



Integração das Redes Globais de Inovação e Produção: Padrões Setoriais da Manufatura Brasileira em 2017

Caroline Giusti de Araújo¹;
Jailson dos Santos Silva²;

Resumo: Este artigo investiga como produção e inovação se articulam na inserção internacional da indústria de transformação brasileira, dialogando com as literaturas de Redes Globais de Produção (GPN/GVC) e Redes Globais de Inovação (GIN). A literatura tem avançado na análise da governança produtiva e das redes globais de conhecimento, mas ainda carece de tipologias integradas que combinem inserção produtiva global, intensidade inovativa e densidade relacional no nível setorial, especialmente em economias de industrialização tardia. Parte-se da hipótese de que, nesses contextos, produção e inovação são dimensões transversais à inserção internacional, porém sua intensidade e articulação variam conforme as características estruturais dos setores, resultando em padrões diferenciados de coordenação produtiva e tecnológica. O objetivo é identificar e qualificar esses padrões na indústria brasileira em 2017, por meio da proposição de uma tipologia denominada GPIN (“Global Production Innovation Network”). A análise utiliza dados da PINTEC 2017 combinados a informações de produção, exportações e investimento direto externo, aplicando análise fatorial para redução dimensional, seguida de clusterização hierárquica para identificação de agrupamentos setoriais e, em etapa complementar, técnicas supervisionadas baseadas em árvores (CART, Random Forest e XGBoost) para qualificar variáveis exógenas associadas aos clusters. Os resultados revelam cinco padrões distintos de articulação entre produção e inovação e indicam que a indústria brasileira apresenta estrutura fragmentada, marcada pela coexistência de setores fortemente integrados a cadeias globais coordenadas por grandes grupos empresariais e setores predominantemente ancorados no mercado doméstico. Nos segmentos intensivos em IDE, o “upgrading” tecnológico mostra-se condicionado à estratégia das matrizes, enquanto outros grupos exibem dinamismo inovativo mais doméstico e incremental. De forma transversal, destacam-se como determinantes relevantes o impacto da inovação na abertura de novos mercados e o uso de redes informatizadas como fontes de informação. O artigo contribui ao integrar empiricamente GVC/GPN e GIN em uma tipologia aplicada ao caso brasileiro, evidenciando que a articulação entre produção e inovação é heterogênea e estruturalmente condicionada, com implicações para o desenho de políticas diferenciadas; como limitação, destaca-se o uso de dados agregados setoriais, sugerindo avanços futuros com microdados de firma.

Palavras-chave: Redes Globais de Inovação; Redes Globais de Produção; Indústria; Inovação; Brasil.

¹ Doutora em Política Científica e Tecnológica (Unicamp/DPCT), Mestre em Desenvolvimento Econômico (Unicamp/IE), graduada em Ciências Econômicas (UFSCar). E-mail: carolgiustiaraujo@gmail.com

² Professor da Universidade do Estado de Santa Catarina - UDESC/CCT - Campus de Joinville. Doutor em Engenharia de Produção (UFSC). E-mail: engjailsonsantos@outlook.com

Código JEL: O36, O54, L14

Área Temática: Comércio internacional, cadeias de valor e internacionalização

Integrating Global Production and Innovation: Sectoral Patterns in Brazilian Manufacturing for 2017

Abstract: This article investigates how production and innovation are articulated in the international insertion of the Brazilian manufacturing industry, engaging with the literature on Global Production Networks (GPN/GVC) and Global Innovation Networks (GIN). The literature has advanced in the analysis of productive governance and global knowledge networks, but still lacks integrated typologies that combine global productive insertion, innovative intensity, and relational density at the sectoral level, especially in late-industrializing economies. The hypothesis is that, in these contexts, production and innovation are transversal dimensions to international insertion, but their intensity and articulation vary according to the structural characteristics of the sectors, resulting in differentiated patterns of productive and technological coordination. The objective is to identify and qualify these patterns in Brazilian industry in 2017, through the proposition of a typology called GPIN (Global Production Innovation Network). This analysis uses data from PINTEC 2017 combined with information on production, exports, and foreign direct investment, applying factor analysis for dimensional reduction, followed by hierarchical clustering to identify sectoral groupings and, in a complementary step, supervised tree-based techniques (CART, Random Forest, and XGBoost) to qualify exogenous variables associated with the clusters. The results reveal five distinct patterns of articulation between production and innovation and indicate that Brazilian industry presents a fragmented structure, marked by the coexistence of sectors strongly integrated into global chains coordinated by large business groups and sectors predominantly anchored in the domestic market. In FDI-intensive segments, technological upgrading is conditioned by the strategy of the parent companies, while other groups exhibit more domestic and incremental innovative dynamism. Across the board, the impact of innovation on opening new markets and the use of computer networks as sources of information stand out as relevant determinants. This article contributes by empirically integrating GVC/GPN and GIN into a typology applied to the Brazilian case, highlighting that the articulation between production and innovation is heterogeneous and structurally conditioned, with implications for the design of differentiated policies; as a limitation, the use of aggregated sectoral data stands out, suggesting future advances with firm microdata.

Keywords: Global Innovation network; Global Production Network; Industry; Innovation; Brazil.

1. Introdução

A integração global tem sido apontada como uma oportunidade para que países em desenvolvimento ampliem suas capacidades produtivas e inovativas, inserindo-se em cadeias globais e redes internacionais de conhecimento. Contudo, a experiência recente demonstra que a participação em fluxos globais de produção e inovação não garante, por si só, aprendizado tecnológico ou reposicionamento estrutural. A natureza dessa inserção depende da articulação entre produção, inovação e capacidades domésticas.

No contexto brasileiro, observa-se simultaneamente um aumento do esforço para desenvolver estrutura produtiva e inovativa e um crescimento das relações de cooperação para inovação. Dados da PINTEC indicam que o percentual de empresas que cooperam em relação às que inovam passou de 3,75% (2003) para 15,56% (2017), evidenciando a intensificação das estratégias colaborativas no período. Esse movimento reforça a relevância de compreender como os setores industriais combinam inserção produtiva global, articulação inovativa e densidade relacional.

Apesar da ampla literatura sobre cadeias globais de valor e redes globais de inovação, ainda há escassez de análises que integrem sistematicamente, no nível setorial, as dimensões produtiva e inovativa da inserção internacional em economias de industrialização tardia. A maior parte dos estudos enfatiza ou a governança produtiva e o “upgrading” nas cadeias globais, ou as redes internacionais de colaboração para inovação, sem explorar de forma integrada a interação entre essas dimensões.

Nesse sentido, este trabalho parte da hipótese de que, em economias de industrialização tardia, produção e inovação constituem dimensões transversais à inserção internacional, mas sua intensidade e articulação variam segundo as características estruturais dos setores, resultando em padrões diferenciados de coordenação produtiva e tecnológica.

O objetivo geral do trabalho é obter e qualificar o padrão global de inovação e produção nos setores selecionados da indústria de transformação brasileira em 2017. Especificamente, busca-se: (i) propor uma taxonomia GPIN (“Global Production Innovation Network”), integrando inserção produtiva global, intensidade inovativa e densidade relacional; e (ii) qualificar os padrões identificados a partir de variáveis exógenas à clusterização — como impacto de mercado, capacidade de absorção, enquadramento regulatório, apoio governamental, pertencimento a grupo empresarial e fontes de informação.

Ao integrar essas dimensões, o artigo pretende contribuir para o debate sobre heterogeneidade estrutural e reposicionamento produtivo em economias emergentes, identificando como diferentes setores combinam escala produtiva, articulação inovativa e inserção internacional.

2. Produção, inovação e inserção internacional e economias de industrialização tardia

A cooperação entre organizações constitui uma das atividades centrais do processo de inovação, elemento-chave da dinâmica competitiva capitalista, conforme destacado por Schumpeter (1961). A relação entre cooperação e inovação tornou-se ainda mais direta no contexto da inovação aberta, na qual a firma interage sistematicamente com seu ambiente externo para acessar e combinar conhecimentos (Chesbrough, 2003). Nesse sentido, a globalização ampliou significativamente as relações interfirmas e interinstitucionais, exigindo que a compreensão do ambiente externo se desse crescentemente em perspectiva internacional (OCDE, 2008).

No campo das Cadeias Globais de Valor (GVC), a literatura demonstrou que a fragmentação internacional da produção reorganizou a divisão internacional do trabalho, permitindo que firmas e países se integrem a etapas específicas das cadeias produtivas (Gereffi, Humphrey e Sturgeon, 2005). A governança dessas cadeias — hierárquica, modular, relacional, cativa ou de mercado — condiciona os fluxos de conhecimento e as possibilidades de “upgrading”. Contudo, a participação em GVC não implica automaticamente aprendizado tecnológico ou captura crescente de valor, podendo resultar tanto em janelas de oportunidade quanto em trajetórias de “lock-in” estrutural. A inserção produtiva, portanto, não é neutra: sua natureza depende da posição ocupada nas cadeias e da capacidade doméstica de internalizar aprendizado.

Historicamente, a literatura de Redes Globais de Inovação (GIN) emerge como desdobramento da literatura de “Global Production Networks” (GPN), uma vez que a globalização da produção precedeu a globalização da inovação. A literatura de GPN enfatiza o espreadimento internacional da produção, caracterizado por dispersão geográfica combinada com integração sistêmica das atividades produtivas.

Nessa dinâmica, a localização das firmas combina atratividade de mercado, estrutura produtiva desenvolvida — especialmente base de fornecedores — e condições adequadas de pesquisa e desenvolvimento (Ernst, 2002; Cooke, 2013; Liu, 2017). Entretanto, a abordagem de GPN tende a privilegiar as relações produtivas e a dimensão do “upgrading” tecnológico nas cadeias, deixando em segundo plano as relações não estritamente produtivas, como a colaboração com universidades e institutos de pesquisa, tanto domésticos quanto internacionais.

A partir dessa lacuna, Barnard e Chaminade (2011) propõem a abordagem de Redes Globais de Inovação (GIN), estruturada em três dimensões: global (G), inovativa (I) e conectividade (N). A taxonomia de GIN demonstra que a inserção em redes globais pode fragilizar a relação tradicional entre munificência da localização e desempenho da firma, permitindo que empresas “fortes” emergjam de contextos institucionais e produtivos considerados “fracos”. Esse resultado dialoga com a literatura de inovação aberta (Chesbrough, 2003) e reforça a centralidade dos canais de colaboração e das redes transnacionais de conhecimento.

Contudo, a literatura de GIN enfatiza predominantemente a dimensão relacional e tecnológica da inserção internacional, sem necessariamente integrar de forma sistemática o peso produtivo, a escala industrial e a densidade estrutural dos setores. É justamente nesse ponto que a articulação entre GVC e sistemas nacionais de inovação se torna fundamental. Conforme discutido por Pietrobelli e Rabelotti (2020), a posição ocupada nas cadeias globais influencia o acesso a aprendizado, mas as capacidades tecnológicas domésticas condicionam o tipo de inserção produtiva possível. Estruturas produtivas nacionais, políticas industriais e sistemas de inovação moldam as possibilidades de “upgrading” e a capacidade de capturar valor nas cadeias globais. Produção e inovação, portanto, não constituem dimensões independentes: sua interação molda as trajetórias de desenvolvimento industrial.

Essa complementaridade é particularmente relevante para economias de industrialização tardia, nas quais a transformação estrutural envolve simultaneamente desafios de escala produtiva, adensamento industrial e construção de capacidades tecnológicas. A literatura de sistemas nacionais de inovação destaca a importância do “learning by doing”, “learning by using” e “learning by interacting” (Lundvall, 1992) como mecanismos centrais de aprendizado. Contudo, como argumentam Chang e Andreoni (2021), tais mecanismos possuem limites quando desvinculados da dinâmica produtiva concreta. A inovação não pode ser analisada de forma dissociada da estrutura produtiva: o “learning in production” constitui dimensão fundamental para sustentar trajetórias industriais robustas.

Nesse sentido, a inserção internacional deve ser compreendida como resultado da articulação entre produção e inovação. Em determinados setores, o desafio central pode residir na coordenação tecnológica e na integração a redes globais de P&D; em outros, o principal obstáculo pode estar associado à escala produtiva, à densidade industrial e ao adensamento doméstico das cadeias. A heterogeneidade estrutural setorial implica que a importância relativa das dimensões produtiva e inovativa não é uniforme ao longo da indústria.

Apesar dos avanços das literaturas de GVC/GPN e GIN, permanece uma lacuna no que se refere à análise integrada, no nível setorial, das dimensões produtiva e inovativa da inserção internacional. Enquanto GVC enfatiza governança produtiva e “upgrading” e GIN privilegia redes de inovação e colaboração, há escassez de tipologias que combinem sistematicamente inserção produtiva global, intensidade inovativa e densidade relacional para analisar padrões diferenciados de coordenação produtiva e tecnológica em economias tardias.

Neste contexto, este artigo propõe a taxonomia GPIN (“Global Production Innovation Network”), integrando três dimensões analíticas: inserção produtiva global, intensidade inovativa e densidade relacional. Ao combinar essas dimensões, busca-se captar padrões diferenciados de articulação entre produção e inovação na inserção internacional setorial.

3. Metodologia

3.1. Formação dos padrões: técnica de clusterização

O agrupamento requer o uso de uma medida de similaridade entre os objetos. Para isto, Hair et al. (2009) apresentam três possibilidades: 1) medidas correlacionais; 2) medidas de associação; e 3) medidas de distância, que são amplamente utilizadas na literatura. O trabalho fará uso de uma combinação de técnicas,

que requerem o cálculo da distância euclidiana, expressa conforme a eq. (1).

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2} \quad (1)$$

onde i e j são duas observações e a distância d_{ij} corresponde à raiz quadrada da soma dos quadrados das diferenças entre os pares de observações de i e j para todas as variáveis (k).

Segundo Hair et al. (2009), diferentes medidas de distância conduzem a diferentes agrupamentos sendo, por isso, aconselhável utilizar diversas medidas e comparar os resultados com padrões teóricos. Na sequência, deve-se escolher o procedimento de partição usado para formar agregados e decidir o número de agrupamentos formados. Para a partição, os autores avaliam que os algoritmos mais utilizados podem ser classificados em hierárquico e não hierárquico.

Os procedimentos **hierárquicos** representam uma série de $n - 1$ decisões de agrupamentos, em que n é o número de observações. Este procedimento é formado a partir de uma matriz de similaridade e pode ser agrupado através do método aglomerativo. Já os procedimentos **não-hierárquicos** designam observações aos agrupamentos pelas suas especificações. Neste caso, deve-se especificar inicialmente o número inicial de agrupamentos. Segundo Favero et al. (2009) entre os procedimentos não-hierárquicos, o mais popular é o k -médias (“ k -means”). Neste procedimento o critério de distância para a formação dos grupos é a distância euclidiana.

O k -médias possui três etapas: 1) partição dos indivíduos em “ k -clusters”, definidos previamente pelo pesquisador; 2) cálculo dos centroides para cada um dos “clusters” e cálculo da distância euclidiana dos centroides a cada observação na base de dados; 3) agrupar as observações aos “clusters” cujos centroides estejam mais próximos e repetir iterativamente o passo 2 até que não ocorra variação significativa na distância mínima de cada observação a cada um dos centroides dos “ k -clusters”.

Hair et al. (2009) sugerem que as técnicas hierárquicas e não-hierárquicas sejam combinadas. Isso é, primeiramente deve-se analisar o número de agrupamentos pelo procedimento hierárquico e, na sequência, utilizar esse número como pontos sementes no procedimento não-hierárquico. Favero et al. (2009) destacam a possível ineficácia do uso do k -médias quando a semente inicial é aleatória, ressaltando a importância de utilizar a combinação dos métodos. Na literatura, destaca-se Castro (2013) que optou pela combinação dos métodos hierárquicos e k -médias. No “software” R é possível utilizar a técnica híbrida por meio do pacote **hkmeans**, que combina o cluster hierárquico com o método “ k -means”. O procedimento envolve: i) calcular o agrupamento hierárquico; ii) cortar “ k -clusters”; iii) calcular o centro (média) de cada cluster; iv) fazer “ k -means” usando o conjunto de centro de “clusters” (definido na etapa anterior) como centros iniciais; v) otimizar o agrupamento.

Dessa forma, o particionamento obtido na etapa final pode ser diferente daquele da primeira etapa. Destaca-se que o procedimento envolve a escolha de um algoritmo, que foi o **Hartingan-Wong**. Esse algoritmo busca minimizar a função objetivo e aumentar a qualidade do agrupamento, em relação aos demais algoritmos disponíveis, ao distribuir os pontos iniciais de forma aleatória, sendo menos propenso a convergir para ótimos locais.

Por fim, sintetizam-se verificações para a clusterização (Kassambra, 2017; Everitt, 2011): Estatística de Hopkins (H): mede a probabilidade de um conjunto de dados ser gerado por uma distribuição uniforme; Coeficiente Silhouette (S_i): mede quão bem uma observação está clusterizada e estima a distância média entre “clusters”; e Índice de Dunn (D): mede a adequação da separação dos objetos entre clusters.

3.1.1. Dados utilizados para formação dos padrões (variáveis endógenas ao cluster)

A taxonomia de GIN foi adaptada, condicionada a disponibilidade de dados, para a construção dos padrões de cooperação para inovação do caso brasileiro, a nível setorial. Uma relação entre as variáveis originalmente propostas por Barnard e Chaminade (2011) e as variáveis utilizadas para a construção do padrão de cooperação brasileiro é apresentada na Quadro 1³.

³ Além das diferenças apresentadas no Quadro 1, Barnard e Chaminade (2011) construíram a taxonomia GIN a partir de um “survey” próprio aplicado em empresas de países desenvolvidos e em desenvolvimento. Já a construção dos padrões de cooperação brasileiro proposta neste trabalho será feita com base em setores considerando apenas o Brasil. Logo, apesar de ser uma inspiração, os resultados não serão comparáveis.

Quadro 1 - Adaptação da taxonomia GIN ao caso brasileiro*

GIN	Taxonomia GIN original	Taxonomia GPIN
Global (G)	Porcentagem das exportações em relação ao total das vendas; maiores mercados das empresas; localização geográfica dos parceiros e com quem as empresas colaboraram para inovar e a localização das empresas terceirizadas.	Investimento direto no país (IDE), categoria de investimento de maior destaque no relacionamento econômico e financeiro do Brasil com o resto do mundo Volume de exportações para o mundo
Inovativo (I)	Se as empresas inovaram entre 2006-2008 em termos de novos produtos, serviços, métodos melhorados de manufatura, logística e atividades de suporte.	O principal processo novo ou substancialmente aprimorado é novo para a empresas (V19_2_2), para o mercado nacional (V19_2_3) e para o mercado mundial (V19_2_4) O principal produto novo ou substancialmente aprimorado é novo para a empresas (V13_2_2), para o mercado nacional (V13_2_3) e para o mercado mundial (V13_2_4)
Network (N)	Como as empresas desempenham diferentes funções; com quem colaborou para o desenvolvimento das suas mais recentes e importantes inovações e se desenvolveu relações formais e informais com uma variedade de organizações externas.	Cooperação por tipos de parceiros & localizações do parceiro: os parceiros são agrupados para que os dados representem apenas se a parceria é nacional ou estrangeira. A cooperação por tipo de parceiro & objeto será designada da seguinte forma: agrupamentos dos parceiros para que o dado final represente apenas a parceira em P&D e ensaios
Production (P)	-	Variável de valor bruto da produção (VBP)

Fonte: PINTEC (2017), COMEX STAT (2017), BACEN (2017), PIA EMPRESA (2017)

Notas*: Todas as variáveis da PINTEC são expressas pelo número de empresas que declaram cada categoria. Além disso, todas as variáveis serão analisadas a nível de setores da CNAE, considerando 23 setores disponíveis em todas as bases.

3.2. Caracterização dos padrões por variáveis exógenas

3.2.1. Robustez e qualificação dos clusters: técnicas supervisionadas e modelo logístico

Com o objetivo de qualificar a interpretação dos clusters obtidos por técnicas não supervisionadas, foram aplicadas quatro abordagens complementares de classificação: (i) árvore de decisão (CART), (ii) florestas aleatórias, (iii) boosting (XGBoost) e (iv) modelo logístico. A lógica é progressiva: parte-se de um classificador interpretável (árvore), avança-se para métodos de comitê que reduzem variância e, por fim, para métodos sequenciais que tendem a elevar desempenho preditivo ao reduzir viés. O modelo logístico é utilizado como contraponto paramétrico, permitindo verificar consistência e sinalizar limitações quando a relação entre preditores e classe não é bem capturada por uma forma funcional linear (James et al., 2021; Wooldridge, 2010).

• Árvore de decisão (CART) e validação cruzada

A árvore de classificação busca segmentar a amostra em regiões internamente mais homogêneas quanto à variável resposta (pertencimento do setor ao cluster). Seguindo James et al. (2021), estimou-se uma árvore completa via divisão binária recursiva e, em seguida, aplicou-se poda por complexidade de custos (“cost-complexity pruning”), selecionando-se a subárvore ótima a partir do parâmetro de penalização α por validação cruzada k-fold. O parâmetro α controla o trade-off entre ajuste aos dados de treinamento e complexidade do modelo: valores mais altos de α penalizam árvores com muitos nós terminais, reduzindo risco de sobreajuste.

Para árvores de classificação, a qualidade da divisão foi avaliada por medidas usuais de impureza e erro. A taxa de erro de classificação em uma região m é dada por:

$$E = 1 - \max_k \widehat{p}_{mk} \quad (2)$$

Em que $\widehat{p}_{mk}$ representa a proporção de observações de treinamento m th regiões que são da k th classe. Adicionalmente, utilizou-se o índice de Gini como critério padrão de impureza dos nós:

$$G = \sum_{k=1}^K \widehat{p}_{mk}(1 - \widehat{p}_{mk}) \quad (3)$$

e a entropia como medida alternativa:

$$D = - \sum_{k=1}^K \widehat{p}_{mk} \log \widehat{p}_{mk} \quad (4)$$

Conforme James et al. (2021), embora diferentes métricas possam ser utilizadas na construção/poda, a **taxa de erro de classificação** é particularmente adequada quando o objetivo é maximizar a precisão

preditiva da árvore podada final.

O desempenho do classificador foi reportado por matriz de confusão e suas métricas derivadas (acurácia, sensibilidade e especificidade), além da curva ROC, que relaciona 1–1–1–especificidade (eixo x) e sensibilidade (eixo y) para diferentes pontos de corte, sintetizada pela área sob a curva (AUC-ROC).

Especificação utilizada. A árvore foi estimada com controle de hiperparâmetros para evitar sobreajuste e garantir comparabilidade entre clusters: custo de complexidade (cp), critério de parada (min) e profundidade máxima. Os valores finais foram definidos pelo melhor desempenho preditivo (James et al., 2021).

- **Florestas aleatórias: redução de variância e ganho de estabilidade**

Como árvores isoladas podem apresentar alta variância e instabilidade fora da amostra, aplicou-se bootstrap no conjunto de treinamento. Esse procedimento gera B amostras reamostradas e estima múltiplas árvores, combinando previsões para reduzir variância (James et al., 2021). No presente trabalho, adotou-se B=100, valor indicado como suficiente para ganhos de estabilidade sem induzir sobreajuste.

Na sequência, estimou-se uma floresta aleatória (random forest), que pode ser interpretada como uma extensão do bagging: além do bootstrap, a cada divisão de nó o algoritmo seleciona aleatoriamente um subconjunto de m preditores dentre ppp disponíveis. Essa aleatorização reduz a correlação entre árvores (sobretudo quando existe um preditor dominante que, no bagging, tenderia a aparecer nas primeiras divisões de muitas árvores), tornando a média das previsões mais estável e confiável (James et al., 2021). O parâmetro mtry (m) e o número de árvores foram definidos com base em desempenho via validação cruzada, com foco em AUC-ROC.

- **Boosting (XGBoost): redução de viés e melhoria de desempenho**

Para avançar além dos ganhos de estabilidade típicos de comitês baseados em bootstrap, aplicou-se boosting, método sequencial no qual cada árvore é ajustada para corrigir erros do modelo anterior. Diferentemente do bagging, em que árvores são estimadas em paralelo e depois agregadas, o boosting constrói árvores em sequência e atualiza iterativamente os resíduos, aprendendo de forma gradual e controlada (James et al., 2021). O parâmetro de redução (λ) desacelera o aprendizado e contribui para controlar sobreajuste, enquanto a profundidade de interação (d) regula a complexidade de cada árvore.

A implementação empregada foi o Extreme Gradient Boosting (XGBoost), calibrado por validação cruzada. Os hiperparâmetros considerados foram: **nrounds** (número de árvores sequenciais), **max_depth** (profundidade máxima), **gamma** (redução mínima de perda para realizar uma quebra), **eta** (taxa de aprendizado), **colsample_bytree** (fração de colunas por rodada), **min_child_weight** (mínimo de observações por folha) e **subsample** (percentual de amostragem no bootstrap). O desempenho foi comparado às técnicas anteriores com base principalmente em AUC-ROC e taxa de erro.

- **Síntese: por que combinar metodologias**

A combinação dessas técnicas é particularmente adequada ao objetivo do trabalho por duas razões. Primeiro, ela permite qualificar os clusters a partir de variáveis externas por diferentes lógicas de aprendizado estatístico: a árvore fornece transparência interpretativa; random forest e bootstrap aumentam estabilidade ao reduzir variância; boosting tende a melhorar desempenho ao reduzir viés e capturar não linearidades (James et al., 2021). Segundo essa estratégia de robustez é consistente com aplicações empíricas na literatura de indústria e inovação, que frequentemente combinam classificadores e modelos paramétricos para sustentar inferências sobre padrões setoriais e seus determinantes, especialmente quando o objetivo é interpretação e não apenas previsão.

3.2.2. Dados utilizados para caracterização dos padrões (variáveis exógenas ao cluster)

Para cada padrão de colaboração encontrado, busca-se qualificá-los a partir de variáveis exógenas, definidas como relevantes pela literatura para explicar colaboração e inovação. Isso é, a partir das variáveis do Quadro 2, busca-se qualificar os padrões de cooperação obtidos na clusterização.

Quadro 2 - Variáveis explicativas da árvore de classificação

Variável	Justificativa	Tratamento na literatura	Referências
Impacto do mercado	Avalia o impacto de mercado das inovações de produto e processo implementadas entre 2012 e 2014 nos três tipos em relação a abertura de novos mercados. São consideradas a média das empresas que declaram impacto alto e médio.	A importância dos impactos mercadológicos da cooperação	Sousa (2012); Laursen e Salter (2014); Herstad et al. (2014); Cassiman e Veugelers (2002)
Enquadramento regulatório	Analisa a importância do obstáculo “dificuldade para se adequar a padrões, normas e regulamentações”	A literatura aponta as dificuldades do enquadramento em regulações e normas padrão relativas ao mercado interno e externo como obstáculo para cooperação.	Haakonsson e Kirkegaard (2016); Zobel e Hagedoorn (2020)
Esforço de inovação	Avalia a capacidade de absorção através da <i>proxy</i> dispêndio em P&D sobre receita líquida de vendas.	A literatura aponta para a relevância da capacidade de absorção no processo de inovação.	Cohen e Levinthal (1990); Ray et al. (2017); Ebersberger e Herstad (2013); Necoechea-Mondragón et al. (2017); Ernst (2010)
Apoio governamental	Analisa o percentual de empresa que fazem uso de apoio governamental para inovar, sem distinção do instrumento utilizado, em relação às empresas que inovam.	O apoio governamental é relevante para a inovação em todos os países, podendo ser ainda mais relevante em países em desenvolvimento.	Nieto e Santamaria (2007); Herstad et al. (2014)
Fontes de informação informatizadas	Analisa uma fonte de informação específica: as redes de informações informatizadas (“Internet”, “Extranet”, “Intranet” etc.).	As redes de informação informatizadas podem ser consideradas como facilitadoras para o estabelecimento de redes colaborativas.	-

Fonte: Dados originais da pesquisa

3.3. Síntese metodológica

A estratégia de qualificação dos clusters com variáveis exógenas a partir das técnicas supracitadas se dá pela aplicação de:

- i) Árvores de decisão: fornecem um modelo interpretável e ajudam a entender hierarquias de decisão;
- ii) “Random Forest”: reduz a variância e melhora a robustez, além de fornecer medidas de importância das variáveis;
- iii) “XGBoost”: supera os demais em termos de performance preditiva, especialmente se houver relação não linear e interações complexas entre as variáveis.

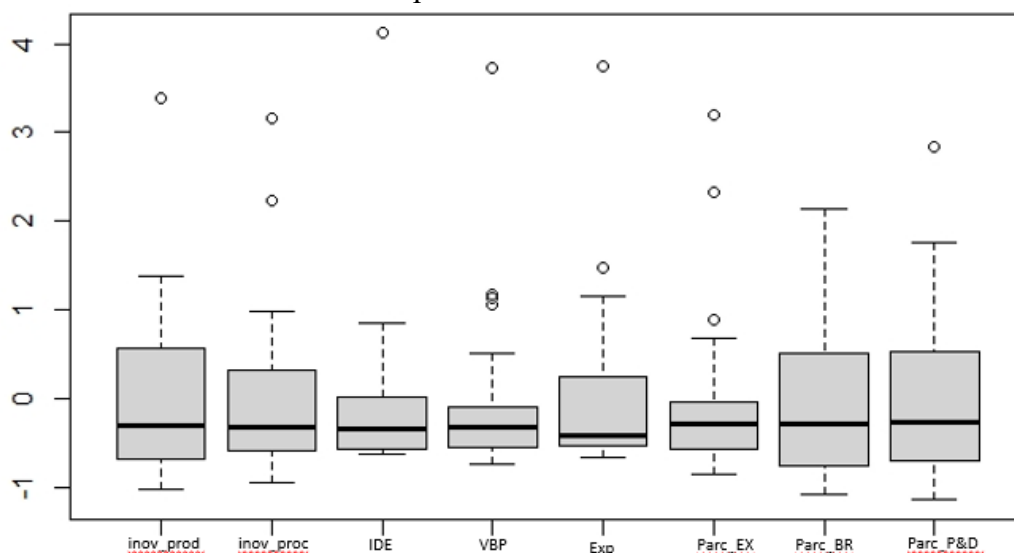
4. Resultados e discussão

4.1. Avaliação da base de dados

No Gráfico 1, apresenta-se a distribuição das variáveis da amostra, já padronizadas, com média zero e desvio padrão um. Nota-se que o IDE é que tem o maior “outlier”, além disso, as variáveis com “boxplot” mais achatado são aquelas com maior dispersão em relação aos setores, é o caso de inovação em produto e em processo, exportação, parceiros no Brasil e parceiros em P&D e ensaios.

Vale o destaque que as variáveis com maior número de valores extremos são aquelas em que a mediana (representada pelo traço preto escuro no Gráfico 1) se diferencia mais da média (igual a zero, dado que as variáveis estão padronizadas).

Gráfico 1 - “Boxplot” das variáveis “clusterizadas”



Fonte: Resultados da pesquisa

Além disso, por um lado, a análise da correlação entre as variáveis da amostra, já padronizadas, sinaliza que o valor bruto da produção (vbp) e as exportações (exp) são altamente correlacionados assim como inovação em produto (in_prod) e processo (in_proce) e os parceiros localizados no Brasil (parceiro_BR) e as parcerias em P&D e testes de ensaios (parceiros_P&D_Ensaio) – variáveis que, em geral, explicam os eventos similares. Por outro lado, não há correlação entre IDE e as inovações, tanto em produto como em processo; baixa correlação entre IDE e parceiros no exterior e no Brasil – uma correlação que, se existente, seria importante para o ecossistema industrial e inovativo. A correlação entre as variáveis é um fator relevante na análise de clusters à medida que influencia na formação dos agrupamentos.

Por fim, a estatística de Hopkins para mostrar que a base de dados pode ser clusterizada, já que sua estatística é de 0,23, inferior àquela apresentada por dados randomizados, de 0,49, que sinaliza a existência de um padrão nos dados. Além disso, vale destacar a aplicação da medida de adequação da MSA⁴ para o conjunto de variáveis resultou em 0,72, considerado mediano. A análise individual dos setores resultou em valores pelo menos medianos. No entanto, o quadro geral releva que reduzir a dimensionalidade dos dados, por meio da aplicação de componentes principais, reescrevendo as oito variáveis em novas, não correlacionadas, pode reduzir o viés do agrupamento dos setores.

4.2. Redução da dimensionalidade das variáveis

A redução da dimensionalidade das variáveis envolve definir critérios para selecionar quantos componentes serão representativos das oito variáveis iniciais. Os critérios utilizados para a definição do número de fatores são: i) teste scree; ii) porcentagem da variância; e iii) raiz latente.

Na Tabela 1 são apresentados os autovalores de cada fator, que é a medida de quanto da variância total das variáveis é explicada pelo fator; percentual da variância por fator; e a variância acumulada. Os fatores vão de 1 a 8, correspondente ao número de variáveis da amostra.

⁴ Medida de adequação da amostra, conhecida como “Kaiser's Measure of Sampling Adequacy” [MSA] em que as faixas para interpretação são: $MSA > 0,90$, admirável; $0,80 \leq MSA < 0,90$, meritório; $0,70 \leq MSA < 0,80$, mediano; $0,60 \leq MSA < 0,70$, medíocre; $0,50 \leq MSA < 0,60$, ruim; $MSA < 0,50$, inaceitável.

Tabela 1 - Autovalores

Fator	Autovalor	Percentual de variância	
		Por fator	Acumulada
1	4,57	57,20%	57,20%
2	1,23	15,39%	72,60%
3	1,06	13,28%	85,88%
4	0,67	8,38%	94,26%
5	0,30	3,73%	98,00%
6	0,10	1,21%	99,21%
7	0,04	0,56%	99,77%
8	0,02	0,23%	100%

Fonte: Dados originais da pesquisa

Nota-se que três fatores explicam, conjuntamente, 85,88% das variáveis (Tabela 1). O quarto fator consegue, sozinho, explicar 8,38% das variáveis, que é um valor elevado comparado ao quinto fator, que explica apenas 3,73%. Logo, por esse critério o número de fatores retidos seria entre 3 e 4. Para a toma de decisão, utiliza-se o critério de Kaiser (1958), no qual são escolhidos fatores com autovalor maior que 1, e escolhem-se três fatores.

Na Tabela 2 apresenta-se; i) autovalor: medida de quanto da variância total das variáveis é explicada por cada componente – indica a importância relativa de cada componente na explicação da variância associada ao conjunto de variáveis analisado; ii) autovetor: obtido para cada autovalor – representa os pesos de cada variável na composição dos fatores; iii) percentual de traço: representa a variação explicada individualmente pelos fatores em porcentagem – esse valor é obtido pela divisão do autovalor pelo número de variáveis, multiplicado por 100.

Tabela 2 - Autovetor, autovalor e percentual do traço dos componentes principais selecionados

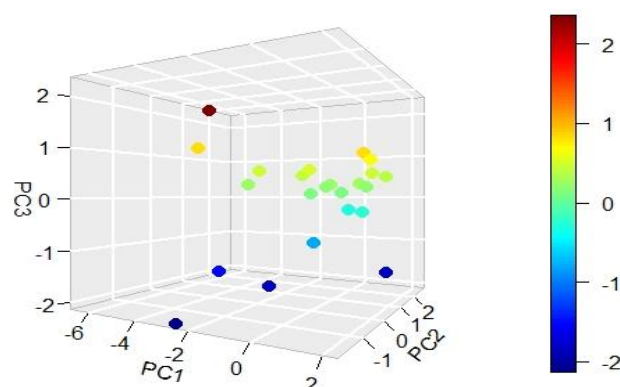
Orientação	Variáveis	PC1	PC2	PC3
Global	exp	-0.348	0.413	0.170
	IDE	-0.105	0.640	-0.449
Production	VBP	-0.365	0.433	0.143
Innovation	Inova_prod	0.424	-0.151	0.291
	Inova_proce	-0.371	-0.173	0.440
Network	Parceiro_BR	-0.425	-0.252	-0.201
	Parceiro_EX	-0.262	-0.319	-0.600
	Parceiros P&D Ensaios	-0.410	-0.123	-0.266
	Autovalor	4.576	1.232	1.063
	Percentual do traço	57%	15%	13%

Fonte: Dados originais da pesquisa

Assim, considerando as 8 variáveis originais da amostra – ou seja, uma matriz de 23 setores x 8 variáveis – e os três componentes principais – ou seja, uma matriz de 23 setores x 3 componentes principais – a multiplicação dessas matrizes 23x8 pela 8x3 resulta em uma matriz de escores fatoriais com os 23 setores para os 3 componentes principais. Essa matriz de escore fatorial reflete, portanto, os dados originais recalculados para os três componentes principais, atingindo o objetivo de redução da dimensionalidade. A análise de cluster será aplicada sobre essa nova matriz, 23x3, cuja distribuição é apresentada no Gráfico 2.

Vale o destaque que pela estatística de Hopkins tanto a matriz 23x8 (original) como a matriz 23x3 (reduzida) não se distribuem uniformemente, ou seja, não estão sendo forçadas a um agrupamento pelo algoritmo. Para os dados randômicos, a estatística de Hopkins é de 0,5; para a matriz original, 0,23; para a matriz reduzida, 0,33. A formação da matriz reduzida a partir de componentes não correlacionados faz com sua distribuição esteja mais próxima àquela dos dados randomizados. Mas ainda sendo considerada passível da formação de grupos.

Gráfico 2 - Distribuição dos 23 setores nos 3 PC

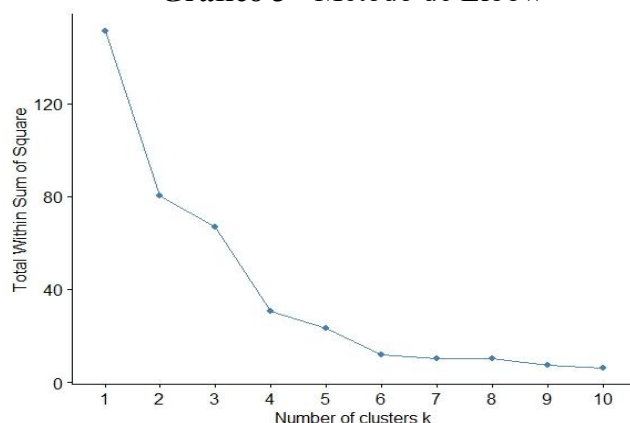


Fonte: Resultados originais da pesquisa

4.3. Seleção do número de clusters

No Gráfico 3, apresenta-se o resultado do método de Elbow – visualização gráfica que cruza a soma dos quadrados internos do cluster com o número de cluster. O método sinaliza que a partir de 6 clusters há menor ganho no ajustamento interno com o aumento do número de clusters.

Gráfico 3 - Método de Elbow



Fonte: Resultados originais da pesquisa

A Tabela 3 apresenta características de diversos clusters, sendo testado de 2 a 6 em função do resultado do método de Elbow.

Pelo “kmeans” e o método hierárquico, os agrupamentos possíveis de serem escolhidos, considerando as estatísticas de validação interna, são 4 e 5, respectivamente. Como o Índice de Dunn e o Silhoutte estão melhorando do “cluster” 4 para o 5, serão considerados cinco “clusters”. Dessa forma, a semente k=5 será aplicada no método híbrido (“hkmeans”).

Tabela 3 - Número de clusters e características para a escolha do número ótimo de clusters

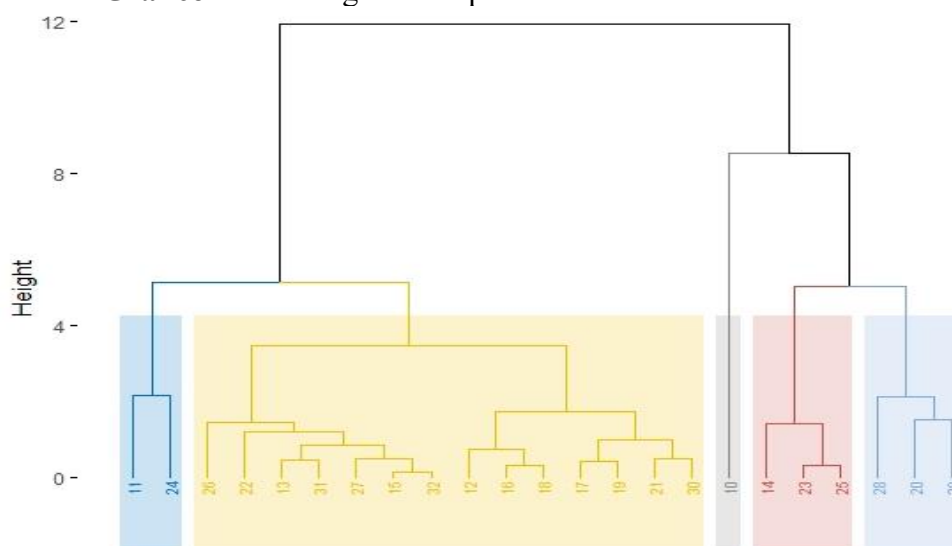
Cluster	Característica do agrupamento	
	k-means	Hierárquico
2	- Tamanho: $n_1=16, n_2=7$ $R^2 = 47\%$ $S_{k=1} = 0,63, S_{k=2} = 0,13, S_{médio} = 0,48$, há 2 observações com $S_i < 0$ - Dunn = 0,149	- Tamanho: $n_1=7, n_2=16$ $S_{k=1} = 0,13, S_{k=2} = 0,63, S_{médio} = 0,48$, há 2 observações com $S_i < 0$ - Dunn = 0,149
3	- Tamanho: $n_1=16, n_2=1, n_3=6$ $R^2 = 71\%$ $S_{k=1} = 0,56, S_{k=2} = 0, S_{k=3} = 0,32, S_{médio} = 0,47$, não há observações com $S_i < 0$ - Dunn = 0,24	- Tamanho: $n_1=1, n_2=16, n_3=6$ $S_{k=1} = 0, S_{k=2} = 0,56, S_{k=3} = 0,32, S_{médio} = 0,47$, não há observações com $S_i < 0$ - Dunn = 0,245
4	Tamanho: $n_1=2, n_2=14, n_3=1, n_4=6$	Tamanho: $n_1=1, n_2=2, n_3=14, n_4=6$

	• $R^2 = 79,8\%$ • $S_{k=1} = 0,56, S_{k=2} = 0, S_{k=3} = 0,32, S_{k=4} = 0,48, S_{médio} = 0,44$, não há observações com $S_i < 0$ • Dunn = 0,30	• $S_{k=1} = 0, S_{k=2} = 0,22, S_{k=3} = 0,58, S_{k=4} = 0,27, S_{médio} = 0,44$, não há observações com coeficiente de Silhouette<0 • Dunn = 0,30
5	• Tamanho: $n_1=2, n_2=1, n_3=3, n_4=10, n_5=7$ • $R^2 = 87,5\%$ • $S_{k=1} = 0,21, S_{k=2} = 0, S_{k=3} = 0,47, S_{k=4} = 0,53, S_{k=5} = 0,21, S_{médio} = 0,38$, há 3 observações com coeficiente de Silhouette<0 • Dunn = 0,376	• Tamanho: $n_1=1, n_2=2, n_3=14, n_4=3, n_5=3$ • $S_{k=1} = 0, S_{k=2} = 0,22, S_{k=3} = 0,49, S_{k=4} = 0,66, S_{k=5} = 0,41, S_{médio} = 0,46$, não há observações com coeficiente de Silhouette<0 • Dunn = 0,44
6	• Tamanho: $n_1=2, n_2=1, n_3=3, n_4=7, n_5=7, n_6=3$ • $R^2 = 92,1\%$ • $S_{k=1} = 0,20, S_{k=2} = 0, S_{k=3} = 0,41, S_{k=4} = 0,40, S_{k=5} = 0,41, S_{k=6} = 0,56, S_{médio} = 0,39$, não há observações com coeficiente de Silhouette<0 • Dunn = 0,29	• Tamanho: $n_1=1, n_2=2, n_3=7, n_4=7, n_5=3, n_6=3$ • $S_{k=1} = 0, S_{k=2} = 0,20, S_{k=3} = 0,40, S_{k=4} = 0,41, S_{k=5} = 0,56, S_{k=6} = 0,41, S_{médio} = 0,39$, não há observações com coeficiente de Silhouette<0 • Dunn = 0,286

Fonte: Resultados originais da pesquisa

A primeira classificação feita pelo método “hkmeans” se dá com a aplicação do método hierárquico. Os resultados podem ser vistos no dendograma apresentado no Gráfico 4, que mostra as distâncias de cada rodada de agrupamento até a formação dos cinco “clusters”.

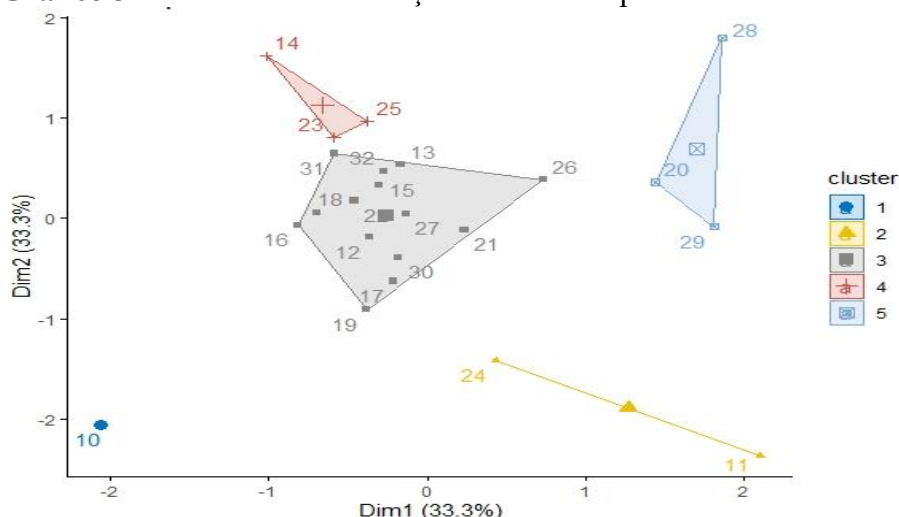
Gráfico 4 - Dendograma da primeira rodada do método híbrido



Fonte: Resultados originais da pesquisa

A classificação apresentada no Gráfico 4, no entanto, não é a final, considerando a etapa de otimização com a aplicação dos pontos médios do método hierárquico no método “kmeans”. No Gráfico 5 é apresentada a distribuição final dos setores nos clusters. Destaca-se que o R^2 do “hkmeans” com cinco cluster foi de 88,2%, maior que quando aplicado apenas o “kmeans” (87,5%).

Gráfico 5 - PCA com clusterização final obtido pela técnica “hkmeans”

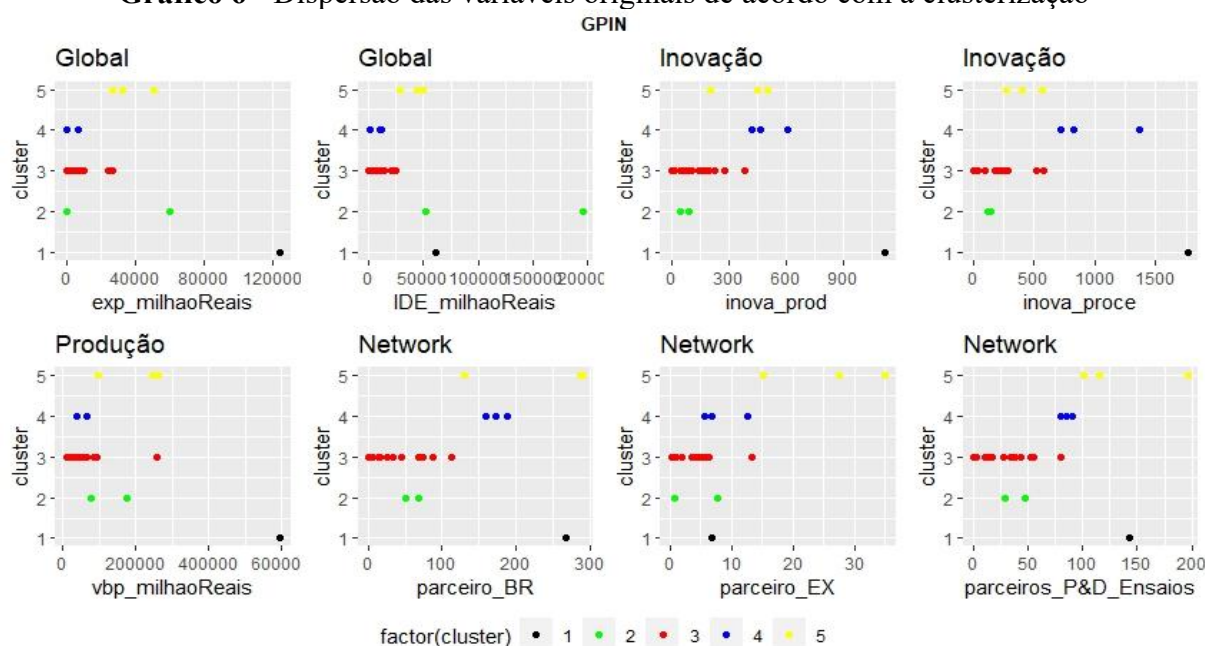


Fonte: Resultados originais da pesquisa

4.4. Análise dos clusters

No Gráfico 6 é apresentada a dispersão das variáveis originais nos “clusters”. Nessa visualização é possível comparar a características dos “clusters”, apontando qual tem maior ênfase em ser global, inovativo, focado em produção ou colaboração. Essa caracterização detalhada será apresentada no Tabela 4.

Gráfico 6 - Dispersão das variáveis originais de acordo com a clusterização



Fonte: Resultados originais da pesquisa

Na Tabela 4 são apresentados os setores agrupados em cada “cluster” e as principais características do “cluster”, calculada como a média das oito variáveis (padronizadas) para cada setor pertencente a um cluster⁵. A partir das principais características é definida uma orientação GPIN:

- “G” é atribuído quando Investimento Direto Externo [IDE] e exportação [exp] são as maiores, “g” quando apenas uma dessas variáveis está entre as maiores;
- “P” é atribuído quando valor bruto da produção [VBP] está entre as maiores;

⁵ Por exemplo: o cluster 2 é composto por dois setores, para os quais é feita uma média dos valores dos dois setores contido em cada uma das oito variáveis e para esse vetor de médias é avaliada qual são as quatro maiores variáveis entre as oito analisadas.

- “I” é atribuído quando inovação em produto e processo são as maiores, “i” quando apenas uma dessas variáveis está entre as maiores;
- “N” é atribuído quando parceiro no Brasil, parceiros no exterior e parceiros em Pesquisa e Desenvolvimento [P&D] e ensaios são as maiores, “n” quando apenas uma dessas variáveis está entre as maiores.

Tabela 4 - Síntese das características dos cluster e orientação GPIN*

Clusters	Setores	Principais características	GPIN	Interpretação
Cluster1 (n=1)	Fabricação de alimentos	1.Exp 2. VBP 3.Inova produto 4. Inova processo	gPI	Competitividade consolidada
Cluster2 (n=2)	Fabricação de bebidas e metalurgia	1.IDE 2. Exp 3.VBP 4. Parceiro P&D e Ensaios	GPI	Integração global forte
Cluster3 (n=14)	Fabricação de produtos do fumo; produtos têxteis; produtos de madeira; celulose, papel e produtos de papel; coque, de produtos derivados do petróleo e de biocombustíveis; produtos farmoquímicos e farmacêuticos; artigos de borracha e plástico; equipamentos de informática, produtos eletrônicos e ópticos; máquinas, aparelhos e materiais elétricos; outros equipamentos de transporte; móveis; produtos diversos; preparação de couros; Impressão e reprodução de gravações	1.VBP 2. Exp 3.IDE 4. Parceiro EX	PGi	Escala produtiva heterogênea
Cluster4 (n=3)	Confecção de artigos de vestuário e acessórios; Fabricação de produtos de minerais não-metálicos e de produtos de metal	1.Inova processo 2. Inova produto 3.Parceiro BR 4. Parceiro P&D e Ensaios	IN	Elevado dinamismo doméstico
Cluster5 (n=3)	Fabricação de produtos químicos; máquinas e equipamentos; e veículos automotores, reboques e carrocerias	1.Parceiro EX 2. Parceiro P&D e Ensaios 3.Parceiro BR 4. VBP	NP	Coordenação tecnológica

Fonte: Resultados originais da pesquisa

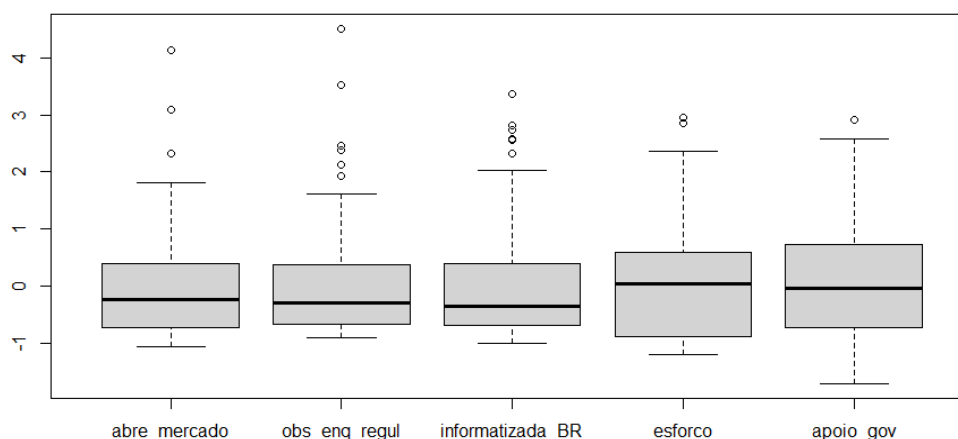
*Nota: Global (G); Production (P); Innovation (I); Network (N)

4.5. Caracterização dos clusters por variáveis exógenas

4.5.1. Características das variáveis exógenas

No Gráfico 7, apresenta-se a correlação entre as variáveis exógenas, já padronizadas. Destaca-se a correlação entre as variáveis fontes de informação por meio de redes informatizadas no Brasil, inovação para abertura de mercado e enquadramento regulatório como obstáculo à inovação. A correlação entre as variáveis, que são preditoras em modelos de árvores de classificação, afetam a escolha da técnica – por exemplo, justificando preferir “Random forest” a árvores de classificação, mesmo tunadas.

Gráfico 7 - “Boxplot” das variáveis exógenas



Fonte: Resultados da pesquisa

Por fim, é importante destacar que a base de dados agrupada, referente a PINTEC de 2017, conta com apenas 23 observações. O baixo número de observações fez com que todos os modelos de árvore de decisão, “random forest” e “boosting” tivessem baixas estatísticas de ajustes. Com isso, optou-se por empilhar observações para as mesmas cinco variáveis para os anos de 2008, 2011, 2014 e 2017 – todos os anos para os quais há PINTEC na mesma agregação (CNAE 2.0). Os setores seguiram o agrupamento atribuído em 2017, ou seja, se o setor de metalurgia está no cluster 2 em 2017, também está nos demais anos. Assim, a base para caracterizar os clusters passou de 23 observações para 92. Destaca-se ainda que o cluster 1 é um “outlier”, composto apenas pelo setor de alimentos e, portanto, não apresenta variabilidade para aplicação das técnicas de classificação. Dessa forma, o cluster não será tratado nesta seção.

4.5.2. Aplicação do xgboosting para caracterizar os cluster por variáveis exógenas

Pela literatura, compreendeu-se que o “xgboosting” é a técnica mais acurada, entre as selecionadas para explorar neste trabalho. Nesse sentido, destaca-se o trabalho de Uribe (2025) que aplica a técnica para lidar com temas de estrutura produtiva e inovação.

Na Tabela 5 a síntese da aplicação do “xgboosting” para cada um dos clusters. Nessa análise, o melhor desempenho está associado a estatística da curva ROC. A base com 92 observações foi separada em 30% treino e 70% teste, percentual com melhor comportamento encontrado para o tamanho da base. Também foi configurado “k-fold” de 10, número de árvores sequências de (nrounds) de 50, 100, 150 e 200, profundidade máxima das árvores (max) de 2, 3 e 4, redução mínima de perda necessária para quebra (gamma) de 0 e 0.001, taxa de aprendizado (eta) de 0.05 e 1, número de colunas sorteadas por ‘round’ (colsample) de 0.6 e 0.8, número mínimo de observação por folha de 2 e 3 e percentual de amostragem para o “bootstrap” de 0.75 e 1. Além disso, o valor atribuído as variáveis referem-se à contribuição média de cada variável para a redução do erro (ou melhoria do modelo), denominado “gain”.

Tabela 5 - Síntese do “xgboosting” para cada cluster*

		Cluster 2	Cluster 3	Cluster 4	Cluster 5
Especificações	Max	2	3	2	2
	Eta	0.1	0.1	0.05	0.1
	Colsample	0.6	0.6	0.6	0.8
XGboosting	abre_mercado	0.49	0.23	0.26	0.08
	apoio_gov	-	0.37	0.04	0.25
	esforco	-	0.05	0.12	0.04
	informatizada_BR	0.30	0.24	0.43	0.52
	obs_enq_regul	0.21	0.11	0.26	0.11
Validação	ROC treino	0.92	0.90	0.96	0.93
	ROC teste	0.75	0.68	0.83	0.65

Fonte: Resultados da pesquisa

*Nota: Em todos os clusters, o nround selecionado pelo algoritmo foi 50, o gama foi zero, o mínimo foi 2 e o subsample foi 0.75.

Na Tabela 5 nota-se que as variáveis uso de redes informatizadas localizadas no Brasil como fonte de informação e impacto causado pela abertura de novos mercados para empresas que implementaram

inovação são variáveis importantes para caracterização de todos os clusters. Já os aspectos específicos a cada cluster são o uso do apoio governamental pelas empresas no cluster 3 e no cluster 5 e as normas e regulamentos serem um obstáculo para inovação no cluster 4.

4.5.3. Validação da análise do xgboosting pela aplicação de outras técnicas

Nesta etapa, propõe-se uma validação dos resultados do xgboosting pela aplicação da árvore de decisão, “random forest” e modelo logístico. Na Tabela 6 apresentam-se os resultados da aplicação da árvore de decisão, em que todos os modelos são aqueles com melhor desempenho. O melhor desempenho está associado a estatística da curva ROC, acurácia e taxa de erro. Além da separação da base em treino e teste, também foi definido “k-folds” padrão igual a 100 e hiperparâmetros, especificamente o custo de complexidade (cp), o critério de parada do algoritmo (min) e a profundidade máxima para a árvore de 30, recomendado na literatura. Também foi aplicado um “set.seed” de 31 para que os resultados possam ser reproduzidos.

As cinco variáveis preditoras de cada um dos clusters são analisadas pela importância total (“overall”), que considera todos os nós em que a variável aparece e a redução de impureza pelo índice de Gini do modelo – quanto maior o valor, maior a importância da variável na predição do modelo.

Tabela 6 - Síntese da árvore de classificação para cada cluster

		Cluster 2	Cluster 3	Cluster 4	Cluster 5
Especificações	Cp	0.001	0	0.001	0.001
	min	3	5	3	3
Árvore de classificação	abre mercado	1.33	8.00	-	0.95
	apoio_gov	-	0.80	-	0.21
	esforco	0.58	2.40	3.36	4.07
	informatizada BR	0.39	6.41	5.04	4.20
	obs_enq_regul	0.39	4	3.36	1.36
Validação	ROC teste	0.64	0.70	0.78	0.52
	ROC treino	1	0.91	0.87	1
	Acurácia	0.81	0.75	0.90	0.84
	Taxa de erro	0.19	0.25	0.10	0.16

Fonte: Resultados da pesquisa

Na Tabela 6 nota-se que o uso de redes de informação e a abertura de mercado continuam sendo relevantes, mas não é um resultado geral como na aplicação do XGBoosting. Além disso, uma diferença relevante é que a variável esforço de inovação aparece como relevante para caracterização dos clusters 2 e cluster 5.

Na sequência, apresenta-se na Tabela 7 a síntese da aplicação do “random forest” para cada um dos clusters. Nessa análise, o melhor desempenho também está associado a estatística da curva ROC. Na base separada em treino e teste, foi configurado o número de árvores para teste de 100 e hiperparâmetro (“mtry”), que representa o número de variáveis selecionadas aleatoriamente para cada divisão da árvore, a ser testado de 2, 3, 4 e 5, considerando que são cinco preditoras, além do “k-fold” de 10 para fazer a validação cruzada. A especificação é padrão para todos os clusters, considerando que possuem o mesmo número de preditoras.

Tabela 7 - Síntese da “random forest” para cada cluster

		Cluster 2	Cluster 3	Cluster 4	Cluster 5
Especificação	mtry	2	3	3	4
Random forest	abre mercado	35.79	100	38.35	100
	apoio_gov	0	0	0	0
	esforco	100	24.95	16.51	72.22
	informatizada BR	27.53	55.34	100	21.94
	obs_enq_regul	24.58	8.98	46.54	4.55
Validação	ROC teste	0.79	0.74	0.90	0.56
	ROC treino	1	1	1	0

Fonte: Resultados da pesquisa

Neste caso, a abertura de mercado se destaca, novamente, como caracterizadora de todos os clusters. As redes informatizadas caracterizam os cluster 3 e cluster 4 e o esforço, o cluster 2 e o cluster 5. O apoio governamental novamente não se destaca como uma variável relevante em nenhum cluster. Uma conclusão

generalizada que se pode tirar de todos os testes feitos é que as variáveis impacto da inovação na abertura de novos mercados e uso de redes informatizadas localizadas no Brasil como fontes de informação são relevantes para caracterização dos clusters.

4.6. Interpretação dos clusters

A tipologia GPIN permite interpretar a heterogeneidade da indústria de transformação brasileira a partir da articulação concreta entre seus padrões setoriais e suas formas de inserção internacional. A composição dos clusters revela que a integração às redes globais de produção e inovação ocorre por mecanismos distintos, associados à natureza tecnológica, intensidade de capital e posição nas cadeias produtivas de cada grupo setorial.

O Cluster 2 (GPi), formado por bebidas e metalurgia, combina elevada presença de IDE, exportações e articulação em P&D. Trata-se de setores intensivos em escala e capital, com forte inserção em cadeias globais coordenadas por grandes grupos empresariais. A metalurgia, em particular, é típica de cadeias hierarquizadas e integradas internacionalmente, nas quais o “upgrading” pode depender da estratégia tecnológica das matrizes.

O Cluster 3 (PGi), que concentra quatorze setores heterogêneos — de têxteis e madeira a eletrônicos e farmacêuticos — apresenta elevada participação no VBP e nas exportações, com presença de IDE, mas menor densidade relacional inovativa. Essa heterogeneidade dificulta uma qualificação uniforme para o cluster. A diversidade interna do grupo indica setores com inserção internacional baseada sobretudo em competitividade produtiva e escala, ainda que com distintos níveis tecnológicos. A presença simultânea de segmentos tradicionais (como madeira e couro) e de média-alta tecnologia (como equipamentos eletrônicos) sugere coexistência de trajetórias: desde especializações baseadas em recursos até segmentos integrados a cadeias tecnológicas mais complexas, ainda que com inovação predominantemente incremental.

O Cluster 4 (IN), composto por vestuário, minerais não-metálicos e produtos de metal, apresenta forte dinamismo inovativo doméstico e articulação com parceiros nacionais de P&D. Esses setores tendem a operar com menor intensidade de IDE e menor projeção exportadora relativa, indicando padrão de inovação ancorado no mercado interno e na adaptação produtiva. Trata-se de atividades com encadeamentos domésticos relevantes, mas com limitada inserção em redes tecnológicas globais.

O Cluster 5 (NP), que reúne químicos, máquinas e equipamentos e veículos automotores, destaca-se pela elevada densidade relacional — tanto doméstica como internacional — associada a significativo peso no VBP. São setores tipicamente estruturantes das cadeias industriais, intensivos em engenharia e articulação tecnológica. A forte presença de parcerias externas sugere inserção via redes globais de inovação mais do que exclusivamente via comércio, aproximando-se de padrões de coordenação tecnológica complexa.

Por fim, o Cluster 1 (gPI), representado pela indústria de alimentos, combina forte presença exportadora e peso produtivo com inovação de produto e processo. Trata-se de setor tradicionalmente competitivo no Brasil, com inserção internacional consolidada e “upgrading” incremental baseado em eficiência produtiva e adaptação tecnológica.

A incorporação das variáveis externas analisadas por meio do XGBoost e reavaliadas por árvore de decisão, *random forest* e modelo logístico não altera essa tipologia, mas contribui para qualificar sua interpretação. Observa-se convergência metodológica em torno de dois fatores transversais: o impacto da inovação na abertura de novos mercados e o uso de redes informatizadas localizadas no Brasil como fonte de informação. Assim, os padrões de inserção internacional descritos anteriormente são complementados por evidências de que a capacidade de acessar mercados e mobilizar fluxos de informação internos ao país desempenha papel estruturante na dinâmica inovativa industrial, operando de forma transversal aos diferentes perfis setoriais.

5. Considerações finais

Este trabalho buscou integrar as dimensões produtiva e inovativa da inserção internacional setorial por meio da proposição da taxonomia GPIN. Ao articular inserção produtiva global, intensidade inovativa e densidade relacional, a tipologia permitiu analisar a heterogeneidade estrutural da indústria de

transformação brasileira sob uma perspectiva integrada entre produção e inovação.

Os resultados sugerem que a indústria brasileira não apresenta um padrão único de inserção internacional, mas sim uma estrutura fragmentada, na qual diferentes grupos setoriais operam sob lógicas distintas de coordenação produtiva e tecnológica. Essa fragmentação confirma a hipótese de que produção e inovação constituem dimensões transversais à inserção internacional, mas sua intensidade e articulação variam segundo o perfil estrutural dos setores.

Nos setores intensivos em IDE e integrados a cadeias globais coordenadas por grandes grupos empresariais, o “upgrading” parece condicionado à estratégia tecnológica das matrizes, sugerindo que a trajetória de avanço produtivo pode depender de decisões externas ao território nacional. Nesses casos, a integração produtiva não implica necessariamente autonomia tecnológica, reforçando o debate da literatura de GVC sobre governança e dependência estrutural.

Por outro lado, clusters caracterizados por forte dinamismo inovativo doméstico e articulação com parceiros nacionais indicam que a inovação pode emergir mesmo com menor projeção exportadora ou menor presença de IDE. Contudo, a limitada inserção em redes tecnológicas globais pode restringir o alcance internacional desses esforços. A coexistência de clusters fortemente integrados a cadeias globais com clusters predominantemente ancorados no mercado doméstico sugere a presença de um dualismo estrutural na indústria brasileira, combinando segmentos com elevada articulação internacional e segmentos cuja dinâmica inovativa permanece essencialmente interna.

A integração das dimensões produtiva e inovativa evidencia que, em economias de industrialização tardia, os desafios de desenvolvimento industrial não são homogêneos. Em determinados setores, o principal obstáculo reside na coordenação tecnológica e na inserção em redes globais de conhecimento; em outros, o desafio central é a escala produtiva, o adensamento industrial e a consolidação de encadeamentos domésticos. A tipologia GPIN contribui ao explicitar essa heterogeneidade, superando abordagens que tratam produção e inovação de forma dissociada.

As evidências obtidas a partir das variáveis exógenas reforçam essa interpretação estrutural. O impacto da inovação na abertura de novos mercados e o uso de redes informatizadas como fonte de informação mostraram-se fatores transversais aos clusters, sugerindo que a capacidade de acessar mercados e mobilizar fluxos de informação internos constitui elemento estruturante da dinâmica inovativa industrial. Ao mesmo tempo, a relevância diferenciada de apoio governamental e de normas regulatórias entre clusters indica que o ambiente institucional interage de forma distinta conforme o perfil setorial.

Do ponto de vista de política industrial, os resultados sugerem que estratégias horizontais podem ser insuficientes diante da heterogeneidade estrutural identificada. Setores fortemente integrados a cadeias globais podem demandar políticas voltadas à internalização de capacidades tecnológicas e maior articulação com atores domésticos, enquanto setores ancorados no mercado interno podem necessitar de políticas de inserção internacional e fortalecimento de redes externas. A diferenciação de instrumentos segundo o padrão estrutural torna-se, portanto, elemento central para estratégias de reposicionamento produtivo.

O trabalho, ao lidar com dados populacionais, tipifica padrões comuns aos setores. Contudo, esse tipo de análise reduz a possibilidade de exploração sob a ótica do comportamento individual das empresas, o que constitui limitação a ser aprofundada por meio de estudos de caso e análises em nível micro. O aprimoramento futuro pode incluir a exploração mais detalhada de dados de patentes, o uso da empresa como unidade de análise — mediante acesso à sala segura do IBGE — e a aplicação de técnicas complementares, como a análise dos resultados do XGBoost por meio de Shapley Additive Explanations (SHAP), para aprofundar a interpretação das variáveis explicativas.

Referências

BARNARD, H.; CHAMINADE, C. Global Innovation Networks: towards a taxonomy. Circle Work. Pap: 1-44, 2011.

CASSIMAN, B.; VEUGELERS, R. R&D Cooperation and Spillovers: Some Empirical Evidence from Belgium. *American Economic Review*, 92 (4): 1169-1184, 2002.
<https://doi.org/10.1257/00028280260344704>

CASTRO, D. F. 2010. Padrões setoriais da inovação tecnológica na indústria brasileira: uma análise de

cluster a partir da PINTEC. Dissertação de Mestrado em Economia Política. Pontifícia Universidade Católica de São Paulo. São Paulo, SP, Brasil.

CHANG, H.; ANDREONI, A. *Industrial Policy in the 21st Century. Development and Change*. International Institute of Social Studies, 2020.
<https://onlinelibrary.wiley.com/doi/full/10.1111/dech.12570>

CHESBROUGH, H. 2003. The era of open innovation. *Sloan Management Review* 44: 35-41.

COHEN, W. M.; LEVINTHAL, D. A. Absorptive Capacity: A New Perspective on Learning and Innovation, 1990. *Administrative Science Quarterly* 35: 128-152.

COOKE, P. Global Production Networks and Global Innovation Networks: Stability Versus Growth. *European Planning Studies*, v. 21, n. 7, p. 1081–1094, 2013.

Ebersberger, B.; Herstad, S. J. The relationship between international innovation collaboration, intramural R&D and SMEs' innovation performance: a quantile regression approach. *Applied Economics Letters* 20: 626–630, 2013.

ERNST, D. Upgrading through innovation in a small network economy: Insights from Taiwan's IT industry. *Economics of Innovation and New Technology* 19: 295–324, 2010.
<https://doi.org/10.1080/10438590802469560>

ERNST, D. Economics of Innovation and New Technology Global production networks and the changing geography of innovation systems. Implications for developing countries. *Econ. Innov. New Techn.*, v. 0, n. 0, p. 497–523, 2002.

EVERITT, B. S.; Landau, S.; Leese, M.; Stahl, D. 2011. *Cluster analysis*. 5ed. Wiley Series in Probability and Statistics. John Wiley and Sons Inc. Chichester. U.K.

FÁVERO, L. P.; Belfiore, P. 2009. *Análise de dados: modelagem multivariada para tomada de decisões*. Elsevier. Rio de Janeiro. Rio de Janeiro. Brasil.

GEREFFI, G.; STURGEON, T.; HUMPHREY, J. 2005. The governance of global value chains. *Review of international political economy*. <https://doi.org/10.1080/09692290500049805>.

GUIMÓN, J.; SALAZAR-ELENA, J. C. 2015. Collaboration in Innovation Between Foreign Subsidiaries and Local Universities: Evidence from Spain. *Industry and Innovation* 22: 445–466.

HAAKONSSON, S. J.; UJJUAL, V. 2015. Internationalization of R & D: New insights into multinational enterprises' R & D strategies in emerging markets. *Management Revue* 26: 101–122.

HAIR, Jr., J.; Black, W. C.; Babin, B. J.; Anderson, R. E.; Tatham, R. L. 2009. *Análise Multivariada de dados*. 6ed. Bookman. Porto Alegre. Rio Grande do Sul. Brasil.

HERSTAD, S. J.; ASLESEN, H. W.; EBERSBERGER, B. 2014. On industrial knowledge bases, commercial opportunities and global innovation network linkages. *Research Policy* 43: 495–504. <https://doi.org/10.1016/j.respol.2013.08.003>

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA [IBGE]. Pesquisa Inovação: PINTEC - Vários anos. Disponível em: <<https://www.ibge.gov.br/estatisticas/multidominio/ciencia-tecnologia-e-inovacao/9141-pesquisa-de-inovacao.html>>. Acesso em: 15 mar. 2024.

JAMES, G.; Witten, D.; Hastie, T.; Tibshirani, R. 2021. *An Introduction to Statistical Learning with*

Applications in R. 2ed. Springer.

KASSAMBARA, A. 2017. Practical guide to cluster analysis in R: unsupervised machine learning. 1ed. Statistical Tools for High-throughput Data Analysis (STHDA).

LAURSEN, K.; SALTER, A. J. 2014. The paradox of openness: Appropriability, external search and collaboration. *Research Policy* 43: 867–878. <https://doi.org/10.1016/j.respol.2013.10.004>

LIU, J. 2017. The global innovation networks and global production networks of firms: Conceptualization and implication. *African Journal of Science, Technology, Innovation and Development*, v. 9, n. 3, p. 229–240.

LUNDVALL, B.-Å. *National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning*. London: Pinter Publishers, 1992.

MINISTÉRIO DO DESENVOLVIMENTO, INDÚSTRIA, COMÉRCIO E SERVIÇOS [MDIC]. Comex Stat. Disponível em: <http://comexstat.mdic.gov.br/pt/home>. Acesso em: 31 abr. 2024.

NECOECHEA-MONDRAGÓN, H.; PINEDA-DOMÍNGUEZ, D., PÉREZ-REVELES, L. E SOTO-FLORES, R., 2017. Critical factors for participation in global innovation networks. Empirical evidence from the Mexican nanotechnology sector. *Technological Forecasting and Social Change* 114: 293–312.

NIETO, M. J.; SANTAMARIA, L. 2007. The importance of diverse collaborative networks for the novelty of product innovation. *Technovation* 27: 367–377. <https://doi.org/10.1016/j.technovation.2006.10.001>

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT [OCDE]. Open innovation in global networks. 2008. Disponível em: <https://www.oecd-ilibrary.org/>. Acesso em: 15 mar. 2024.

PIETROBELLI, C.; RABELLOTTI, R. Global Value Chains Meet Innovation Systems: Are There Learning Opportunities for Developing Countries? *World Development*, 2011. <https://doi.org/10.1016/j.worlddev.2010.05.013>.

RAY, P. K.; RAY S.; KUMAR, V. 2017. Internationalization of latecomer firms from emerging economies: The role of resultant and autonomous learning. *Asia Pacific Journal of Management* 34: 851–873. <https://doi.org/10.1007/s10490-017-9527-y>

SOUSA, A. G. 2012. Impactos da cooperação em atividades de P&D das empresas brasileiras. Tese de Doutorado em Economia Aplicada. Universidade Federal de Viçosa, Viçosa, MG, Brasil.

SCHUMPETER, J. A. 1961. Capitalismo, socialismo e democracia (Tradução Ruy Jungmann). Editora Fundo de Cultura. Rio de Janeiro. Rio de Janeiro. Brasil.

URIBE, H, N, Investment in intangible assets and economic complexity. *Research Policy*, 2025. <https://doi.org/10.1016/j.respol.2024.105133>

ZOBEL, A. N.; HAGEDOORN, J. 2020. Implications of open innovation for organizational boundaries and the governance of contractual relations. *Academy of Management Perspectives* 34: 400–423. <https://doi.org/10.5465/amp.2016.0175>