adicionalmente, para entender como cada atributo afeta o processo de criação das árvores do RF, obtiveram-se as taxas de ganho de informação para todas as variáveis (de modo a alcançar

o objetivo 3 do artigo). Na Figura 2, percebe-se que os principais atributos para segmentação da informação são a idade do migrante, seu nível de escolaridade, a composição demográfica do município de destino, a raça/cor e o nível de renda per capita do migrante. Ao observar que quatro dos cinco principais atributos são características do migrante, é possível inferir que boa parte da seletividade migratória (Lee, 1980) e dos fatores de atração e de expulsão (Singer, 1973) está no perfil do migrante: pessoas migram por haver oportunidades no destino, contanto que satisfaçam um possível 'perfil desejado' - o que também dialoga com as proposições de Todaro (1980).

Figura 2: Taxa de ganho de informação para atributos relativos aos fluxos migratórios intermunicipais, Brasil (2005-2010).



Fonte: IBGE - Censo Demográfico de 2010. Tabulação e elaboração do autor.

Por último, foi aplicado o algoritmo *k-medoids* para analisar como as transações (e, portanto, os municípios que 'recebem' e/ou 'enviam' migrantes) se agrupam e quais as diferenças entre os grupos[5]. Esta análise complementa os resultados da Figura 1 e nos auxilia a completar os objetivos 1 e 2 desta pesquisa. A escolha do número 'ideal' de agrupamentos foi feita pela soma da distância interna dos clusters, adotando como critério de parada o momento em que a redução deixasse de ser a taxas monotônicas, o que levou à definição de 9 grupos.

Os centroides obtidos (Figura 3) mostram que todos os fluxos têm maior participação de pessoas de baixa renda. Com poucas exceções, os fluxos são de pessoas jovens-adultas (salvo nos grupos 8 e 9), com distribuição semelhante por sexo. Exceto o grupo 9, o nível de pobreza/vulnerabilidade no município de destino é sempre menor ou parecido ao encontrado no município de origem. O fluxo historicamente mais relevante e mais estudado (Nordeste-Sudeste) aparece apenas duas vezes entre os 3 principais (2º mais frequente no cluster 1 e 3º no cluster 9), enquanto a maior parte dos fluxos de grande magnitude tem ocorrido em escala intrarregional.

Estes elementos recolocam em pauta o fato de as migrações internas estarem assumindo contornos mais locais, com espaços migratórios (Baeninger, 1999) em nível micro e mesorregional. Adicionalmente, tais resultados dão suporte à noção de rotatividade migratória (Baeninger, 2012): se as trocas migratórias têm ocorrido em nível local, há uma maior propensão a pouquíssimos municípios (geralmente, algumas capitais como São Paulo, Rio de Janeiro e Brasília) serem puramente receptores durante muito tempo.

[5]Para a análise, retiraram-se os fluxos com menos de 50 migrantes, uma vez que, em termos estatísticos, são altamente sensíveis a variações na sua composição sociodemográfica. Com isto, restaram 49.544 instâncias.

Figura 3: Centroides dos agrupamentos para os fluxos migratórios intermunicipais e principais fluxos regionais, Brasil (2005-2010).

Variáveis		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
Do migrante	Sexo	+ mulheres	+ homens	+ mulheres	+ mulheres	+ homens	+ mulheres	+ homens	+ mulheres	+ mulheres
	Raça/cor	++ negros	+ negros	++ negros	++ brancos	+ brancos	+ negros	++ brancos	++ brancos	+ negros
	Escolaridade	Muito baixa	Muito baixa	Muito baixa	Baixa	Muito baixa	Baixa	Média	Muito baixa	Muito baixa
	Idade	Jovem-adulto	Jovem-adulto	Jovem-adulto	Jovem-adulto	Jovem-adulto	Jovem-adulto	Jovem-adulto	Jovem	Jovem
	Renda PC	0-1 SM	0-1 SM	0-1 SM	1-2 SM	1-2 SM	1-2 SM	1-2 SM	0-1 SM	0-1 SM
Do destino	Comp. 1	Muito baixo	Muito baixo	Muito baixo	Muito baixo	Muito baixo	Muito baixo	Muito baixo	Muito baixo	Muito alto
	Comp. 2	Muito alto	Muito alto	Médio	Muito alto	Muito alto	Muito alto	Muito alto	Muito alto	Médio
	Comp. 3	Muito alto	Médio	Médio	Alto	Médio	Muito alto	Muito alto	Muito alto	Médio
	Comp. 4	Médio	Médio	Médio	Alto	Médio	Muito alto	Muito alto	Muito alto	Médio
Da origem	Comp. 1	Muito alto	Médio	Muito baixo	Muito baixo	Muito baixo	Muito baixo	Muito baixo	Muito baixo	Muito baixo
	Comp. 2	Médio	Muito alto	Muito alto	Muito alto	Muito alto	Alto	Muito alto	Muito alto	Muito alto
	Comp. 3	Alto	Médio	Baixo	Muito alto	Médio	Alto	Muito alto	Muito baixo	Muito alto
	Comp. 4	Médio	Médio	Médio	Muito alto	Médio	Médio	Muito alto	Médio	Muito alto
Volume de fluxo		Maioria baixo	Maioria baixo	Maioria baixo	Maioria baixo	Maioria baixo	Maioria médio	Maioria médio	Maioria baixo	Maioria baixo
Casos		21,1%	13,1%	8,7%	13,5%	12,6%	8,9%	9,8%	5,2%	7,0%
Principais fluxos identificados		NE-NE	NE-NE	SE-SE	SE-SE	SE-SE	SE-SE	SE-SE	NO-NO	NE-NE
		NE-SE	SE-SE	SU-SU	SU-SU	SU-SU	NE-NE	SU-SU	SE-SE	SE-SE
		NO-NO	NO-NO	CO-CO	NE-NE	CO-CO	SU-SU	NE-NE	SU-SU	SE-NE
		SE-SE	CO-CO	NO-NO	CO-CO	NO-NO	CO-CO	SE-NE	NE-SE	NO-NO
		NE-CO	SU-SU	NE-NE	NO-NO	NE-NE	NE-SE	CO-CO	CO-CO	CO-CO

Nota: Excluídos fluxos inter-municipais com menos de 50 migrantes.

Fonte: IBGE - Censo Demográfico de 2010. Tabulação e elaboração do autor.

Finalmente, combinamos os resultados já apresentados (especialmente das Figuras 1 e 3) a dois indicadores municipais, a fim de aprimorar as interpretações dos achados apresentados sobre o perfil migratório do período 2005-2010: (1) a eficácia migratória (Baeninger, 1999, 2012) do município[6]; e (2) o tamanho do município[7]. Com isto, é possível ter uma avaliação mais profunda dos resultados observados na Figura 3, de modo a alcançar, especialmente, o objetivo 2 do trabalho. A Tabela 2 apresenta os resultados para o Índice de Eficácia Migratória (IEM), enquanto o Quadro 3 traz resultados por tamanho de município.

Em termos de IEM (Tabela 2), se observa que os fluxos migratórios no Brasil, independente do volume, são majoritariamente de áreas de evasão para áreas de absorção. Entretanto, há diferenças sutis (que aparecem nos dados abaixo) que variam conforme o volume do fluxo. As principais diferenças podem ser assim resumidas:

- Os principais fluxos migratórios têm duas direções: (1) saem de áreas de evasão (geralmente, municípios pequenos e/ou periféricos) para grandes aglomerações urbanas (municípios de alta absorção); ou (2) partem de tradicionais municípios receptores (geralmente, capitais ou grandes cidades) para novos polos regionais (em expansão nos últimos anos), geralmente próximos aos primeiros;
- Os fluxos de menor intensidade, aqueles que geralmente ocorrem por todo o interior do país, se dividem em dois grupos: (1) os que se dirigem para municípios de maior rotatividade (possivelmente, polos locais ou regionais, mas que não são capitais ou grandes centros estaduais); e (2) aqueles que migram 'na contramão' (em direção a municípios de evasão), apontando possíveis situações de migrações de retorno.

[6]O Índice de Eficácia Migratória (IEM) é a razão entre o saldo migratório (imigrantes menos emigrantes) e a migração bruta (soma dos dois primeiros), variando entre -1 e $+1$, e mensura quanto um município absorve ou repele migrantes. Neste trabalho, esta categorização foi assim revista: alta evasão (-1.000 a -0.301); evasão moderada (-0.300 a -0.121); tendência à evasão (-0.120 a -0.601); rotatividade (-0.600 a +0.600); tendência à absorção (+0.601 a +0.120); absorção moderada (+0.121 a +0.300); e alta absorção (+0.301 a +1.000).

[7]Esta classificação foi elaborada a partir de duas divisões prévias, de Martine (1994) e Silva (1946). A classificação dividiu os municípios em 6 grupos: vilas (até 20.000 habitantes); cidades pequenas (20.001 a 50.000); cidades médio-pequenas (50.001 a 100.000); cidades médias (100.001 a 500.000); cidades médio-grandes (500.001 a 1.000.000); e cidades grandes (acima de 1.000.000 de habitantes).

Tabela 2: Distribuição (%) dos índices de eficácia migratória por tamanho do fluxo e municípios de origem e de destino, Brasil (2005-2010).

IEM	Origem				Destino			
	Muito alto	**Alto**	**Médio**	**Baixo**	**Muito alto**	**Alto**	**Médio**	**Baixo**
Alta absorção	8,7	12,8	9,8	11,3	30,8	26,7	22,4	23,7
Absorção moderada	13,5	13,9	14,4	15,4	18,8	20,8	20,7	20,5
Tendência à absorção	6,3	7,3	7,2	7,6	7,7	8,7	8,1	8,3
Rotatividade	17,1	14,6	17,5	17,8	14,5	15,5	17,8	17,3
Tendência à evasão	3,4	6,0	7,1	7,1	3,0	4,3	5,6	5,5
Evasão moderada	41,0	31,4	28,0	24,6	22,8	20,8	19,9	18,0
Alta evasão	10,0	14,0	15,9	16,1	2,3	3,2	5,6	6,8
Total (N)	**994**	**1.545**	**19.692**	**27.313**	**994**	**1.545**	**19.692**	**27.313**

Fonte: IBGE - Censo Demográfico de 2010. Tabulação e elaboração do autor.

Observando os fluxos por tamanho do município (Quadro 3), além do que já foi pontuado, se percebe um fluxo menos esperado em transações acima de 500 migrantes: pessoas saem de municípios grandes em direção a outros de médio porte (100 a 500 mil habitantes). Embora existam fluxos expressivos rumo às grandes cidades, esta potencial 'contra-tendência' converge com as hipóteses de reversão da polarização (Azzoni, 1986) e de desconcentração econômica, industrial (Pacheco, 1996; Matos e Baeninger, 2008) e demográfica (Matos, 1995; Baeninger, 1998), os quais refletem diversos "espaços industriais obsoletos" (Faria, 1991). Tais fluxos se direcionam para fora dos limites metropolitanos (Lencioni, 2011), mas em municípios próximos às capitais estaduais - pelo menos no caso de São Paulo (Martine e McGranahan, 2010).

Quadro 3: Principais tamanhos de municípios de origem e de destino para fluxos com pelo menos 100 migrantes, Brasil (2005-2010).

	Muito alto	**Alto**	**Médio**
1º	Grande → Média (18,6%)	Média → Média (12,0%)	Média → Média (6,2%)
2º	Média → Média (11,8%)	Grande → Média (6,7%)	Pequena → Média (6,0%)
3º	Média → Grande (11,5%)	Média → Grande (6,6%)	Vila → Média (5,7%)
4º	Grande → Grande (7,8%)	Pequena → Grande (6,4%)	Vila → Pequena (5,0%)
5º	Grande → Pequena (6,0%)	Pequena → Média (6,0%)	Pequena → Pequena (4,1%)

Fonte: IBGE - Censo Demográfico de 2010. Tabulação e elaboração do autor.

Partindo destas peculiaridades, analisamos o perfil dos fluxos com pelo menos 1.000 migrantes (994 das quase 300 mil transações), a fim de atingir o objetivo 4 deste artigo. Estes fluxos foram submetidos a uma análise descritiva (Quadro 4) e na sequência, devido ao volume reduzido de transações, optou-se por aplicar o algoritmo *Expectation-Maximization* (EM)[8] para clusterização (Tabela 4). Em termos gerais, se observa que estes fluxos são bastante expressivos (em média, cerca de 500 pessoas por ano), com maior saída dos municípios maiores e destino para os de médio e médio-grande porte. Os principais resultados observáveis são:

- Os fluxos dos mais pobres saem da capital do estado em que residem rumo à região metropolitana. Esta é uma tendência generalizada até 3 salários mínimos em 2010;
- Fluxos para Campinas, Curitiba e São Paulo aparecem para quem recebia entre 3 e 5 salários mínimos, sendo uma situação intermediária entre os mais abastados e os pobres;

[8]Esta é uma extensão do algoritmo *k-means* original, para determinar o número ideal de agrupamentos através de um procedimento iterativo de estimação da máxima verossimilhança. O algoritmo foi originalmente apresentado por Dempster et al. (1977); o software Weka® permite a aplicação do algoritmo também para dados categóricos.

- Quando a renda *per capita* era acima de 5 salários mínimos, os fluxos se direcionavam para 5 capitais: São Paulo, Rio de Janeiro, Belo Horizonte, Brasília e Goiânia (SRBBG);
- Os municípios que fogem desta última tendência estão próximos às capitais de estado, em municípios de expansão imobiliária de alto padrão (como Santana do Parnaíba).

Quadro 4: Principais fluxos migratórios registrados entre os de maior intensidade, segundo renda *per capita* declarada pelo migrante na data do Censo, Brasil (2005-2010).

#	**Até 1 salário mínimo**	**De 1 a 3 salários mínimos**
1	Belo Horizonte → Ribeirão das Neves	São Paulo → Guarulhos
2	Brasília → Águas Lindas de Goiás	Goiânia → Aparecida de Goiânia
3	São Paulo → Itaquaquecetuba	Belo Horizonte → Contagem
4	Aracaju → Nossa Senhora do Socorro	Belém → Ananindeua
5	Fortaleza → Maracanaú	Recife → Jaboatão dos Guararapes
6	Natal → São Gonçalo do Amarante	Natal → Parnamirim
7	São Paulo → Francisco Morato	São Paulo → Praia Grande
8	Santarém → Manaus	Fortaleza → Caucaia
9	Brasília → Novo Gama	São Paulo → Osasco
10	São Luís → Paço do Lumiar	Brasília → Valparaíso de Goiás

#	**De 3 a 5 salários mínimos**	**Acima de 5 salários mínimos**
1	São Paulo → São Bernardo do Campo	Rio de Janeiro → São Paulo
2	Salvador → Lauro de Freitas	Rio de Janeiro → Niterói
3	São Paulo → Santo André	São Paulo → Rio de Janeiro
4	São Paulo → Cotia	Rio de Janeiro → Brasília
5	Salvador → São Paulo	Goiânia → Brasília
6	São Paulo → Campinas	São Paulo → Santos
7	São Paulo → Curitiba	São Paulo → Santana do Parnaíba
8	São Bernardo do Campo → Santo André	Belo Horizonte → São Paulo
9	São Gonçalo → Niterói	São Paulo → Brasília
10	São Paulo → São Caetano do Sul	Brasília → Rio de Janeiro

Nota: O salário mínimo em 2010 era de R$ 510, definido pela Lei 12.255/2010.

Fonte: IBGE - Censo Demográfico de 2010. Tabulação e elaboração do autor.

Tais resultados auxiliam na análise da divisão destes fluxos em grupos baseados, novamente, na composição sociodemográfica destes[9]. A clusterização pelo algoritmo EM retornou 6 classes, detalhadas na Tabela 4. Para além das diferenças numericamente visíveis, a análise dos fluxos em cada grupo trouxe informações bastante relevantes, as quais podem ser assim resumidas:

- Grupo 1 - Fluxos de caráter intra-metropolitano saindo das principais capitais de estado, envolvendo municípios de São Paulo e da região Sul. Fluxos predominantemente masculinos, de baixa escolaridade e renda *per capita* médio-baixa;
- Grupo 2 - Fluxos de caráter intra-metropolitano saindo das capitais, envolvendo destinos em São Paulo, no Nordeste e no Centro-Oeste. Fluxos majoritariamente femininos, de pessoas brancas, com renda *per capita* entre 2 e 3 salários mínimos;
- Grupo 3 - Fluxos interestaduais no polígono SRBBG e fluxos intra-estaduais direcionados para áreas de expansão urbana 'nobre'. Fluxos em sua grande maioria de pessoas brancas, com ensino médio ou superior e renda *per capita* acima dos 5 salários mínimos;

[9] A importância de realizar novamente esta etapa, agora apenas com fluxos mais expressivos, se deve ao fato de tais fluxos serem numericamente inexpressivos; isto justifica porque, na Figura 3, os mesmos não tiveram influência nos resultados. É também nestes fluxos que aparecem migrantes com rendas *per capita* mais elevadas.

- Grupo 4 - Fluxos de caráter intra-metropolitano saindo das capitais, envolvendo cidades do Nordeste, Sudeste e Centro, bem como fluxos interestaduais na região Norte. Fluxos predominantemente de pessoas negras, pobres e com baixos níveis de escolaridade;
- Grupo 5 - Fluxos de caráter intra-metropolitano, sobretudo de origem paulistana, com outros fluxos interestaduais para São Paulo e Paraná. Fluxos majoritariamente de pessoas brancas da classe média, com ensino médio completo ou frequentando a universidade;
- Grupo 6 - Fluxos de caráter intra-metropolitano saindo das capitais, envolvendo as regiões Norte e Nordeste, bem como o estado do Rio de Janeiro. Fluxos em sua maioria de pessoas negras e mulheres, com baixos níveis de renda e de escolaridade.

Tabela 4: Proporção (%) de fluxos com mais de 1.000 migrantes segundo variáveis sociodemográficas, por cluster de pertencimento, Brasil (2005-2010).

	C1	**C2**	**C3**	**C4**	**C5**	**C6**
Mais homens	80,6	26,1	32,8	40,8	37,7	20,8
Mais mulheres	19,4	73,9	67,2	59,2	62,3	79,2
Muito mais brancos	18,2	14,2	60,6	0,7	29,5	10,7
Mais brancos	33,0	48,6	35,1	2,8	46,0	21,3
Mais negros	39,8	31,6	3,4	51,5	23,7	54,8
Muito mais negros	8,9	5,6	0,9	44,9	0,8	13,2
Muito mais com ensino médio	1,0	0,5	32,0	0,6	1,2	0,4
Mais com ensino médio	3,3	33,8	65,0	1,3	82,0	3,9
Mais sem ensino médio	84,9	65,0	1,7	31,8	16,2	85,3
Muito mais sem ensino médio	10,9	0,6	1,3	66,3	0,6	10,4
Acima de 5 SM	2,3	1,6	67,9	0,8	6,1	0,6
De 3 a 5 SM	7,6	8,9	27,6	1,4	71,9	2,3
De 2 a 3 SM	12,0	77,1	2,4	0,9	19,0	2,8
Até 2 SM	78,1	12,4	2,1	96,9	3,0	94,3

Fonte: IBGE - Censo Demográfico de 2010. Tabulação e elaboração do autor.

Estes resultados convergem com ponderações colocadas por Cunha (2005, 2012) acerca da relevância numérica dos fluxos intra-metropolitanos. Adicionalmente, Cunha (2012) aponta que estes não são, em termos estritos, deslocamentos que alteram o espaço de vida da população que migra, sendo bastante parecidos, pelo menos neste aspecto, com os movimentos pendulares. Neste sentido, a migração da população de mais baixa renda é uma potencial busca por melhores condições de vida, sem alterar o ambiente no qual tal população se insere (as regiões metropolitanas), mas 'fugindo' da capital, possivelmente, por conta do elevado custo de vida e da violência.

Embora as trajetórias migratórias da população de mais alta renda sejam completamente distintas e de majoritariamente inter-estaduais, estas também podem ser vistas como manutenção de espaços de vida, ainda que com diferenças. De um lado, a manutenção se observa à medida que as origens e os destinos são, invariavelmente, municípios com perfil de cidades globais (Sassen, 2005). De outro, as diferenças se processam em termos dos motivos que mais provavelmente levaram ao deslocamento: para os mais pobres, a migração aparenta ser uma necessidade econômica; para os mais abastados, parte da trajetória pessoal-profissional, ou até mesmo uma opção.

4. Considerações Finais

O propósito deste artigo foi introduzir técnicas de aprendizado de máquina para analisar uma fonte de dados muito conhecida e utilizada nas Ciências Sociais Aplicadas: o Censo Demográfico. Neste trabalho, empregando os dados do Censo de 2010, os olhares se voltaram para a análise da dinâmica migratória intermunicipal brasileira, buscando superar o uso clássico dos

dados demográficos e incluir informações sobre os próprios migrantes e os municípios de origem e destino. Mais do que apenas utilizar as informações sobre a população migrante e as municipalidades para traçar um perfil estatístico, o artigo buscou ampliar os horizontes de análise, ao procurar, literalmente, minerar tais dados para obter novos conhecimentos.

Estudar a migração interna no Brasil demanda considerar múltiplas dimensões e trabalhar os dados a partir de perspectivas de aprendizagem que nos coloquem em uma posição na qual nada é óbvio. Quatro questões relevantes surgem a partir dos resultados, com potencial de direcionar os estudos sobre migração interna. Primeiro, há uma dificuldade intrínseca em identificar e classificar corretamente as instâncias positivas. No caso da migração intermunicipal, estas correspondem aos migrantes de renda/escolaridade mais elevada. Todavia, como foram levantados os dados mais exaustivos possíveis e observou-se uma predominância de fluxos de baixa renda, coloca-se o questionamento acerca de como fazer isto.

Segundo, o excesso de fluxos migratórios de pequeno porte (menos de 100 pessoas) dificulta o reconhecimento de padrões e a descoberta de conhecimento, por serem mais suscetíveis a variações unitárias na composição sociodemográfica. 83.4% das transações registradas no período 2005-2010 são de fluxos com menos de 50 pessoas, enquanto menos de 1.0% tem 500 ou mais casos registrados. Por mais intensa que possa parecer a troca migratória entre municípios, a grande maioria dos fluxos se restringe a poucas cidades: de um lado, fluxos em direção aos grandes centros; de outro lado, migrações entre municípios pequenos/locais. Nesta direção, ao estudar os 50 municípios com maiores fluxos, o pesquisador terá um quadro bastante próximo do real.

Terceiro, observou-se a existência de seletividades migratórias combinadas com espaços de migração predominantemente intra-regionais (e, no caso dos fluxos de pessoas mais pobres, até mesmo intra-metropolitanos). Neste sentido, se infere que os fluxos migratórios recentes assumiram um caráter mais rotativo e de distâncias menores, embora alguns grandes centros (como São Paulo, Rio de Janeiro e Brasília) continuem atraindo migrantes de todo o país. Por outro lado, pessoas de mais alta renda desenvolvem trajetórias e espaços migratórios *sui generis*, que guardam relação com a dinâmica global. Com isto, a inclusão de dados sobre rotatividade migratória e a análise dos fluxos por tempo de residência podem auxiliar na compreensão desta dinâmica.

Quarto, uma inovação do artigo foi o estudo da migração em uma perspectiva multiescalar, analisando indivíduos e municípios. Esta combinação contribuiu para o melhor entendimento do que é migração. Trabalhar com diversas escalas permite observar o fenômeno migratório não apenas sob a ótica do fluxo, mas também a partir de uma perspectiva de causa-efeito (em termos local, nacional e global) que tem potencial para superar explicações limitadas do tipo 'migra porque é pobre'. Esta abordagem, além do mais, motiva o levantamento de dados mais acurados sobre os municípios de origem e destino, assim como sobre o perfil e as opiniões dos migrantes.

Sobre as limitações da pesquisa, as análises foram feitas utilizando dados do momento do recenseamento (isto é, no destino migratório). Seria importante dispor de dados que permitam analisar tanto o perfil socioeconômico e demográfico do migrante quanto o perfil municipal no começo do processo migratório (no caso, 5 anos antes do Censo). Este tipo de dado permitiria analisar: (1) o perfil do migrante na origem; (2) os motivos e as causas que levaram alguém a migrar; e (3) as diferenças por nível de renda. Além disso, tais dados auxiliariam na compreensão de por que determinados migrantes voltarem atrás (e, com isto, não serem captados no Censo).

Com base nas ponderações apresentadas, propõem-se três recomendações relevantes para os estudos de migração sob a ótica do aprendizado de máquina. Primeiro, o uso de mais informações/atributos se faz fundamental para observar adequadamente os padrões migratórios nacionais, sendo estes relativos ao migrante e à sua condição antes e depois do fluxo, mas também aos municípios de origem e de destino. Segundo, trabalhar com microrregiões pode agregar mais conhecimento do que os dados municipais, uma vez que essas são formadas por municípios com similaridades em termos econômicos e sociais. Terceiro, dada a grande importância das migrações de pessoas de baixa renda e a peculiaridade dos fluxos dos mais abastados, sugere-se a desagregação das transações por nível socioeconômico, para evitar potenciais efeitos de confusão.

Para estudos futuros envolvendo a análise das migrações a partir do Aprendizado de Máquina, os quais estão sendo desenvolvidos como continuidade desta pesquisa, é importante incluir outras questões relevantes, a saber: (1) a análise comparativa do perfil de quem é 'nativo' de um município *versus* quem migrou para lá; (2) a avaliação das diferenças entre quem saiu de um município e quem não migrou; (3) a distribuição espacial da população em um município (ou região metropolitana), segundo condição e tempo de migração; e (4) a análise das condições de vida da população migrante em localidades semelhantes, diferenciando-os por tempo de migração.

Agradecimentos. O autor agradece as contribuições feitas pelo Dr. Stanley Robson de Medeiros Oliveira (EMBRAPA - Informática Agropecuária) e por dois pareceristas anônimos a este artigo.

Referências

Agrawal, R. e Skirant, R. Fast algorithms for mining association rules. In: *Proceedings of the XX International Conference on Very Large Data Bases*. San Francisco. Morgan Kaufmann, 1994. p. 487–499.

Azzoni, C. R. *Indústria e reversão da polarização no Brasil*. São Paulo: USP-IPE, 1986.

Baeninger, R. Deslocamentos populacionais, urbanização e regionalização. *Revista Brasileira de Estudos de População*, v. 15, n. 2, p. 67–81, 1998.

Baeninger, R. *Região, Metrópole e Interior: espaços ganhadores e espaços perdedores nas migrações recentes no Brasil - 1980/1996*. Tese (Doutorado em Ciências Sociais) - Instituto de Filosofia e Ciências Humanas, Universidade Estadual de Campinas, Campinas-SP, 1999.

Baeninger, R. Rotatividade migratória: um novo olhar para as migrações internas no Brasil. *REMHU - Revista Interdisciplinar de Mobilidade Humana*, v. 20, n. 39, p. 77–100, 2012.

Barchiesi, D., Preis, T., Bishop, S. e Moat, H. S. Modelling human mobility patterns using photographic data shared online. *Royal Society Open Science*, v. 2, n. 8, p. 1–8, 2015.

Billari, F. e Zagheni, E. Big data and population processes: a revolution? In: Petrucci, A. e Verde, R. (ed.), *Statistics and Data Science: new challenges, new generations*, p. 167-178. Florença: Firenze University Press, 2017.

Brandão, C. A. *Território e Desenvolvimento: as múltiplas escalas entre o local e o global*. Campinas: Editora da UNICAMP, 2007.

Breiman, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.

Campos, M. B. Seletividade e migração. In: Bruno, M. (org.), *População, espaço e sustentabilidade: contribuições para o desenvolvimento do Brasil*, p. 187-202. Rio de Janeiro: IBGE, 2015.

Cunha, J. M. P. Migração e urbanização no Brasil: alguns desafios metodológicos para análise. *São Paulo em Perspectiva*, v. 19, n. 4, p. 3–20, 2005.

Cunha, J. M. P. Retratos da mobilidade espacial no Brasil: os censos demográficos como fonte de dados. *REMHU - Revista Interdisciplinar de Mobilidade Humana*, v. 20, n. 39, p. 29–50, 2012.

Dempster, A. P., Laird, N. M. e Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 39, n. 1, p. 1–38, 1977.

Faria, V. E. Cinqüenta anos de urbanização no Brasil: tendências e perspectivas. *Novos Estudos CEBRAP*, n. 29, p. 98–119, 1991.

Foster, I., Ghani, R., Jarim, R. S., Kreuter, F. e Lane, J. *Big Data and Social Science: a Practical Guide to Methods and Tools*. New York: Chapman and Hall/CRC, 2016.

Franco-Arcega, A., Franco-Sánchez, K. D., Castro-Espinoza, F. A. e García-Islas, L. H. Data mining for discovering patterns in migration. In: *Nature-Inspired Computation and Machine Learning - Proceedings of the 13th Mexican International Conference on Artificial Intelligence, Part II*. Basel. Springer, 2014. p. 285–295.

Frank, E., Hall, M. A. e Witten, I. H. *The WEKA Workbench*. Burlington: Morgan Kaufmann, 2016.

González-Bailón, S. Social science in the era of big data. *Policy and Internet*, v. 5, n. 2, p. 147–160, 2013.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., e Tatham, R. L. *Análise multivariada de dados*. Porto Alegre: Bookmann, 2009.

Han, J., Kamber, M. e Pei, J. *Data mining: concepts and techniques*. Burlington: Morgan Kaufmann, 2012.

Hastie, T., Tibshirani, R. e Friedman, J. H. *The elements of statistical learning: data mining, inference and prediction*. New York: Springer, 2008.

James, G., Witten, D., Hastie, T. e Tibshirani, R. *An introduction to statistical learning with applications in R*. New York: Springer, 2015.

Jolliffe, I. T. Discarding variables in a principal component analysis I: Artificial data. *Journal of the Royal Statistical Society: Series C (Applied Statisticis)*, v. 21, n. 2, p. 160–173, 1972.

Jolliffe, I. T. Discarding variables in a principal component analysis II: Real data. *Journal of the Royal Statistical Society: Series C (Applied Statistics),*, v. 22, n. 1, p. 21–31, 1973.

Jolliffe, I. T. e Cadima, J. Principal component analysis: a review and recent developments. *Philosophical Transactions: Series A (Mathematical, Physical and Engineering Sciences)*, v. 374, n. 2065, 2016.

Lammers, L. *Forecasting European population movement and country relations using tone and migration data*. Dissertação (Master of Science in Data Science) - School of Humanities and Digital Sciences, Tilburg Univesity, Tilburg-Netherlands, 2017.

Lee, E. S. Uma teoria sobre a migração. In: Moura, H. (e.), *Migração interna: textos selecionados*, p. 95-114. Fortaleza: Banco do Nordeste do Brasil S.A., 1980.

Lencioni, S. A metamorfose de São Paulo: o anúncio de um novo mundo de aglomerações difusas. *Revista Paranaense de Desenvolvimento*, v. 32, n. 120, p. 133–148, 2011.

Letouzé, E. *Demography, meet big data; big data, meet demography: Reflections on the data-rich future of population science*. 2015. Disponível em: https://goo.gl/xeWDDt. Acesso em: 15/12/2017.

Lindström, A. *Predicting internal migration on individual level in Sweden using micro data: a performance comparison of logistic regression and neural networks*. Dissertação (Master of Science in Economics) - Faculty of Social Sciences, Umea Univesity, Umea-Sweden, 2017.

Magalhães, J. C. R. e Miranda, R. B. Dinâmica da renda per capita, longevidade e educação nos municípios brasileiros. *Estudos Econômicos (São Paulo)*, v. 39, n. 3, p. 539–569, 2009.

Maria, P. F. *Rotina para construção das matrizes migratórias municipais pelo quesito data-fixa e das tabelas de fluxos direcionados, não direcionados e mistos - censos de 1991, 2000 e 2010*. 2017. Disponível em: https://goo.gl/QVRuCb. Acesso em: 15/12/2017.

Martine, G. *A redistribuição espacial da população brasileira durante a década de 80*. Texto para Discussão nº 329, IPEA - Instituto de Pesquisa Econômica Aplicada, 1994.

Martine, G. e McGranahan, G. A transição urbana brasileira: trajetória, dificuldades e lições aprendidas. In: Baeninger, R. (org.), *População e Cidades: subsídios para o planejamento e para as políticas sociais*, p. 11-24. Campinas: NEPO/UNICAMP e UNFPA, 2010.

Matos, R. Questões teóricas acerca dos processos de concentração e desconcentração da população no espaço. *Revista Brasileira de Estudos de População*, v. 12, n. 1/2, p. 35–58, 1995.

Matos, R. e Baeninger, R. Migração e urbanização no Brasil: processos de concentração e desconcentração espacial e o debate recente. *Cadernos do Leste*, Edição Especial 2000 a 2008 , p. 342–386, 2008.

McCaa, R. e Ruggles, S. The census in global perspective and the coming microdata revolution. *Scandinavian Population Studies*, v. 13, p. 7–30, 2002.

Pacheco, C. A. Desconcentração econômica e fragmentação da economia nacional. *Economia e Sociedade*, v. 5, n. 1, p. 113–140, 1996.

Pande, N. e Rajan, K. S. Spatio-temporal analysis for finding migration patterns in Andhra Pradesh using RWeka. In: *Proceedings of the FOSS4G India 2015*. Dehradun. OSGeo-India, 2015. p. 1–5.

Ravenstein, E. G. As leis das migrações. In: Moura, H. (e.), *Migração interna: textos selecionados*, p. 25-88. Fortaleza: Banco do Nordeste do Brasil S.A., 1980.

Robinson, C. e Dilkina, B. A machine learning approach to modeling human migration. In: *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. Menlo Park and San Jose, CA, USA. ACM, 2018. p. 1–8.

Ruggles, S. Big microdata for population research. *Demography*, v. 51, n. 1, p. 287–297, 2014.

Sassen, S. The global city: introducing a concept. *Brown Journal of World Affairs*, v. 11, n. 2, p. 27–43, 2005.

Silva, M. M. F. Tentativa de classificação das cidades brasileiras. *Revista Brasileira de Geografia*, v. 8, n. 3, p. 283–316, 1946.

Simini, F., González, M. C., Maritan, A. e Barabási, A. L. A universal model for mobility and migration patterns. *Nature*, v. 484, p. 96–100, 2012.

Singer, P. Migrações internas: considerações teóricas sobre o seu estudo. In: Singer, P. (e.), *Economia Política da Urbanização*, p. 29-60. São Paulo: Editora Brasiliense / CEBRAP, 1973.

Todaro, M. A migração da mão de obra e o desemprego urbano em países subdesenvolvidos. In: Moura, H. (e.), *Migração interna: textos selecionados*, p. 145-172. Fortaleza: Banco do Nordeste do Brasil S.A., 1980.

Vainer, C. B. As escalas do poder e o poder das escalas: o que pode o poder local? *Cadernos IPPUR*, v. 16, n. 1, p. 13–32, 2002.

Vignoli, J. R. *Migración interna y desarrollo: el caso de américa latina*. Avance de investigación - Proyecto BID/CEPAL SF-9157-RG, CELADE - Centro Latinoamericano y Caribeño de Demografía, 2007.

Witten, I. H. e Frank, E. *Data mining: practical machine learning tools and techniques*. Amsterdam: Elsevier, 2005.

Apêndices

Apêndice 1: Lista de variáveis utilizadas para execução do PCA.

Variável	**Descrição**	**Fonte**
V001	% de pessoas naturais do município	A
V002	% de pessoas naturais da unidade da federação	A
V003	Esperança de vida ao nascer	B
V004	Taxa de mortalidade infantil	B
V005	Taxa de fecundidade total	B
V006	Razão de dependência	C
V007	Taxa de envelhecimento	B
V008	Grau de urbanização	B
V009	População total	B
V010	Razão de sexo	C
V011	Densidade demográfica	D
V012	Área total do município	E
V013	% de pessoas de 10 anos ou mais sem ensino fundamental completo	A
V014	% de pessoas de 10 anos ou mais com ensino superior completo	A
V015	% de pessoas de 5 a 6 anos na escola	B
V016	Taxa de analfabetismo - 18 anos ou mais	B
V017	% de 25 anos ou mais com superior completo	C
V018	Expectativa de anos de estudo	C
V019	% de 6 a 17 anos na escola	C
V020	% de 6 a 17 anos no básico com 2 anos ou mais de atraso	C
V021	% de pessoas de 10 anos ou mais casadas	A
V022	% de pessoas de 10 anos ou mais solteiras	A
V023	% de famílias do tipo "Casal com filhos"	A
V024	% de famílias do tipo "Casal sem filhos"	A
V025	% de famílias do tipo "Mãe com filhos"	A
V026	% de famílias do tipo "Pai com filhos"	A
V027	% de pessoas de 10+ anos com renda nominal mensal de até 1 SM	A
V028	% de pessoas de 10+ anos com renda nominal mensal de mais de 5 SM	A
V029	Índice de Gini	C
V030	Índice de Vulnerabilidade Social (IVS)	B
V031	Índice de Theil - L	C
V032	Índice de Desenvolvimento Humano Municipal (IDHM)	C
V033	% de extremamente pobres	C
V034	% de pobres	C
V035	Renda per capita dos extremamente pobres (R$)	C

Continuação...

Variável	Descrição	Fonte
V036	Renda per capita dos pobres (R$)	C
V037	Renda per capita	C
V038	% da renda proveniente de rendimentos do trabalho	C
V039	Taxa de atividade - 10 anos ou mais	C
V040	Taxa de desocupação - 10 anos ou mais	C
V041	% de empregados com carteira - 18 anos ou mais	C
V042	% de empregados sem carteira - 18 anos ou mais	C
V043	Grau de formalização dos ocupados - 18 anos ou mais	C
V044	Renda per capita (R$), exceto renda nula	C
V045	% dos ocupados com superior completo - 18 anos ou mais	C
V046	Participação do setor Agrícola no VAB - 2010 (%)	F
V047	Participação do setor Industrial no VAB - 2010 (%)	F
V048	Participação do setor de Serviços no VAB - 2010 (%)	F
V049	Variação da participação do setor Agrícola no VAB - 2005/2010 (p.p.)	F
V050	Variação da participação do setor Industrial no VAB - 2005/2010 (p.p.)	F
V051	Variação da participação do setor de Serviços no VAB - 2005/2010 (p.p.)	F
V052	PIB per capita de 2010 a preços de 2010 (R$) - deflator implícito do PIB	F
V053	PIB per capita de 2005 a preços de 2010 (R$) - deflator implícito do PIB	F
V054	Crescimento do PIB per capita a preços de 2010 (2005-2010, % a.a.)	F
V055	Participação (%) do PIB municipal no PIB do estado - 2010	F

Fonte: Organização do autor. As seguintes fontes foram utilizadas para extração das variáveis:
A: IBGE - Censo Demográfico de 2010 - Sistema IBGE de Recuperação de Dados
B: IPEA - Atlas da Vulnerabilidade Social
C: PNUD-IPEA - Atlas do Desenvolvimento Humano no Brasil
D: IBGE - Censo Demográfico de 2010 e Área Territorial Brasileira
E: IBGE - Área Territorial Brasileira
F: IBGE - Produto Interno Bruto dos Municípios

Apêndice 2: Código para extrair dados socioeconômicos e demográficos dos migrantes.

```
data censo2010a; /* Criando um banco de dados novo */
  set censo2010; /* Banco de dados original */
  keep /* Mantendo variaveis de interesse */
    muniat V6264 municod V0010 /* Origem, destino e peso */
    V0601 SexoM SexoF /* Sexo */
    V6036 /* Idade */
    V0606 CorB CorP /* Cor */
    V6400 SemE FunM MedS SupP /* Escolaridade */
    RDPC; /* Renda per capita */

  muniat = v0001*100000+v0002; /* Criando o atributo "municipio" */
  if v6264 = . then delete; /* Removendo os que nunca migraram */
  if v6264 = muniat then delete; /* Removendo origem = destino */
  municod = substrn(v6264,3,5)*1; /* Retendo codigo municipal sem UF */
  if municod=99999 then v6264=7777777 /* Origem ignorada */;
  if v6264=7777777 then muniat=5555555; /* Destino irrelevante */

  if v0601=1 then SexoM=1; else SexoM=0; /* Homens */
  if SexoM=0 then SexoF=1; else SexoF=0; /* Mulheres */

```

```
21    if v0606 in(1,3) then CorB=1; else CorB=0; /* Brancos */
22    if v0606 in(2,4,5,9) then CorP=1; else CorP=0; /* Negros */
23
24    if v6400 in(1,5) then SemE=1; else SemE=0; /* Analfabeto funcional */
25    if v6400=2 then FunM=1; else FunM=0; /* Com ensino fundamental */
26    if v6400=3 then MedS=1; else MedS=0; /* Com ensino medio */
27    if v6400=4 then SupP=1; else SupP=0; /* Com ensino superior */
28
29    RDPC = V6531;
30  run;
31
32  proc sql;
33    create table Fluxo2010 as select
34      muniat as mun2010, /* Municipio de destino */
35      V6264 as mun2005, /* Municipio de origem */
36      sum(v0010) as fluxo, /* Fluxo total */
37      sum(SexoM*v0010) as Homens, /* Numero de homens */
38      sum(SexoF*v0010) as Mulheres, /* Numero de mulheres */
39      sum(CorB*v0010) as Brancos, /* Numero de brancos */
40      sum(CorP*v0010) as Negros, /* Numero de negros */
41      sum(SemE*v0010) as SemEF,
42      sum(FunM*v0010) as SemEM,
43      sum(MedS*v0010) as SemES,
44      sum(SupP*v0010) as ComES,
45      sum(v6036*v0010)/sum(v0010) as idadeM, /* Idade media */
46      sum(RDPC*v0010)/sum(v0010) as RDPC /* Renda per capita media */
47    from censo2010a
48    group by mun2010, mun2005; /* Agrupando origens por destino */
49  quit;
50
51  proc export data=Fluxo2010 /* Exportando o banco de dados */
52    outfile="C:\Users\pf\Desktop\Banco.csv"
53    dbms=csv replace; /* Trocar pasta */
54  run;
```

Fonte: Elaboração e implementação do autor em SAS®, baseado em Maria (2017).