r b e
REVISTA BRASILEIRA DE ECONOMIA
DOI 10.5935/0034-7140.20230014
ISSN 1806-9134 (online)
FGV
FUNDAÇÃO GETULIO VARGAS

Migração seletiva de retorno e distribuição salarial: Evidências para população migrante em São Paulo

Juliane do Carmo Duarte Magalhães*
Hilton Martins de Brito Ramalho†
Aléssio Tony Cavalcanti de Almeida‡

Sumário

Palavras-chave
migração de retorno, autosseleção, desigualdade salarial

JEL Codes
R23, J24, C14

Resumo · Abstract

O objetivo deste trabalho é investigar o impacto da migração seletiva de retorno sobre a distribuição salarial de migrantes residentes em São Paulo. Para tanto, foram utilizados os dados do Censo Demográfico de 2010 e um método de estimação que leva em conta a seleção em não observáveis. Os achados deste estudo apontam para seleção negativa em observáveis e não observáveis no fluxo dirigido ao estado de São Paulo. Considerando a seletividade, a população migrante estaria recebendo mais se todos permanecessem, e a desigualdade dentro do grupo diminuiria substancialmente entre os percentis extremos.

1. Introdução

Segundo dados dos censos demográficos no Brasil, houve uma expansão da migração interestadual de retorno nas últimas décadas. Em termos absolutos, a população de remigrados subiu de 1,14 milhões entre 1995 e 2000 para 1,23

*Universidade Federal da Paraíba, Centro de Ciências Sociais Aplicadas (UFPB/CCSA), Departamento de Economia. Cidade Universitária, Campus I, João Pessoa, PB, CEP 58051-900, Brasil. 0000-0003-0005-3092

†Universidade Federal da Paraíba, Centro de Ciências Sociais Aplicadas (UFPB/CCSA), Departamento de Economia. 0000-0002-1218-2687

‡Universidade Federal da Paraíba, Centro de Ciências Sociais Aplicadas (UFPB/CCSA), Departamento de Economia. 0000-0003-0436-359X

juliane.ada@gmail.com hiltonmbr@gmail.com alessio@ccsa.ufpb.br

Recebido em Fevereiro de 2020
Aceitação final em Junho de 2021

milhões no quinquênio de 2005/2010, representando 22,02% e 24,52% do total de migrantes interestaduais do país, respectivamente (IBGE, 2012).

Esse regresso, no entanto, não ocorre de forma aleatória dentro do processo migratório. Segundo Biavaschi (2016), a escolha de residência ótima de um migrante racional está condicionada à decisão inicial de migração. Borjas e Bratsberg (1996), por sua vez, argumentam que a dinâmica de retorno reforça a seleção que caracterizou o fluxo de migração inicialmente. Isso significa que se o fluxo inicial de migrantes for positivamente selecionado em termos de suas habilidades inatas, então os retornados serão aqueles com habilidades inferiores dentro do grupo inicial de partida A volta de indivíduos negativamente selecionados, no entanto, levaria a uma redução no bem-estar da população, isso porque esses indivíduos tenderiam a ser desmotivados e menos favorecidos no mercado de trabalho (Siqueira, 2006).

A hipótese de autosseleção é, portanto, uma questão central nos movimentos populacionais. Ela sugere que os migrantes não são pessoas aleatórias na população de origem isto é, esses indivíduos dispõe de características não observáveis como agressividade, empreendedorismo, ambição, propensão ao risco, entre outras (Chiswick, 1999). Além das características pessoais, os custos de migração e os diferencias de salários líquidos entre as regiões podem influenciar a seleção dos migrantes (Cattaneo, 2007; McKenzie & Rapoport, 2010; Sjaastad, 1962).

Existem evidências de que os migrantes interestaduais no Brasil são positivamente selecionados, isto é, possuem características não observáveis favoráveis ao seu sucesso no mercado de trabalho tanto em relação aos seus conterrâneos quanto em relação aos não migrantes dos estados que os recebem (Freguglia & Procópio, 2013; Gama & Machado, 2014; Santos Jr, Menezes-Filho, & Ferreira, 2005). Ramalho e Queiroz (2011) identificam seleção positiva dos migrantes não retornados em relação aos não migrantes em fatores não observados. O grupo de retornados, por sua vez, também registra vantagens frente aos não migrantes em termos de características inatas. A conclusão dos autores foi que a migração gera benefícios mesmo quando o individuo opta por retornar devido aos ganhos de capital humano. Gama e Machado (2014), por sua vez, corroboram esse resultado, os autores identificam seleção positiva em habilidades não observáveis para os remigrados e migrantes permanentes em relação aos não migrantes, mas não comparam relações entre permanentes e retornados.

Em se tratando do progresso do migrante no destino, Silveira Neto e Magalhães (2004), ao estudar a migração para o estado de São Paulo, verificam que há uma lacuna significativa entre nativos e migrantes perdurando ao longo do tempo, favorecendo o grupo de paulistas. Portanto, para os migrantes, a valoração de seus atributos é feita de forma diferenciada pelo mercado, em

termos de características observáveis como educação e estado de origem. Outros estudos apontam para a mesma característica de migração. Segundo Batista e Cacciamali (2009) esse resultado é presente entre entre migrantes e não migrantes na própria região Sudeste. Os autores Assis, Costa, e Mariano (2012), ao analisar a migração entre os estados de São Paulo e Bahia identificam, mais uma vez, um hiato salarial entre baianos e paulistas, favorável ao último grupo.

O progresso econômico do migrante no destino, no entanto, pode sofrer efeitos da saída de trabalhadores. A depender da natureza de seleção desses indivíduos a desigualdade salarial entre os grupos pode ser mitigada ou intensificada. Com isso, a assimilação do migrante no destino pode ser subestimada ou superestimada se não levarmos em conta a migração seletiva de retorno (Ambrosini, Mayr, Peri, & Radu, 2010; Cohen & Haberfeld, 2001). Isso posto, surgem duas questões relevantes a serem investigadas: como se comportaria a distribuição de salários dos migrantes interestaduais na ausência da migração de retorno? A migração de retorno é um mecanismo que facilita ou dificulta a assimilação dos migrantes na região de destino?

Portanto, o objetivo geral deste trabalho é analisar o impacto da migração interestadual de retorno na distribuição de salários dos migrantes que permaneceram na região de destino. De forma específica, procura-se: (i) analisar o tipo de seletividade em características inatas dos migrantes; (ii) investigar o efeito da migração de retorno considerando grupos por níveis educacionais; (iii) verificar diferenciais salariais dentro do grupo de migrantes permanentes e entre migrantes e naturais da região de destino mediante a ausência de migração de retorno.

Para alcance dos objetivos citados serão feitas estimativas de distribuição de salários mediante dois cenários: com e sem migração de retorno, considerando o estado de São Paulo como região de destino dos fluxos de migração. O estado de São Paulo possui o maior Produto Interno Bruto (PIB) frente as demais unidades federativas do Brasil, tem um mercado de trabalho dinâmico e competitivo, além de se caracterizar como destino principal dos migrantes brasileiros (A. T. R. d. Oliveira, Ervatti, & O'Neill, 2011; Ramalho, Figueiredo, & Netto, 2016; Santos, 2006).

Nesse contexto, procura-se recuperar a distribuição de salários na ausência de migração de retorno, conforme método proposto por Biavaschi (2016). Esse método amplia aquele apresentado por DiNardo, Fortin, e Lemieux (1995) e aplicado em Chiquiar e Hanson (2005) apenas para características observáveis, ao considerar a seleção dos migrantes em características não observadas. Também parte do pressuposto de que o viés de seleção amostral desaparece para migrantes com alta probabilidade de permanência na região de destino. Esse procedimento é conhecido na literatura como identificação no infinito (Chamberlain, 1986) e foi defendido por Heckman (1990) na estimação

do termo constante em modelos de seleção de amostra semiparamétrica. A seleção amostral é tratada através de um modelo de regressão de cópula. Esse método assegura flexibilidade na distribuição dos dados e eficiência do modelo paramétrico, além de lidar com relações de resposta não linear das covariadas do modelos.

Destarte, esse trabalho avança em relação à literatura nacional ao considerar o efeito da migração seletiva de retorno sobre a distribuição salarial dos migrantes remanescentes no destino. Bem como pela utilização de um método ainda não explorado que considera características não observadas, identificação no infinito e a utilização de função cópula na modelagem de dependência entre o processo salarial e de migração de retorno.

Entre os principais achados temos que a população migrante em São Paulo estaria ganhando mais caso não houvesse migração de retorno. Além disso, a desigualdade entre os próprios migrantes diminuiria substancialmente no cenários em que todos os trabalhadores não naturais decidem permanecer em São Paulo.

Afora essa introdução, esta pesquisa está organizada da seguinte forma: (ii) resumo de fatos observados na migração interestadual brasileira; (iii) detalhamento do modelo empírico e uma descrição dos dados e variáveis utilizadas no estudo; (iv) resultados empíricos e discussões; (v) considerações finais.

2. Migração interestadual no Brasil: Fatos observados

Nesta seção são apresentados dados mais recentes sobre a migração interestadual no Brasil, considerando o cruzamento das questões de estado de residência na data de entrevista, estado de residência anterior e estado de naturalidade. Conforme dados do Censo Demográfico Brasileiro de 2010, entende-se o migrante permanente o indivíduo que declarou residência em uma unidade federativa na data da da entrevista, tendo como residência anterior o estado de nascimento. O migrante retornado, por sua vez é o indivíduo declarou residir no seu estado de nascimento, tendo anteriormente residido em outro estado. Tais conceitos permitem estimar estoques de migrantes para o período de 2000 a 2010.

A Tabela 1 contém dados do quantitativo de migrantes interestaduais segundo a unidade federativa de residência em 2010: migrantes não retornados na coluna (1); migrantes remigrados de São Paulo na coluna (3); e unidade federativa de residência anterior — migrantes retornados — na coluna (2). Observa-se que o estado de São Paulo concentra não somente a maior parcela de migrantes permanentes (23%), mas também é o estado de origem da maior parcela de migrantes retornados, ou seja, 24,28% dos remigrados para seu estado

Tabela 1. Distribuição dos migrantes interestaduais segundo estado de residência na data censitária e estado de residência anterior – 2010

Unidade Federativa	(1) Migrantes Permanentes (Residência em 2010)	(2) Migrantes Retornados (Residência anterior)	(3) Retornados de SP (Residência em 2010)	(4) Migrantes/Habitantes da UF(%)
Rondônia (RO)	117.844	12.154	418	7,54
	(1,61)	(1,43)	(0,20)	
Acre (AC)	23.566	1.976	139	3,21
	(0,32)	(0,23)	(0,07)	
Amazonas (AM)	127.458	10.090	162	3,66
	(1,74)	(1,18)	(0,08)	
Roraima (RR)	53.511	2.749	15	11,88
	(0,73)	(0,32)	(0,01)	
Pará (PA)	322.553	34.775	1.418	4,25
	(4,41)	(4,08)	(0,69)	
Amapá (AP)	74.006	4.834	68	11,05
	(1,01)	(0,57)	(0,03)	
Tocantins (TO)	142.891	14.840	502	10,33
	(1,95)	(1,74)	(0,24)	
Maranhão (MA)	142.511	19.911	3.684	2,17
	(1,95)	(2,34)	(1,78)	
Piauí (PI)	92.141	11.786	5.864	2,95
	(1,26)	(1,38)	(2,84)	
Ceará (CE)	137.176	17.325	12	1,62
	(1,87)	(2,03)	(5,94)	
Rio Grande do Norte (RN)	90.603	8.827	4.959	2,86
	(1,24)	(1,04)	(2,40)	
Paraíba (PB)	113.011	14.808	8.190	3,00
	(1,54)	(1,74)	(3,96)	
Pernambuco (PE)	186.934	27.123	17.210	2,13
	(2,55)	(3,18)	(8,32)	
Alagoas (AL)	73.259	12.245	7.088	2,35
	(1,00)	(1,44)	(3,43)	
Sergipe (SE)	75.160	7.914	2.848	3,63
	(1,03)	(0,93)	(1,38)	
Bahia (BA)	304.219	49.448	36.197	2,17
	(4,16)	(5,81)	(17,50)	
Minas Gerais (MG)	492.874	61.498	42.122	2,52
	(6,74)	(7,22)	(20,37)	
Espírito Santo (ES)	209.894	21.207	1.970	5,97
	(2,87)	(2,49)	(0,95)	
Rio de Janeiro (RJ)	471.007	58.553	6.649	2,95
	(6,44)	(6,88)	(3,22)	
São Paulo (SP)	1.684.486	206.792	–	4,08
	(23,02)	(24,28)	–	
Paraná (PR)	368.095	55.499	35.417	3,52
	(5,03)	(6,52)	(17,13)	
Santa Catarina (SC)	454.487	44.925	5.259	7,27
	(6,21)	(5,28)	(2,54)	
Rio Grande do Sul (RS)	114.216	22.770	5.468	1,07
	(1,56)	(2,67)	(2,64)	
Mato Grosso do Sul (MS)	159.592	22.189	4.028	6,52
	(2,18)	(2,61)	(1,95)	
Mato Grosso (MT)	304.940	34.758	4.028	10,05
	(4,17)	(4,08)	(0,75)	
Goiás (GO)	612.654	31.174	3.288	10,20
	(8,37)	(3,66)	(1,59)	
Distrito Federal (DF)	368.097	41.420	–	14,32
	(5,03)	(4,86)	–	
População estimada	7.317.185	851.592	206.791	
	(100)	(100)	(100)	

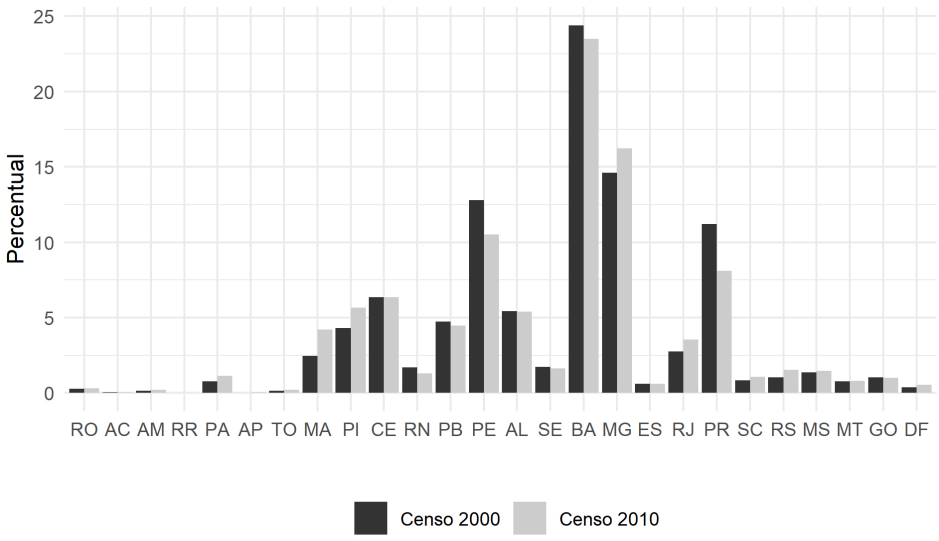
Nota: Valores expandidos para população. Percentual entre parênteses.

Fonte: Elaboração própria com base nos dados do Censo Demográfico de 2010.

de nascimento residiam anteriormente no estado. Destarte, mesmo atraindo muitos migrantes, São Paulo também é um grande emissor para outros estados.

Dada a grande relevância do estado de São Paulo na atração de fluxos migratórios, em particular, como residência anterior de quase 1/4 dos remigrados no período analisado, a coluna (3) da tabela em destaque apresenta a distribuição dos migrantes retornados de São Paulo segundo seu estado de nasciment. Destaca-se que do total de remigrados provenientes de São Paulo, cerca de 47% tiveram como destino os estados da região Nordeste e o estado de Minas Gerais, 20,37%.

Os dados apresentados na Figura 1 mostram entre que entre os Censos Demográficos de 2000 e 2010 não houve mudanças significativas nos padrões de representatividade dos estados de origem dos migrantes residentes em São Paulo (SP). Os cinco estados com maior população migrante em SP são os mesmos nos dois censos. Note-se que o estado da Bahia é o maior fornecedor nos dois períodos. O fluxo São Paulo–Bahia tem sido observado como recorrente nos estudos sobre mobilidade populacional. Assis et al. (2012), por exemplo, ressaltam que 60% dos migrantes retornados baianos residiam anteriormente em São Paulo, segundo dados da PNAD de 2009. Minas Gerais é o segundo estado mais representativo entre os migrantes, no censo de 2010 há um ligeiro aumento de envio. No entanto, no período de 2005 a 2010 apresentou equilíbrio na entrada



Nota: Valores expandidos para população.*Fonte:* Elaboração própria a partir dos dados do Censo Demográfico de 2000 e 2010.

Figura 1. Distribuição de migrantes em São Paulo por estado de origem – Censos de 2000 e 2010

e saída de migrantes, passando para categoria de estado com rotatividade migratória conforme (Ramalho et al., 2016).

São Paulo tem se sobressaído como absorvedor de migrantes há décadas e as tendências de origem dos migrantes também não sofreram significativas mudanças ao longo do tempo. O que vem se destacando é a grande perda de população sofrida pelo estado, fruto especialmente da migração de retorno de nordestinos e de mineiros. O fluxo de migrantes oriundos de São Paulo para a região Nordeste constitui-se basicamente de migrantes retornando aos seus estados, o que vinha contribuindo para saldos migratórios positivos. Os dados apresentados acima sugerem que esse movimento continua na mesma direção.

Em se tratando de capital humano, a Tabela 2 relata os percentuais de migrantes interestaduais residentes no estado de São Paulo em 2010 e de acordo com a região de nascimento (residência anterior) e o nível de escolaridade. Observa-se uma maior concentração de migrantes com baixa instrução de origem na região Nordeste, 69,50% dos migrantes sem instrução são oriundos dessa região. Esse padrão se repete para os níveis fundamentais e médio com 67,14% e 56,27%, respectivamente. Nos níveis de instrução mais altos encontram-se em destaque os sulistas com 43% dos migrantes com ensino superior e aqueles oriundos da região Centro-Oeste do país (23,63%). Dos migrantes que possuem pós-graduação, 43,97% são oriundos da região Sul. No geral, a população migrante em São Paulo é de baixa instrução, com cerca 53% de

Tabela 2. Distribuição dos migrantes permanentes em São Paulo segundo faixa de instrução e por região de residência anterior (nascimento)

	Região de origem do migrante					
Nível de instrução	Norte	Nordeste	Sudeste	Sul	Centro Oeste	Total
Sem-instrução	29.345	616.956	13.706	151.099	76.555	893.472
%	3,31	69,50	1,54	17,02	8,62	53,04
Nível Fundamental	10.058	202.769	6.309	53.271	29.594	303.580
%	3,33	67,14	2,09	17,64	9,80	18,02
Nível Médio	15.976	204.166	9.159	85.817	45.792	363.873
%	4,43	56,57	2,54	23,78	12,69	21,06
Nível superior	7.530	21.012	3.755	42.656	23.753	101.739
%	7,63	21,29	3,80	43,22	24,06	6,04
Pós-Graduação	1.017	2.131	700	5.221	2.806	12.298
%	8,56	17,95	5,89	43,97	23,63	0,73
Indefinido	324	6.099	232	2.028	812	9.524
%	3,41	6,23	2,44	21,36	8,55	0,56

Nota: Valores expandidos para população.
Fonte: Elaboração própria com base nos dados do Censo Demográfico de 2010.

pessoas no nível mais baixo, apenas 6,04% dessa população declarou possuir nível superior e 0,73% disse possuir pós-graduação.

As informações trazidas nesse capítulo são úteis para identificar as tendências e padrão de migração interestadual no Brasil. Chama atenção o estado de São Paulo enquanto grande polo de absorção e de residência anterior de remigrados. No entanto, a influência da migração de retorno assumiu papel relevante nos movimentos populacionais do Brasil. Uma questão relevante é investigar os efeitos da migração de retorno sobre a desigualdade salarial do grande quantitativo de migrantes residentes em São Paulo. As seções a seguir procuram responder como esse problema pode ser investigado.

3. Metodologia

3.1 Modelo Empírico

Seguindo Biavaschi (2016), a decisão de ficar na região de destino ou retornar para a região de nascimento vai depender do valor do benefício líquido da permanência. Caso ele seja maior do que zero, o trabalhador deve permanecer na região de destino, caso contrário deve remigrar. Dessa forma, façamos $S_i = 1$ um indicador que o trabalhador migrante permaneceu em São Paulo e $S_i = 0$ caso tenha retornado para o estado de nascimento. A decisão em destaque será determinada por

$$S = \begin{cases} 1, & \text{se } Z_i'\alpha > \epsilon_i \\ 0, & \text{se } Z_i'\alpha \leq \epsilon_i \end{cases} \qquad \text{para } i = 1, \dots, r+n, \tag{1}$$

onde r é o número de retornados; n de migrantes permanentes; Z_i uma matriz de variáveis que influenciam na decisão de ficar no destino; α um vetor de parâmetros (inclusive intercepto) e ϵ_i um termo randômico. Portanto, $\left(Z_i'\alpha - \epsilon_i\right)$ mensura o benefício líquido da decisão de permanência.

A equação de determinação de salários do tipo *minceriana* para um migrante selecionado aleatoriamente na população de migrantes presentes em São Paulo é dada por

$$Y_i^* = X_i'\beta + u_i^* \qquad i = 1, \dots, r+n, \tag{2}$$

onde Y_i^* é o logaritmo do salário-hora dos migrantes; X_i é o conjunto de variáveis observáveis que determinam o processo de pagamentos de salários; β um vetor de parâmetros (inclusive intercepto) e u_i^* é o termo aleatório que mensura atributos produtivos não observáveis na determinação desses rendimentos.

Alguns pressupostos devem ser admitidos em relação às equações (1) e (2). Primeiro, $\left(Y_i^*, S_i, X_i, Z_i\right)$ são variáveis aleatórias observadas. Segundo, devemos

assumir que $(X_i, Z_i, u_i^*, \epsilon_i)$ são variáveis aleatórias independentemente e identicamente distribuídas (i.i.d) e (X_i, Z_i) são variáveis aleatórias exógenas. Terceiro, os termos randômicos u_i^* e ϵ_i podem ser correlacionados, pois características não observáveis podem influenciar tanto a decisão de permanência do migrante quanto o seu salário.

Cabe ressaltar que o salário Y_i é observado apenas para os migrantes permanentes. Logo,

$$Y_i = S_i Y_i^* \qquad i = 1, \dots, r+n. \tag{3}$$

Desejamos obter a distribuição de salários Y_i^* para todos os migrantes, incluindo salários dos remigrados caso tivessem permanecido na região de destino, isto é, $f(Y_i^*)$. No entanto, tal distribuição é dada por um deslocamento da função de densidade $f(u_i^*)$ pelo valor médio de salários preditos na região de destino $X_i'\beta$ (Biavaschi, 2016). Portanto, precisamos de métodos para estimativas não tendenciosas para a função de densidade de probabilidade $f(u_i^*)$ e dos vetores de parâmetros β.

Usamos duas estratégias de identificação conforme discutido a seguir. No primeiro caso, usamos o método de identificação no infinito para a distribuição contrafatual (Chamberlain, 1986; Heckman, 1990) $f(u_i^*)$, uma vez que não observamos u_i^* para os trabalhadores remigrados. Na segunda etapa, estimamos o modelo (1) e (2) relaxando a hipótese de normalidade com especificação semiparamétrica e distribuição conjunta modelada por funções cópulas. Também introduzimos uma restrição de exclusão, isto é, uma variável instrumental que supostamente não determina diretamente a formação de salários, mas se relaciona com a decisão de permanência na região de destino — local de nascimento do filho do migrante (Biavaschi, 2016).

Cabe ressaltar que a estratégia de identificação causal no infinito proposta por Heckman (1990) não requer o uso de variáveis instrumentais. Ao contrário, as próprias características do grupo de controle composto por indivíduos com elevada probabilidade de autosseleção permitem estimar uma distribuição salarial contrafactual livre de viés de seleção amostral em variáveis não observadas (Biavaschi, 2016; Mulligan & Rubinstein, 2008). Por outro lado, conforme discutido a seguir, o modelo empírico baseado em cópulas minimiza a dependência de identificação de parâmetros com o uso de instrumentos (Marra & Radice, 2013; Wojtyś, Marra, & Radice, 2016).

3.2 Estimação dos parâmetros

O problema de seleção amostral acontece quando as observações disponíveis não provêm de uma amostra aleatória da população. Há evidências de que o grupo de migrantes brasileiros são positivamente selecionados em relação aos não

migrantes (Freguglia & Procópio, 2013; Gama & Machado, 2014; Santos Jr et al., 2005). Da mesma forma, temos que os migrantes de retorno se autosselecionam na decisão de retornar e isto deve ser considerado para que estimativas não tendenciosas do resultado no mercado de trabalho do migrante interestadual sejam produzidas (Biavaschi, 2016; Borjas & Bratsberg, 1996; Chiswick, 1999).

O modelo bivariado proposto por Heckman (1979), conhecido como *Tobit-2*, é um dos modelos mais utilizados em problemas de seleção amostral. Ele pode ser representado pela equação de seleção (1) e a equação de resultado potencial (2). Sua estimativa por Máxima Verossimilhança pressupõe os termos aleatórios u_i^* e ϵ_i seguem uma distribuição normal bivariada. No entanto, como essa suposição está sujeita a erros de especificação de distribuição, modelos que relaxem essa hipótese têm sido sugeridos na literatura especializada (Cameron & Trivedi, 2005; Toomet & Henningsen, 2008; Wojtyś et al., 2016).

O uso de funções cópulas, por exemplo, assegura flexibilidade distributiva ao relaxar a hipótese de normalidade conjunta de u_i^* e ϵ_i, e garante eficiência na estimação das equações (1) e (2). Além disso, a opção oferecida pela abordagem de cópula é útil sempre que a precisão das estimativas dos parâmetros estruturais for a prioridade, de modo que não necessariamente precisamos de uma identificação de parâmetros por meio de restrição de exclusão (Wiesenfarth & Kneib, 2010).

Como visto na equação (3), a variável S_i controla se a variável de resultado Y_i^* é observada ou não. Seja F_i a função de distribuição cumulativa (FDC) conjunta de (S_i, Y_i^*) e F_{1i} e F_{2i} a FDC marginal de (S_i, Y_i^*), respectivamente. Assumimos a normalidade das distribuições marginais, a dependência entre elas, por sua vez, é modelada usando a abordagem de cópula. Logo, $S_i \sim \mathcal{N}(\mu_{1i}, 1)$ (modelo probit) e $Y_i^* \sim \mathcal{N}(\mu_{2i}, 1)$, onde μ_{1i} e μ_{2i} são preditores lineares e $\sigma > 0$. F_{1i} e F_{2i} estão relacionadas à equação de seleção e de salários respectivamente. Assim, a distribuição conjunta é determinada da seguinte forma:

$$F_i\left(S_i, Y_i^*\right) = C\left(F_{1i}\left(S_i\right), F_{2i}\left(Y_i^*\right); \theta\right), \tag{4}$$

onde θ é um parâmetro de dependência e C é uma função cópula bidimensional. As famílias de cópulas implementadas neste estudo são as arquimedianas: *Clayton*, *Frank*, *Ali–Mikhail–Haq* (AMH), *Farlie–Gumbel–Morgenstern* (FGM), *Gumbel* e *Joe*, bem como as versões rotacionadas (90°, 180° e 270°) para *Clayton*, *Joe* e *Gumbel*. A cópula elíptica Gaussiana juntamente com covariadas lineares ou pré-especificadas não lineares corresponde ao modelo de Heckman (1979) usando o método de máxima verossimilhança (Wojtyś et al., 2016). Mais detalhes sobre as características das famílias selecionadas encontram-se na Tabela 8 do Apêndice, que contém as equações e os intervalos do parâmetro de dependência para cada família de cópula.

Para as cópulas normais, *Frank*, *FGM* e *AMH* o teste para viés de seleção de amostra pode ser baseado no parâmetro de dependência θ, pois a ausência de viés de seleção de amostra é equivalente à condição $\theta = 0$ e $\theta = 1$ para a cópula de *Gumbel*. As cópulas *Clayton*, *Joe* e *Gumbel* não têm uma interpretação direta de θ. Nesses casos a estatística de *Kendall*, que apresenta uma relação matemática com θ, pode facilitar a interpretação e comparação do grau de dependência obtido por diferentes cópulas.[1] Esse parâmetro geralmente situa-se no intervalo $[-1, 1]$, e indica independência entre variáveis aleatórias modeladas na cópula. Quanto mais próximo τ estiver de -1, mais forte será a associação negativa, valores mais próximos de 1 indicam forte dependência positiva (Brechmann & Schepsmeier, 2013; Marra & Wyszynski, 2016).

Outra vantagem das cópulas é que elas permitem uma especificação de modelo por partes, isto é, as distribuições marginais não são limitadas a pertencer à mesma família da distribuição de cópula bivariada escolhida. Utilizar uma versão rotacionada, disponível para *Clayton*, *Joe* e *Gumbel*, é útil quando a cobertura da cópula não é completa. A dependência negativa entre as equações só pode ser capturada pelas versões rotacionadas a 90° e 270° (Marra & Wyszynski, 2016). Quando rotacionadas a 180° obtêm-se as cópulas de sobrevivência (Brechmann & Schepsmeier, 2013).

Com a função cópula incorporada à estimação dos parâmetros das equações (1) e (2) temos que a equação log-verossimilhança a ser maximizada será a seguinte (Wojtyś et al., 2016):

$$\ell = \sum_{i=1}^{n} \left\{ (1 - S_i) \log F_{1i}(0) + S_i \log \left(f_{2i}\left(Y_i^*\right) - \frac{\partial}{\partial Y} F_i\left(0, Y^*\right) \Big|_{Y^* \to Y_i^*} \right) \right\}. \quad (5)$$

Essa função é uma soma de dois conjuntos disjuntos da amostra: um para aqueles valores não observados do salários, ou seja, que representam os migrantes retornados, e outro para as demais observações. Utilizando 4, temos:

$$\ell = \sum_{i=1}^{n} \left\{ (1 - S_i) \log F_{1i}(0) + S_i \log \left(f_{2i}\left(Y_i^*\right)(1 - z_i) \right) \right\}, \quad (6)$$

onde

$$z_i = \frac{\partial}{\partial v} C\left(F_{1i}(0), v; \theta\right) \Big|_{v \to F_{2i}\left(Y_i^*\right)}$$

e F_{1i} e F_{2i} são funções de distribuição cumulativas marginais[2] de (S_i, Y_i^*), respectivamente.

[1] Por exemplo, para cópula normal o parâmetro τ de *Kendall* em termos de θ é obtido através da expressão $(2/\pi)\arcsin(\theta)$. Mais detalhes da relação entre os parâmetros τ e θ podem ser consultados em Marra e Wyszynski (2016, p.6).

[2] A normalidade dessas funções implica em

$$F_{1i}(0) = \Phi\left(-\mu_{1i}\right) \quad \text{e} \quad f_{2i}\left(Y_i^*\right) = \sigma^{-1}\phi\left(\left(Y_i^* - \mu_{2i}\right)\sigma^{-1}\right).$$

Em muitos estudos variáveis contínuas como idade e anos de estudo podem guardar uma relação não linear com as variáveis de interesse, já que envolvem questões como produtividade e ciclo de vida. Nesse casos, a relação é feita de forma pré-especificada o que não garante que a complexidade dessa relação seja devidamente capturada e entendida. Dessa forma, o modelo bivariado de seleção amostra com cópula pode flexibilizar a modelagem de covariadas contínuas utilizando funções suaves que são representadas por *splines*[3] de regressão (Marra & Radice, 2010; Marra & Wyszynski, 2016; Wojtyś et al., 2016). Portanto, as matrizes dos modelos de seleção e salários podem ser construídas como a soma de componentes paramétricos (intercepto, *dummies* e variáveis categóricas) e funções suaves desconhecidas das covariáveis contínuas $K_{1,2}$. Logo, as equações (1) e (2), na abordagem semiparamétrica, podem ser reescritas da seguinte forma:

$$S = \begin{cases} 1 & \text{se } Z_i'\alpha + \sum_{k_1=1}^{K_1} s_{1k_1}\left(z_{1k_1 i}\right) > \epsilon_i \\ 0 & \text{se } Z_i'\alpha + \sum_{k_1=1}^{K_1} s_{1k_1}\left(z_{1k_1 i}\right) \leq \epsilon_i \end{cases} \qquad \text{para } i = 1, \ldots, r+n, \quad (7)$$

$$Y_i^* = X_i'\beta + \sum_{k_2=1}^{K_2} s_{2k_2}\left(z_{2k_2 i}\right) + u_i^* \qquad i = 1, \ldots, r+n, \quad (8)$$

onde s_{1k_1} e s_{2k_2} são os termos suavizados desconhecidos dos regressores contínuos das equações de seleção e salário, respectivamente. Tanto a abordagem totalmente paramétrica quando esta abordagem são discutidas neste estudo.

A definição da cópula utilizada é feita através de testes que consistem em estimar vários modelos e selecionar aquele com melhor critério de informação. Utiliza-se o critério de penalização por parcimônia *Bayesiano Akaike* ou *Schwarz* (AIC e BIC, respectivamente), o que permite identificar a cópula e o modelo que melhor se ajusta aos dados (Wojtyś et al., 2016).[4]

---

[3] Funções suaves são utilizadas com o intuito de flexibilizar o relacionamento entre as variáveis de interesse e variáveis explicativas contínuas e assim evitar as desvantagens da abordagem paramétrica. Essas funções suaves são representadas por *splines* (funções polinomiais particionadas). O *spline* de regressão de uma determinada variável explicativa é composto por um combinação linear de funções básicas conhecidas $b_{jk}(x_j^*)$ e parâmetros de regressão desconhecidos δ_{jk}. Logo, $s_j(x_j^*) = \sum_{k=1}^{q_j} \delta_{jk} b_{jk}(x_j^*)$, onde j indica o termo suave da j-ésima variável explicativa e q_j é o numero de parâmetros de regressão (bases do *spline)* usados para representar o j-ésimo termo suave, o cálculo dessa expressão para cada j produzirá q_j curvas que multiplicadas pelo vetor de parâmetros e somadas produzirão estimativas lineares ou não lineares para $s_j(x_j^*)$. A base q escolhida para uma função suave determina o grau de flexibilidade máxima para um termo suave e bases devem possuir propriedades matemáticas e numéricas adequadas (Marra & Radice, 2010; Wojtyś et al., 2016).

[4] Os critérios AIC e BIC podem ser calculados, respectivamente, por AIC $= -2\widehat{\ell} + 2k$ e BIC $= -2\widehat{\ell} + \log(n)k$, onde $\widehat{\ell}$ é o valor maximizado da função de log-verossimilhança e k o grau de

3.3 Estimação de densidade contrafatual

A estratégia de recuperação da densidade contrafatual de salários $f(Y_i^*)$, isto é, da densidade salarial dos migrantes caso não houvesse migração de retorno, foi dividida em duas etapas. Na primeira foi recuperada a distribuição das características não observáveis, isto é, $f(u_i^*)$. Nesse ponto apenas as características produtivas não mensuráveis são levadas em conta. Note que não é possível recuperar essa distribuição diretamente da equação de salários (2), tendo em vista que os resíduos u_i^* apenas poderiam ser calculados para os trabalhadores que decidiram permanecer em São Paulo. Logo, não podemos determinar, a partir dos dados, o componente não observável da equação de salários ao incluir remigrados na amostra ("devolvê-los para São Paulo"). Para superar essa dificuldade utilizamos a técnica proposta por Biavaschi (2016).

Considerando que a distribuição $f(u_i^*)$ pode ser escrita a partir do pressuposto da Lei da Probabilidade Total como a soma ponderada da distribuição dos termos de erro nas subamostras de permanentes e retornados cujos pesos são dados pela probabilidade de estar em qualquer subamostra, temos que:

$$f\left(u_i^*|Z_i'\alpha\right) = f\left(u_i^*|S_i = 1, Z_i'\alpha\right)\Pr\left(S_i = 1|Z_i'\alpha\right) + f\left(u_i^*|S_i = 0, Z_i'\alpha\right)\Pr\left(S_i = 0|Z_i'\alpha\right). \quad (9)$$

Como explicado, o componente não observável $\left(f\left(u_i^*|S_i = 0\right), Z_i'\alpha\right)$ dessa soma é desconhecido. No entanto, se a probabilidade de ficar em São Paulo $\Pr\left(S_i = 1|Z_i'\alpha\right)$, condicionada às variáveis que influenciam essa decisão, for próxima de 1, então[5]

$$f\left(u_i^*|Z_i'\alpha\right) = f\left(u_i^*\right) \approx f\left(u_i^*|S = 1, Z_i'\alpha\right) = f\left(u_i^*|S_i = 1\right).$$

Então o viés desaparece e a distribuição contrafatual de fatores não observáveis pode ser estimada diretamente nessa subamostra na qual quase todos os indivíduos são permanentes. A intuição por trás dessa estratégia é que a seleção desaparece no limite para aqueles indivíduos no conjunto de alta probabilidade (Chamberlain, 1986). Devemos adicionar dois pressupostos básicos para que essa técnica garanta uma estimativa consistente da quantidade de interesse, são eles: o conjunto de características que seleciona a amostra (Z) deve conter variáveis exógenas e as observações precisam ser i.i.d.

Essa estratégia é conhecida como identificação no infinito e tem sido utilizada na literatura para estimar o termo constante em modelos de seleção

liberdade do modelos de seleção e salário. Ambos os critérios penalizam a adição de variáveis nas estimações (Wojtyś et al., 2016).

[5] Assumindo $f\left(u_i^*|S = 1, Z_i'\alpha\right) = f\left(u_i^*|S = 1\right)$ e $f\left(u_i^*|Z_i'\alpha\right) = f\left(u_i^*\right)$.

semiparamétrica.[6] O trabalho de Biavaschi (2016) foi o primeiro a aplicar essa estratégia para recuperação de distribuição contrafatual com objetivo de considerar a seleção em características não observáveis em uma densidade contrafatual. O presente estudo buscou aplicar essa estratégia na estimativa da distribuição das características não observáveis baseando-se em uma subamostra selecionada nas características observadas (Z), na qual quase todos os indivíduos permanecem no estado de São Paulo.

Seja H_i o indicador se a observação faz parte do conjunto de alta probabilidade e seja $H_i = 1\left[\Pr\left(S_i = 1|Z_i'\hat{\alpha}\right) > \overline{p}_n\right]$, o estimador proposto para $f(u_i^*)$ será

$$\widehat{f\left(u_i^*\right)} = \frac{\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{u-u_i^*}{h}\right) S_i H_i}{\sum_{i=1}^{n} S_i H_i}, \tag{10}$$

onde $K(\cdot)$ é um estimador de densidade de *Kernel* da variável aleatória u_i^* para um conjunto de observações cuja probabilidade de estar na amostra selecionada é próxima de 1 e h é o parâmetro *bandwidth* ou largura dos intervalos de classe. Seguindo DiNardo et al. (1995), Chiquiar e Hanson (2005) e Biavaschi (2016) será utilizado o *Kernel* gaussiano com parâmetro de suavização ótimo dado por $h = 1{,}06\,\hat{\sigma}\,N^{-1/5}$ (Silverman, 2018), onde $\hat{\sigma}$ é o desvio-padrão da distribuição gaussiana, N o total de observações e $\overline{p}_n$ é o valor que define o percentil que limita os indivíduos no grupo de alta probabilidade. Seguindo Biavaschi (2016), esse limite será o percentil 95º calculado a partir da predição de probabilidade pela equação de seleção (1).

Cabe observar que para determinar H_i precisamos antes da estimativa consistente de $\hat{\alpha}$ para determinar as observações com alta probabilidade de permanência por meio da predição $\Pr\left(S_i = 1|Z_i'\hat{\alpha}\right)$ com a equação (2).

A segunda etapa da estimação contrafatual da densidade de salários dos migrantes consistiu em recuperar a distribuição completa dos salários somando a $f(u_i^*)$ pela média do valor predito ($X_i'\hat{\beta}$). Em outras palavras, os salários previstos reais e contrafactuais que foram calculados como o produto do retorno sobre as habilidades incluídas na equação de salários e as características dos migrantes permanentes (população total migrante) para distribuição real (contrafatual) prevista dos salário, isto é, $\hat{\beta}X_j$, onde j = apenas permanentes (população total migrante — devolvendo os remigrados para São Paulo). Nesse ponto são levadas em conta tanto caraterísticas não observáveis quanto as observáveis na análise das diferenças entre retornados e permanentes. Essa separação é útil para identificar em que tipo de características se concentram as diferenças entre os migrantes.

[6] Em trabalhos como os de Andrews e Schafgans (1998); Chzhen e Mumford (2011); Heckman (1990); Liu, Hsiao, Matsumoto, e Chou (2009); Shen (2013).

3.4 Base de dados e tratamento

Neste estudo foram utilizados os microdados do Censo Demográfico 2010 (IBGE, 2012). Essa base de dados permite identificar as diferentes categorias de migrantes a partir do cruzamento dos dados sobre lugar de residência, lugar de nascimento, lugar de última residência e duração de residência.

A amostra selecionada para a análise empírica é composta de homens com idade entre 18 e 70 anos que estavam empregados no momento do recenseamento. Em relação à posição do domicilio foram incluídos na amostra os chefes de domicílio e cônjuges[7] que não estivessem frequentando algum curso, essa estratégia foi aplicada afim de excluir migrantes agregados. Também foram excluídos da análise os indivíduos que não declaram cor e que não responderam o questionário referente à migração, além dos funcionários públicos, militares, sem remuneração e empregadores, essa exclusão foi feita de modo que a análise esteja voltada para decisões de migração e mercado de trabalho (Gama & Machado, 2014; Lima, Simões, & Hermeto, 2015).

Após os cortes mencionados a amostra final totalizou 28.049 migrantes interestaduais, sendo deste total 24.175 (86,18%) migrantes que ficaram no estado de São Paulo e 3.874 (13,81%) migrantes retornados.

Define-se migrante de retorno aquele que morava, na data do recenseamento, em seu estado de naturalidade, tendo declarado residência anterior no estado de São Paulo. O migrante não retornado é definido como aquele que, na data da entrevista, declarou residir fora de seu estado de naturalidade (residência em São Paulo). Existem dois critérios para definir a condição migratória: última etapa e data fixa (curto prazo). No critério de data fixa o indivíduo é questionado sobre o local de residência há cinco anos da entrevista. Com o objetivo de captar um fluxo de retornados em maior espaço de tempo foi utilizado o critério última etapa combinado ao quesito lugar de nascimento para identificação dos migrantes.

O modelo de seleção foi construído a partir de quatro grupos de variáveis. O primeiro grupo refere-se às características pessoais do indivíduo como idade, raça, posição e nível de instrução. O segundo grupo diz respeito aos vínculos familiares como: posição no domicílio (chefe), chefe de família casado, e chefe casado com cônjuge nascido em SP, chefe com filho menor de 14 anos e número de pessoas no domicílio, como essas variáveis estão relacionadas aos custos de migração e vínculos à pessoas no destino elas devem funcionar bem como preditores da decisão de permanecer. Outro grupo de covariadas utilizado é

[7] A partir dos dados do Censo Demográfico de 2010 não é possível identificar o parentesco direto de um indivíduo cuja posição no domicílio seja filho ou enteado. Dessa forma a amostra é limitada aos responsáveis e cônjuges de modo que se torne viável construir a variável de restrição de exclusão (filho nascido em SP).

o relacionado às variáveis de residência, como setor do domicílio e região de origem do migrante.[8]

A variável filho nascido no destino, isto é, no estado de São Paulo, foi incluída como restrição de exclusão, e deve ser incluída na equação de seleção e excluída da equação de salários. É esperado que ter um filho nascido em São Paulo seja um forte indutor da decisão de permanência já que se trata de *proxy* para vínculo social no destino e perspectivas de investimento em capital humano, o que aumenta o custo de oportunidade da migração de retorno. Existem diversas evidências na literatura especializada de que os vínculos familiares são decisivos no campo da migração. Mincer (1978) considerou os laços familiares como importantes para decisões de migração. O tamanho da família é importante porque o retorno da migração aumenta menos que os custos com a presença de filhos e cônjuges. Dustmann (2003), por exemplo, apresenta evidências de que as preocupações de pais altruístas sobre a criança podem levar a um aumento ou a uma diminuição na tendência de voltar ao local de origem dependendo da percepção deles em relação ao bem-estar e ganhos econômicos futuros dos filhos no destino. Nesse ponto é importante destacar que o local de destino objeto deste estudo (São Paulo) está entre os primeiros em qualidade na educação básica segundo os resultados do IDEB (Índice de Desenvolvimento da Educação Básica), aumentando o retorno futuro do investimento em capital humano da criança.[9]

Desse modo, quando o migrante inicia uma nova família ou altera a composição familiar na região de destino são criados novos vínculos e perspectivas para escolhas futuras. Preocupações dos pais no tocante à qualidade de vida, oferta e qualidade de serviços educacionais, saúde, lazer e cultura podem reduzir a chance dessa família retornar para uma região onde tais características são precárias. Portanto, uma vez controlada a composição familiar (cônjuge, outros filhos), o nascimento de um filho na região de destino é um evento que pode induzir novos vínculos sociais e parâmetros para escolhas futuras, o que pode se correlacionar com a decisão de permanência.[10]

Por outro lado, conforme defendido por Biavaschi (2016), não é provável que existam canais que liguem o local de nascimento do filho ao salário do chefe

[8] Uma limitação dos dados do Censo Demográfico consiste na impossibilidade de identificação do momento (antes ou após o retorno) em que as característica dos indivíduos como foram adquiridas.

[9] Os indíces por de 2007 a 2019 podem ser consultados em https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb/resultados

[10] O trabalho de Biavaschi (2016) considerou que o efeito de ter um filho nascido nos EUA não deve prever o salário de um indivíduo após controlar os efeitos de vínculo e rede por meio da variável cônjuge nascido nos EUA e do tempo de permanência nos EUA. Variável semelhante ao tempo de permanência na UF não está disponível na base de dados do Censo demográfico brasileiro de 2010 para o caso dos migrantes retornados.

de família do sexo masculino, uma vez que mulheres tendem a sacrificar mais horas de trabalho para os cuidados dos filhos. Portanto, ao usar uma amostra só de migrantes homens, o nascimento de um filho no local destino pode não ter influência sobre a produtividade do trabalho do migrante, mitigando a endogeneidade da variável instrumental.[11]

O modelo de salários[12] também foi construído a partir dessas variáveis educacionais, socioeconômicas e de estrutura familiar (exceto pela restrição de exclusão), incluindo a variável possuir esposa nascida no destino, pois essa variável deve capturar os efeitos das redes de sociais (Biavaschi, 2016). A Tabela 7 no Apêndice contém um resumo das variáveis disponíveis na base de dados relacionadas à migração, às condições socioeconômicas e educacionais que foram utilizadas utilizadas nas regressões. Vale ressaltar que esse conjunto de variáveis está de acordo com a literatura nacional e internacional especializada conforme a disponibilidade do Censo Demográfico de 2010 (Biavaschi, 2016; Gama & Machado, 2014; Justo & Ferreira, 2012; Mincer, 1978; Ramalho & Queiroz, 2011; Santos, 2006).

A Tabela 3 contém as estatísticas descritivas das características socioeconômicas e demográficas amostra de trabalhadores selecionados de migrantes permanentes, retornados e dos não migrantes naturais do estado de São Paulo. Há importantes diferenças entre esses grupos que merecem destaque. Primeiro, os migrantes permanentes são, em média, mais jovens e menos escolarizados que os migrantes retornados, fato que pode indicar uma seleção positiva dos migrantes de retorno. Em relação aos não migrantes, tanto migrantes retornados quanto permanentes são, em média, mais jovens e menos escolarizados.

As variáveis demográficas indicam que para permanentes e retornados há grande predominância da origem no Nordeste, principalmente para o grupo de permanentes. Diversos trabalhos já destacaram não só a preferência dos nordestinos pelo destino São Paulo, mas também que há um crescente movimento de retorno para essa região (A. T. R. d. Oliveira et al., 2011; Ramalho et al., 2016; Ramalho & Queiroz, 2011; Sachsida, de Castro, de Mendonça, & Albuquerque, 2009; Santos, 2006).

[11] Ainda é preciso reconhecer que o uso de variáveis instrumentais não é trivial por si só. O caso em que necessitamos de uma variável que seja exógena em relação à variável de resultado e endógena em relação à escolha do indivíduo representa um desafio na formulação de modelos e cenários preditivos adequados, limitado ainda pela indisponibilidade de dados. Na grande maioria das pesquisas publicadas tais limitações são reconhecidas, porém contrapostas com hipóteses e argumentações teóricas em linha com aquelas que já apresentamos. Todavia, foram aplicadas metodologias e testes com diferentes especificações que permitem contornar a dependência de variável instrumental.

[12] A variável de salário-hora é construída como rendimentos do trabalho principal divididos por horas de trabalho. O salário em São Paulo para os retornados não é observado.

Tabela 3. Médias das características socioeconômicas e demográficas observáveis de migrantes e não migrantes (Amostra)

Variável	Permanentes	Migrantes Retornados	Não Migrantes
Não branco (%)	59,94	52,14***	31,71
Idade	32,97	39,33***	42,39
Sem instrução (%)	55,46	49,27***	36,64
Nível fundamental (%)	20,09	19,64	18,51
Nível médio (%)	21,43	24,85***	32,68
Nível superior (%)	2,88	5,97***	11,67
Pós graduação (%)	0,14	0,27	0,49
Chefe (%)	79,64	81,60***	81,32
Chefe e vive com cônjuge (%)	61,30	68,62***	71,23
Filho de 14 anos (%)	62,67	70,25***	72,22
Filho de 14 anos nascido em SP (%)	37,72	28,78***	71,52
Chefe e vive cônjuge nascido em SP (%)	14,58	7,38***	79,42
Nº de pessoas no domicílio	3,47	3,36***	3,45
Renda no trabalho principal (R$)	1074.26	1109.75**	1596,45
Salário-hora (R$)	6,31	6,50**	9,36
Setor urbano (%)	93,31	87,71***	91,09
Setor rural (%)	6,69	12,29***	8,91
Origem – Sudeste (%)	19,43	25,29***	–
Origem – Nordeste (%)	65,44	40,72***	–
Origem – Centro-Oeste (%)	2,52	3,86***	–
Origem – Norte (%)	1,10	0,82*	–
Origem – Sul (%)	11,51	29,30***	–
Amostra total	24.175	3.874	140.166

Nota: Níveis de significância: *10%, **5%, ***1% para um teste *t* para diferenças nas médias entre os grupos de permanentes e retornados. A amostra consiste em homens de 18 a 70 anos, empregados em atividade remuneratória. Amostra não expandida, isto é, não foi considerado o peso amostral.

Fonte: Elaboração própria com base nos dados do censo 2010.

4. Resultados

4.1 Determinantes da decisão de remigrar e salários

Uma série de especificações empíricas para o modelo de determinação conjunta da decisão de remigração e salários (equações (1) e (2)) foram previamente estimados considerando vários tipos de cópulas, covariadas e formas paramétricas e semiparamétricas. A Tabela 9 no Apêndice contém os valores do critério de informação AIC[13]. Os resultados sugerem que a cópula *Joe*[14] com rotação

[13] A partir de simulações, Wojtyś et al. (2016) mostram que a escolha da cópula verdadeira é melhor executada a parir o critério AIC.

[14] Especificações totalmente paramétricas, utilizando a cópula Joe 180° e a cópula Normal (equivalente ao modelo de Heckman (1979) em dois estágios estimado por máxima

180° se ajusta melhor aos dados em todas as especificações. O modelo mais completo, contendo variáveis referentes às características pessoais, familiares, trabalho e residência obteve o menor valor de referência. A especificação semiparamétrica não favorece uma cópula diferente dos outros modelos, mas apresenta menor valor AIC em relação à especificação totalmente paramétrica. O parâmetro de dependência das distribuições marginais de (1) e (2) registra valor $\theta = 2{,}44$, estatisticamente diferente de zero, cujo parâmetro τ de *Kendall* é 0,439, sugerindo que fatores não observados que afetam a decisão de permanência também afetam os salários positivamente.

Os resultados da Tabela 4 referem-se ao modelo de seleção amostral bivariado semiparamétrico baseado na cópula *Joe* 180°. Em se tratando da restrição de exclusão: trabalhador que reside com filho nascido em São Paulo, expectativa teórica é que tal variável tenha uma relação importante na predição da decisão de permanência do indivíduo, pois se trata de um forte vínculo familiar no destino. Segundo os resultados do modelo de seleção essa variável é significativa a 1% e indica que ter um filho nascido em São Paulo pode aumentar a probabilidade de o trabalhador migrante permanecer no estado em relação a ter um filho migrante. Ter cônjuge nascido em São Paulo também deve aumentar essa probabilidade, evidenciando o papel dos laços familiares formados no destino.

As variáveis refentes ao nível de instrução indicam uma relação inversa com a decisão do trabalhador não remigrar. Comparados àqueles sem instrução (categoria omitida), indivíduos com nível médio e superior tem menos chances de permanecer. A grande predominância de indivíduos sem instrução na amostra (55,46%) pode indicar que esse resultado está relacionado à frustração das expectativas de salários de migrantes com alto nível de instrução. Ao se deparar com um mercado de trabalho acirrado e seletivo esse grupo pode sofrer frustração em relação às suas expectativas de pagamento do capital humano, facilitando o retorno de trabalhadores mais instruídos. Já os trabalhadores sem instrução podem acessar uma maior oferta de empregos de baixa qualidade. Essa suposição é consistente com Ambrosini et al. (2010), onde defende-se que as decisões de migração e retorno estão relacionadas ao prêmio salarial oferecido para uma classe específica de habilidades na origem e no destino.

Migrantes com origem no Norte e Nordeste apresentam maior propensão de permanecer em São Paulo. Migrantes nascidos na região Sul ou no Centro-Oste, por sua vez, têm mais chances de retornar. Apesar de um movimento mais intenso de retorno para o Nordeste, a migração ainda é fortemente relacionada aos níveis de pobreza e desigualdade de renda nas regiões tradicionalmente

verossimilhança, Tobit-2), são disponibilizadas no apêndice. A partir da Figura 9 (no Apêndice) verifica-se que os resultados são muito próximos aos da abordagem semiparamétrica, mantendo-se as conclusões do modelo selecionado.

Tabela 4. Estimativas das equações de seleção e salários

Covariadas	Equação de Seleção (probit, $S = 1$) (1)	Equação de Salários (2)
Filho migrante (*categoria omitida*)		
Filho de 14 anos nascido em SP	0,3153***	–
	(0,0253)	
Sem filhos (*categoria omitida*)		
Filho de 14 anos	−0,4755***	−0,0012
	(0,0394)	(0,0046)
Branco (*categoria omitida*)		
Não Branco	−0,0050	−0,0193***
	(0,0212)	(0,0031)
Sem instrução (*categoria omitida*)		
Nível Fundamental	−0,1684***	0,0512***
	(0,0274)	(0,0039)
Nível Médio	−0,2576***	0,1058***
	(0,0266)	(0,0039)
Nível superior	−0,3196***	0,3996***
	(0,0527)	(0,0095)
Pós-Graduação	−0,2146	0,5621***
	(0,2259)	(0,0388)
Cônjuge (*categoria omitida*)		
Chefe	0,2977***	0,0342***
	(0,0460)	(0,0054)
Chefe solteiro (*categoria omitida*)		
Chefe e vive com cônjuge	−0,4535***	−0,0089*
	(0,0424)	(0,0049)
Chefe e vive cônjuge nascido em SP	0,5121***	0,0074*
	(0,0370)	(0,0045)
Setor rural (*categoria omitida*)		
Setor Urbano	0,2475***	0,1004***
	(0,0344)	(0,0060)
Origem – Sudeste (*categoria omitida*)		
Origem – Norte	0,1972*	0,0001
	(0,1051)	(0,0144)
Origem – Nordeste	0,2983***	−0,0400***
	(0,0259)	(0,0040)
Origem – Sul	−0,3435***	0,0060
	(0,0307)	(0,0056)
Origem – Centro-Oeste	−0,1525***	0,0240**
	(0,0585)	(0,0099)
Intercepto	1,1667***	1,4407***
	(0,0535)	(0,0082)
Idade (*spline*)	sim	sim
Total de pessoas no domicílio (*spline*)	sim	sim
Observações	28.049	24.175

Notas: Parâmetro de Dependência (Intervalo de confiança 95%): $\theta = 2{,}44\ (2{,}3; 2{,}6)$ τ de *Kendall*: $\tau = 0{,}439\ (0{,}415; 0{,}464)$. Erros-Padrão entre parêntese. O modelo bivariado é do tipo cópula *Joe* rotacionada a 180°. Níveis de significância: *10%; **5%; ***1%.

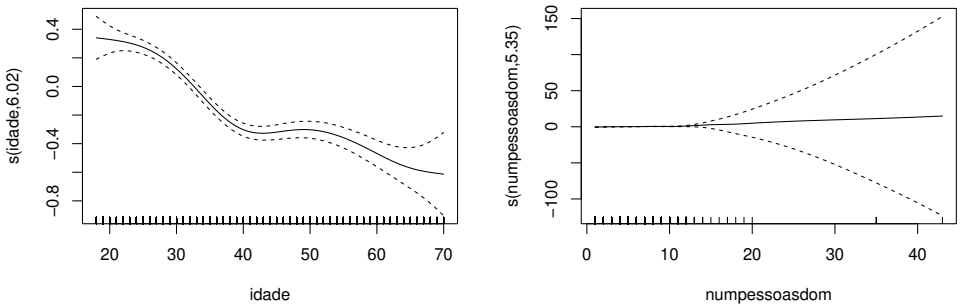
Fonte: Elaboração própria a partir dos microdados do Censo Demográfico de 2010.

emissoras, sendo esse um fator de expulsão populacional (Gama & Machado, 2014). Paralelamente, a dinâmica migratória brasileira sofreu significativas mudanças, a região Centro-Oeste passou a ser uma região de grande absorção e retenção de migrantes, e São Paulo passou a apresentar perdas migratórias para essa região e para o Sul do país (Baeninger, 2008).

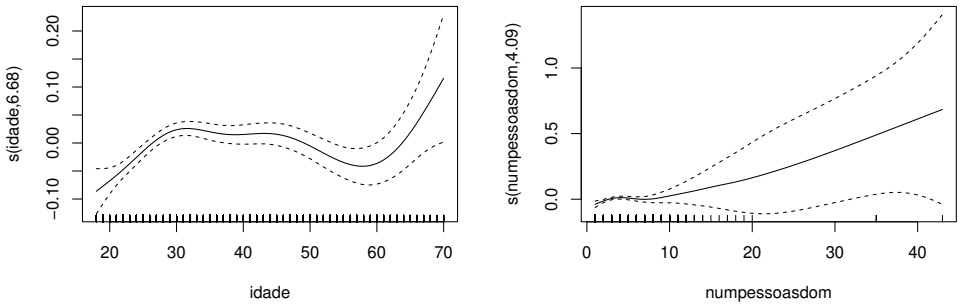
Em relação à equação de salários: o logaritmo do salário-hora responde positivamente ao nível de instrução. Quanto maior esse nível, maiores os coeficientes da equação de salários. Note-se ainda que a variável referente à raça aponta para discriminação no mercado de trabalho, com uma diferença desfavorável aos não brancos de 1,91%. Em termos de vínculos familiares temos que, um chefe de família cujo cônjuge é natural de SP aufere salários maiores em relação a um chefe de família solteiro. Esse fato pode estar relacionado às redes sociais construídas no destino (Biavaschi, 2016).

Residir em área urbana aumenta os rendimentos em 10,56%. Com relação à origem do migrante foram significativas as variáveis de nascimento no Nordeste e Centro-Oeste. Migrantes com origem no Nordeste (Centro-Oeste) auferem salários inferiores (superiores) em relação aos de origem Sudeste. Esse resultado pode ser explicado em termos de atributos educacionais tendo em vista a concentração de migrantes nordestinos com níveis de escolaridade mais baixos. Como visto na Tabela 2, 69,5% dos migrantes sem instrução são oriundos dos estados do Nordeste, enquanto o Centro-Oeste se destaca nos níveis mais elevados de instrução.

A Figura 2 registra os resultados não paramétricos das equações de seleção e salários para o modelo cópula selecionado. Os resultados para variável idade são consistentes com a suposição de tanto rendimentos quando a probabilidade de permanência respondem não linearmente à medida que as pessoas envelhecem. A estimativa suavizada da equação seleção indica que a probabilidade de permanecer no destino tende a diminuir à proporção que a idade aumenta, até por volta dos 40 anos, quando há uma ligeira mudança para uma tendência positiva até por volta dos 50 anos, voltando a diminuir. O resultado não paramétrico para os salários sugere que o salário aumenta com a idade até por volta dos 30 anos, se tornando quase constante dos 30 aos 40 anos e declinando após essa faixa até o os 60 anos. A partir desse ponto a relação é crescente. Os intervalos do resultado suavizado para o número de pessoas no domicílio (numpessoasdom) na equação de seleção e salários contêm linha zero para a maior parte da faixa de valores da covariável, indicando que esse pode ser um preditor fraco da permanência e dos rendimentos.



(a) Equação de Seleção



(b) Equação de Salários

Notas: Estimativas das funções suaves do modelo bivariado cópula Joe 180°. Intervalos de confiança a 95%. O "tapete" na parte inferior dos gráficos representa os valores das covariadas. Os números mostrados no eixo y em cada gráfico indicam os graus estimados de liberdade das curvas suaves. P-valores para os termos idade e número de pessoas no domicílio da equação de seleção (a): 0,000, 0,000, respectivamente; equação de salários (b): 0,000, 0,003.

Fonte: Elaboração própria com base nos dados do Censo Demográfico de 2010.

Figura 2. Densidade real e contrafatual do salário-hora para migrantes – abordagem paramétrica

5. Perfil do grupo de controle

Este estudo usa a estratégia de identificação no infinito para recuperar a distribuição de salários contrafatual em características não observadas, isto é, a distribuição que vigoraria caso não houvesse migração de retorno (Andrews & Schafgans, 1998; Biavaschi, 2016; Heckman, 1990). Conforme discutido nas equações (9) e (10), essa distribuição pode ser estimada livre de viés de seleção amostral usando um grupo de controle formado por migrantes com Alta Probabilidade de Permanência (GAP) na região de destino. A Tabela 5 contém as médias das características observáveis para os indivíduos do GAP, isto é, aqueles trabalhadores migrantes cuja probabilidade predita de permanência em São Paulo se encontra acima do percentil 95º (vide equação (1) e Tabela 4). Os dados permitem identificar as características que compõem o perfil desses trabalhadores.

Tabela 5. Características médias dos migrantes do GAP versus demais migrantes

Variável	GAP ($H_i = 1$)	Outros migrantes ($H_i = 0$)
Não Branco (%)	66,07	58,82***
Idade	23,74	34,33***
Sem Instrução (%)	69,14	54,23***
Nível Fundamental (%)	19,81	20,04
Nível Médio (%)	11,04	22,30***
Nível Superior (%)	0,00	3,26***
Pós-Graduação (%)	0,00	0,16***
Chefe (%)	95,29	79,09***
Chefe e vive com cônjuge (%)	40,55	63,42***
Chefe e vive cônjuge nascido em SP (%)	38,42	12,21***
Filho até 14 anos (%)	33,50	65,34***
Filho até 14 anos nascido em SP (%)	32,86	36,59***
Total de pessoas no domicílio	4,42	3,41***
Urbano (%)	99,07	92,14***
Origem – Sudeste (%)	2,49	21,08***
Origem - Nordeste (%)	96,79	60,52
Origem – Centro-oeste (%)	0,00	2,84***
Origem – Norte (%)	0,71	1,08
Origem – Sul (%)	0,00	14,48***
Amostra total	1.393	26.656

Notas: H_i indica se a observação encontra-se acima do percentil 95º. Níveis de significância: *10%, **5%, ***1% para um teste de diferenças nas médias entre os grupos.

Fonte: Elaboração própria a partir dos microdados do Censo Demográfico de 2010.

O primeiro ponto importante a ser destacado é a grande concentração de indivíduos no GAP cuja origem é a região Nordeste. Ou seja, 96,79% daqueles migrantes com alta probabilidade de permanência são oriundos dessa região. Esses indivíduos também apresentam nível de instrução baixo, com maior proporção de indivíduos na categoria sem instrução, 69,14%, nenhum migrante com nível superior ou pós graduação foi identificado, nessa amostra, como pertencente ao conjunto. Eles ainda são mais jovens em relação àqueles migrantes fora do grupo de controle, com uma diferença de idade de ≈ 10 anos. Os vínculos familiares também se destacam nas diferenças entre esses grupos. Os indivíduos no GAP são, em média, mais propensos a ter um cônjuge nascido em São Paulo, em relação aos outros. Em contrapartida, são, na maioria, chefes de família solteiros e sem filhos.

Conforme colocado na seção 3.3, calculamos os resíduos da equação de salários (vide Tabela 4) para grupo de controle (GAP) e usamos o estimador de núcleo *Kernel* (vide (10)) para estimar a distribuição contrafatual de salários em características não observadas. Em seguida, conforme Biavaschi (2016),

modificamos tal densidade pela média da distribuição de salários previstos para toda população migrante, o que corresponde à estimativa completa da distribuição dos salários dos migrantes interestaduais de São Paulo caso não houvesse migração de retorno. A seção seguinte relata as densidades dessas distribuições em ambos os cenários.

6. Efeitos da migração de retorno sobre os salários dos imigrantes

A Tabela 6 resume a comportamento das densidades de salários por horas de trabalho estimadas para os migrantes conforme as métricas de decis, médias e diferenças de percentis (valores em logaritmo). A coluna (1) — fatual — foi obtida a partir da amostra de migrantes permanentes (residentes em São Paulo) com a aplicação do estimador de densidade *Kernel*. Sua primeira parte corresponde à densidade de salários-hora preditos em características observadas $f(X'\hat{\beta})$. A segunda resulta da estimativa de densidade sobre a distribuição dos resíduos $f(Y_i - X'\hat{\beta})$, que corresponde aos fatores não observáveis. Na terceira parte, tem-se a estimativa de densidade de salários totais, isto é, considerando a distribuição dos resíduos e o deslocamento pelos salários médios previstos. A coluna (2) — contrafatual — foi calculada no cenário em que não há migração de retorno. Assim, no primeiro painel foi utilizada a amostra completa de migrantes interestaduais caso todos tivessem permanecido (salários previstos devolvendo os retornados para São Paulo) para recuperação da distribuição em variáveis observadas. No segundo painel, os valores foram obtidos a partir do cálculo de resíduos de salários para grupo de alta probabilidade de permanência (grupo de controle), como visto na seção 3. Já o terceiro painel refere-se aos salários totais, considerando a distribuição dos resíduos e o deslocamento pelos salários médios previstos. Em outras palavras, essa distribuição corresponde à soma das características observáveis médias ($\hat{Y}_i$) com o componente não observado $\hat{u}_i$ da população migrante em cada decil.

Em termos de características observáveis os migrantes interestaduais em São Paulo estariam, em média, ganhando mais caso não houvesse migração de retorno. Note-se que, na Tabela 6 a diferença logarítmica ao longo dos decis é quase sempre positiva, exceto pelos primeiros decis daqueles indivíduos na parte inferior da distribuição, para esses indivíduos há evidências de seleção negativa em observáveis. No cenário contrafatual, considerando apenas diferenças observáveis, a população migrante estaria recebendo 0,29% a mais, em média, e 0,14% na mediana.

O último painel da Tabela 6 mostra os decis da distribuição total de salários-hora. É possível ver que há uma influência mais evidente das características não observadas no efeito da migração de retorno para os decis extremos.

Tabela 6. Decis das distribuições de salários dos migrantes com e sem migração de retorno

	(1) Fatual	(2) Contrafatual	Diferença (2)−(1)
Observáveis		$\widehat{Y}_i$	
1º decil	1,3197	1,1730	−0,1467
2º decil	1,4825	1,4773	−0,0052
3º decil	1,5099	1,5081	−0,0017
4º decil	1,5302	1,5298	−0,0004
5º decil	1,5487	1,5501	0,0014
6º decil	1,5739	1,5767	0,0028
7º decil	1,6019	1,6077	0,0058
8º decil	1,6458	1,6549	0,0091
9º decil	2,1795	2,4397	0,2602
Média	1,5659	1,5688	0,0029
Não observáveis		$\hat{u}_i$	
1º decil	−0,8587	−0,4376	0,4210
2º decil	−0,2423	−0,2049	0,0374
3º decil	−0,1553	−0,1137	0,0416
4º decil	−0,0869	−0,0505	0,0365
5º decil	−0,0256	−0,0044	0,0213
6º decil	0,0374	0,0421	0,0047
7º decil	0,1188	0,1074	−0,0114
8º decil	0,2519	0,2160	−0,0359
9º decil	0,9117	0,8859	−0,0257
Média	0,0000	0,0147	0,0147
Total		Y_i	
1º decil	0,7072	1,1311	0,4240
2º decil	1,3236	1,3639	0,0403
3º decil	1,4105	1,4550	0,0445
4º decil	1,4789	1,5183	0,0394
5º decil	1,5402	1,5644	0,0242
6º decil	1,6033	1,6109	0,0076
7º decil	1,6847	1,6762	−0,0085
8º decil	1,8177	1,7848	−0,0329
9º decil	2,4775	2,4547	−0,0228
Média	1,5659	1,5835	0,0176
Diferenciais Salariais			
10º percentil – 90º percentil	1,7703	1.3226	−0,4477
10º percentil – 50º percentil	0,833	0,4333	−0,3997
50º percentil – 90º percentil	0,9373	0,8903	−0,047

Notas: As predições de salários foram baseadas nos parâmetros estimados para o modelo bivariado do tipo cópula Joe rotacionada a 180°.

Fonte: Elaboração própria a partir dos microdados do Censo Demográfico de 2010.

Quando reintroduzidos no mercado de trabalho paulista, os retornados acabam aumentando o salario médio da população migrante em quase toda distribuição. Quando não consideramos características inatas, somente a parte média e superior da distribuição mostram melhora. As características não observáveis acabam melhorando o desempenho da cauda inferior e piorando parte da causa superior, sugerindo que aqueles com rendimentos mais baixos se diferem dos demais em termos de caraterísticas como: ambição, empreendedorismo, agressividade, propensão ao risco, motivação e etc. Destarte, a parcela de migrantes na parte inferior da distribuição estaria ganhando mais caso não houvesse migração de retorno. Ademais, a população total migrante estaria ganhando 1,76% a mais na média e 2,42% na mediana, devido à diferenças não observáveis entre migrantes permanentes e retornados.

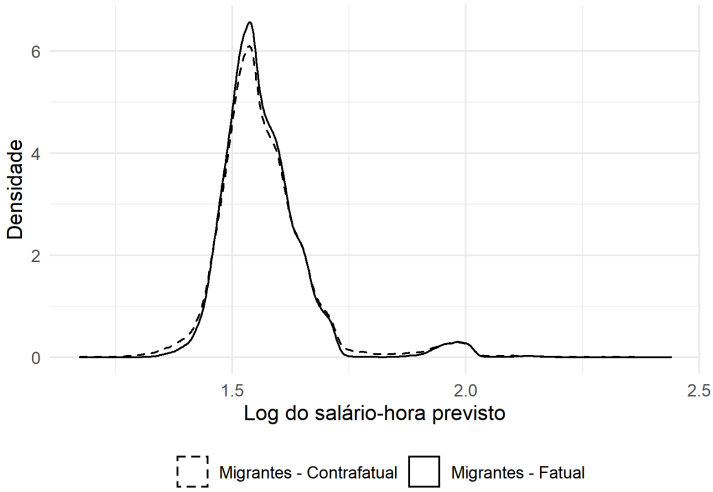
As diferenças entre os grupos de migrantes permanentes e migrantes retornados mudam com a introdução de uma análise baseada em características não mensuráveis. Apesar do grupo de trabalhadores de baixa qualificação ser menos favorecido no mercado de trabalho, suas habilidades não mensuráveis parecem compensar essa falha. O que não ocorre para parte do grupo com habilidades observáveis superiores.

As Figuras 3 e 4 descrevem os resultados das distribuições de salários reais (linha sólida) e contrafatuais (linha pontilhada) graficamente. As diferenças estatisticamente não nulas entre essas linhas representam os efeitos da migração de retorno sobre a densidade dos salários dos imigrantes (De Coulon & Piracha, 2005; DiNardo et al., 1995).

A Figura 3 mostra as distribuições condicionadas ao conjunto de variáveis observáveis. É possível observar que essas distribuições são muito próximas, indicando que retornados e permanentes são bastante semelhantes em termos das suas habilidades mensuráveis. A distribuição contrafatual da população total migrante mostra mais massa na parte inferior e menos massa na parte superior da distribuição quando comparada a distribuição real dos residentes em SP.

As figuras 4(a) e 4(b), por sua vez, refere-se às estimativas de distribuição de salários totais. Elas evidenciam maiores diferenças entre as distribuições. Na ausência de migração de retorno mais migrantes apareceriam na metade superior e topo da distribuição, além de haver um deslocamento à direita na parte esquerda da distribuição, aumentando o salário médio dessa população. Como constatado na análise dos decis, parte da cauda superior dessa distribuição possui menos massa no cenário em que não há migração de retorno, indicando uma piora salarial. Essa figura dá maior suporte à seleção positiva em não observáveis dos migrantes de retorno oriundos de São Paulo.

Há evidências de que uma análise baseada apenas em características observáveis subestima o efeito da migração de retorno sobre a distribuição dos



Nota 1: A distribuição real representa a distribuição dos salários dos trabalhadores migrantes em São Paulo, considerando apenas as características observáveis. A distribuição contrafatual representa a distribuição dos salários caso todos os migrantes tivessem permanecido em SP, isto é, da população total migrante.

Nota 2: Um teste K-S rejeitou a hipótese nula de que as distribuições sejam iguais, a 1% de significância: estatística $D = 0{,}022$ e p-valor $= 0{,}0000$.

Nota 3: As predições de salários foram baseadas nos parâmetros estimados para o modelo bivariado do tipo cópula *Joe* rotacionada a 180°.

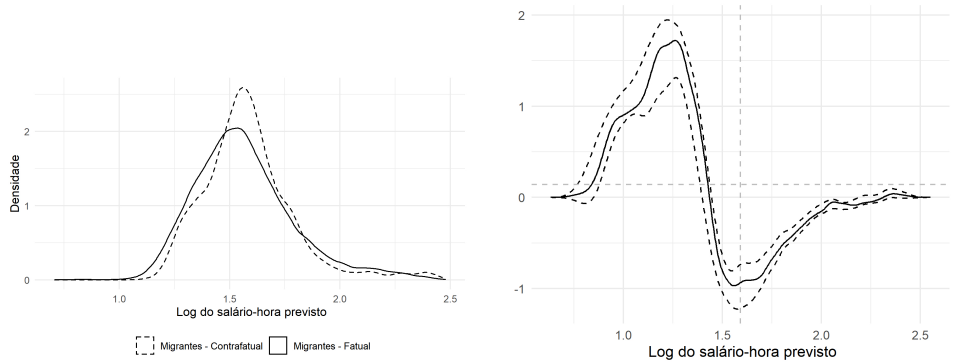
Fonte: Elaboração própria a partir dos microdados do Censo Demográfico de 2010.

Figura 3. Densidade real e contrafatual do salário-hora para migrantes e paulistas – observáveis

salários da população migrante. Além disso, observa-se que ao contar com não mensuráveis a distância entre as distribuições de migrantes de paulistas diminui em relação à análise em características observáveis.

As diferenças entre as distribuições reais e contrafatuais são melhor visualizadas a partir da Figura 4(b). Observa-se que as diferenças entre as densidades são maiores nos níveis mais baixos e médios de rendimento, indicando um efeito maior da migração de retorno. Abaixo da média salarial total (linha pontilhada mais clara) a diferença é em grande parte positiva, enquanto acima da média, por volta do salário-hora de R$1,7, a diferença é sempre negativa.

A Tabela 6 mostra que a migração de retorno também afeta a desigualdade salarial entre os migrantes que permanecem no estado, a partir da diferença entre os percentis 10–90, 10–50 e 50–90 das distribuições reais e contrafatuais. A diferença salarial entre os percentis extremos diminuiria significativamente caso não houvesse migração de retorno, representando uma variação de $-44{,}77\%$ nos diferenciais de salários. Na parte inferior da distribuição também há um movimento de arrefecimento substancial das desigualdades, caso não houvesse migração de retorno a diferença entre os percentis 90 e 50 seria $-39{,}97\%$. Já na



(a) Distribuição real e contrafatual dos salários

(b) Diferença entre as densidades

Nota 1: (a) A distribuição real corresponde à distribuição dos salários totais (características observáveis e não observáveis) dos migrantes permanentes (residentes em SP), o contrafatual corresponde a distribuição dos salários totais dos migrantes interestaduais caso não houvesse migração de retorno, i.e., caso todos perecessem no destino.

Nota 2: (b) Intervalo de confiança representado pelas linhas pontilhadas, supondo um nível de 95% de confiança. A estimação dos erros-padrão foi feita usando a técnica de reamostragem *bootstrap*.

Nota 3: O teste *Kolmogorov–Smirnov* rejeitou-se a hipótese nula de que as densidades reais e contrafatuais sejam iguais, ao nível de 1% de significância: estatística $D = 0{,}08$ e p-valor $= 0{,}0000$.

Nota 4: As predições de salários foram baseadas nos parâmetros estimados para o modelo bivariado do tipo cópula *Joe* rotacionada a 180°.

Fonte: Elaboração própria com base nos dados do Censo Demográfico de 2010.

Figura 4. Densidade reais e contrafatuais do salário-hora total para migrantes

parte superior da distribuição (10–50) há um efeito mais discreto da migração de retorno sobre as diferenças salariais (−4,7%).

No geral, a desigualdade salarial entre a população migrante diminuiria se todos os migrantes decidissem ficar em São Paulo. Tendo em vista que sem migração de retorno, a média salarial dos migrantes aumenta na parte inferior e diminui na parte superior. Portanto, há uma tendência de aprofundamento da desigualdade entre os migrantes permanentes devido à saída de trabalhadores positivamente selecionados e com maiores níveis de instrução, aumentando a dispersão dos salários, considerando paralelamente a grande proporção de trabalhadores sem instrução que tendem a permanecer no destino.

Em relação à população não migrante natural de São Paulo é fato que há uma diferença substancial em relação aos migrantes não só devido à características observáveis e não observáveis entre aqueles que optam por não de deslocar, já evidenciada na literatura, mas também pelas grandes disparidades regionais que são identificadas no país. Segundo Batista e Cacciamali (2009) há um distanciamento entre migrantes e não migrantes na região sudeste. Silveira Neto e Magalhães (2004), por sua vez, apontam para magnitude dos diferenciais de renda dentro do universo de migrantes e não migrantes no estado de São Paulo devido aos diferenciais de capital humano. Já os autores

Assis et al. (2012), ao analisar a migração entre os estados de São Paulo e Bahia identificam, mais uma vez, um hiato salarial entre baianos e paulistas, favorável ao último grupo. No que se refere ao estudo em questão, na Figura 10 no Apêndice observa-se uma ligeira aproximação e, portanto, uma diminuição da desigualdade entre migrantes e não migrantes caso não houvesse migração de retorno.

6.1 Efeitos da Migração de Retorno considerando heterogeneidades

6.1.1 Mercado de trabalho formal e informal

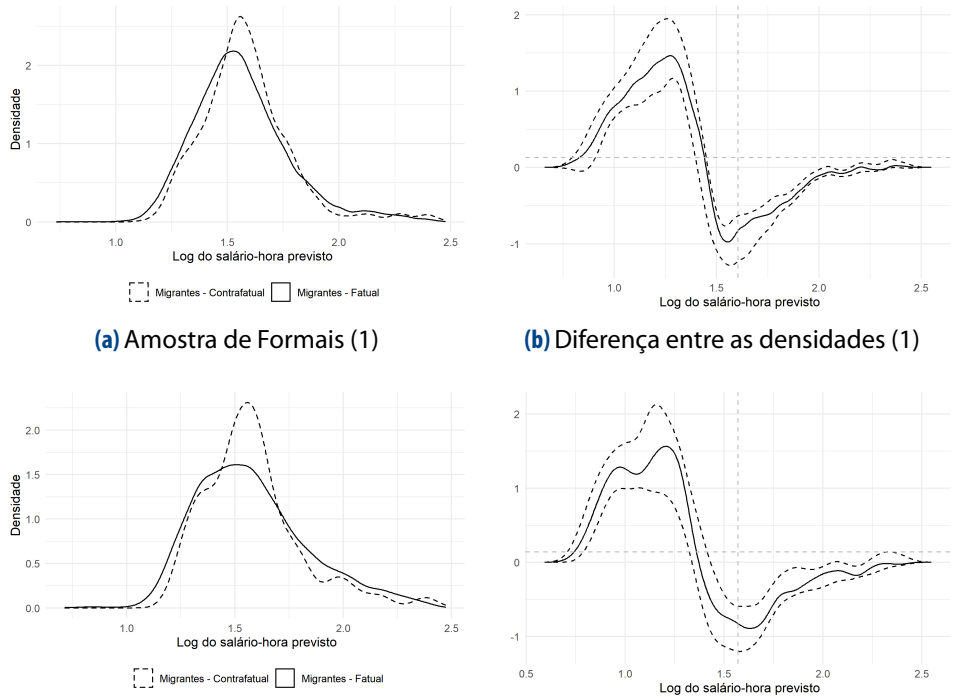
O grupo de trabalhadores formais representa 71,93% da amostra total utilizada neste estudo. Desses, 68,35% são migrantes permanentes. Há também predominância desse grupo no conjunto de alta probabilidade de permanência com 82,91%. Esses dados indicam que embora a população de migrantes em São Paulo seja de baixa instrução é também, em sua maioria, de trabalhadores com carteira de trabalho assinada, em regime de formalidade. Os trabalhadores do setor informal, por sua vez, são aqueles sem carteira de trabalho assinada, na maioria autônomos (conta-própria).

Sem migração de retorno, o grupo de migrantes do setor formal estaria ganhando mais na metade inferior e grande parte da metade superior da distribuição, resultado próximo ao da análise da amostra total. As diferenças reforçam esse resultado (Figura 5(b)). No caso do grupo na informalidade e autônomos há certa irregularidade no comportamento da distribuição contrafatual (tracejada). A metade inferior, próximo à cauda inferior mostra aumento da média salarial sem migração de retorno, enquanto na metade superior as diferenças entre as distribuições são quase sempre negativas, mesmo abaixo da média do salário-hora, indicando piora salarial.

Uma característica interessante desse processo é que migrantes vivendo em formalidade, embora com nível escolar baixo e, portanto, rendimentos mais baixos na distribuição são beneficiados, caso não houvesse migração de retorno, de forma mais contundente quando levamos em conta suas características inatas.

6.1.2 Migração Nordeste–São Paulo

Ao longo do trabalho foi observada a predominância da população nordestina como principal fonte de emissão, população permanente, grupo de alta probabilidade e com níveis de escolaridade mais baixa. Nesse sentido é interessante observar as distribuições considerando apenas a população dos estados do nordeste. Portanto, no cenário real temos migrantes nordestinos vivendo em São Paulo. No cenário em que nordestinos não retornam, podemos observar



(a) Amostra de Formais (1)

(b) Diferença entre as densidades (1)

(c) Amostra de Informais (2)

(d) Diferença entre as densidades (2)

Nota 1: O teste *Kolmogorov–Smirnov* rejeitou-se a hipótese nula de que as densidades reais e contrafatuais sejam iguais, ao nível de 1% de significância: estatística $D = 0{,}09$ e p-valor $= 0{,}0000$.

Nota 2: As predições de salários foram baseadas nos parâmetros estimados para o modelo bivariado do tipo cópula *Joe* rotacionada a 180°.

Fonte: Elaboração própria com base nos dados do Censo Demográfico de 2010.

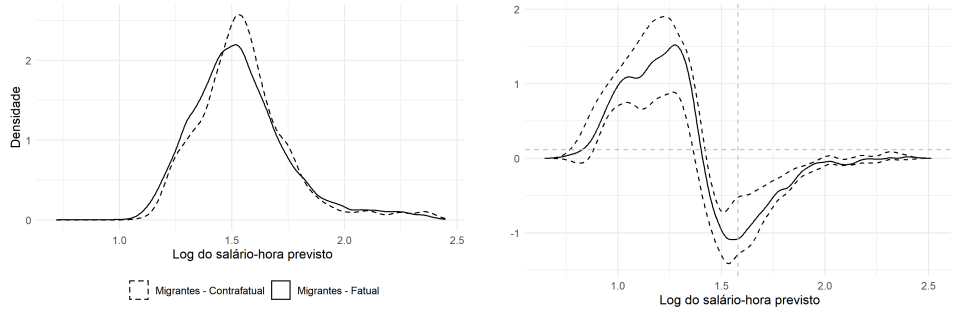
Figura 5. Densidade reais e contrafatuais do salário-hora total para migrantes – (1) Setor Formal, (2) Setor Informal

que as distribuição são mais próximas que aquelas observadas na amostra total, indicando que a população nordestina tenha características observáveis e não observáveis menos divergentes. No entanto, permanece a análise de um ligeiro deslocamento à direita na metade inferior e topo da distribuição. Na parte superior, daqueles decis mais altos, não há afastamento significativo.

6.1.3 Escolaridade

Foi possível verificar os resultados para os níveis de escolaridade: Sem Instrução, Fundamental e Médio, não existem indivíduos com nível superior ou pós-graduação no GAP, o que impossibilita a construção do cenário contrafactual nesses casos.

Na Figura 7 temos as densidades reais e contrafatuais nesses níveis educacionais. Para o nível sem instrução, 7(a) e 7(b), há uma grande aproximação



(a) Distribuição e real e contrafactual **(b)** Diferença entre as densidades

Nota 1: O teste *Kolmogorov–Smirnov* rejeitou-se a hipótese nula de que as densidades reais e contrafatuais sejam iguais, ao nível de 1% de significância: estatística $D = 0{,}07$ e p-valor $= 0{,}0000$.

Nota 2: As predições de salários foram baseadas nos parâmetros estimados para o modelo bivariado do tipo cópula *Joe* rotacionada a 180°.

Fonte: Elaboração própria com base nos dados do Censo Demográfico de 2010.

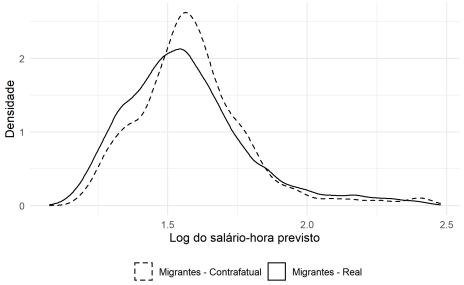
Figura 6. Densidade reais e contrafatuais do salário-hora total para migrantes – Região Nordeste

do resultado para amostra total, devido a grande proporção de indivíduos com esse nível de escolaridade. As densidades 7(b) e 7(c), para o nível fundamental, indicam um resultado semelhante, embora menos significativo, há um ligeiro deslocamento à direita da na parte esquerda e topo da distribuição contrafatual. As densidades referentes ao ensino médio indicam um deslocamento à esquerda na meta inferior da distribuição, bem como menos migrantes na metade superior, mostrando que as diferenças entre as densidades para valores mais altos de rendimento são fortemente negativas. Destarte, para o nível médio o retorno representa uma piora salarial.

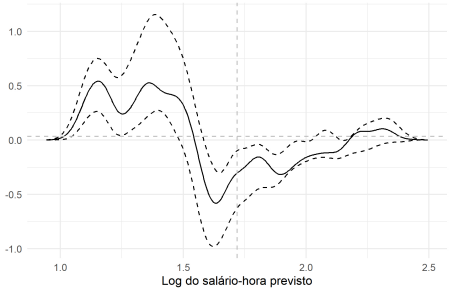
7. Análise de robustez

7.1 Modelo sem restrição de exclusão e validade do instrumento

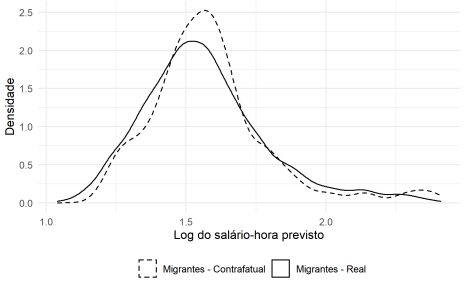
Muitas aplicações empíricas apoiam-se no uso de restrições de exclusões com o objetivo de auxiliar na correta identificação do modelo com viés de seleção amostral (Toomet & Henningsen, 2008). No modelo de Heckman (1979), por exemplo, além da premissa restritiva de normalidade conjunta das equações de seleção e salários, a correlação dos seus termos randômicos requer o uso de restrição de exclusão (variável instrumental) para garantir a correta identificação dos parâmetros estimados. O modelo semi-paramétrico determinação conjunta de remigração e salários tem sua distribuição conjunta baseada em cópulas (vide equações (1) e (2)). Isso assegura maior flexibilidade ao relaxar a hipótese



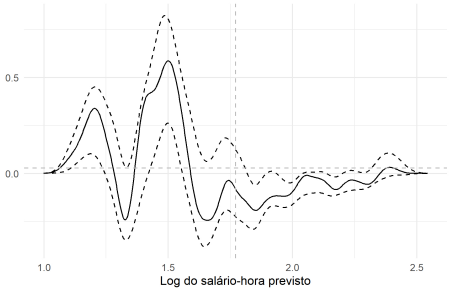
(a) Sem Instrução



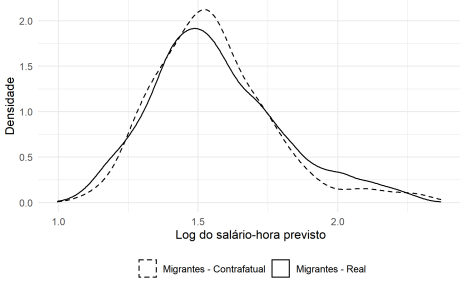
(b) Diferença das densidades – Sem Instrução



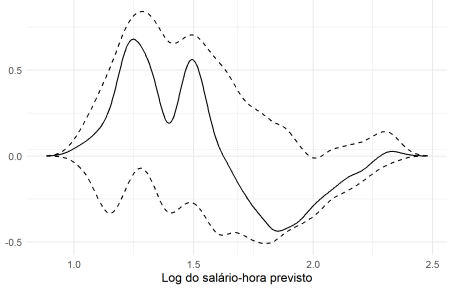
(c) Nível Fundamental



(d) Diferença das densidades – Nível Fundamental



(e) Nível Médio



(f) Diferença das densidades – Nível Médio

Nota 1: A distribuição real foi construída a partir da amostra de migrantes residentes em SP condicionada aos três níveis de instrução supracitados. A distribuição contrafatual corresponde a distribuição dos salários totais no cenário em que não há migração de retorno condicionada aos mesmos níveis educacionais, calculada pela mudança dos resíduos no GAP pela média dos salários previstos dos migrantes caso todos tivessem permanecido.

Nota 2: O teste K-S indicou diferença entre a distribuição real e contrafatual significativa a 1% apenas no nível sem instrução com estatística $D = 0{,}4669$ e p-valor $= 0{,}0000$. No nível fundamental: p-valor $= 0{,}1547$. Nível médio: p-valor $= 0{,}4644$.

Nota 3: As predições de salários foram baseadas nos parâmetros estimados para o modelo bivariado do tipo cópula *Joe* rotacionada a 180°.

Fonte: Elaboração própria com base nos dados do Censo Demográfico de 2010.

Figura 7. Densidades reais e contrafatuais do salário-hora total para migrantes e paulistas por nível educacional

de distribuição normal dos resíduos e não necessariamente precisamos de uma identificação de parâmetros por meio de restrição de exclusão (Wiesenfarth & Kneib, 2010). Todavia, podemos obter estimativas mais consistentes quando ao menos um regressor a mais é incluído na equação de seleção (Marra & Radice, 2013; Wojtyś et al., 2016).

A variável filho nascido em São Paulo foi usada neste estudo entrando na equação de seleção e excluída da equação de salários (instrumento de exclusão), sob a premissa de que a convivência com filhos na região de destino cria um vínculo familiar que pode dificultar a migração de retorno (Biavaschi, 2016). A Figura 8(e) mostra as densidades reais e contrafatuais para um modelo[15] de seleção amostral semelhante ao empregado na análise principal que não usa essa variável como restrição de exclusão.[16] Verifica-se que que os resultados são consistentes com a análise principal. Ademais, para verificar a validade desse instrumento, no entanto, tendo em vista que os pressupostos requeridos ao usar essa estratégia de identificação, foram aplicados os testes propostos por Huber e Mellace (2014) e os métodos de Chen e Szroeter (2014) e Bennett (2009) para testar estatisticamente restrições de desigualdade.

Huber e Mellace (2014) propõem testar a validade do instrumento e a separabilidade aditiva do termo de erro na equação de seleção conjuntamente.[17] A intuição do teste é que duas restrições de desigualdade surgem da identificação pontual e delimitação da distribuição dos indivíduos sempre selecionados (aqueles selecionados independente do instrumento – *always takers*) na subpopulação selecionada que não recebe o instrumento, considerando que a população que recebe o instrumento inclui *compliers*, cuja seleção reage ao instrumento. O ponto identificado na ausência do instrumento deve estar dentro dos limites na presença dele, gerando duas restrições testáveis:

$$H_0 : \begin{pmatrix} E\left(Y|Z=1,S=1,Y\leq y_q\right)-E(Y|Z=0,S=1) \\ E(Y|Z=0,S=1)-E\left(Y|Z=1,S=1,Y\geq y_{1-q}\right)\end{pmatrix} \equiv \begin{pmatrix}\theta_1^m \\ \theta_2^m\end{pmatrix} \leq \begin{pmatrix}0\\0\end{pmatrix}, \quad (11)$$

onde q é a proporção de migrantes sempre selecionados na população mista de indivíduos com e sem filho nascido em São Paulo e y_q é o quantil da variável de

[15] Na Tabela 9 do Apêndice estão dispostos os valores do critério de informação AIC para o modelo sem restrição de exclusão. A cópula *Joe* rotacionada a 180° obteve o menor valor AIC.

[16] O mesmo vetor de variáveis foi usado nas duas equações (1) e (2).

[17] A restrição de exclusão é independente de $u_i^*, \epsilon_i|X$. Em termos de monotonicidade isso implica que o estado potencial de seleção de cada indivíduo aumenta ou diminui fracamente monotonicamente com o instrumento. Dessa forma, denotando por $S(z)$ o estado potencial de seleção do instrumento, temos que $\Pr(S(1) \geq S(0)|X) = 1$ (monotonicidade positiva) e $\Pr(S(1) \leq S(0)|X) = 1$ (monotonicidade negativa) (Huber & Mellace, 2014).

resultado (logaritmo do salário-hora) condicionado à probabilidade q. Para testar essas restrições conjuntamente Huber e Mellace (2014) utilizam os métodos de Chen e Szroeter (2014). Foram também utilizados os métodos de Bennett (2009) baseados em reamostragem por *bootstrap* para obtenção de p-valores robustos em amostras finitas.[18]

Os resultados dos testes das restrições em 11 estão dispostos na Tabela 10, no Apêndice. São reportados os resultados para várias subamostras condicionadas à diferentes subconjuntos das covariadas utilizadas neste estudo. A terceira coluna contém a porcentagem de *compliers*[19] em cada subamostra, a quarta coluna contém a distância padronizada, que indica o grau de violação das restrições de desigualdade em 11 (Huber & Mellace, 2014). A partir da quinta coluna são dispostos os p-valores obtidos a partir dos métodos de Chen e Szroeter (2014) e Bennett (2009) com recentramento completo e parcial, respectivamente.

A partir dos resultados[20] verificamos que a distância padronizada é negativa em grande parte das subamostras. Ademais, todas as estatísticas de p-valor são consideravelmente maiores que 10%. Assim posto, no geral, não há evidências para rejeitar a validade do instrumento utilizado neste estudo, especialmente nos casos em que o poder do teste é maior pela menor presença de *compliers*.

7.2 Análise incluindo mulheres

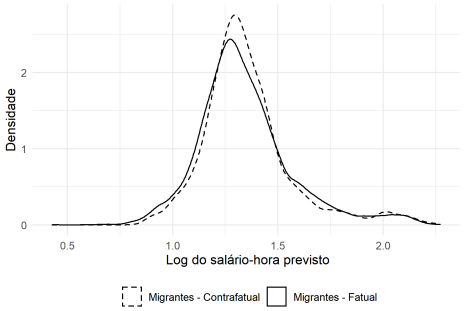
As mulheres representam 36,36% da amostra total caracterizada pelos recortes adotados na seção 3. Ao incluir esse grupo na amostra, no entanto, dois problemas podem surgir na análise. Primeiro, em muitos estudos sobre migração foi mostrado que as mulheres têm maiores chances de serem migrantes agregados (Aguiar, 2017; Avelino, 2010; K. F. d. Oliveira & Jannuzzi, 2004; Siqueira, 2006). Segundo, a restrição de exclusão (residente com filho nascido em São Paulo) no modelo de seleção tem relação com a produtividade da mulher no mercado de trabalho. Há evidências de que a maternidade cria uma série de barreiras para a participação da mulher no mercado de trabalho, principalmente no curto prazo. Apesar disso, a participação das mulheres vem aumentando, enquanto as desigualdades e a problemática da jornada dupla persistem (Costa, 2007; Fontoura & Gonzalez, 2009; Pazello & Fernandes, 2004; Queiroz & Aragón, 2015).

Assim posto, ao longo deste estudo a análise empírica foi realizada a partir de uma amostra composta apenas por trabalhadores homens. Essa estratégia foi

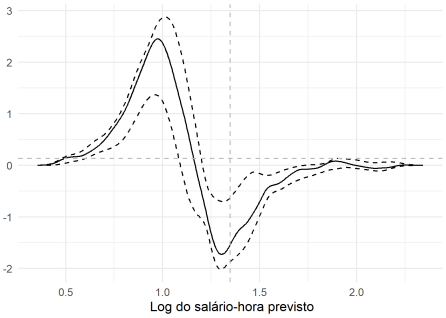
[18] Detalhes sobre algoritmos em Chen e Szroeter (2014) e Bennett (2009).

[19] Huber e Mellace (2014) destacam que o poder do teste diminui quanto maior o percentual desse grupo.

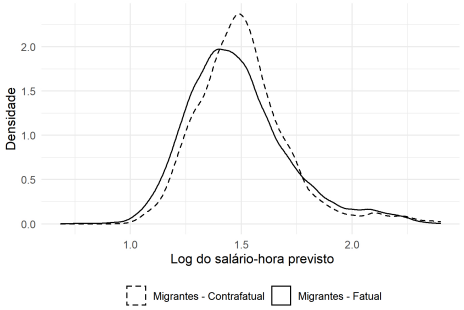
[20] O valor p associado é 1, portanto, não rejeitamos a hipótese nula de validade e monotonicidade nos níveis de significância convencionais.



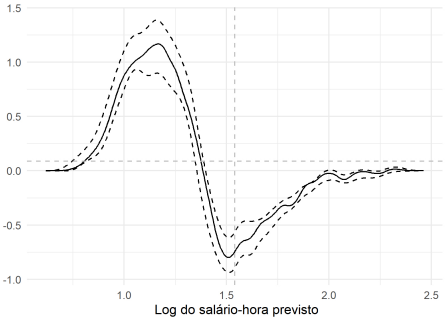
(a) Amostra de mulheres (1)



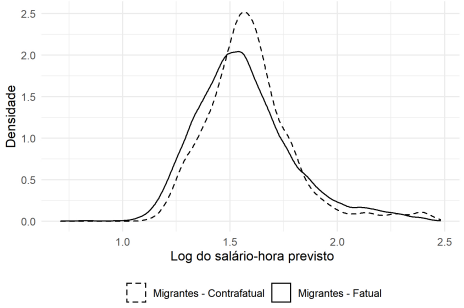
(b) Diferença entre as densidades (1)



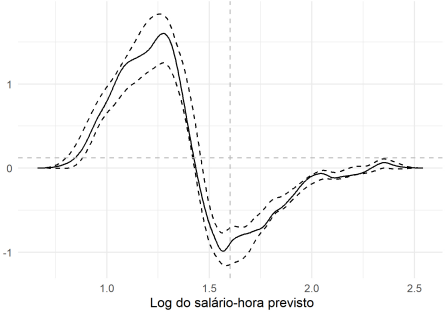
(c) Modelo incluindo homens e mulheres (2)



(d) Diferença entre as densidades (2)



(e) Modelo sem restrição de exclusão (3)



(f) Diferença entre as densidades (3)

Nota 1: (2)(a) Amostra incluindo homens e mulheres.

Nota 2: (1)(b) e (2)(d) Intervalos de confiança representado pelas linhas pontilhadas, supondo um nível de 95% de confiança. A estimação dos erros-padrão foi feita usando a técnica de reamostragem *bootstrap*.

Nota 3: O teste *Kolmogorov–Smirnov* rejeitou-se a hipótese nula de que as densidades reais e contrafatuais sejam iguais, ao nível de 1% de significância em todos os modelos conforme os seguintes dados: (1) p-valor $= 0{,}0000$; (2) p-valor $= 0{,}0000$; (3) p-valor $= 0{,}0000$.

Nota 4: As predições de salários foram baseadas nos parâmetros estimados para o modelo bivariado do tipo cópula *Joe* rotacionada a 180° em todos os modelos.

Fonte: Elaboração própria com base nos dados do Censo Demográfico de 2010.

Figura 8. Densidade reais e contrafatuais do salário-hora total para migrantes – (1) Amostra de mulheres; (2) Modelo incluindo homens e mulheres; (3) Modelo sem restrição de exclusão

implementada afim de possibilitar maior confiança na exogeneidade da variável de restrição de exclusão em relação à equação de resultados (rendimentos), isto é, ao reduzir risco de correlação entre cuidado com os filhos e produtividade do trabalho. No entanto, pode haver uma preocupação de que uma amostra que inclua as mulheres altere os resultados. Para verificar se os resultados são consistentes na presença de mulheres foram estimados dois modelos: o primeiro modelo considera apenas mulheres, enquanto o segundo inclui ambos homens e mulheres.

Considerando somente as mulheres, a Figura 8(a) relata as densidades reais e contrafatuais do logaritmo do salário-hora. Verifica-se que os resultados são muito próximos aos da análise principal, embora mais discretos. Note-se que há um deslocamento à direita da distribuição de salários sem migração de retorno, principalmente no topo e metade e inferior, mais uma vez sugerindo que migrantes com rendimentos mais baixos e médios estariam auferindo maiores salários caso não houvesse migração de retorno. A Figura 8(c) representa amostra em que são considerados homens e mulheres conjuntamente. Esses resultados são, mais uma vez, consistentes com as conclusões da amostra principal, o deslocamento dessa vez se destaca de forma mais perceptível na parte inferior e topo da distribuição.

8. Considerações finais

Conforme resultado teórico de Borjas e Bratsberg (1996), os efeitos do fenômeno da migração de retorno podem se rebater na população migrante que não optou pela remigração, uma vez que o fluxo de migração inicial tenha se autosselecionado em atributos produtivos não observados. Tal questão permanecia pouco explorada na literatura nacional.

Este estudo procurou preencher a lacuna reportada ao investigar empiricamente qual seriam os efeitos da migração de retorno sobre a distribuição de salários dos migrantes interestaduais que optaram pela permanência no estado de São Paulo. Tal estado, além de representar a maior economia dentre as demais unidades federativas do Brasil, tem recebido historicamente o maior contingente de migrantes, bem como se diferenciado como origem principal dos remigrados. Além de características mensuráveis como escolaridade, idade, entre outras, foram também consideradas características inatas na construção dessas distribuições, de forma a verificar se o impacto da migração de retorno sobre a população migrante pode ser explicado por questões intrínsecas ao trabalhador.

Os achados deste estudo revelam que o fluxo de migração interestadual dirigido ao maior polo econômico do Brasil parece ser formado por trabalhadores negativamente selecionados, não apenas em atributos produtivos observáveis

(baixa instrução), mas também em características produtivas não observadas. O último fato é reforçado pelas evidências de que os migrantes retornados têm características produtivas não observadas melhores quando comparadas aos migrantes permanentes, as quais, em geral, aumentariam o salário médio da população migrante e reduziria a desigualdade caso não houvesse migração de retorno.

Referências bibliográficas

Aguiar, M. A. S. d. (2017). *Autosseleção e impacto da migração de retorno sobre a distribuição salarial: Análise para os migrantes da Região Nordeste* [Dissertação de Mestrado]. Fortaleza, CE.

Ambrosini, J. W., Mayr, K., Peri, G., & Radu, D. (2010). *The selection of migrants and returnees: Evidence from Romania and implications*. World Bank.

Andrews, D. W. K., & Schafgans, M. M. A. (1998). Semiparametric estimation of the intercept of a sample selection model. *The Review of Economic Studies*, *65*(3), 497–517.

Assis, R. S. d., Costa, E. M., & Mariano, J. L. (2012, dezembro). Impacto da migração de não naturais e da migração de retorno sobre a distribuição de renda dos estados da Bahia e de São Paulo: Um olhar sobre a inserção desses indivíduos no mercado de trabalho local. *In 40º Encontro Nacional de Economia da ANPEC*, Porto de Galinhas, PE.

Avelino, R. R. G. (2010). Self-selection and the impact of migration on earnings. *Brazilian Review of Econometrics*, *30*(1), 69–89.

Baeninger, R. (2008, setembro). Rotatividade migratória: Um novo olhar para as migrações no século XXI. *In XVI Encontro Nacional de Estudos Populacionais, da Associação Brasileira de Estudos Populacionais, ABEP*, Caxambu, MG.

Batista, N. N. F., & Cacciamali, M. C. (2009). Diferencial de salários entre homens e mulheres segundo a condição de migração. *Revista Brasileira de Estudos de População*, *26*(1), 97–115.

Bennett, C. J. (2009). *Consistent and asymptotically unbiased minP tests of multiple inequality moment restrictions* (Working Paper Nº 09-W08). Nashville, TN: Vanderbilt University Department of Economics.

Biavaschi, C. (2016). Recovering the counterfactual wage distribution with selective return migration. *Labour Economics*, *38*, 59–80.

Borjas, G. J., & Bratsberg, B. (1996). *Who leaves? The outmigration of the foreign-born* (Working Paper Nº 4913). National Bureau of Economic Research (NBER).

Brechmann, E. C., & Schepsmeier, U. (2013). Modeling dependence with C- and D-Vine Copulas: The R package CDVine. *Journal of Statistical Software*, *52*(3), 1–27.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press.

Cattaneo, C. (2007, fevereiro). *The self-selection in the migration process: What can we learn?* (LIUC Paper Nº 199). Castellanza, VA: Università Carlo Cattaneo.

Chamberlain, G. (1986). Asymptotic efficiency in semi-parametric models with censoring. *Journal of Econometrics*, *32*(2), 189–218.

Chen, L.-Y., & Szroeter, J. (2014). Testing multiple inequality hypotheses: A smoothed indicator approach. *Journal of Econometrics*, *178*(Part 3), 678–693.

Chiquiar, D., & Hanson, G. H. (2005). International migration, self-selection, and the distribution of wages: Evidence from Mexico and the United States. *Journal of Political Economy*, *113*(2), 239–281.

Chiswick, B. (1999). Are immigrants favorably self-selected? *American Economic Review*, *89*(2), 181–185.

Chzhen, Y., & Mumford, K. (2011). Gender gaps across the earnings distribution for full-time employees in Britain: Allowing for sample selection. *Labour Economics*, *18*(6), 837–844.

Cohen, Y., & Haberfeld, Y. (2001). Self-selection and return migration: Israeli-born Jews returning home from the United States during the 1980s. *Population Studies*, *55*(1), 79–91.

Costa, J. S. M. (2007). *Determinantes da participação feminina no mercado de trabalho brasileiro* (Tese de Mestrado, Universidade de Brasília, Brasília).

De Coulon, A., & Piracha, M. (2005). Self-selection and the performance of return migrants: The source country perspective. *Journal of Population Economics*, *18*(4), 779–807.

DiNardo, J., Fortin, N. M., & Lemieux, T. (1995, abril). *Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach* (Working Paper Nº 5093). National Bureau of Economic Research (NBER).

Dustmann, C. (2003). Children and return migration. *Journal of Population Economics*, *16*(4), 815–830.

Fontoura, N. d. O., & Gonzalez, R. (2009). Aumento da participação de mulheres no mercado de trabalho: Mudança ou reprodução da desigualdade? *Mercado de Trabalho*, *41*(Nota Técnica – Ipea), 21–26.

Freguglia, R. d. S., & Procópio, T. S. (2013). Efeitos da mudança de emprego e da migração interestadual sobre os salários no Brasil formal: Evidências a partir de dados em painel. *Pesquisa e Planejamento Econômico*, *43*(2), 255–278.

Gama, L. C. D., & Machado, A. F. (2014). Migração e rendimentos no Brasil: Análise dos fatores associados no período intercensitário 2000–2010. *Estudos Avançados*, *28*(81), 155–174.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *1*(47), 153–161.

Heckman, J. J. (1990). Varieties of selection bias. *The American Economic Review*, *80*(2), 313–318.

Huber, M., & Mellace, G. (2014). Testing exclusion restrictions and additive separability in sample selection models. *Empirical Economics*, *47*(1), 75–92.

IBGE – Instituto Brasileiro de Geografia e Estatística. (2012). *Censo Demográfico 2010: Resultados gerais da amostra*. Rio de Janeiro: IBGE.

Justo, W. R., & Ferreira, R. d. A. (2012, novembro). Migração interestadual no Brasil – perfil do retornado: Evidências para o período de 1998–2008. *In XVIII Encontro Nacional de Estudos Populacionais*.

Lima, A. C. d. C., Simões, R., & Hermeto, A. M. (2015). Privação relativa e deslocamentos da mão de obra no Brasil entre 1980 e 2010: Evolução das interações entre pobreza, desigualdade de renda e migração. *Pesquisa e Planejamento Econômico*, *45*(1), 7–36.

Liu, E., Hsiao, C., Matsumoto, T., & Chou, S. (2009). Maternal full-time employment and overweight children: Parametric, semi-parametric, and non-parametric assessment. *Journal of Econometrics*, *152*(1), 61–69.

Marra, G., & Radice, R. (2010). Penalised regression splines: Theory and application to medical research. *Statistical Methods in Medical Research*, *19*(2), 107–125.

Marra, G., & Radice, R. (2013). Estimation of a regression spline sample selection model. *Computational Statistics & Data Analysis*, *61*, 158–173.

Marra, G., & Wyszynski, K. (2016). Semi-parametric copula sample selection models for count responses. *Computational Statistics & Data Analysis*, *104*, 110–129.

McKenzie, D., & Rapoport, H. (2010). Self-selection patterns in Mexico–US migration: The role of migration networks. *The Review of Economics and Statistics*, *92*(4), 811–821.

Mincer, J. (1978). Family migration decisions. *Journal of Political Economy*, *86*(5), 749–773.

Mulligan, C. B., & Rubinstein, Y. (2008). Selection, investment, and women's relative wages over time. *The Quarterly Journal of Economics*, *123*(3), 1061–1110.

Oliveira, A. T. R. d., Ervatti, L. R., & O'Neill, M. M. V. C. (2011). O panorama dos deslocamentos populacionais no Brasil: PNADs e censos demográficos. *In* L. A. P. de Oliveira & A. T. R. de Oliveira (Orgs.), *Reflexões sobre os deslocamentos populacionais no Brasil* (Vol. 1). Rio de Janeiro: IBGE.

Oliveira, K. F. d., & Jannuzzi, P. d. M. (2004, setembro). Motivos para migração no Brasil: Padrões etários, por sexo e origem/destino. *In XIV Encontro Nacional de Estudos Populacionais, ABEP*, Caxambu, MG.

Pazello, E. T., & Fernandes, R. (2004, dezembro). A maternidade e a mulher no mercado de trabalho: Diferença de comportamento entre mulheres que têm e mulheres que não têm filhos. *In 32º Encontro da Associação Nacional de Pós-Graduação em Economia (ANPEC)*, João Pessoa, PB.

Queiroz, V. d. S., & Aragón, J. A. O. (2015). Alocação de tempo em trabalho pelas mulheres brasileiras. *Estudos Econômicos*, *45*(4), 787–819.

Ramalho, H. M. d. B., Figueiredo, E., & Netto, J. L. d., Jr. (2016). Determinantes das migrações interestaduais no Brasil: Evidências a partir de um modelo gravitacional. *Pesquisa e Planejamento Econômico*, *46*(1), 67–112.

Ramalho, H. M. d. B., & Queiroz, V. d. S. (2011). Migração interestadual de retorno e autosseleção: Evidências para o Brasil. *Pesquisa e Planejamento Econômico*, *41*(3), 369–396.

Sachsida, A., de Castro, P. F., de Mendonça, M. J. C., & Albuquerque, P. H. (2009). *Perfil do migrante brasileiro* (Texto para Discussão Nº 1410). Brasília, DF: Instituto de Pesquisa Econômica Aplicada (IPEA).

Santos, C. A. R. (2006). *Migração e distribuição regional de renda no Brasil* (Dissertação de Mestrado, Fundação Getulio Vargas, Rio de Janeiro, RJ).

Santos Jr, E. d. R. d., Menezes-Filho, N., & Ferreira, P. C. (2005). Migração, seleção e diferenças regionais de renda no Brasil. *Pesquisa e Planejamento Econômico*, *35*(3), 299–331.

Shen, C. (2013). Determinants of health care decisions: Insurance, utilization, and expenditures. *Review of Economics and Statistics*, *95*(1), 142–153.

Silveira Neto, R. d. M., & Magalhães, A. M. (2004). O progresso econômico do migrante em São Paulo: Evidências a partir dos censos demográficos de 1991 e 2000. *Economia e Desenvolvimento*, *5*(2), 245–281.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

Siqueira, L. B. O. d. (2006). *Uma análise do fluxo migratório brasileiro: Migração para regiões pobres e migração de retorno* (Tese de Doutorado, Universidade Federal de Pernambuco, Recife, PE).

Sjaastad, L. A. (1962). The costs and returns of human migration. *Journal of Political Economy*, *70*(5, Part 2), 80–93.

Toomet, O., & Henningsen, A. (2008). Sample selection models in R: Package sampleSelection. *Journal of Statistical Software*, *27*(7), 1–23.

Wiesenfarth, M., & Kneib, T. (2010). Bayesian geoadditive sample selection models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *59*(3), 381–404.

Wojtyś, M., Marra, G., & Radice, R. (2016). Copula regression spline sample selection models: The R package SemiParSampleSel. *Journal of Statistical Software*, *71*(6).

Wood, S. (2015). *Package 'mgcv'* [R package].

Apêndice.

Tabela 7. Descrição das variáveis utilizadas nos modelos de seleção e salários

	Descrição
Variáveis Pessoais	
Idade	Em anos completos
Não Branco	Variável binária: 1 indivíduo que declarou cor ou raça diferente da branca; 0 indivíduo declarou raça ou cor branca
Nível de Instrução	Categórica: define se o indivíduo não possui instrução ou tem o fundamental incompleto, se possui o fundamental completo ou médio incompleto, se possui o médio completo ou superior incompleto, se possui o superior completo ou se possui pós graduação.
Variáveis familiares	
Chefe	Variável binária: 1 posição de responsável pelo domicílio; 0 posição de cônjuge no domicílio
Chefe e vive com cônjuge	Variável Binária: 1 chefe vive com cônjuge; 0 caso contrário
Chefe e vive com cônjuge nascido em SP	Variável binária: 1 chefe vive com cônjuge natural de São Paulo; 0 c.c.
Filho	Variável binária: 1 indivíduo possui filho com idade até 14 anos; 0 c.c.
Filho nascido em SP	Variável binária: 1 indivíduo com filho nascido em SP com idade até 14 anos; 0 c.c.
Total de pessoas no domicílio	Número de pessoas morando no domicílio em 31 julho de 2010
Emprego e Ocupação	
Rendimento	Rendimento no trabalho principal
Salário-Hora	Rendimento do trabalho principal dividido pela média de horas trabalhadas por mês
Residência	
Urbana	Variável binária: 1 reside na zona urbana; 0 c.c.
Norte	Variável Binária: 1 se o indivíduo nasceu na região Norte; 0 c.c.
Nordeste	Variável Binária: 1 se o indivíduo nasceu na região Nordeste; 0 c.c.
Sudeste	Variável Binária: 1 se o indivíduo nasceu na região Sudeste; 0 c.c.
Sul	variável Binária: 1 se o indivíduo nasceu na região Sul; 0 c.c.
Centro-oeste	Variável Binária: 1 se o indivíduo nasceu na região centro-Oeste; 0 c.c.

Fonte: Elaboração própria com base nos dados do Censo Demográfico de 2010.

Tabela 8. Características das famílias de cópulas selecionadas

Cópula	Forma Funcional $C(u,v;\theta)$	Parâmetro de Dependência
Gaussiana	$\Phi_2\left(\Phi^{-1}(u),\Phi^{-1}(v);\theta\right)$	$\theta \in [-1,1]$
Clayton	$\left(u^{-\theta}+v^{-\theta}-1\right)^{-1/\theta}$	$\theta \in (0,\infty)$
Joe	$1-\left[(1-u)^{\theta}+(1-v)^{\theta}-(1-u)^{\theta}(1-v)^{\theta}\right]^{1/\theta}$	$\theta \in (1,\infty)$
Frank	$-\theta^{-1}\log\left[1+\left(e^{-\theta u}-1\right)\left(e^{-\theta v}-1\right)/\left(e^{-\theta}-1\right)\right]$	$\theta \in \mathbb{R}\backslash\{0\}$
Gumbel	$\exp\left\{-\left[(-\log u)^{\theta}+(-\log v)^{\theta}\right]^{1/\theta}\right\}$	$\theta \in [1,\infty)$
FGM	$uv[1+\theta(1-u)(1-v)]$	$\theta \in [-1,1]$
AMH	$uv/[1-\theta(1-u)(1-v)]$	$\theta \in [-1,1]$

Nota: $\Phi_2(\cdot,\cdot;\theta)$ denota a função de distribuição cumulativa de uma distribuição normal bivariada padrão com coeficiente de correlação θ; Φ^{-1} é a inversa da função acumulada normal padrão; u e v são distribuições marginais.

Fonte: Wojtyś et al. (2016).

Tabela 9. Critério de informação AIC: seleção do modelo segundo a função cópula – modelo principal

Cópula	(1) AIC	(2) AIC	(3) AIC	(4) AIC	(5) AIC
Joe 180	51195,86	50901,53	50901,53	50177,72	49667,60
Clayton	51211,43	50918,81	50918,81	50201,46	49692,93
Gumbel 180	51238,02	50953,47	50953,47	50247,69	–
Joe 270	51838,35	51480,04	51480,04	50523,38	50082,30
Clayton 90	51845,53	51487,22	51487,22	50533,24	–
FGM	51935,79	51572,46	51572,46	50642,76	50153,87
Frank	51944,88	51580,55	51580,55	50649,57	50167,79
AMH	52026,47	51660,86	51660,86	50728,96	50228,31
Gumbel 270	52075,39	51707,49	51707,49	50749,00	–
Joe	52085,35	51717,44	51717,44	50780,46	50258,41
Clayton 180	52090,99	51723,64	51723,64	50788,46	–
Gumbel	52107,66	51740,18	51740,18	50805,63	–
Normal	52115,21	51748,08	51748,08	50811,05	49916,25
Clayton 270	52115,22	51748,08	51748,08	50812,56	–
Gumbel 90	52115,22	51748,08	51748,08	50812,57	–
Joe 90	52115,22	51748,08	51748,08	50812,57	–
Especificação					
Variáveis Pessoais	sim	sim	sim	sim	sim
Variáveis Familiares	não	sim	sim	sim	sim
Variáveis de Residência	não	não	não	sim	sim
Restrição de exclusão	sim	sim	sim	sim	sim
Splines	não	não	não	não	sim

Nota: Amostra de trabalhadores homens entre 18 e 70 anos em atividade remuneratória. Os modelos incluem duas partes: equação de seleção e equação de salários. Variáveis: Filho nascido em SP (restrição de exclusão), filho menor de 14 anos, cônjuge nascido em SP, variável binária para chefe casado, quantitativo do total pessoas no domicílio, e número de pessoas no domicílio ao quadrado. Variáveis de residência: setor do domicílio, variáveis binárias por região de origem do migrante. O modelo (5) utiliza, além dos componentes paramétricos, funções suavizadas (splines) das variáveis idade e número de pessoas no domicílio afim de capturar relações não lineares mais complexas. A tabela está ordenada de acordo com a coluna (5).

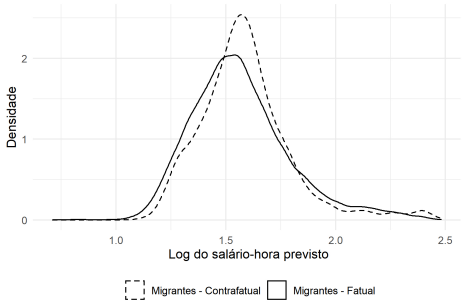
Fonte: Elaboração própria a partir dos microdados do Censo Demográfico de 2010.

Tabela 10. Teste para restrição de exclusão (filho nascido em SP) – Huber e Mellace (2014)

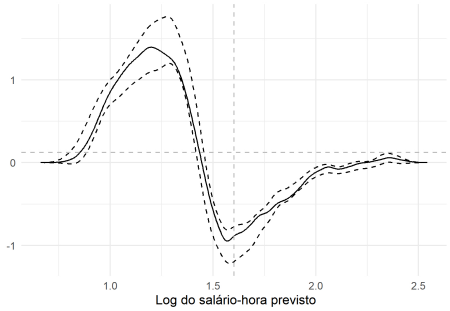
				P-valor		
Controles	Observações	% *compliers*	DP	(1)	(2)	(3)
Amostra Completa	28049	0,0526	0,0016	1,0000	1,0000	1,0000
Sem Instrução	15420	0,0376	−0,0354	1,0000	1,0000	1,0000
Nível Fundamental	5619	0,0575	−0,0565	1,0000	1,0000	0,5970
Nível Médio	6097	0,0675	−0,0178	1,0000	1,0000	1,0000
Nível Superior	869	0,1104	0,0966	0,9970	1,0000	0,6195
Pós-graduação	44	0,1818	–	0,9834	0,9855	0,8960
Idade entre 18 e 29	11659	0,0550	0,0831	0,7307	0,9595	0,5195
Idade entre 30 e 49	13931	0,0568	−0,0205	1,0000	1,0000	1,0000
Idade entre 50 e 70	2459	0,0795	−0,1335	0,9480	0,9850	0,5005
Jovem entre 25 e 35 sem Instrução	6713	0,0463	0,0113	0,9997	1,0000	0,5395
Região de Origem – NE	17484	0,0243	−0,0579	0,9457	0,9950	0,5105
Região de Origem – SE	5651	0,0798	0,0940	0,9047	0,9840	0,5095
Região de Origem – SUL	3858	0,1476	0,1339	0,9915	1,0000	0,5175
Região de Origem – CO	757	0,1183	0,0713	0,9944	0,9965	0,5605
Região de Origem – NO	299	0,1266	0,3314	0,1803	0,2365	0,1160
Formais	20175	0,0335	0,0031	1,0000	1,0000	0,7085
Informais	7874	0,1236	0,0056	1,0000	1,0000	1,0000

Nota: DP (distância padronizada) das restrições de desigualdade em 11. P valores: (1) baseado no teste de Chen e Szroeter (2014); (2) e (3) – Baseados em Bennett (2009) com recentramento completo e parcial de restrições de desigualdade em 11, respectivamente. Total de 999 reamostragens *bootstrap*.

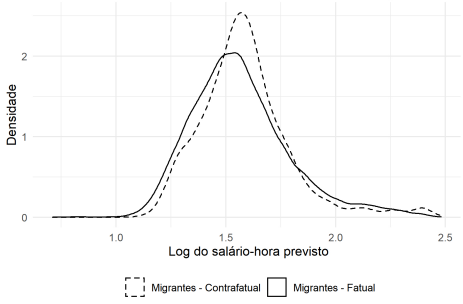
Fonte: Elaboração própria a partir dos microdados do Censo Demográfico de 2010.



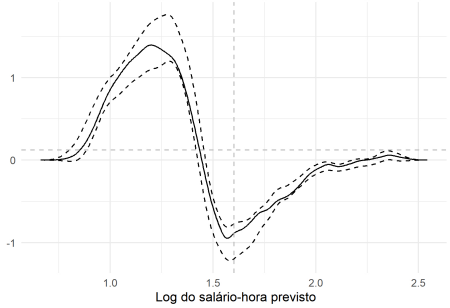
(a) Densidade Real e Contrafatual – Cópula Normal



(b) Diferença das Densidades – Normal



(c) Densidade Real e Contrafatual – Cópula Joe 180



(d) Diferença das Densidades – Joe 180

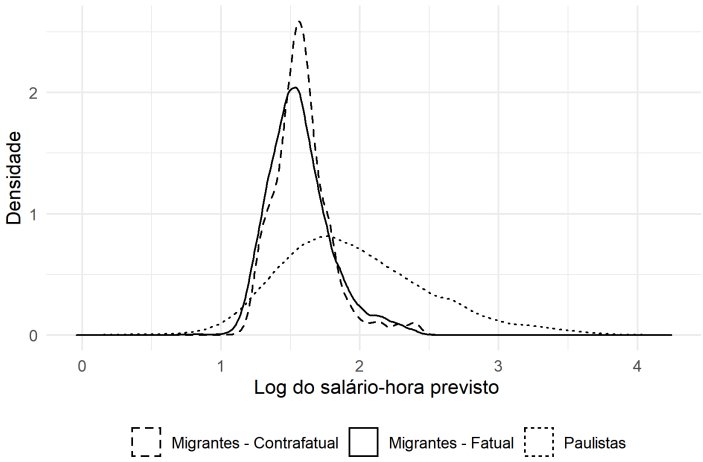
Nota 1: (a) e (b): as predições dos salários foram baseadas na cópula Normal com especificação paramétrica, equivalente ao modelo de Heckman (1979) em dois estágio – Tobit2. (c) e (d): predições de salários baseadas nos parâmetros estimados para o modelo bivariado paramétrico do tipo cópula *Joe* rotacionada a 180°.

Nota 2: Intervalos de confiança representado pelas linhas pontilhadas, supondo um nível de 95% de confiança. A estimação dos erros-padrão foi feita usando a técnica de reamostragem *bootstrap*.

Nota 3: O teste *Kolmogorov–Smirnov* rejeitou-se a hipótese nula de que as densidades reais e contrafatuais sejam iguais a 1% de significância em ambos os modelos.

Fonte: Elaboração própria com base nos dados do Censo Demográfico de 2010.

Figura 9. Densidade real e contrafatual do salário-hora para migrantes – abordagem paramétrica



Nota 1: Os parâmetros das equações salariais para os paulistas foram estimados a partir de modelos aditivos generalizados com estimativa integrada de suavidade, mais detalhes em Wood (2015). As predições de salários de migrantes foram baseadas nos parâmetros estimados pelo modelo bivariado do tipo cópula *Joe* rotacionada a 180°.

Nota 2: Um teste K-S rejeitou a hipótese nula de que as distribuições sejam iguais, a 1% de significância: p-valor $= 0{,}0000$.

Fonte: Elaboração própria a partir dos microdados do Censo Demográfico de 2010.

Figura 10. Densidade real e contrafatual do salário-hora para migrantes e paulistas