

# CONVERGÊNCIA DE HABILIDADES NA INDÚSTRIA: UM ABORDAGEM DE REGIONALIZAÇÃO VIA MACHINE LEARNING

Gustavo Henrique Leite de Castro<sup>1</sup>, Carlos Roberto Azzoni<sup>2</sup> e André Luis Squarize Chagas<sup>3</sup>

**Resumo:** O objetivo deste estudo é avaliar se há evidências de convergência das habilidades nas ocupações formais na indústria de transformação privada no Brasil, utilizando métodos de regionalização via *machine learning*. Para identificar as heterogeneidades espaciais foram utilizados além das técnicas de ML, o método de clusters hierárquicos com restrição de contiguidade e uma análise de quintil. As análises foram estimadas por meio dos métodos de econometria espacial. Os principais resultados mostram que as habilidades motoras convergem mais rápido que as habilidades sociais e cognitivas nas indústrias das regiões brasileiras. Além disso, as regiões menos industrializadas estão convergindo dentro dos seus clusters. Entretanto, as regiões mais industrializadas não estão convergindo dentro dos seus clusters.

**Palavras-chave:** indústria, habilidades, econometria espacial, máquina de aprendizado

## SKILLS CONVERGENCE IN MANUFACTURING INDUSTRY: A REGIONALIZATION APPROACH THROUGH MACHINE LEARNING

**Abstract:** The aim of this study is to assess if there is evidence of skill convergence in formal occupations in the private manufacturing industry in Brazil, using regionalization approaches through machine learning. To identify spatial heterogeneities, in addition to ML techniques, the hierarchical cluster method with contiguity constraint and a quintile analysis were used. The heterogeneities were estimated using spatial econometric approach. The main results indicate that motor skills converge faster than social and cognitive skills in the manufacturing industry of the Brazilian regions. In addition, less industrialized regions are converging within their clusters. However, the more industrialized regions are not converging within their clusters.

**Keywords:** manufacturing industry, skills, spatial econometrics, machine learning

JEL: C33, C38, J24

### 3. Localização e concentração das atividades econômicas

## 1 INTRODUÇÃO

Os métodos e técnicas de análise regional e urbana cresceram de maneira significativa desde a metade do século XX. É possível justificar esse aumento a partir das pesquisas de Isard (1960) e Isard et al., (1998), em que os autores apresentaram uma amostra significativa dos métodos e técnicas de análise regional desenvolvidas com o propósito de cobrir o estado da arte da ciência regional e auxiliar seus possíveis *stakeholders*. Concomitante a esse processo, os métodos econométricos avançaram e ganharam robustez para estimar relações causais e aumentar a acurácia das predições. Nesse interim, de acordo com Almeida (2012), as buscas por modelar as relações espaciais nos modelos econométricos ganharam espaço e os métodos de econometria espacial surgiram como uma forma de estimar essas relações espaciais.

Entretanto, umas das barreiras está em identificar as relações espaciais, as suas estruturas e dimensões quando não há relação teórica *a priori*. Um exemplo é em relação aos clubes de convergência, baseado em modelos que produzem múltiplos equilíbrios. Os diferentes equilíbrios que uma região alcançará dependerá de sua posição inicial ou de algum outro atributo. Um grupo de regiões poderá ter um determinado equilíbrio se eles compartilham a mesma localização inicial ou algum atributo correspondente a esse equilíbrio. Isso produz clubes de convergências. A barreira está em identificar esse atributo na

---

<sup>1</sup> Departamento de economia da Universidade de São Paulo. E-mail: gustavocastro@usp.br

<sup>2</sup> Departamento de economia da Universidade de São Paulo. E-mail: cazoni@usp.br

<sup>3</sup> Departamento de economia da Universidade de São Paulo. E-mail: achagas@usp.br

ausência de uma base teórica, o exercício é ainda mais complexo se houver alguma estrutura espacial envolvida.

Assim, as técnicas de *machine learning* (ML) é uma estratégia para identificar essas relações espaciais e identificar esses efeitos heterogêneos no espaço. Segundo Kopczewska (2022), o aprendizado de máquina usa algoritmos automatizados para descobrir padrões de dados e permitir previsões de alta qualidade, embora as relações entre os dados de entrada não tenham sido estudadas *a priori*. O ML tem três propósitos principais: agrupamento de dados em grupos *a priori* desconhecidos, classificação de dados em grupos conhecidos com base em um modelo treinado e, por fim, a previsão. Os modelos de ML podem ser supervisionados, não supervisionados ou mistos.

O aprendizado não supervisionado, foco deste estudo, é a coleção de métodos de aprendizado de máquina que são equivalentes às estatísticas. Assim como a mineração de dados, ela não estuda relações ou causalidade, mas procura padrões de dados desconhecidos, mas significativos (Kopczewska, 2022). O aprendizado não supervisionado abrange principalmente agrupamento, redução de dimensão e regras de associação de maneira não *ad hoc*. O estudo de VoPham et al. (2018) identificou que cerca de 80% de todos os dados podem ter uma dimensão geográfica, e muitos desses dados podem ser georreferenciados, reforçando a utilização das técnicas de ML na análise espacial. Desta forma, o ML tem um grande potencial na análise de *Big Data* e não *Big Data* em termos dos seus métodos serem usados como complementos à estatística espacial e econometria. Logo, o objetivo desta pesquisa é utilizar métodos de regionalização via *machine learning* no contexto espacial. O contexto espacial proposto refere-se as análises de convergência das habilidades nas ocupações formais na indústria de transformação privada no Brasil.

A justificativa para este foco se deve a evidência do processo de desconcentração regional que a indústria estaria passando no Brasil, como apontado por Silva e Diniz (2020). Porém, pouco se sabe a respeito da dinâmica das habilidades requeridas no próprio exercício da ocupação da indústria de transformação nas regiões brasileiras. As habilidades requeridas nas ocupações da indústria podem variar a depender das características do mercado de trabalho nas regiões onde há ou não a concentração de emprego formal e estabelecimentos da indústria de transformação. As regiões industriais deverão apresentar importantes diferenças institucionais e estruturais em suas escalas econômicas, mercado de trabalho e infraestrutura, que podem impactar de forma heterogênea na análise de convergência das habilidades.

Algumas regiões industriais podem ser beneficiadas, dentre outros fatores, pelas economias de aglomeração, podendo servir como fonte para os retornos crescente de escala via, por exemplo, o *learning by doing* à la Arrow (1971) e Romer (1986) ou via efeito *sharing, matching e learning* de Duranton e Puga (2005). Esses efeitos podem fazer com que a indústria usufrua de retornos crescentes em comparação com a indústria de outras regiões. Por isso, se faz necessário uma análise de heterogeneidade (clubes) de convergência das habilidades da indústria. Entretanto, não se sabe a estrutura e as relações espaciais envolvidas *a priori* para definir essa heterogeneidade e por esse motivo se faz necessário as técnicas de regionalização.

Para atender o objetivo proposto este estudo encontra-se dividido em 5 seções, sendo a primeira está introdução; a segunda seção aborda o *framework* teórico e empíricos sobre as habilidades e a indústria; a terceira seção apresenta a estratégia empírica utilizada, os métodos de agrupamentos e o banco de dados utilizados; a quarta seção se ateve em apresentar os resultados da pesquisa e, por fim, a última seção apresenta as considerações finais.

## 2 REFERENCIAL TEÓRICO

Pela ótica espacial, percebe-se um espraiamento da indústria de transformação nas regiões brasileiras, saindo das regiões tradicionalmente industrializadas para o interior do país. De acordo com o IBGE (2022), em 2006, aproximadamente 41,2% das RGI (Regiões Geográficas Imediatas), cerca de 210 RGI, concentravam 90% dos estabelecimentos formais da indústria de transformação. Em 2019, esse percentual passou para 43,0%, correspondendo a 220 RGI. Pelo lado do emprego, em 2006, aproximadamente 33,3% das RGI, cerca de 170 RGI, concentravam 90% do emprego formal na indústria de transformação. Já em 2019, esse percentual passou para 44,71%, correspondendo a 228 RGI.

Embora a indústria seja concentrada espacialmente em poucas RGI. Os dados apresentados estão de acordo com o estudo de Silva e Diniz (2020) em que os autores constataram que a indústria brasileira enfrenta um processo de desconcentração geográfica pelo índice de aglomeração Ellison-Glaeser de aproximadamente 2,7% ao ano. Entretanto, os autores ressaltam que essa desconcentração varia entre setores e macrorregiões brasileiras. Por exemplo, este processo é mais acelerado nas regiões Norte e Nordeste (regiões com menores níveis absolutos de estabelecimentos indústrias) e menos acelerado nas regiões Sul e Sudeste (regiões com maiores níveis absolutos de estabelecimentos indústrias), acentuando o processo de disparidade regional entre as indústrias.

Uma possível explicação, dentre outras, para a desconcentração das atividades industriais em algumas regiões pode estar relacionada aos níveis de produtividade dessas regiões. De acordo com Schettini e Azzoni (2013), existem evidências de concentração de altos níveis de produtividade nas regiões tradicionalmente mais importante em termos industriais no país (Sudeste e Sul). Por outro lado, aparecem casos de alta produtividade em regiões não tradicionais, provavelmente resultantes da especialização em um ou poucos setores ligados, via de regra, ao setor agropecuário como é o caso do setor de Alimentos no Centro-Oeste e Extremo Oeste Baiano e o setor de Madeira no Pará. Assim, observa-se uma expansão das atividades agrícolas para o Oeste do país. Todavia, o quadro geral da indústria replica as competitividades relativas tradicionais das regiões brasileiras.

Uma das explicações referentes as diferenças de produtividade das indústrias nas regiões se deve, entre outros fatores, as diferenças de capital humano entre elas. Assim, a educação, determinada pelo nível de qualificação dos trabalhadores, surge como alternativa para a redução das disparidades industriais e ao fortalecimento da competitividade das economias regionais, influenciando ganhos à população, elevando a produtividade do capital humano e também o nível de produtividade do capital físico na indústria (Castro e Azzoni, 2022).

Entretanto, os estudos geralmente destacam a qualificação da população regional, em termos de níveis educacionais. Essa é uma variável relevante para indicar o potencial produtivo. Porém, existem outras variáveis tão relevantes quanto a qualificação e que são pouco exploradas, como o caso do nível de habilidades demandada pelas firmas, como apontado em Neves, Azzoni e Chagas (2021). Logo, processos produtivos mais complexos envolvem a contratação de pessoas para ocupações que requerem maiores habilidades dos trabalhadores. Assim, determinada região pode ter um conjunto grande de pessoas com alto nível de educação, mas as indústrias ali localizadas demandam trabalhadores pouco qualificados, não aproveitando os recursos existentes. Um aspecto relevante do desenvolvimento da indústria na região, portanto, baseia-se na complexidade das atividades que sedia, complexidade essa que se manifesta nos requerimentos de habilidades dos seus trabalhadores (Castro e Azzoni, 2022).

Argumento semelhante é apontado por Dawid et al. (2013), em que os autores evidenciam que a produção é caracterizada por uma complementaridade entre o estoque de capital e as habilidades específicas dos trabalhadores que utilizam o estoque de capital. Assim, as mudanças de produtividade em uma região são influenciadas tanto pelo investimento de empresas locais em bens de capital quanto pela evolução da distribuição de habilidades específicas dos trabalhadores empregados na região. Logo, a mesma analogia pode ser feita para a indústria, ou seja, o crescimento da indústria depende da sua produtividade que depende, dentre outros fatores, das habilidades dos trabalhadores ocupados nas indústrias dessas regiões<sup>4</sup>.

Portanto, se as evidências indicam para uma desconcentração da indústria no Brasil e reforçam o papel da produtividade via qualificação da mão de obra para o avanço industrial nas regiões brasileiras, porém pouco se explora a respeito da demanda de habilidades da indústria. Ou seja, a indústria pode estar avançando para outras regiões, as regiões podem estar convergindo em termos do número de emprego e estabelecimento. Porém, o nível de habilidades demandadas pela indústria pode não estar seguindo essa

---

<sup>4</sup> A mesma analogia pode ser feita para a inovação industrial. Há ampla evidência empírica de que, em muitos casos, a difusão de inovações requer habilidades adequadas da força de trabalho das empresas que adotam a inovação. (Ver, por exemplo, Bassanini e Scarpetta 2002; Griffith et al. 2004).

dinâmica e mais, podem estar acentuando as disparidades regionais e, conseqüentemente, o prêmio salarial<sup>5</sup> da indústria. Logo, é possível que o emprego e o número de indústrias podem estar crescendo em algumas regiões, porém com ocupações que requerem baixos níveis de habilidades e, conseqüentemente, salários. Esse é o ponto que será explorado nesta pesquisa.

### 3 ESTRATÉGIA EMPÍRICA

A estratégia empírica a ser empregada neste estudo parte da hipótese de convergência da renda per capita entre as economias, testada por meio do estudo pioneiro de Baumol (1986). Após este estudo, diversas outras pesquisas surgiram testando a abordagem do  $\beta$ -convergência (Barro 1991; Barro; Sala-I-Martin, 1991, 1992). A hipótese do  $\beta$ -convergência sugere que o crescimento econômico de um país (ou região) depende do seu nível inicial. Logo, se coeficiente de convergência for significativo então os países (ou regiões) menos desenvolvidas crescem mais do que países (ou regiões) mais desenvolvidas. A mesma abordagem será aplicada, porém com o foco nas habilidades das ocupações na indústria de transformação privada.

#### 3.1 Estratégia de identificação

O  $\beta$  convergência condicional infere o tempo de convergência das regiões e como isso acontece (Arbia, 2006). Reformulando em um contexto de habilidades da indústria em dados em painel, o modelo é dado por:

$$\log\left(\frac{s_{i,t}}{s_{i,t-1}}\right) = \alpha_i - \beta \log(s_{i,t-1}) + \mu_i + \varphi_i + u_{i,t} \quad (1)$$

Em que  $i$  indica a região (unidade espacial), com  $i = 1, 2, \dots, N$  e  $t$  é o período de tempo,  $t = 1, 2, \dots, T$ . O vetor  $s_{i,t}$  contém as observações sobre o nível médio de habilidades da indústria de transformação para a região  $i$  no período  $t$ ;  $\alpha_i$  é o termo constante,  $\mu_i$  e  $\varphi_i$  são, respectivamente, o efeito fixo específico que considera as características individuais não observáveis e invariantes no tempo; e o efeito fixo de tempo que são choques comuns às indústrias ao longo do tempo, ambos os efeitos podem contribuir para a variabilidade da variável dependente e  $u_{i,t}$  é o termo de erro independente e identicamente distribuído para  $i$  e  $t$  com média zero e variância constante.

Em alguns casos, os autores assumem que o termo de erro tem média zero e a mesma variância para todas as observações ( $E(uu') = \sigma^2 I$ ). Essa suposição subjacente é particularmente crucial e restritiva se as observações forem organizadas espacialmente. De fato, se existe *spillovers* espaciais entre as regiões, ocasionado pela presença da autocorrelação espacial, a suposição formulada seria violada. Segundo LeSage e Pace (2009), ao investigar as unidades espaciais, há grande probabilidade da amostra apresentar autocorrelação espacial. Ao se tratar de atividade econômicas como a indústria que usualmente se concentra espacialmente, exige a incorporação desse efeito no modelo no sentido de evitar problemas advindos de omissão de variável relevante (Arbia, 2006).

Para saber qual especificação é a mais apropriada para a convergência condicional para um determinado processo de geração de dados, considerou-se os a estratégia descrita em LeSage e Pace (2009) e Elhorst (2010), partindo de uma especificação geral para modelos de painel espacial:

$$\log\left(\frac{s_{i,t}}{s_{i,t-1}}\right) = \beta \log(s_{i,t-1}) + \rho \sum_{i=1}^n w_{ij} \log\left(\frac{s_{i,t}}{s_{i,t-1}}\right) + \theta \sum_{i=1}^n w_{ij} \log(s_{i,t-1}) + \mu_i + \varphi_t + u_{i,t} \quad (2)$$

com

$$u_{i,t} = \lambda \sum_{i=1}^n w_{ij} u_{i,t} + \varepsilon_{i,t} \quad i = 1, \dots, N \quad t = 1, \dots, T \quad (3)$$

e

<sup>5</sup> Ver: Neves, Azzoni e Chagas (2021).

$$\varepsilon_{i,t} \sim i. i. d. (0, \sigma_\varepsilon^2) \quad (4)$$

Em que  $s_{i,t}$  contém as observações sobre o nível médio de habilidades da indústria de transformação privada para a região  $i$  no período  $t$ ;  $\rho$  é um parâmetro escalar espacial. Logo,  $\rho$  captura a sensibilidade da variável endógena para a variável defasada espacialmente (taxa de crescimento do nível médio de habilidades da indústria de transformação privada dos vizinhos);  $\theta$  representa as variáveis exógenas defasadas espacialmente (nível médio de habilidades da indústria de transformação privada dos vizinhos) e  $\lambda$  é o parâmetro de erro espacial, ou seja, é um coeficiente escalar que indica a intensidade da autocorrelação espacial entre os resíduos da equação observada. Os demais parâmetros seguem a especificação da equação (1). Se  $\lambda = 0$ , tem-se o modelo Spatial Durbin Model (SDM), se  $\lambda = 0$  e  $\theta = 0$ , tem-se o modelo espacial autoregressivo (SAR) que capta o efeito de interação de mudanças nas habilidades industriais dos vizinhos, medida a partir de uma média ponderada de todos os vizinhos da região  $i$ . Se  $\rho = 0$  e  $\theta = 0$ , tem-se o modelo de Erro Espacial (SEM) que mensura o efeito médio dos erros dos vizinhos em relação na região em questão. Os modelos espaciais serão estimados pelo método de máxima verossimilhança. O teste de restrição de parâmetros ( $\lambda$ ,  $\theta$  e  $\rho$ ) será utilizado para escolher o modelo com melhor ajuste da representação do processo espacial para fins de reporte dos resultados das estimações.

### 3.2 Heterogeneidade Espacial

A ideia de clube de convergência é baseada em modelos que produzem múltiplos equilíbrios. Qual desses diferentes equilíbrios uma região alcançará depende de sua posição inicial ou de algum outro atributo. Um grupo de regiões pode ter um determinado equilíbrio se eles compartilham a mesma localização inicial ou a algum atributo correspondente a esse equilíbrio. Isso produz clubes de convergências (Islam, 2003).

Para a análise de heterogeneidade espacial, considerou-se também a estratégia de Arbia (2006), LeSage e Pace (2009) e Elhorst (2010), partindo de uma especificação geral para modelos de painel espacial com uma adaptação com dummies para captar a heterogeneidade espacial de acordo com os agrupamentos citados na próxima seção. O modelo ficou da seguinte forma:

$$\log\left(\frac{s_{i,t}}{s_{i,t-1}}\right) = \beta_i \log(s_{i,t-1}) \times d_1^k + \rho \sum_{i=1}^n w_{ij} \log\left(\frac{s_{i,t}}{s_{i,t-1}}\right) + \theta_i \sum_{i=1}^n w_{ij} \log(s_{i,t-1}) \times d_1^k + \mu_i + \varphi_t + u_{i,t} \quad (5)$$

com

$$u_{i,t} = \lambda \sum_{i=1}^n w_{ij} u_{i,t} + \varepsilon_{i,t} \quad i = 1, \dots, N \quad t = 1, \dots, T \quad (6)$$

e

$$\varepsilon_{i,t} \sim i. i. d. (0, \sigma_\varepsilon^2) \quad (7)$$

Em que  $s_{i,t}$  contém as observações sobre o nível médio de habilidades na indústria de transformação privada para a região  $i$  no período  $t$ ;  $\rho$  é um parâmetro escalar espacial que captura taxa de crescimento do nível do nível médio habilidades na indústria de transformação privada dos vizinhos;  $\theta$  representa o nível médio de habilidades na indústria de transformação privada dos vizinhos e  $\lambda$  é o parâmetro de erro espacial;  $d_1^k$  são variáveis dummies multiplicativas que qualificam o nível médio de habilidades da indústria de transformação privada para as regiões pertencentes a cada clube. Esta especificação permite que o processo de convergência e os efeitos das variáveis condicionantes sejam diferentes entre os clusters. Os demais parâmetros seguem a especificação da equação (1).

Para este estudo, os clubes de convergência serão determinados por meio das técnicas de *machine learning* não supervisionadas, pelo método de cluster com restrição de contiguidade e pela análise do quintil de emprego formal na indústria de transformação. Esse procedimento permite verificar se a dinâmica do  $\beta$  convergência muda a depender das regiões.

### 3.2.1 Clubes via machine learning

Para a técnica de *machine learning*, segue-se a abordagem matemática de Sousa (2019), considerando que as variáveis<sup>6</sup> formem um conjunto de dados  $D$  com  $K$  pontos, definido da seguinte forma:

$$D = \{x_i\}_{i=1}^n \quad (8)$$

Um cluster, particiona esse conjunto  $D$  em  $K$  subconjuntos:

$$C = \{C_1, C_2, \dots, C_k\} \quad (9)$$

sendo  $C_i$  o cluster do conjunto  $D$ . Cada  $C_i$  está associado a um centroide ( $y_i$ ), que representa um cluster e é comparado com todos os outros elementos de  $D$ . Logo,  $y_i$  é determinado por meio da média de todos os pontos pertencentes ao cluster:

$$y_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j \quad (10)$$

em que  $n_i$  é o número de elementos de  $C_i$ .

Para calcular a dispersão dos elementos de um cluster e verificar quão próximos eles estão, compare-se a distância dos pontos ao centroide correspondente, para isso, foi utilizado a função semelhança:

$$F_s(P) = \sum_{i=1}^k \sum_{x_j \in C_i} E(x_j, y_i) \quad (11)$$

em que  $E$  é a distância euclidiana. Ou seja,  $F_s$  define a soma de todas as distâncias entre cada elemento e o centroide do cluster, ou seja, mede quão bem o centroide representa o seu grupo.

Logo, o algoritmo *k-means* tem como objetivo minimizar  $F_s$ , convergindo para a solução local, a mais viável, não levando em consideração se esse resultado é a solução ótima do problema. Assim, o algoritmo consiste nas seguintes etapas a partir do banco de dados ( $D$ ): i) distribui todos os pontos do conjunto  $D$  de forma aleatória em  $k$  clusters; ii) calcula o centroide de cada cluster ( $y_i$ ); iii) associa cada ponto  $x_j \in P$  a um cluster  $C_i^*$ , do centroide  $y_i^*$  mais próximo ao ponto, ou seja:

$$i^* = \operatorname{argmin}_{i=1,2,\dots,k} \|x_j - y_i\|_2^2 \quad (12)$$

Significa que um ponto  $x_j$  qualquer será agrupado a  $C_i$ , quando este ponto possuir a menor distância ao centroide desse cluster, comparado com a distância aos outros centroides do resto do clusters; iv) faz-se necessário a atualização dos centroides de cada cluster, dado que o passo anterior tenha alterado os pontos entre os grupos, assim será encontrado um novo centroide  $y_i$  para o cluster  $C_i$ ; v) esse processo será repetido até que os respectivos centroides não se alterem ou satisfaçam uma precisão estabelecida  $\epsilon$ ; vi) o teste de parada é analisando pela soma das diferenças dos centroides (resíduo) da iteração atual pela anterior.

$$\sum_{i=1}^k \|y_i^t - y_i^{t-1}\|_2 \leq \epsilon \quad (13)$$

em que  $y_i^t$  representa o centroide da iteração atual,  $\epsilon > 0$  é a precisão determinada e o limite da convergência. Para essa análise foi determinado um número máximo de cluster ( $K_{max}$ ), igual a 10, sabe-se que quando o número de cluster tende ao infinito, o resíduo tende a zero. Porém, a ideia do *machine learning* não supervisionado é tentar resumir em poucas variáveis, ou seja, em poucos clusters, a tomada de decisão do agrupamento.

A validação do algoritmo de *machine learning* é uma etapa indispensável para a confiabilidade dos resultados e é por meio dela que é demonstrado o quão representativo são os clusters induzidos por algoritmo de agrupamento como o *k-means*. Para a validação do algoritmo deste estudo foram utilizadas algumas métricas internas (índices internos) baseadas em medidas estatísticas de Davies-Bouldin (Davies e Boudien, 1979), que mede a similaridade média e a dispersão dos grupos e Silhouette (Rousseeuw, 1987) que é uma medida de qualidade dos clusters, ou seja, busca avaliar quão ajustados os dados estão em relação aos clusters.

<sup>6</sup>As variáveis utilizadas para a formação do conjunto de dado serão o número de empregos e estabelecimentos formais na indústria de transformação privada nas RGI. Essas variáveis representam a escala (tamanho) da indústria privada nas RGI.

### 3.2.2 Clubes por similaridade com restrição de contiguidade

O agrupamento com dados nos procedimentos individuais apresentados acima pode ser vinculado com agrupamentos espaciais. De acordo com Kopczewska (2022) há na literatura, alguns exemplos de agrupamentos espacialmente restritos podem ser encontrados. Todos eles tratam da questão da integração de aspectos espaciais e não espaciais. Em geral, eles adotam duas abordagens: ordem de agrupamento – questões espaciais primeiro e depois dados (generalização dominada por dados espaciais) ou o oposto (generalização dominada por dados não espaciais) ou avaliando uma troca misturando ou ponderando a dissimilaridade matrizes de dados e espaço.

Entre as soluções mais recentes está o ClustGeo (Chavent et al. 2018) que examina o potencial agrupamento de dados e locais estudando a inércia do agrupamento hierárquico paralelo de espaço e valores. Ele deriva duas funções de inércia (para espaço e valores) dependendo da divisão. Logo, quando ambas as funções de inércia se cruzam, define a proporção de ambos os agrupamentos expressa pela mistura do parâmetro  $\alpha$ . Ele pondera ambas as matrizes de dissimilaridade,  $D_0$  para valores e  $D_1$  para localizações, a fim de aumentar a coerência espacial dos clusters.

Para o conjunto de dados ( $D_0$ ) foram utilizados os dados de empregos formais e número de estabelecimentos, ambos para a indústria de transformação privada. Com base nesse conjunto foram estimados os cluster hierárquicos por similaridade. Para o conjunto espacial ( $D_1$ ) foram utilizadas uma matriz de distância entre os centroides das RGI. Em seguida é calculado o parâmetro de mistura  $\alpha[0,1]$  que define a importância de  $D_0$  e  $D_1$  no processo de agrupamento. Quando  $\alpha = 0$  não são consideradas as diferenças geográficas e quando  $\alpha = 1$  são as distâncias do emprego e do número de estabelecimentos que não são consideradas e os clusters são obtidos apenas com as distâncias geográficas. O objetivo é então calcular separadamente a homogeneidade das variáveis ( $Q_0$ ) e a homogeneidade geográfica ( $Q_1$ ) das partições obtidas para um intervalo de diferentes valores de  $\alpha$  e um determinado número de clusters. A homogeneidade  $Q_0$  (resp.  $Q_1$ ) é a proporção da inércia explicada calculada com  $D_0$  (resp.  $D_1$ ). Logo, os clubes serão determinados pela parcimônia destas proporções.

### 3.3 Escolha da matriz espacial

A dinâmica de distribuição da dimensão espacial das variáveis também é relevante. Apesar da recomendação de uso da teoria nos modelos de econometria espacial, especialmente relacionados a estrutura teórica *a priori* na escolha da matriz espaciais sugerido por Corrado e Fingleton (2012). Este estudo considera o procedimento de seleção de matrizes baseado no critério “*goodness-of-fit*”, sugerido por Stakhovych e Bijmolt (2009) e Zhang e Yu (2018) para encontrar a verdade especificação da matriz espacial. Logo, os modelos serão estimados com  $N$  diferentes matrizes espaciais. O valor do critério de informação AIC de cada modelo é calculado, então o modelo que exibe o menor AIC tem a maior probabilidade de conter a estrutura espacial correta. Esse resultado é equivalente à maximização do valor da log-verossimilhança. Para este estudo os modelos serão estimados utilizando a matriz de pesos espaciais K-vizinhos, com o número de vizinhos variando de 3 a 20.

### 3.4 Base de Dados

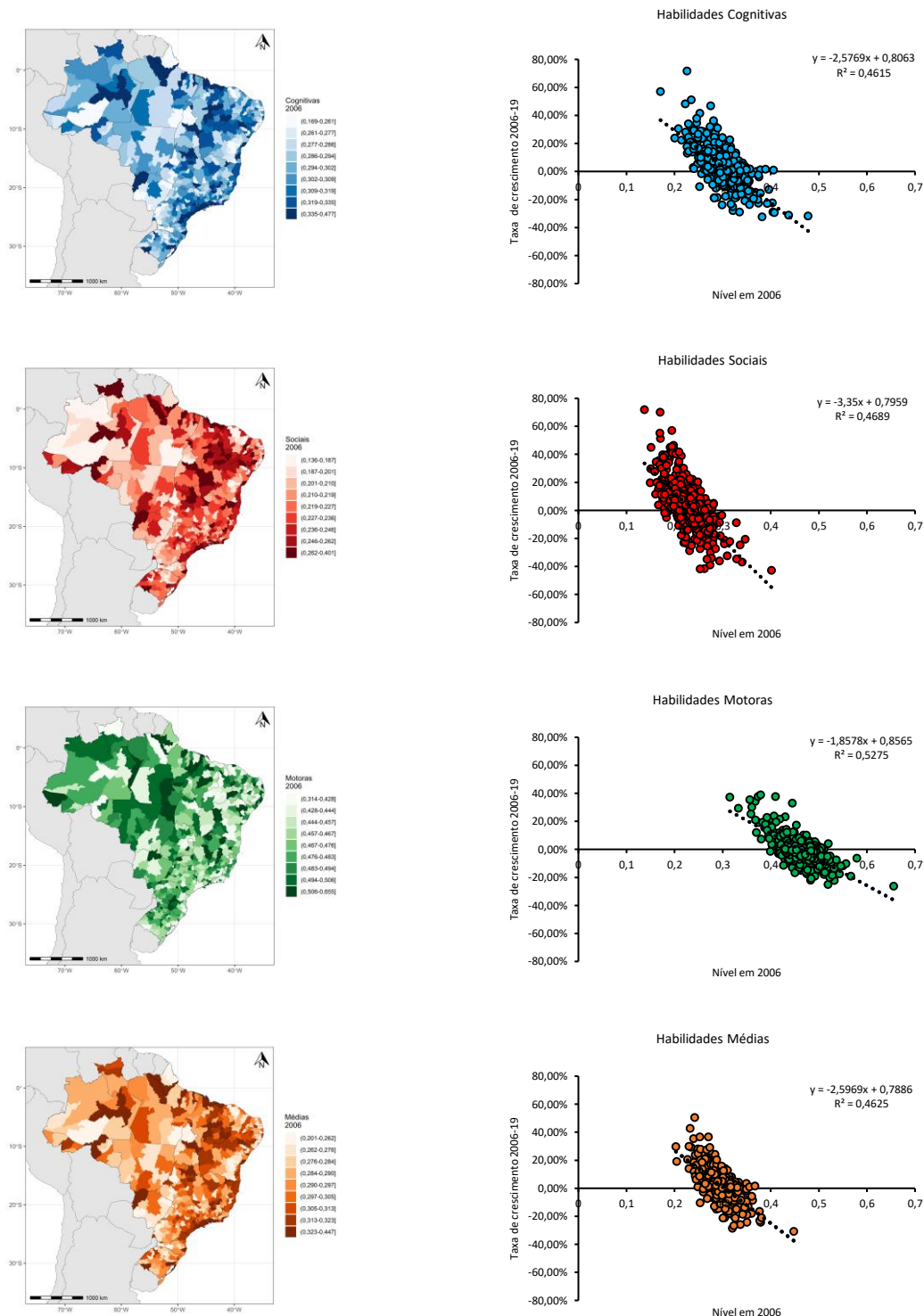
As regressões espaciais são do tipo painel de dados, ao nível das 510 Regiões Geográficas Imediatas (RGI) do país, segundo a classificação do IBGE (2017), sendo estimada entre os anos de 2006 a 2019. O período inicial foi escolhido devido as alterações da Classificação Nacional de Atividade Econômica (CNAE) 1.0 para 2.0 que começou a valer em 2006 para os dados de emprego formal da Relações Anuais de Informações Sociais (RAIS). O período final escolhido se ateu em não modelar o choque exógeno da COVID-19 que ocorreu nos meados de 2020. A base de dados refere-se a média de habilidades das ocupações formais na indústria de transformação privada<sup>7</sup>.

---

<sup>7</sup> A escolha da indústria privada se deve as evidências que o setor privado remunera seus trabalhadores com base na produtividade esperada e as notas das habilidades representam essas produtividades, conforme apontado por Neves, Azzoni e Chagas (2021). Por isso, optou-se por não analisar as indústrias de natureza jurídica pública, em que essas evidências podem ou não ocorrer.

O indicador é extraído da seguinte forma: i) número de ocupações formais da indústria de transformação da RAIS (CNAE 2.0, Div.10 a 33) para o setor privado (Natureza jurídica: não pública); ii) *matching* entre a CBO das ocupações e a nota de 0 a 1 para as habilidades cognitivas, sociais, motoras e médias<sup>8</sup>, com base em Maciente (2013) e Neves, Azzoni e Chagas (2021) (Apêndice A1); iii) cálculo da média das habilidade para cada RGI e para cada ano.

Figura 1 – Distribuição espacial das habilidades e as taxas de crescimento (2006-19) da indústria de transformação privada



Fonte: elaborado pelos autores (2022).

<sup>8</sup> Utilizaram-se como pesos os retornos salariais do setor privado estimadas por Neves, Azzoni e Chagas (2021): 0,55 para habilidades cognitivas; 0,33 para habilidades sociais e 0,12 para habilidades médias.



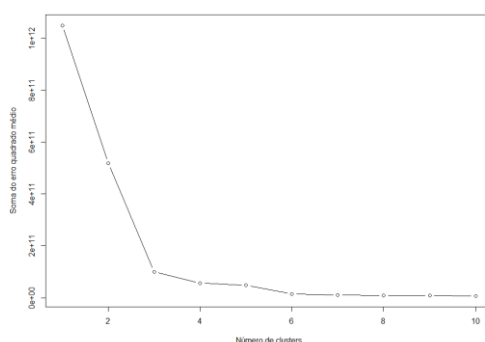
Estratégia semelhante utilizada no estudo de Castro e Azzoni (2022). O passo seguinte foi calcular a taxa de crescimento das habilidades para as RGI em cada ano do banco de dados. Para fins de reporte e análise a Figura 1 apresenta a distribuição espacial das habilidades em 2006 e as suas respectivas taxas de crescimento no período. De maneira geral, as regiões tradicionalmente industrializadas (Sudeste, Sul e ZNF de Manaus) estão alocadas nos maiores percentis para as habilidades cognitivas, sociais e médias. Entretanto, regiões pouco industrializadas (Noroeste e Norte da região Norte e interior da região Nordeste) também apresentam notas elevadas das habilidades, indicando possível erro de medida, reforçando o uso dos métodos de econometria espacial e do controle de efeito fixo. A Figura 1 também indica possível convergência de habilidades entre as regiões, com as habilidades sociais convergindo mais rápido e as motoras mais devagar. Percebe-se também que o nível médio em 2006 é maior para as habilidades motoras, seguido das cognitivas, sociais e, por último, das habilidades médias. Esse resultado é semelhante ao apresentando por Castro e Azzoni (2022) para as habilidades de todos os setores, exceto o setor público, nas regiões do Estado do Paraná. Esse resultado implica que o retorno salarial das habilidades motoras das ocupações da indústria é menor em comparação com as cognitivas e sociais.

## 4 RESULTADOS E DISCUSSÃO

### 4.1 Clubes via *machine learning*

Para essa análise foi determinado um número máximo de cluster ( $K_{max}$ ), igual a 10, sabe-se que quando o número de cluster tende ao infinito, o resíduo tende a zero. Porém, o objetivo do *machine learning* não supervisionado é tentar resumir em poucas variáveis, ou seja, em poucos clusters, a tomada de decisão do agrupamento. Assim, o ponto principal da aprendizagem é decidir o menor número de clusters para o menor valor do resíduo ( $\hat{k}$ ). Para essa decisão foi utilizado a análise gráfica (Figura 2) e o número de cluster escolhidos foi igual a 3, entretanto também foi testado as regressões para o cluster igual a 6.

Figura 2 - Erro quadrado médio dos clusters



Fonte: elaborado pelos autores (2022).

A distribuição espacial dos clusters  $k = 3$  e  $k = 6$  para as 510 regiões e as suas respectivas distribuições podem ser observadas na Figura 3 e na Tabela 1.

Tabela 1 – Clusters de emprego e estabelecimentos da indústria em 2019 via *machine learning*

| C     | $k = 3$ |     |         |      |       |      | $k = 6$ |      |         |      |       |      |
|-------|---------|-----|---------|------|-------|------|---------|------|---------|------|-------|------|
|       | RGI     | %   | Emp.    | %    | Ind.  | %    | RGI     | %    | Emp.    | %    | Ind.  | %    |
| 1     | 491     | 9,6 | 4.073,0 | 56,5 | 232,1 | 59,6 | 105     | 20,6 | 1.529,9 | 21,3 | 77,1  | 19,8 |
| 2     | 1       | 0,2 | 852,6   | 11,8 | 46,2  | 11,8 | 1       | 0,2  | 852,6   | 11,8 | 46,2  | 11,8 |
| 3     | 18      | 3,4 | 2.281,4 | 31,7 | 111,6 | 28,6 | 13      | 2,5  | 1.374,7 | 19,1 | 66,5  | 17,1 |
| 4     |         |     |         |      |       |      | 4       | 0,8  | 837,6   | 11,6 | 41,3  | 10,6 |
| 5     |         |     |         |      |       |      | 342     | 67,1 | 933,8   | 13,0 | 68,4  | 17,5 |
| 6     |         |     |         |      |       |      | 45      | 8,8  | 1.678,4 | 23,3 | 90,3  | 23,2 |
| Total | 510     | 100 | 7.207,1 | 100  | 390,1 | 100  | 510     | 100  | 7.207,1 | 100  | 390,1 | 100  |

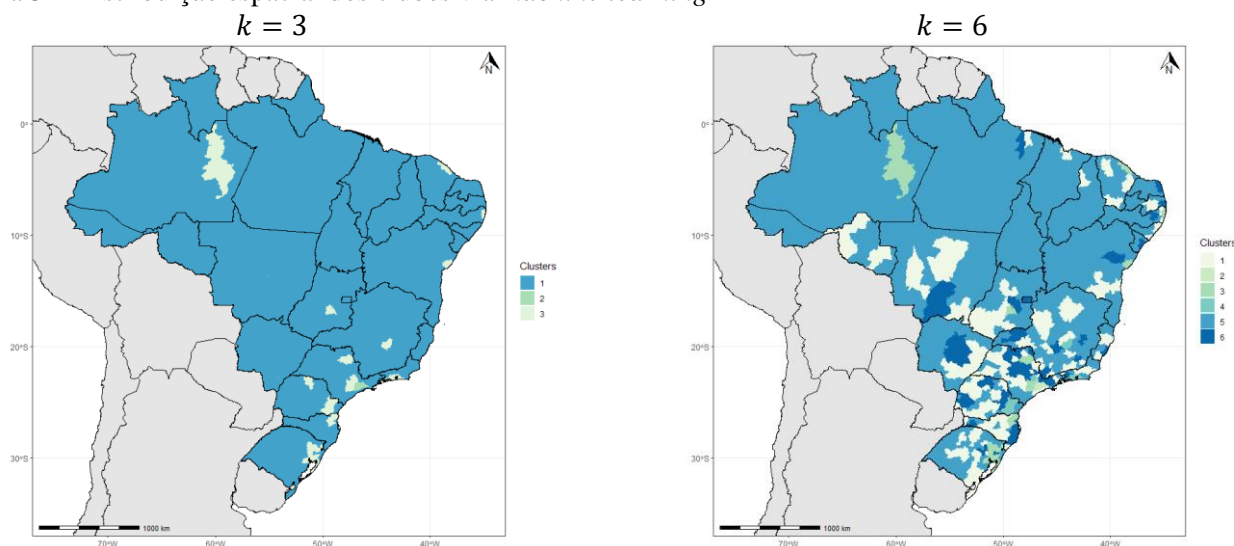
Fonte: elaborados pelos autores com base nos microdados da RAIS (2022).

Nota: o número de emprego e estabelecimento estão divididos por mil.

Para  $k = 3$  percebe-se a RGI de SP ficou isolado no cluster 2, no cluster 3 estão 18 RGI que pertencem há algumas capitais e também algumas fora dos centros metropolitanos (como o caso de Ribeirão Preto, Campinas, Jundiaí, Sorocaba, Londrina, Joinville, Blumenau, Novo Hamburgo/São Leopoldo e Caxias do Sul), as demais RGI, cerca de 491, ficaram no cluster 1.

Para  $k = 6$  a RGI de São Paulo também ficou isolada no cluster 2, no cluster 3 foram alocados as RGI de: Manaus, Fortaleza, Recife, Salvador, Sorocaba, Ribeirão Preto, Jundiaí, Joinville, Blumenau, Porto Alegre, Novo Hamburgo/São Leopoldo, Caxias do Sul e Goiânia. As RGI de Belo Horizonte, Rio de Janeiro, Campinas e Curitiba estão no cluster 4. As demais RGI foram estão nos cluster 1, 5 e 6 e estão distribuídas em grande parte no interior do país

Figura 3 – Distribuição espacial dos clubes via *machine learning*



Fonte: elaborado pelos autores.

A validação dos resultados do algoritmo *k-means* pelas estatísticas internas estão apresentadas na Tabela 2. Percebe-se que para a amostra os valores e a média do indicador de Silhueta foram maiores do que zero<sup>9</sup>, indicando que as observações foram bem agrupadas. Em relação ao índice Davies-Bouldin, os valores encontrados para a base de treinamento são próximos de zero demonstrando os clusters são compactos e distantes entre si. Estas estatísticas estão condizentes com a literatura.

Tabela 2 - Estatísticas de validação interna

| Clusters       | Silhueta |         |
|----------------|----------|---------|
|                | $k = 3$  | $k = 6$ |
| 1              | 0,91     | 0,48    |
| 2              | 0,00     | 0,00    |
| 3              | 0,52     | 0,64    |
| 4              |          | 0,78    |
| 5              |          | 0,74    |
| 6              |          | 0,46    |
| Média          | 0,83     | 0,66    |
| Davies-Bouldin | 0,35     | 0,45    |

Fonte: elaborado pelos autores (2022).

Os resultados das regressões para  $k = 3$  estão apresentados na Tabela 3. Para cada habilidade será apresentado o modelo econométrico espacial que obteve o melhor ajuste. Percebe-se que para as habilidades cognitivas (SDM), as regiões do cluster 1 estão convergindo entre si no que se refere as habilidades cognitivas da indústria de transformação privada e não há diferenças estatisticamente significativas entre a convergência com as demais regiões do demais clusters.

<sup>9</sup> Exceto para o cluster 2 para  $k = 3$  e  $k = 6$ . Nesse cluster o algoritmo detectou apenas uma região (RGI de São Paulo) para compor esse grupo, sendo essa RGI um possível *outlier*.

Tabela 3 – Análise de convergência condicional das habilidades das ocupações da indústria, clubes, *machine learning* ( $k = 3$ )

| Var.                                | Cognitivas         |                    |                    |                    | Sociais            |                    |                    |                    | Motoras             |                   |                    |                    | Médias             |                    |                    |                    |
|-------------------------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|---------------------|-------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
|                                     | OLS                | SDM                | SAR                | SEM                | OLS                | SDM                | SAR                | SEM                | OLS                 | SDM               | SAR                | SEM                | OLS                | SDM                | SAR                | SEM                |
| $\beta$                             | -0,403***<br>0,027 | -0,406***<br>0,027 | -0,402***<br>0,027 | -0,404***<br>0,027 | -0,448***<br>0,028 | -0,445***<br>0,028 | -0,448***<br>0,028 | -0,448***<br>0,028 | -0,506***<br>0,037  | -0,506***<br>0,36 | -0,506***<br>0,037 | -0,507***<br>0,037 | -0,406***<br>0,026 | -0,410***<br>0,026 | -0,405***<br>0,026 | -0,407***<br>0,026 |
| $\beta C_2$                         | 0,411***<br>0,15   | -0,479<br>0,351    | 0,422***<br>0,15   | 0,501***<br>0,17   | 0,806***<br>0,155  | 0,102<br>0,293     | 0,808***<br>0,154  | 0,798***<br>0,153  | 0,595***<br>0,134   | -0,396<br>0,254   | 0,603***<br>0,136  | 0,568***<br>0,127  | 0,601***<br>0,148  | -0,410<br>0,306    | 0,608***<br>0,147  | 0,646***<br>0,153  |
| $\beta C_3$                         | 0,647***<br>0,149  | 0,280<br>0,204     | 0,644***<br>0,149  | 0,590***<br>0,154  | 1,122***<br>0,186  | 0,785***<br>0,208  | 1,125***<br>0,186  | 1,076***<br>0,191  | 0,703***<br>0,16563 | 0,679***<br>0,203 | 0,706***<br>0,165  | 0,695***<br>0,162  | 0,759***<br>0,149  | 0,294<br>0,199     | 0,758***<br>0,149  | 0,691***<br>0,155  |
| $\theta$                            |                    | 0,047***<br>0,022  |                    |                    |                    | 0,038<br>0,028     |                    |                    |                     | 0,046*<br>0,025   |                    |                    |                    | 0,044*<br>0,023    |                    |                    |
| $\theta C_2$                        |                    | -7,661**<br>3,599  |                    |                    |                    | -1,634<br>2,20     |                    |                    |                     | -2,575<br>3,117   |                    |                    |                    | -5,875**<br>2,818  |                    |                    |
| $\theta C_3$                        |                    | 1,795***<br>0,691  |                    |                    |                    | 1,68**<br>0,822    |                    |                    |                     | 0,321<br>0,622    |                    |                    |                    | 2,091***<br>0,649  |                    |                    |
| $\rho$                              |                    | 0,050***<br>0,016  | 0,033**<br>0,015   |                    |                    | 0,039**<br>0,017   | 0,017<br>0,01452   |                    |                     | 0,041*<br>0,021   | 0,019<br>0,036     |                    |                    | 0,051***<br>0,015  | 0,035**<br>0,014   |                    |
| $\lambda$                           |                    |                    |                    | 0,052***<br>0,016  |                    |                    |                    | 0,029*<br>0,015    |                     |                   |                    | 0,083**<br>0,033   |                    |                    |                    | 0,052***<br>0,015  |
| Obs.                                | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552                | 6552              | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               |
| R2                                  | 0,234              | 0,221              | 0,218              | 0,219              | 0,247              | 0,234              | 0,232              | 0,233              | 0,264               | 0,255             | 0,253              | 0,253              | 0,234              | 0,221              | 0,218              | 0,219              |
| LogLik                              | 13843,33           | 13853,61           | 13845,98           | 13848,86           | 11762,31           | 11767,31           | 11763,02           | 11764,08           | 13649,4             | 13653,23          | 13649,52           | 13651,52           | 15388,47           | 15399,2            | 15391,37           | 15394,05           |
| AIC                                 | -27658,66          | -27691,23          | -27681,97          | -27687,72          | -23496,67          | -23518,61          | -23516,04          | -23518,16          | -27270,81           | 27290,46          | -27289,05          | -27293,04          | -30748,94          | -30782,4           | -30772,74          | -30778,11          |
| BIC                                 | -27563,64          | -27636,93          | -27648,03          | -27653,78          | -23401,65          | -23464,31          | -23482,1           | -23484,22          | -27175,78           | -27236,16         | -27255,11          | -27259,11          | -30653,92          | -30728,1           | -30738,8           | -30744,14          |
| LMI                                 | 10,203***          |                    |                    |                    | 6,156**            |                    |                    |                    | 1,0763              |                   |                    |                    | 12,664***          |                    |                    |                    |
| RLMI                                | 2,7603*            |                    |                    |                    | 0,83655            |                    |                    |                    | 4,258**             |                   |                    |                    | 2,9157*            |                    |                    |                    |
| LMe                                 | 0,83655            |                    |                    |                    | 9,4938***          |                    |                    |                    | 4,3742**            |                   |                    |                    | 20,889***          |                    |                    |                    |
| RLMe                                | 17,489***          |                    |                    |                    | 4,1743***          |                    |                    |                    | 7,5559***           |                   |                    |                    | 11,141***          |                    |                    |                    |
| Restrição                           | Prob>chi2          |                    |                    |                    | Prob>chi2          |                    |                    |                    | Prob>chi2           |                   |                    |                    | Prob>chi2          |                    |                    |                    |
| $\theta = 0 \text{ e } \rho \neq 0$ | 0,0035             |                    |                    |                    | 0,0458             |                    |                    |                    | 0,0238              |                   |                    |                    | 0,001              |                    |                    |                    |
|                                     | SAR $\neq$ SDM     |                    |                    |                    | SAR $\neq$ SDM     |                    |                    |                    | SAR $\neq$ SDM      |                   |                    |                    | SAR $\neq$ SDM     |                    |                    |                    |
| $\theta = -\lambda \beta$           | 0,0114             |                    |                    |                    | 0,1005             |                    |                    |                    | 0,1078              |                   |                    |                    | 0,0032             |                    |                    |                    |
|                                     | SEM $\neq$ SDM     |                    |                    |                    | SEM = SDM          |                    |                    |                    | SEM = SDM           |                   |                    |                    | SEM $\neq$ SDM     |                    |                    |                    |

Fonte: elaborado pelos autores. Nota: Nível de significância \*\*\*<0.01; \*\*<0.05% e \*<10%. Erro padrão robusto.

Tabela 4 – Análise de convergência condicional das habilidades das ocupações da indústria, clubes, *machine learning* ( $k = 6$ )

| Var.                                | Cognitivas         |                   |                   |                   | Sociais           |                   |                   |                   | Motoras            |                    |                    |                    | Médias             |                    |                    |                     |
|-------------------------------------|--------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|---------------------|
|                                     | OLS                | SDM               | SAR               | SEM               | OLS               | SDM               | SAR               | SEM               | OLS                | SDM                | SAR                | SEM                | OLS                | SDM                | SAR                | SEM                 |
| $\beta$                             | -0,420***<br>0,030 | -0,422***<br>0,03 | -0,419***<br>0,03 | -0,421***<br>0,03 | -0,466***<br>0,03 | -0,467***<br>0,03 | -0,465***<br>0,03 | -0,465***<br>0,03 | -0,523***<br>0,039 | -0,523***<br>0,038 | -0,523***<br>0,038 | -0,523***<br>0,038 | -0,409***<br>0,027 | -0,412***<br>0,027 | -0,408***<br>0,027 | -0,410***<br>0,0271 |
| $\beta C_1$                         | 0,614***<br>0,232  | 0,579***<br>0,232 | 0,613***<br>0,232 | 0,607***<br>0,231 | 1,022***<br>0,221 | 0,957***<br>0,215 | 1,020***<br>0,222 | 1,010***<br>0,222 | 0,847***<br>0,163  | 0,823***<br>0,164  | 0,847***<br>0,163  | 0,837***<br>0,163  | -0,183*<br>0,104   | -0,173*<br>0,104   | -0,181*<br>0,105   | -0,180*<br>0,105    |
| $\beta C_2$                         | 0,627***<br>0,182  | 0,263<br>0,323    | 0,636***<br>0,182 | 0,704***<br>0,193 | 1,081***<br>0,188 | 0,026<br>0,036    | 1,082***<br>0,188 | 1,084***<br>0,188 | 0,730***<br>0,153  | 0,273<br>0,389     | 0,739***<br>0,154  | 0,739***<br>0,153  | -0,320***<br>0,089 | 0,399*<br>0,234    | -0,326***<br>0,089 | -0,036***<br>0,095  |
| $\beta C_3$                         | 0,809***<br>0,182  | 0,607***<br>0,233 | 0,804***<br>0,182 | 0,765***<br>0,189 | 1,431***<br>0,216 | 0,858***<br>0,237 | 1,432***<br>0,216 | 1,398***<br>0,221 | 0,790***<br>0,204  | 0,638***<br>0,22   | 0,793***<br>0,204  | 0,794***<br>0,204  | -0,535***<br>0,132 | -0,297*<br>0,164   | -0,533***<br>0,132 | -0,492***<br>0,136  |
| $\beta C_4$                         | 0,749***<br>0,214  | 0,363<br>0,262    | 0,754***<br>0,205 | 0,709***<br>0,206 | 1,066***<br>0,309 | 0,370<br>0,275    | 1,072***<br>0,308 | 1,057***<br>0,308 | 0,935***<br>0,163  | 0,695***<br>0,24   | 0,940***<br>0,165  | 0,959***<br>0,161  | -0,354**<br>0,164  | -0,038<br>0,128    | -0,365**<br>0,156  | -0,347**<br>0,149   |
| $\beta C_6$                         | 0,597***<br>0,215  | 0,512***<br>0,197 | 0,590***<br>0,214 | 0,566***<br>0,213 | 1,225***<br>0,251 | 1,045***<br>0,235 | 1,222***<br>0,251 | 1,201***<br>0,251 | 0,497***<br>0,153  | 0,418***<br>0,146  | 0,470***<br>0,153  | 0,474***<br>0,152  | -0,464***<br>0,142 | -0,433***<br>0,132 | -0,466***<br>0,141 | -0,459***<br>0,138  |
| $\rho$                              |                    | 0,064***<br>0,017 | 0,033***<br>0,015 |                   |                   | 0,023<br>0,015    | 0,016<br>0,145    |                   |                    | 0,036*<br>0,02     | 0,020<br>0,036     |                    |                    | 0,050***<br>0,015  | 0,035**<br>0,014   |                     |
| $\lambda$                           |                    |                   |                   | 0,049***<br>0,017 |                   |                   |                   | 0,024<br>0,015    |                    |                    |                    | 0,073**<br>0,033   |                    |                    |                    | 0,052***<br>0,015   |
| $\theta$                            | Não                | Sim               | Não               | Não               | Não               | Sim               | Não               | Não               | Não                | Sim                | Não                | Não                | Não                | Sim                | Não                | Não                 |
| Obs.                                | 6552               | 6552              | 6552              | 6552              | 6552              | 6552              | 6552              | 6552              | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552                |
| R2                                  | 0,237              | 0,226             | 0,223             | 0,224             | 0,251             | 0,240             | 0,238             | 0,238             | 0,269              | 0,261              | 0,259              | 0,260              | 0,234              | 0,222              | 0,219              | 0,220               |
| LogLik                              | 13856,87           | 13867,14          | 13859,46          | 13861,87          | 11779,65          | 11785,56          | 11780,24          | 11780,82          | 13669,45           | 13673,58           | 13669,6            | 13671,08           | 15391,61           | 15400,99           | 15394,5            | 15397,17            |
| AIC                                 | -27679,74          | -27706,27         | -27702,92         | -27707,74         | -23525,3          | -23543,13         | -23544,48         | -23545,64         | -27304,91          | -27319,15          | -27323,2           | -27326,17          | -30749,21          | -30773,98          | -30773             | -30778,33           |
| BIC                                 | -27564,35          | -27611,25         | -27648,62         | -27653,44         | -23409,91         | -23448,1          | -23490,18         | -23491,33         | -27189,52          | -27224,13          | -27268,9           | -27271,87          | -30633,82          | -30678,95          | -30718,7           | -30724,03           |
| LMI                                 | 10,18***           |                   |                   |                   | 5,9784**          |                   |                   |                   | 1,168              |                    |                    |                    | 20,228***          |                    |                    |                     |
| RLMI                                | 1,737              |                   |                   |                   | 0,013289          |                   |                   |                   | 2,3901             |                    |                    |                    | 9,9852***          |                    |                    |                     |
| LMe                                 | 16,16***           |                   |                   |                   | 7,2927***         |                   |                   |                   | 3,622*             |                    |                    |                    | 12,644***          |                    |                    |                     |
| RLMe                                | 7,7171***          |                   |                   |                   | 1,3276            |                   |                   |                   | 4,8441**           |                    |                    |                    | 2,402              |                    |                    |                     |
| Restrição                           | Prob>chi2          |                   |                   |                   | Prob>chi2         |                   |                   |                   | Prob>chi2          |                    |                    |                    | Prob>chi2          |                    |                    |                     |
| $\theta = 0 \text{ e } \rho \neq 0$ | 0,1169             |                   | SAR = SDM         |                   | 0,0203            |                   | SAR $\neq$ SDM    |                   | 0,3963             |                    | SAR = SDM          |                    | 0,0058             |                    | SAR = SDM          |                     |
| $\theta = -\lambda \beta$           | 0,3014             |                   | SEM = SDM         |                   | 0,0257            |                   | SEM $\neq$ SDM    |                   | 0,544              |                    | SEM = SDM          |                    | 0,0147             |                    | SEM = SDM          |                     |

Fonte: elaborado pelos autores. Nota: Nível de significância \*\*\*<0.01; \*\*<0.05% e \*<10%. Erro padrão clusterizado (RGI).

Para as habilidades sociais na indústria (SEM) apenas as regiões do cluster 1 estão convergindo entre si, já os cluster 2 e 3 não convergem e são diferentes do cluster 1. Nas habilidades motoras (SEM) os resultados são semelhantes aos das habilidades sociais. Por último, os resultados das habilidades médias (SDM) são semelhantes aos resultados das habilidades cognitivas.

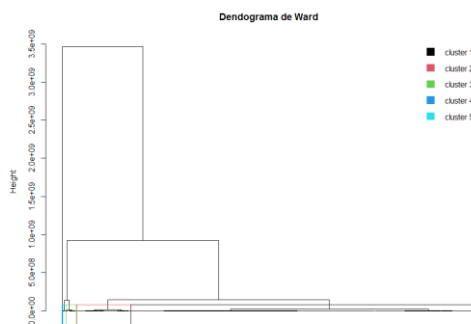
Quanto maior o parâmetro  $\beta$ , maior é a velocidade de convergência, logo os resultados as habilidades estimadas para o cluster 1 mostram que as habilidades das ocupações formais da indústria de transformação privada estão convergindo mais rápido para as habilidades motoras, seguido das habilidades sociais e cognitivas. Com base em Elhorst (2010), os modelos de regressão espacial exploram a estrutura de dependência entre as unidades espaciais (RGI), o efeito da mudança de uma variável explicativa para uma RGI específica afetará a própria RGI e, potencialmente, todas as outras RGI indiretamente, por isso é necessário reportar os efeitos marginais (diretos, indiretos e totais) dos  $\beta$ . Esses resultados estão no Apêndice A2.

A Tabela 4 contém os resultados das regressões para os clubes com  $k = 6$ , em geral, as RGI do cluster 5 estão convergindo para as habilidades cognitivas (SAR) e para as outras habilidades, as demais RGI de cada cluster não convergem entre si e são estatisticamente diferentes das RGI do cluster 5, exceto para o cluster 4 para as habilidades sociais (SDM), o cluster 6 para as habilidades motoras (SAR) e todos os clusters para as habilidades médias das ocupações formais da indústria de transformação privada (SDM).

## 4.2 Clubes por similaridade com restrição de contiguidade

Com base no conjunto de dados ( $D_0$ ) foram estimados os cluster hierárquicos por similaridade, conforme o dendrograma de Ward na Figura 4, foram encontrados 5 cluster.

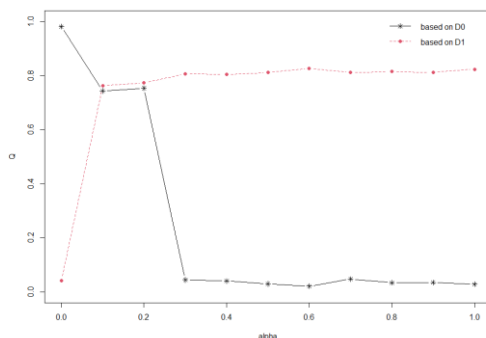
Figura 4 – Cluster hierárquicos



Fonte: elaborado pelos autores.

O gráfico das curvas de  $Q_0$  e  $Q_1$  (Figura 5) é uma ferramenta para escolher um valor de  $\alpha$  que é uma relação entre a perda de homogeneidade da variáveis de emprego e estabelecimento e o ganho de coesão geográfica.

Figura 5: Curva  $Q_0$  e  $Q_1$  e Alpha

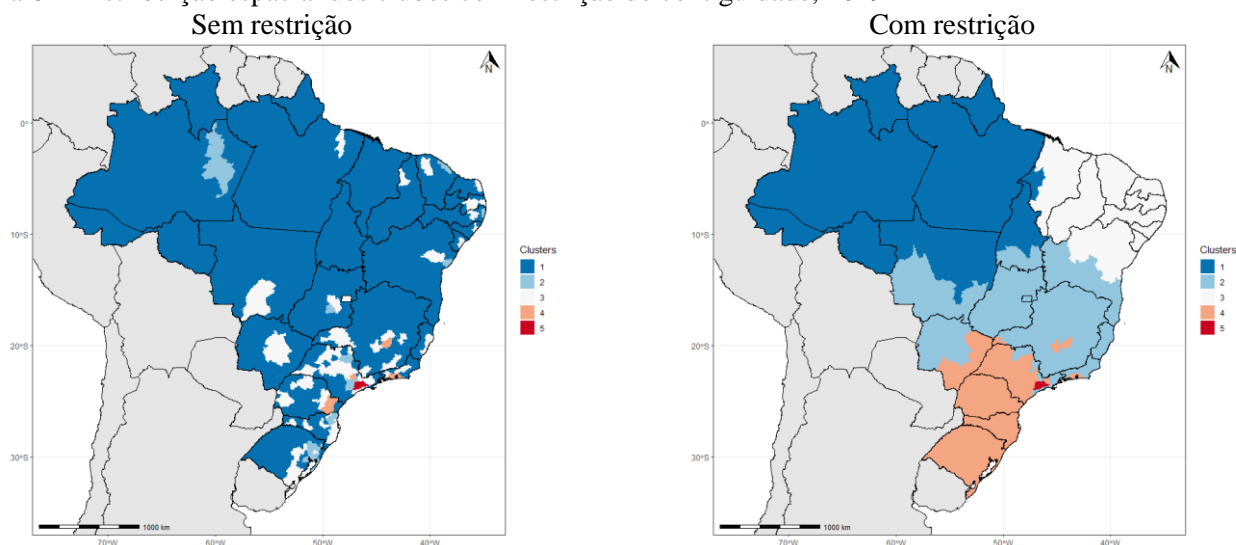


Fonte: elaborado pelos autores.

Percebe-se que a proporção de inércia explicada calculada com  $D_0$  (as distâncias das variáveis) é igual a 0,98 quando  $\alpha = 0$  e diminui quando  $\alpha$  aumenta (linha preta). Pelo contrário, a proporção da inércia explicada calculada com  $D_1$  (as distâncias geográficas) é igual a 0,81 quando  $\alpha = 1$  e diminui quando  $\alpha$  diminui (linha vermelha). Logo, o gráfico sugere escolher  $\alpha = 0,2$  o que corresponde a uma perda de homogeneidade das variáveis de 24% e um ganho de homogeneidade geográfica considerável.

A distribuição espacial dos cluster está disponível na Figura 6, percebe-se que o mapa de cluster sem restrição de contiguidade é semelhante aos mapas da Figura 3 com  $k = 3$  e com o mapa da quintis (Figura 7). Assim, os cluster com os maiores números de RGI estão localizados nas regiões menos industrializadas do Brasil. Os demais clusters estão localizados nas regiões tradicionalmente industrializadas Sudeste e Sul e em algumas capitais.

Figura 6 – Distribuição espacial dos clubes com restrição de contiguidade, 2019



Fonte: elaborado pelos autores.

A distribuição espacial dos cluster com restrição de contiguidade divide as RGI em 5 grandes grupos que visualmente representam 3 grandes faixas:

1ª faixa: Contém 2 grupos, sendo o primeiro referente as RGI que estão localizadas nos estados da região Norte e parte do Centro Oeste. O segundo grupo representa em grande medida as RGI localizadas nos estados da região Nordeste. De acordo com a Tabela 5 essa faixa representa 44 % das RGI do Brasil, 24% do emprego e 27% das indústrias.

2ª faixa: representam as RGI que estão localizadas em parte na região Centro Oeste, Sudeste e Nordeste. Essa faixa apresenta 27% das RGI do Brasil e 12,7% e 12,4% do emprego e das indústrias, respectivamente.

3ª faixa: Contém 2 grupos, sendo o primeiro referente as RGI que estão localizadas em parte na região Centro Oeste, Sudeste e Sul. O segundo grupo representa a RGI de São Paulo. Esses grupos representam 28% das RGI do Brasil, 63% do emprego industrial e 60% das indústrias.

Tabela 5 – Clusters de empregos e de indústria com restrição de contiguidade, 2019

| Clusters | RGI | %    | Emprego | %    | Indústrias | %    |
|----------|-----|------|---------|------|------------|------|
| 1        | 71  | 13,9 | 303,4   | 4,2  | 14,6       | 3,7  |
| 2        | 154 | 30,2 | 1.450,4 | 20,1 | 91,2       | 23,4 |
| 3        | 140 | 27,5 | 917,3   | 12,7 | 48,2       | 12,4 |
| 4        | 144 | 28,2 | 3.683,4 | 51,1 | 189,8      | 48,7 |
| 5        | 1   | 0,2  | 852,6   | 11,8 | 46,2       | 11,9 |
| Total    | 510 | 100  | 7.207,2 | 100  | 390,1      | 100  |

Fonte: elaborados pelos autores com base nos microdados da RAIS (2019).

Nota: o número de emprego e estabelecimento estão divididos por mil.

Tabela 6 – Análise de convergência condicional das habilidades das ocupações da indústria, clubes com restrição de contiguidade ( $k = 5$ )

| Var.                                | Cognitivas         |                    |                    |                    | Sociais            |                    |                    |                    | Motoras           |                   |                    |                    | Médias             |                    |                    |                    |
|-------------------------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|-------------------|-------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
|                                     | OLS                | SDM                | SAR                | SEM                | OLS                | SDM                | SAR                | SEM                | OLS               | SDM               | SAR                | SEM                | OLS                | SDM                | SAR                | SEM                |
| $\beta$                             | -0,355***<br>0,029 | -0,353***<br>0,029 | -0,355***<br>0,029 | -0,355***<br>0,029 | -0,368***<br>0,031 | -0,369***<br>0,031 | -0,368***<br>0,031 | -0,368***<br>0,031 | -0,418***<br>0,04 | -0,417***<br>0,04 | -0,418***<br>0,039 | -0,417***<br>0,039 | -0,355***<br>0,029 | -0,354***<br>0,029 | -0,355***<br>0,029 | -0,355***<br>0,029 |
| $\beta C_1$                         | -0,367<br>0,325    | -0,037<br>0,324    | -0,365<br>0,325    | -0,366<br>0,324    | -1,028***<br>0,381 | -1,018***<br>0,381 | -1,026***<br>0,382 | -1,029***<br>0,38  | -0,496*<br>0,268  | -0,496*<br>0,267  | -0,500*<br>0,268   | 0,506*<br>0,268    | -0,465<br>0,307    | -0,463<br>0,308    | -0,459<br>0,307    | -0,461<br>0,306    |
| $\beta C_3$                         | -0,248<br>0,259    | -0,276<br>0,258    | -0,247<br>0,258    | -0,258<br>0,259    | -0,370<br>0,335    | -0,372<br>0,336    | -0,370<br>0,335    | -0,379<br>0,336    | -0,325<br>0,197   | -0,331*<br>0,197  | -0,325*<br>0,197   | -0,331*<br>0,197   | -0,235<br>0,273    | -0,262<br>0,274    | -0,233<br>0,273    | -0,244<br>0,274    |
| $\beta C_4$                         | -0,067<br>0,211    | -0,136<br>0,218    | -0,063<br>0,211    | -0,080<br>0,214    | 0,012<br>0,237     | -0,145<br>0,245    | 0,015<br>0,237     | -0,006<br>0,238    | 0,137<br>0,134    | 0,116<br>0,137    | 0,140<br>0,133     | 0,133<br>0,134     | -0,035<br>0,209    | -0,130<br>0,213    | -0,030<br>0,209    | -0,052<br>0,211    |
| $\beta C_5$                         | -0,287<br>0,174    | 0,243<br>0,181     | 0,300*<br>0,174    | 0,357*<br>0,187    | 0,540***<br>0,195  | 0,384**<br>0,19    | 0,542***<br>0,195  | 0,521***<br>0,191  | 0,336**<br>0,142  | 0,303**<br>0,139  | 0,346**<br>0,144   | 0,319**<br>0,136   | 0,462***<br>0,171  | -0,376**<br>0,172  | 0,472***<br>0,171  | 0,487***<br>0,173  |
| $\rho$                              |                    | 0,049***<br>0,017  | 0,033**<br>0,015   |                    |                    | 0,026<br>0,016     | 0,016<br>0,014     |                    |                   | 0,040*<br>0,021   | 0,024<br>0,036     |                    |                    | 0,050***<br>0,016  | 0,034**<br>0,014   |                    |
| $\lambda$                           |                    |                    |                    | 0,051***<br>0,017  |                    |                    |                    | 0,027*<br>0,015    |                   |                   |                    | 0,089***<br>0,033  |                    |                    |                    | 0,051***<br>0,016  |
| $\theta$                            | Não                | Sim                | Não                | Não                | Não                | Sim                | Não                | Não                | Não               | Sim               | Não                | Não                | Não                | Sim                | Não                | Não                |
| Obs.                                | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552              | 6552              | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               |
| R2                                  | 0,235              | 0,223              | 0,220              | 0,221              | 0,253              | 0,241              | 0,239              | 0,239              | 0,268             | 0,259             | 0,258              | 0,259              | 0,236              | 0,223              | 0,221              | 0,221              |
| LogLik                              | 13849,07           | 13858,82           | 13851,72           | 13854,58           | 11788,63           | 11794,64           | 11789,22           | 11790,12           | 13667,73          | 13671,15          | 13667,93           | 13670,2            | 15396,99           | 15407,03           | 15399,75           | 15402,35           |
| AIC                                 | -27666,14          | -27693,63          | -27689,45          | -27695,16          | -23545,26          | -23565,27          | -23564,44          | -23566,25          | -27303,46         | -27318,29         | -27321,87          | -27326,39          | -30761,98          | -30790,06          | -30785,5           | -30790,7           |
| BIC                                 | -27557,54          | -27612,18          | -27641,94          | -27647,65          | -23436,66          | -23483,82          | -23516,93          | -23518,73          | -27194,86         | -27236,84         | -27274,35          | -27278,88          | -30653,38          | -30708,61          | -30737,99          | -30743,19          |
| LMI                                 | 10,22***           |                    |                    |                    | 5,8401**           |                    |                    |                    | 1,30090           |                   |                    |                    | 12,256***          |                    |                    |                    |
| RLMI                                | 1,9538             |                    |                    |                    | 0,0014757          |                    |                    |                    | 4,4901**          |                   |                    |                    | 1,4804             |                    |                    |                    |
| LMe                                 | 16,597***          |                    |                    |                    | 6,8262***          |                    |                    |                    | 5,1166**          |                   |                    |                    | 18,568***          |                    |                    |                    |
| RLMe                                | 8,33***            |                    |                    |                    | 0,3117             |                    |                    |                    | 8,3064***         |                   |                    |                    | 7,7924***          |                    |                    |                    |
| Restrição                           | Prob>chi2          |                    |                    |                    | Prob>chi2          |                    |                    |                    | Prob>chi2         |                   |                    |                    | Prob>chi2          |                    |                    |                    |
| $\theta = 0 \text{ e } \rho \neq 0$ | 0,0516             |                    | SAR $\neq$ SDM     |                    | 0,0493             |                    | SAR $\neq$ SDM     |                    | 0,6442            |                   | SAR = SDM          |                    | 0,0492             |                    | SAR $\neq$ SDM     |                    |
| $\theta = -\lambda \beta$           | 0,1653             |                    | SEM = SDM          |                    | 0,0683             |                    | SEM $\neq$ SDM     |                    | 0,8946            |                   | SEM = SDM          |                    | 0,1514             |                    | SEM = SDM          |                    |

Fonte: elaborado pelos autores. Nota: Nível de significância \*\*\*<0.01; \*\*<0.05% e \*<10%. Erro padrão clusterizado (RGI).

Na Tabela 6 estão os resultados das regressões espaciais para os cluster com restrição de contiguidade. Percebe-se que na média as RGI estão convergindo com seus respectivos clusters em habilidades cognitivas (SEM) e não diferença entre os clusters, exceto para o cluster 5.

Já para as habilidades sociais (SDM), as RGI do cluster 3 estão convergindo entre si e a convergência das RGI dos clusters 2 e 4 não são diferentes das RGI dos cluster 3. Entretanto, as RGI do cluster 1 estão convergindo entre si e sua convergência é diferente da região de referência. O cluster 5 não converge e é diferente do cluster de referência. Para as habilidades motoras (SAR) os resultados são iguais aos resultados das habilidades sociais, exceto para o cluster 2 que é estatisticamente diferente do cluster de 3. No que se refere as habilidades médias (SEM) os resultados são análogos aos das habilidades sociais (SDM).

### 4.3 Clubes via quintil de emprego

A separação dos clubes de acordo com os quintis de emprego formal na indústria de transformação privada, foram divididos em cinco quintis em que cada quintil representam as regiões que contemplam cerca de 20% do emprego em 2019. Logo, o número de emprego formal na indústria é fixo, porém o número de RGI varia de acordo com cada quintil. A Tabela 7 mostra a distribuição de RGI por quintil.

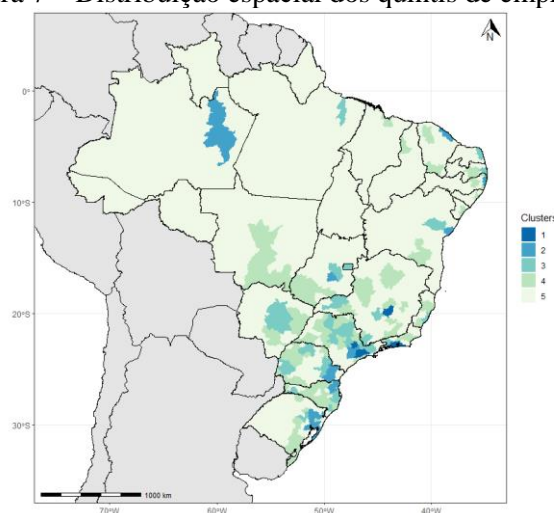
Tabela 7 – Quintil de emprego formal na indústria privada

| Cluster | RGI | (%)  | Acumulado (%) | Emprego   | (%)  | Acumulado (%) |
|---------|-----|------|---------------|-----------|------|---------------|
| 1       | 4   | 0,78 | 0,78          | 1.497.425 | 20,8 | 20,8          |
| 2       | 13  | 2,55 | 3,33          | 1.482.051 | 20,6 | 41,3          |
| 3       | 35  | 6,86 | 10,19         | 1.462.381 | 20,3 | 61,6          |
| 4       | 77  | 15,1 | 25,26         | 1.442.890 | 20,0 | 81,6          |
| 5       | 381 | 74,7 | 100,00        | 1.322.420 | 18,4 | 100           |
| Total   | 510 | 100  |               | 7.207.167 | 100  | -             |

Fonte: elaborados pelos autores com base nos microdados da RAIS (2019).

A análise do quintil mostra como o emprego na indústria é concentrado no Brasil, aproximadamente 25% das RGI concentram 82% do emprego formal da indústria de transformação. Por outro lado, 75% das RGI concentram apenas 18% do emprego. A distribuição espacial é mostrada na Figura 7.

Figura 7 – Distribuição espacial dos quintis de emprego



Fonte: elaborado pelos autores.

A título de comparação com os outros métodos de regionalização, os clubes foram definidos com base no quintil de emprego formal da indústria de transformação privada. Os resultados são semelhantes aos apresentados via *machine learning*, ou seja, para as habilidades cognitivas (SEM) as RGI do cluster 4 e 5 convergem entre si, as demais RGI não convergem entre si e são diferentes estatisticamente em relação ao cluster de referência.



Tabela 8 – Análise de convergência condicional das habilidades das ocupações da indústria, clubes, quintil

| Var.                                | Cognitivas         |                    |                    |                    | Sociais            |                    |                    |                    | Motoras            |                    |                    |                    | Médias             |                    |                    |                    |
|-------------------------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
|                                     | OLS                | SDM                | SAR                | SEM                | OLS                | SDM                | SAR                | SEM                | OLS                | SDM                | SAR                | SEM                | OLS                | SDM                | SAR                | SEM                |
| $\beta$                             | -0,410***<br>0,029 | -0,413***<br>0,029 | -0,409***<br>0,029 | -0,411***<br>0,029 | -0,458***<br>0,029 | -0,459***<br>0,029 | -0,458***<br>0,029 | -0,458***<br>0,029 | -0,514***<br>0,038 | -0,514***<br>0,038 | -0,514***<br>0,038 | -0,516***<br>0,038 | -0,415***<br>0,028 | -0,418***<br>0,028 | -0,414***<br>0,028 | -0,416***<br>0,028 |
| $\beta C_1$                         | 0,559**<br>0,222   | 0,121<br>0,287     | 0,570***<br>0,211  | 0,569***<br>0,207  | 0,978***<br>0,278  | 0,342<br>0,274     | 0,983***<br>0,276  | 0,970***<br>0,276  | 0,844***<br>0,17   | 0,666***<br>0,238  | 0,852***<br>0,172  | 0,877***<br>0,168  | 0,659**<br>0,26    | 0,067<br>0,244     | 0,667***<br>0,252  | 0,656***<br>0,251  |
| $\beta C_2$                         | 0,736***<br>0,168  | 0,487**<br>0,23    | 0,731***<br>0,168  | 0,686***<br>0,176  | 1,333***<br>0,208  | 0,913***<br>0,232  | 1,335***<br>0,208  | 1,297***<br>0,212  | 0,747***<br>0,192  | 0,677***<br>0,2161 | 0,750***<br>0,192  | 0,744***<br>0,193  | 0,886***<br>0,168  | 0,552**<br>0,218   | 0,883***<br>0,168  | 0,829***<br>0,175  |
| $\beta C_3$                         | 0,459**<br>0,226   | 0,386*<br>0,212    | 0,452**<br>0,225   | 0,429*<br>0,224    | 0,997***<br>0,234  | 0,847***<br>0,23   | 0,994***<br>0,234  | 0,970***<br>0,234  | 0,479***<br>0,159  | 0,424***<br>0,152  | 0,481***<br>0,159  | 0,475***<br>0,156  | 0,548**<br>0,222   | 0,449**<br>0,209   | 0,540**<br>0,221   | 0,510**<br>0,22    |
| $\beta C_4$                         | 0,333<br>0,206     | 0,326<br>0,259     | 0,332<br>0,259     | 0,330<br>0,26      | 0,879***<br>0,241  | 0,817***<br>0,237  | 0,876***<br>0,241  | 0,870***<br>0,241  | 0,630***<br>0,16   | 0,612***<br>0,159  | 0,630***<br>0,16   | 0,630***<br>0,16   | 0,454*<br>0,241    | 0,433*<br>0,239    | 0,452*<br>0,241    | 0,448*<br>0,241    |
| $\rho$                              |                    | 0,050***<br>0,0168 | 0,033**<br>0,0154  |                    |                    |                    | 0,025*<br>0,015    | 0,016<br>0,014     |                    |                    | 0,037*<br>0,02     | 0,020<br>0,036     |                    | 0,050***<br>0,015  | 0,034*<br>0,014    |                    |
| $\lambda$                           |                    |                    |                    | 0,051***<br>0,0169 |                    |                    |                    | 0,026<br>0,0154    |                    |                    |                    | 0,082**<br>0,033   |                    |                    |                    | 0,050***<br>0,016  |
| $\theta$                            | Não                | Sim                | Não                | Não                | Não                | Sim                | Não                | Não                | Não                | Sim                | Não                | Não                | Não                | Sim                | Não                | Não                |
| Obs.                                | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               | 6552               |
| R2                                  | 0,235              | 0,223              | 0,219              | 0,220              | 0,249              | 0,238              | 0,236              | 0,236              | 0,266              | 0,257              | 0,256              | 0,256              | 0,235              | 0,223              | 0,220              | 0,221              |
| LogLik                              | 13847,01           | 13856,34           | 13849,62           | 13852,00           | 11771,66           | 11776,74           | 11772,29           | 11773,09           | 13657,28           | 13660,97           | 13657,42           | 13659,37           | 15394,06           | 15404,08           | 15396,87           | 15399,31           |
| AIC                                 | -27660,01          | -27688,67          | -27685,23          | -27690,65          | -23509,33          | -23529,47          | -23530,57          | -23532,18          | -27280,56          | -27297,93          | -27300,84          | -27304,74          | -30754,13          | -30784,16          | -30779,75          | -30784,62          |
| BIC                                 | -27544,62          | -27607,22          | -27637,72          | -27643,13          | -23393,94          | 23448,02           | -23483,06          | -23484,67          | -27165,17          | -27216,48          | -27253,33          | -27257,22          | -30638,74          | -30702,71          | -30732,23          | -30737,11          |
| LMI                                 | 10,188***          |                    |                    |                    | 6,0577**           |                    |                    |                    | 1,1300             |                    |                    |                    | 12,553***          |                    |                    |                    |
| RLMI                                | 2,5679             |                    |                    |                    | 0,28856            |                    |                    |                    | 3,5362*            |                    |                    |                    | 2,3905             |                    |                    |                    |
| LMe                                 | 17,195***          |                    |                    |                    | 8,3937***          |                    |                    |                    | 4,1441**           |                    |                    |                    | 20,058***          |                    |                    |                    |
| RLMe                                | 9,5752***          |                    |                    |                    | 2,6245             |                    |                    |                    | 6,5504**           |                    |                    |                    | 9,8949***          |                    |                    |                    |
| Restrição                           | Prob>chi2          |                    |                    |                    | Prob>chi2          |                    |                    |                    | Prob>chi2          |                    |                    |                    | Prob>chi2          |                    |                    |                    |
| $\theta = 0 \text{ e } \rho \neq 0$ | 0,0833             | SAR $\neq$ SDM     |                    |                    | 0,0887             | SAR $\neq$ SDM     |                    |                    | 0,6326             | SAR = SDM          |                    |                    | 0,0338             | SAR $\neq$ SDM     |                    |                    |
| $\theta = -\lambda \beta$           | 0,2168             | SEM = SDM          |                    |                    | 0,1185             | SEM = SDM          |                    |                    | 0,8558             | SEM = SDM          |                    |                    | 0,0800             | SEM $\neq$ SDM     |                    |                    |

Fonte: elaborado pelos autores. Nota: Nível de significância \*\*\*<0.01; \*\*<0.05% e \*<10%. Erro padrão clusterizado (RGI).

Para as habilidades sociais (SEM) há convergência nas RGI do cluster 5 e para os demais clusters não. No que se refere às habilidades motoras (SAR), os resultados são os mesmos dos anteriores, exceto para as regiões do cluster 3 e, por fim, para as habilidades médias (SDM) os resultados também são semelhantes aos apresentados para as habilidades sociais.

Em suma, as regiões menos industrializadas estão convergindo dentro dos seus clusters com velocidades maiores para as habilidades motoras, seguidas das sociais e das cognitivas. Entretanto, as regiões mais industrializadas não estão convergindo dentro dos seus clusters. Quando se compara as diferenças entre os clusters percebe-se que existem diferenças entre as regiões menos industrializadas para as tradicionalmente industrializadas. Esses resultados se alteram quando se compara os clusters *via machine learning* e quintil com os clusters com restrição de contiguidade. Não há diferenças significativas entre os clusters *via machine learning* e *via quintil* para a amostra desse estudo. Esses resultados eram previstos uma vez que os clusters *via machine learning* são mais eficientes com volumes significantes de dados (*Big Data*). Logo, um simples método em um conjunto de dados reduzidos pode ser uma estratégia mais eficiente. Caso contrário, recomenda-se o uso das técnicas de *machine learning*, inclusive para detectar *outliers*, como o caso da RGI de SP para este estudo.

Embora as indústrias estejam se desconcentrando espacialmente e apresente algumas heterogeneidades, como apontado em Souza e Diniz (2020), o nível de habilidades da indústria também segue um processo heterogêneo. Para os clusters *via machine learning* e quintil, as evidências mostram que as regiões menos industrializadas estão ficando mais parecidas entre si, por outro lado, as demais não estão convergindo. Para o cluster com restrição de contiguidade, apenas a RGI de São Paulo não está convergindo, as demais estão convergindo e ficando iguais estatisticamente. Logo, percebe-se a disparidade da demanda por habilidades na indústria entre as RGI. As possíveis causas podem estar relacionadas ao tipo de indústria nessas regiões e suas respectivas produtividades como apontado em Schettini e Azzoni (2013). Em geral as evidências indicam que a indústria está convergindo mais rápido para habilidades motoras, resultado semelhante ao estudo de Castro e Azzoni (2022) para todos os setores, isso implica em ocupações menos complexas, com produtividade menor e retorno salarial menor, conforme apontado também por Neves, Azzoni e Chagas (2018).

Se há evidência de que a complementariedade entre habilidades e estoque de capital influenciam a produtividade, como apontado por Dawid et al. (2013), a produtividade da indústria nas RGI poderia estar sendo influenciada por habilidades menos complexas como apontam as evidências deste estudo. Esses resultados poderiam estar transbordando também para o processo de inovação na indústria, conforme as evidências empíricas indicam (Bassanini e Scarpetta 2002; Griffith et al. 2004). Entretanto, é necessária uma investigação mais profunda para esses casos. De qualquer forma, a dinâmica de convergência das habilidades na indústria de transformação privada depende das relações espaciais e não espaciais e os métodos de regionalização apresentados neste estudo podem servir como um guia para encontrar formas de regionalização da indústria de maneira mais eficiente a nível *stricto sensu*.

## CONSIDERAÇÕES FINAIS

Este estudo analisou a convergência das habilidades nas ocupações formais na indústria de transformação privada no Brasil, utilizando os métodos de regionalização *via machine learning*, bem como, uma combinação de cluster por similaridade com restrição de contiguidade. Também foram utilizados os clusters dos quintis de emprego na indústria.

Os resultados mostraram que as regiões menos industrializadas estão convergindo dentro dos seus clusters e as regiões mais industrializadas não estão convergindo dentro dos seus clusters. Quando se compara as diferenças entre os clusters percebe-se que existem diferenças entre as regiões menos industrializadas para as tradicionalmente industrializadas. Esses resultados se alteram quando se compara os clusters *via machine learning* e quintil com os clusters com restrição de contiguidade.

Há de se ressaltar que os métodos de regionalização apresentados neste estudo podem servir como um guia para encontrar formas de agrupamento da indústria de maneira mais eficiente quando não há uma estrutura teórica *a priori*. Para pesquisas futuras, recomenda-se testar novos métodos de regionalização a

partir de uma amostra espaço-temporal e avaliar a influência das habilidades na produtividade e nas inovações da indústria de transformação nas regiões brasileiras.

## REFERÊNCIAS

- ALMEIDA, Eduardo. *Econometria espacial*. Campinas–SP. Alínea, 2012.
- ARBIA, G. Spatial econometrics: statistical foundations and applications to regional convergence. Springer Science & Business Media, 2006.
- ARROW, K. J. The economic implications of learning by doing. In: *Readings in the Theory of Growth*. Palgrave Macmillan, London, 1971. p. 131-149.
- BARRO, R. J. Economic growth in a cross section of countries. *The quarterly journal of economics*, v. 106, n. 2, p. 407-443, 1991.
- BARRO, R. J.; SALA-I-MARTIN, X.; BLANCHARD, O. J.; HALL, R. E. Convergence across states and regions. *Brookings papers on economic activity*, 107-182, 1991.
- BARRO, R. J.; SALA-I-MARTIN, X. Convergence. *Journal of political Economy*, v. 100, n. 2, p. 223-251, 1992.
- CASTRO, G. H. L.; AZZONI, C. R. Sofisticação da demanda por mão de obra nas regiões do estado do Paraná. *Revista Paranaense de Desenvolvimento*, v. 42, n. 140, 2021.
- CHAVENT, M.; KUENTZ-SIMONET, V.; LABENNE, A.; SARACCO, J. ClustGeo: an R package for hierarchical clustering with spatial constraints. *Computational Statistics*, 33(4), 1799-1822, 2018.
- CORRADO, L.; FINGLETON, B. Where is the economics in spatial econometrics?. *Journal of Regional Science*, v. 52, n. 2, p. 210-239, 2012.
- DAVIES, D.; BOULDIN, D. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 1(2): 224-227, 1979.
- DAWID, H.; GEMKOW, S.; HARTING, P.; NEUGART, M. Labor market integration policies and the convergence of regions: the role of skills and technology diffusion. In: *The Two Sides of Innovation*. Springer, 2013. p. 167-186.
- DURANTON, G. E PUGA, D. From sectoral to functional urban specialisation. *Journal of urban Economics*, 57(2):343–370, 2005.
- BASSANINI, A.; SCARPETTA, S. Does human capital matter for growth in OECD countries? A pooled mean-group approach. *Economics letters*, v. 74, n. 3, p. 399-405, 2002.
- BAUMOL, W.J. Productivity growth, convergence, and welfare: what the long-run data show. *The american economic review*, p. 1072-1085, 1986.
- ELHORST, J. P. Applied spatial econometrics: raising the bar. *Spatial economic analysis*, v. 5, n. 1, p. 9-28, 2010.
- GRIFFITH, R.; REDDING, S.; REENEN, J. V. Mapping the two faces of R&D: Productivity growth in a panel of OECD industries. *Review of economics and statistics*, v. 86, n. 4, p. 883-895, 2004.
- IBGE (2017) Divisão regional do Brasil em regiões geográficas imediatas e regiões geográficas intermediárias/ IBGE, Coordenação de Geografia. - Rio de Janeiro: 82 p.
- ISARD, W et al. *Methods of regional and interregional analysis*, Ashgate: Aldershot, 1998.
- ISARD, W. *Methods of regional analysis*. Cambridge: MIT Press, 1960.
- ISLAM, N. What have we learnt from the convergence debate? *Journal of economic surveys*, v. 17, n. 3, 2003.
- KOPCZEWSKA, K. Spatial machine learning: new opportunities for regional science. *The Annals of Regional Science*, v. 68, n. 3, p. 713-755, 2022.
- LESAGE, J.; PACE, R. K. *Introduction to spatial econometrics*. Chapman and Hall/CRC, 2009.
- MACIENTE, A. N. *The determinants of agglomeration in Brazil: input-output, labor and knowledge externalities*. 2013. 282 p. PhD Dissertation, University of Illinois at Urbana Champaign, 2013.
- NEVES, E. C.; AZZONI, C. R.; CHAGAS, A. S. Habilidades e prêmio salarial urbano. *Pesquisa e Planejamento Econômico*, v. 51, n. 2, 2021.
- ROMER, P. M. Increasing returns and long-run growth. *Journal of political economy*, v. 94, n. 5, p. 1002-1037, 1986.
- ROUSSEUW, P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Computational Applied Mathematics*. 20(1): 53-65, 1987.
- SCHETTINI, D.; AZZONI, C. R. Diferenciais regionais de competitividade industrial do Brasil no século 21. *Economia*, v. 14, n. 1B, 2013.
- SILVA, A. L. F.; DINIZ, M. B. Padrões de concentração regional da indústria de transformação brasileira. *Nova Economia*, v. 30, p. 407-454, 2020.

SOUZA, M. C. C. *Uma análise do algoritmo K-Means como introdução ao aprendizado de máquinas*. Trabalho de conclusão de curso (Graduação). Universidade Federal do Tocantins. Palmas- TO, Brasil, 2019.

STAKHOVYCH, S; BIJMOLT, T. H.A. Specification of spatial models: A simulation study on weights matrices. *Papers in Regional Science*, v. 88, n. 2, p. 389-408, 2009.

VOPHAM, T.; HART, J.E.; LADEN, F.; CHIANG, Y.Y. Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology. *Environ Health* 17(1):1–6, 2018.

ZHANG, X; YU, J. Spatial weights matrix selection and model averaging for spatial autoregressive models. *Journal of Econometrics*, v. 203, n. 1, p. 1-18, 2018.

## APÊNDICE

Tabela A1 – Atributos para as habilidades queridas para as ocupações (CBO)

| Cognitivas  | Ocupação*                            | Sociais                              | Ocupação*                            | Motoras   | Ocupação*                                       |
|---|--------------------------------------|--------------------------------------|--------------------------------------|---|---|
| -Analisar dados;<br>-Resolução de problemas complexos;<br>... | -Físico;<br>-Médico Cirurgião<br>... | -Coordenação;<br>-Negociação;<br>... | -Educação;<br>-Diretor de RH;<br>... | -Força dinâmica;<br>-Manuseio de objetos;<br>...  | -Preparador de Estruturas;<br>-Mecânico;<br>... |
| -Raciocínio matemático;<br>-Ordenamento de informação;        | -Atendente;<br>-Garçom;              | -Instrução;<br>-Serviços;            | -Tecelão;<br>-Costureiro;            | -Flexibilidade do corpo;<br>-Equilíbrio corporal; | -Ouvidor;<br>-Matemático;                       |

Fonte: adaptado de Neves, Azzoni e Chagas (2021).

Nota: \* as notas variam de 0 a 1 (Ex: a nota para a habilidade cognitiva é igual a 1 para a ocupação de Físico e 0 para a ocupação de Garçom).

Tabela A2 - Efeitos marginais (direto, indireto e total) das regressões

|                 | ML (k = 3)         |                    | ML (k = 6)         |                    |                    | Restrição de Contiguidade |                     | Quartil            |                    |                    |
|-----------------|--------------------|--------------------|--------------------|--------------------|--------------------|---------------------------|---------------------|--------------------|--------------------|--------------------|
|                 | Cognitivo (SDM)    | Médias (SDM)       | Cognitivo (SAR)    | Sociais (SDM)      | Motoras (SAR)      | Média (SDM)               | Sociais (SDM)       | Motoras (SAR)      | Motoras (SAR)      | Médias (SDM)       |
| <i>Direto</i>   |                    |                    |                    |                    |                    |                           |                     |                    |                    |                    |
| C <sub>1</sub>  | -0,405***<br>0,027 | -0,409***<br>0,027 | 0,603***<br>0,23   | 0,949***<br>0,216  | 0,839***<br>0,164  | -0,177*<br>0,100          | -1,0356***<br>0,377 | -0,510*<br>0,26    | 0,842***<br>0,171  | 0,049<br>0,239     |
| C <sub>2</sub>  | -0,592*<br>0,356   | -0,500<br>0,310    | 0,645***<br>0,179  | 0,033<br>0,355     | 0,745***<br>0,154  | 0,483**<br>0,231          | -0,368**<br>0,032   | -0,416**<br>0,041  | 0,759***<br>0,189  | 0,578***<br>0,211  |
| C <sub>3</sub>  | 0,314<br>0,194     | 0,331*<br>0,189    | 0,808***<br>0,188  | 0,873***<br>0,241  | 0,796***<br>0,217  | -0,313*<br>0,159          | -0,350<br>0,334     | -0,314<br>0,195    | 0,477***<br>0,163  | 0,456**<br>0,212   |
| C <sub>4</sub>  |                    |                    | 0,758***<br>0,201  | 0,395<br>0,272     | 0,945***<br>0,164  | -0,041<br>0,132           | -0,146<br>0,251     | 0,134<br>0,139     | 0,627***<br>0,157  | 0,433*<br>0,233    |
| C <sub>5</sub>  |                    |                    | -0,419**<br>0,031  | 0,466**<br>0,031   | -0,521***<br>0,039 | -0,411**<br>0,027         | 0,380**<br>0,193    | 0,345**<br>0,143   | -0,513***<br>0,039 | -0,416**<br>0,029  |
| C <sub>6</sub>  |                    |                    | 0,599***<br>0,212  | 1,056**<br>0,231   | 0,477***<br>0,152  | -0,424***<br>0,132        |                     |                    |                    |                    |
| <i>Indireto</i> |                    |                    |                    |                    |                    |                           |                     |                    |                    |                    |
| C <sub>1</sub>  | 0,266<br>0,02      | 0,023<br>0,021     | 0,029*<br>0,016    | 0,573<br>0,384     | 0,019<br>0,034     | -0,116<br>0,21            | 0,209<br>0,452      | -0,016<br>0,024    | 0,020<br>0,033     | -0,380<br>1,359    |
| C <sub>2</sub>  | -7,989*<br>3,714   | -6,137**<br>2,912  | 0,031**<br>0,015   | -3,918**<br>1,791  | 0,019<br>0,034     | 4,806***<br>1,519         | -0,036<br>0,031     | -0,012<br>0,016    | 0,018<br>0,03      | 1,486**<br>0,609   |
| C <sub>3</sub>  | 1,926***<br>0,697  | 2,235***<br>0,655  | 0,039**<br>0,018   | 1,182***<br>0,693  | 0,018<br>0,031     | -1,433***<br>0,434        | 0,318<br>0,256      | -0,008<br>0,014    | 0,011<br>0,019     | 0,995*<br>0,563    |
| C <sub>4</sub>  |                    |                    | 0,037**<br>0,0179  | 1,498<br>1,033     | 0,019<br>0,034     | -1,014<br>1,047           | 0,788***<br>0,256   | 0,004<br>0,008     | 0,014<br>0,024     | 0,133<br>0,308     |
| C <sub>5</sub>  |                    |                    | -0,020***<br>0,007 | 0,004<br>0,024     | -0,012<br>0,021    | 0,024<br>0,022            | -0,573<br>1,96      | 0,011<br>0,015     | -0,012<br>0,019    | 0,019<br>0,022     |
| C <sub>6</sub>  |                    |                    | 0,029*<br>0,015    | 0,933*<br>0,565    | 0,011<br>0,02      | -0,003<br>0,436           |                     |                    |                    |                    |
| <i>Total</i>    |                    |                    |                    |                    |                    |                           |                     |                    |                    |                    |
| C <sub>1</sub>  | -0,378***<br>0,033 | 0,385***<br>0,032  | 0,632***<br>0,033  | 1,521***<br>0,478  | 0,858***<br>0,043  | -0,293<br>0,0237          | -0,827<br>0,602     | -0,526*<br>0,272   | 0,863***<br>0,0179 | -0,331<br>1,401    |
| C <sub>2</sub>  | -8,58**<br>3,89    | -6,637**<br>3,073  | 0,632***<br>0,242  | -3,88*<br>1,989    | 0,763***<br>0,165  | 5,289***<br>1,643         | -0,404***<br>0,044  | -0,428***<br>0,044 | 0,863***<br>0,179  | 2,065***<br>0,583  |
| C <sub>3</sub>  | 2,240***<br>0,66   | 2,565***<br>0,614  | 0,847***<br>0,199  | 2,689***<br>0,657  | 0,815***<br>0,219  | -1,746***<br>0,452        | -0,033<br>0,363     | -0,322<br>0,200    | 0,489***<br>0,168  | 1,452**<br>0,595   |
| C <sub>4</sub>  |                    |                    | 0,795***<br>0,213  | 1,89*<br>1,053     | 0,968***<br>0,176  | -1,055<br>1,091           | 0,641**<br>0,312    | 0,137<br>0,143     | 0,641***<br>0,161  | 0,565<br>0,396     |
| C <sub>5</sub>  |                    |                    | -0,439**<br>0,033  | -0,461***<br>0,039 | -0,533<br>0,043    | -0,387***<br>0,036        | -0,193<br>2,028     | 0,356**<br>0,15    | -0,525***<br>0,043 | -0,398***<br>0,038 |
| C <sub>6</sub>  |                    |                    | 0,628***<br>0,223  | 1,989**<br>0,607   | 0,488***<br>0,158  | -0,428<br>0,425           |                     |                    |                    |                    |

Fonte: elaborado pelos autores. Nota: Nível de significância \*\*\*<0.01; \*\*<0.05 e \*<10%. Erro padrão clusterizado (RGI).