

Análise de Competitividade no E-Commerce Brasileiro via Correlação Cruzada em Séries Temporais de Preços[‡]

Pedro H. J. de Carvalho¹, Alice Sternberg², Myrna Amorim¹,
Gabriela Ruberg³, Rebecca Sales⁴, Eduardo Ogasawara¹

¹Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ)

²Petrobras

³Banco Central do Brasil

⁴INRIA

{pedro.jaoulack, myrna.amorim}@cefet-rj.br

gabriela.ruberg@bcb.gov.br, rebecca.sales@inria.fr, eogasawara@ieee.org

Abstract. *Brazilian e-commerce moved R\$196.1 billion in 2023, driven by dynamic pricing strategies such as price matching. However, reproducible quantitative methods for formally identifying temporal influence patterns, such as leadership and response, across platforms remain scarce. This work combines cross-correlation and sliding windows to detect such patterns in price time series. We analyze real-world data from 30 products across 21 platforms over 500 days (441k records). The results reveal non-stationary correlations, highlighting seasonal behavior and episodes of price leadership in which marketplaces act as benchmarks for pharmacy chains.*

Resumo. *O comércio eletrônico brasileiro movimentou R\$196,1 bilhões em 2023, impulsionado por estratégias de precificação dinâmica, como o price matching. No entanto, ainda são escassas abordagens quantitativas reproduzíveis para explorar padrões de influência temporal, como liderança e resposta, entre plataformas. Este trabalho combina correlação cruzada e janelas deslizantes como ferramenta exploratória para analisar esses padrões em séries temporais de preços. O estudo analisa dados reais de 30 produtos em 21 plataformas ao longo de 500 dias (441 mil registros). Os resultados revelam correlações não estacionárias, evidenciando comportamento sazonal e episódios compatíveis com liderança de preço.*

1. Introdução

O comércio eletrônico consolida-se como um vetor fundamental da economia, com faturamento projetado superior a R\$200 bilhões em 2024 [ABComm, 2024]. Nesse cenário, plataformas monitoram concorrentes continuamente e reajustam preços por regras manuais ou algorítmicas de *price matching* [Nalca et al., 2018], em um ambiente que a literatura

*Os autores agradecem à CAPES, CNPq e FAPERJ pelo apoio parcial à pesquisa.

[‡]Ferramentas de IA generativa foram utilizadas apenas na revisão linguística, com foco em clareza. Os autores revisaram todas as sugestões e permaneceram integralmente responsáveis pelo conteúdo.

descreve como altamente dinâmico [Gerpott e Berends, 2022; Spann et al., 2025] e marcado por estratégias multicanal, como o *showrooming* [Rejón-Guardia e Luna-Nevarez, 2017]. Ainda assim, não é trivial explorar, com base em dados reais, quais plataformas parecem reagir, quais parecem liderar, com qual atraso e em quais períodos essa interação se intensifica.

Medidas tradicionais de correlação avaliam associação simultânea, mas não capturam a defasagem temporal inerente à reação competitiva [Lock et al., 2020]. Por outro lado, modelos causais e autorregressivos mais sofisticados geralmente exigem séries mais estáveis, longas e densas do que as encontradas em bases reais de preços coletadas por *web scraping*, tipicamente heterogêneas, esparsas e sujeitas a ruídos operacionais.

Propõe-se uma abordagem exploratória baseada em correlação cruzada com janela deslizante (*sliding window*) para analisar séries temporais de preços diários. A proposta transforma registros heterogêneos de múltiplas plataformas em um painel temporal comparável, mede a dependência com defasagem em uma faixa limitada de *lags* e acompanha sua evolução ao longo do tempo. O método busca distinguir sincronização de curto prazo, resposta com atraso e regimes sazonais de maior competição [Bungert, 2003].

O estudo utiliza uma base proprietária com mais de 441 mil registros de preços de 30 produtos, 10 marcas e 21 plataformas, observados ao longo de 500 dias entre julho de 2023 e março de 2025. Com base nesses dados, investiga-se se há evidências quantitativas de liderança e resposta entre plataformas de *e-commerce* brasileiras quando se consideram defasagens temporais e variação local ao longo do tempo.

As principais contribuições deste trabalho são: (i) um fluxo metodológico de Descoberta de Conhecimento em Banco de Dados (KDD) para transformar dados brutos de preços em séries comparáveis; (ii) uma estratégia em duas escalas, que combina análise global por *lag* e análise local por janela deslizante para captar a não estacionariedade competitiva; (iii) critérios interpretativos para explorar sincronização, liderança e resposta; e (iv) evidências empíricas sobre a competitividade de preços no *e-commerce* brasileiro.

2. Revisão da Literatura

A literatura sobre precificação competitiva em mercados digitais estabelece que a resposta a preços rivais é um componente central da competição online. Chen et al. [2020] discutem como a escolha entre modelos de revenda e agência afeta preços e promoções em contextos de competição entre varejistas. Gerpott e Berends [2022] observam que o campo avançou na descrição de mecanismos de precificação dinâmica, mas ainda carece de evidência empírica suficientemente granular sobre a interação competitiva em mercados reais. Bungert [2003] analisam as condições sob as quais guerras de preços se iniciam e se encerram, mostrando que o comportamento de sinalização entre rivais desempenha papel central na dinâmica competitiva e sustenta a análise de padrões de liderança e resposta.

No eixo empírico, Chen et al. [2016] mostram que a precificação algorítmica é amplamente utilizada no *Amazon Marketplace* e tende a aumentar a volatilidade de preços sem necessariamente produzir guerra de preços permanente. Isso indica que a competição digital não deve ser lida apenas pela amplitude das variações, mas também pela temporalidade da resposta entre vendedores e plataformas. Na mesma direção, Hanspach et al.

[2024] documentam padrões similares em plataformas europeias, reforçando a generalidade do fenômeno.

Para analisar relações de liderança e resposta entre plataformas, destacam-se duas famílias de abordagens. A primeira reúne modelos econométricos e autorregressivos, como o estimador de mínimos quadrados generalizados factíveis (FGLS) discutido em [Bai et al., 2021] e o VAR com penalização Lasso aplicado ao domínio de *e-commerce* em [Faehnle e Guidolin, 2021]. Essas abordagens são potentes para a modelagem multivariada, mas costumam exigir hipóteses mais fortes sobre a estrutura dos dados, a estabilidade dos coeficientes e a disponibilidade de séries densas, premissas que raramente se sustentam em dados reais de *web scraping*. A segunda família envolve medidas de dependência temporal, entre elas a correlação cruzada, cuja formulação clássica quantifica a similaridade entre duas séries em função de um deslocamento temporal [Sarwate e Pursley, 1980]. Para dados de preços diários com cobertura assimétrica, essa medida oferece uma alternativa interpretável e computacionalmente leve para a triagem exploratória de relações de liderança-resposta, sem exigir estacionariedade global das séries [Lock et al., 2020]. Em contrapartida, estudos empíricos descrevem bem a volatilidade e a automação de preços, mas raramente operacionalizam a identificação explícita de *lags* de resposta entre plataformas em séries incompletas, heterogêneas e não estacionárias, lacuna à qual este trabalho procura responder.

3. Metodologia

A metodologia segue um fluxo de KDD, estruturado em coleta, transformação e análise de séries temporais. A proposta organiza-se em dois passos complementares. O primeiro, detalhado na Seção 3.1, transforma registros brutos de preços em séries temporais comparáveis por meio de ETL, diferenciação e normalização. O segundo, apresentado na Seção 3.2, aplica correlação cruzada com janelas deslizantes para explorar padrões de sincronização, liderança e resposta ao longo do tempo.

3.1. Coleta e Tratamento de Dados (ETL)

O processo de ETL (*Extract, Transform, Load*) envolve as seguintes etapas: (i) ingestão, (ii) harmonização, (iii) transformação temporal, (iv) normalização e (v) estruturação. Na ingestão, realizam-se a consolidação de arquivos diários e o tratamento de falhas de coleta, como páginas indisponíveis ou alterações de estrutura. Na etapa de harmonização, realizam-se a padronização de nomes de produtos, plataformas e datas, a remoção de duplicidades e a correção de inconsistências.

Na transformação temporal, converte-se a série de níveis de preço em variações diárias ($\Delta P_t = P_t - P_{t-1}$), reduzindo o efeito de patamar e aproximando a análise do comportamento de ajuste competitivo [Hyndman e Athanasopoulos, 2018]. Datas sem observação permanecem ausentes, sem interpolação, e a diferenciação é realizada entre registros consecutivos disponíveis de cada série produto-plataforma. Na normalização, aplica-se *Z-score* por série produto-plataforma para permitir a comparação entre produtos de diferentes faixas de preço, em que $\mu_{\Delta P}$ representa a média e $\sigma_{\Delta P}$ o desvio-padrão das variações diárias, resultando na série normalizada $Z_t = \frac{\Delta P_t - \mu_{\Delta P}}{\sigma_{\Delta P}}$ [Lock et al., 2020].

Finalmente, na estruturação, os dados são pivotados para criar séries temporais alinhadas por data para cada par de competidores, com controle explícito de dados faltantes.

Sem a etapa de transformação temporal, duas plataformas poderiam parecer fortemente associadas apenas por operarem na mesma faixa de preço, sem relação dinâmica entre elas. Ao privilegiar variações normalizadas e alinhamento temporal, o método busca captar resposta competitiva em vez de mera semelhança de posicionamento comercial.

3.2. Correlação Cruzada com Janela Deslizante

A correlação cruzada mede a similaridade entre duas séries deslocadas uma em relação à outra por um atraso τ . Para o mesmo produto e um par de plataformas concorrentes, sejam $X = \{Z_t^{(a)}\}$ e $Y = \{Z_t^{(b)}\}$ as séries normalizadas de variações diárias observadas nas plataformas a e b , respectivamente. A correlação no *lag* τ é definida em [Sarwate e Pursley, 1980], conforme a Equação 1, em que x_t e y_t denotam observações válidas das séries X e Y , $\bar{x}$ e $\bar{y}$ são suas médias e o somatório ocorre apenas sobre índices temporais com sobreposição válida após o deslocamento:

$$r_{XY}(\tau) = \frac{\sum_t (x_t - \bar{x})(y_{t+\tau} - \bar{y})}{\sqrt{\sum_t (x_t - \bar{x})^2} \sqrt{\sum_t (y_{t+\tau} - \bar{y})^2}} \quad (1)$$

Para capturar a dinâmica temporal, aplica-se a técnica de *sliding window* [Ogasawara et al., 2025]. Define-se uma janela de tamanho fixo w e, para a i -ésima janela, denotada por $W_i = \{t_i, \dots, t_i + w - 1\}$, calcula-se a correlação $r_{XY}^{(i)}(\tau)$ em uma faixa limitada de atrasos $\tau \in [-L, +L]$. Esse desenho produz uma leitura global, obtida pelo perfil médio de correlação por *lag*, e outra local, obtida pela evolução da correlação ao longo do tempo. Adota-se $w = 7$ e $\tau \in [-7, +7]$ como escolha operacional compatível com o ciclo comercial semanal e com o objetivo exploratório de triagem.

Para evitar interpretações espúrias, três restrições são adotadas. Primeiro, o cálculo só é realizado quando há sobreposição temporal mínima entre as séries alinhadas (ao menos duas observações válidas por janela e ao menos três na análise estática agregada), limiar tratado aqui como escolha operacional de triagem exploratória. Segundo, séries com variância nula na janela são descartadas. Por fim, a interpretação do *lag* ótimo é sempre contextual: um pico em *lag* positivo sugere precedência temporal de X sobre Y , mas não é tratado como causalidade, apenas como evidência compatível com um padrão de liderança-resposta. A interpretação dos resultados baseia-se em três critérios: *lag* 0 sugere reação simultânea, compatível com alta automação ou monitoramento em tempo real; *lag* positivo indica que a plataforma X antecede Y em τ dias; e a intensidade do coeficiente expressa o grau de acoplamento entre as estratégias.

Em comparação com modelos VAR [Faehnle e Guidolin, 2021], a proposta abre mão de parte da inferência estrutural para ganhar interpretabilidade e robustez em dados incompletos; em comparação com a correlação estática, incorpora direção temporal e sensibilidade à sazonalidade competitiva. Sua aplicação empírica pode combinar leitura agregada do perfil médio de correlação por *lag*, inspeção de casos específicos com maior cobertura temporal e verificação de consistência externa em pares de plataformas com relação competitiva conhecida.

4. Estudo de Caso e Resultados

O estudo utiliza uma base proprietária com 441 mil registros de preços de 30 produtos, dez marcas e 21 plataformas de *e-commerce*, observados ao longo de 500 dias entre julho

de 2023 e março de 2025. O foco recai sobre a categoria de cosméticos e perfumaria, escolhida por combinar alta frequência de reajustes, presença multicanal e comparabilidade entre itens. Na instanciação empírica do método, adota-se janela deslizante de tamanho $w = 7$ dias e faixa de atrasos $\tau \in [-7, +7]$. A escolha de $w = 7$ é coerente com o ciclo comercial semanal e com a escala de reação investigada, enquanto janelas maiores ou menores envolvem o compromisso usual entre sensibilidade local e suavização [Chen et al., 2016]. A análise empírica organiza-se em três estudos complementares: (i) uma leitura agregada do mercado, por meio da média das correlações cruzadas de dez produtos representativos, selecionados pela maior densidade de observações comparáveis entre plataformas; (ii) a inspeção de casos micro para examinar padrões de liderança de preço, convergência e sazonalidade competitiva; e (iii) uma verificação de consistência externa com pares de plataformas pertencentes ao mesmo grupo econômico.

A Figura 1(a) apresenta a distribuição das correlações médias por *lag* para dez produtos representativos, selecionados pela maior densidade de observações comparáveis entre plataformas. Observa-se um comportamento simétrico, com pico em *lag* 0 e decaimento suave nos *lags* adjacentes (+1 e -1). Esse padrão sugere sincronização agregada de curto prazo: em média, as plataformas tendem a ajustar seus preços no mesmo dia ou no dia seguinte em resposta aos concorrentes [Lock et al., 2020]. O resultado, porém, deve ser lido como evidência exploratória agregada [Chen et al., 2016; Wang e Li, 2020; Aparicio et al., 2024].

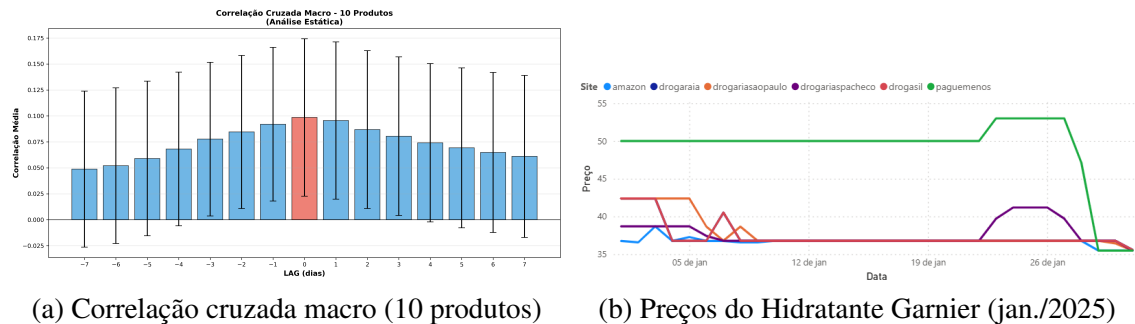


Figura 1. Visão agregada e exemplo micro de liderança-resposta no e-commerce.

A Figura 1(b) ilustra a dinâmica de preços do produto *Hidratante Garnier* em janeiro de 2025. Observa-se convergência: no início do mês, há dispersão entre R\$35 e R\$50; a partir do dia 12, a maioria das plataformas, em especial redes farmacêuticas, ajusta seus preços para convergir ao valor praticado pela Amazon, que surge como candidata plausível a *price leader* nesse período. A aplicação da correlação cruzada nesse intervalo revela coeficientes elevados ($r > 0,8$) com *lags* entre 0 e 2 dias, indicando reação não imediata entre competidores e sugerindo diferenças na sofisticação tecnológica de monitoramento [Bungert, 2003]. A convergência ocorre com atraso pequeno, porém mensurável, o que torna a leitura útil para estudos exploratórios de competição digital [Chen et al., 2016; Hanspach et al., 2024].

A análise de janela deslizante mostra que a correlação não é estática: a Figura 2(a) evidencia aumento significativo em períodos promocionais. Durante a preparação para a *Black Friday* (outubro de 2024), observa-se uma “corrida para o fundo” (*race to the bottom*) no produto *Elseve Óleo Extraordinário*, com correlações cruzadas sustentadas acima de 0,9. Em contrapartida, em períodos pós-sazonais, como janeiro, a correlação

diminui, indicando acomodação de preços, na qual fatores de estoque interno prevalecem sobre a pressão competitiva. Uma correlação calculada sobre toda a série mistura períodos de competição agressiva e de estabilidade, reduzindo a capacidade analítica do método [Wang e Li, 2020; Aparicio et al., 2024; González e Miles-Touya, 2018].

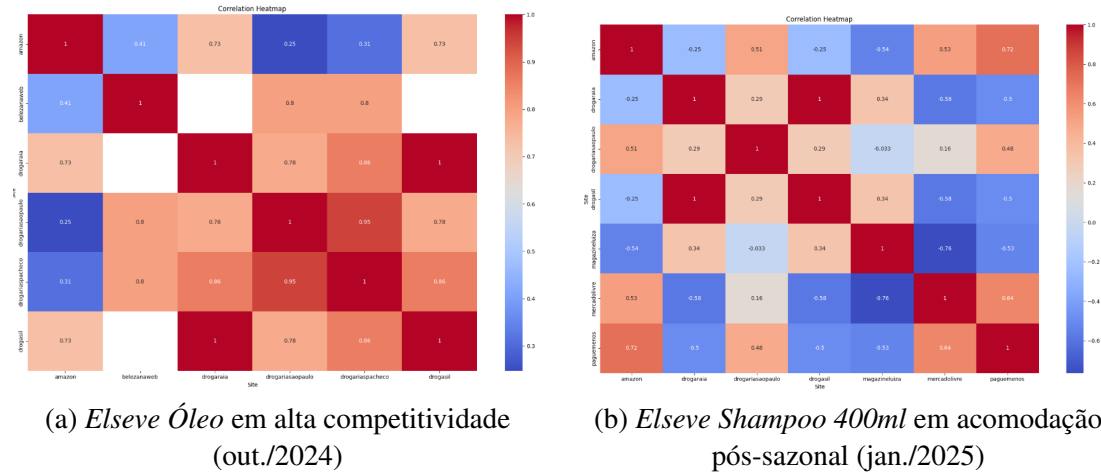


Figura 2. Contraste entre regimes competitivos identificados.

A identificação de correlações persistentemente elevadas entre plataformas pertencentes ao mesmo grupo econômico, como Droga Raia e Drogasil (grupo RD Saúde) e Drogarias Pacheco e Drogaria São Paulo (grupo DPSP), funciona como validação de sanidade (*sanity check*): ao recuperar relações conhecidas de coordenação comercial intragrupo, o método aumenta a confiança na interpretação de relações menos óbvias entre concorrentes distintos. Observa-se heterogeneidade na resposta entre tipos de plataforma, com *marketplaces* amplos tendendo a propagar movimentos mais rapidamente do que redes especializadas [Bungert, 2003].

A Figura 2(b) ilustra um cenário de menor acoplamento competitivo. Em janeiro de 2025, o produto *Elseve Shampoo 400ml* apresenta correlações positivas mais moderadas e menos homogêneas entre plataformas. Isso indica que a liderança de preço não é constante nem uniforme entre produtos: ela emerge com mais força quando o ambiente comercial cria incentivos para monitoramento intenso, como em campanhas promocionais ou momentos de reposição de estoque.

5. Conclusão

A análise com correlação cruzada e janelas deslizantes permite explorar padrões de liderança e resposta na competitividade do *e-commerce* brasileiro que leituras estáticas tendem a ignorar. Os resultados sugerem três padrões recorrentes: (i) episódios compatíveis com liderança de preço; (ii) heterogeneidade de resposta, com atrasos de 0 a 3 dias entre tipos de plataforma; e (iii) não estacionariedade competitiva, com intensificação das correlações em períodos promocionais e atenuação em fases de acomodação. A proposta se apresenta, portanto, como alternativa exploratória intermediária entre a simplicidade da correlação estática e a complexidade de modelos causais completos, adequada a bases reais e imperfeitas de *e-commerce*. Como limitações, destacam-se o uso de dados proprietários, a presença de observações ausentes e a sensibilidade das evidências às escolhas de parametrização e agregação.

Referências

- ABComm (2024). Crescimento do E-commerce Brasileiro. Technical report, <https://dados.abcomm.org/crescimento-do-ecommerce-brasileiro>.
- Aparicio, D., Metzman, Z., e Rigobon, R. (2024). The pricing strategies of online grocery retailers. *Quantitative Marketing and Economics*, 22:1 – 21.
- Bai, J., Choi, S. H., e Liao, Y. (2021). Feasible generalized least squares for panel data with cross-sectional and serial correlations. *Empirical Economics*, 60:309 – 326.
- Bungert, M. (2003). *Termination of Price Wars: A Signaling Approach*. Deutscher Universitätsverlag.
- Chen, L., Mislove, A., e Wilson, C. (2016). An empirical analysis of algorithmic pricing on amazon marketplace. In *25th International World Wide Web Conference, WWW 2016, WWW '16*, pages 1339 – 1349. International World Wide Web Conferences Steering Committee.
- Chen, P., Zhao, R., Yan, Y., e Li, X. (2020). Promotional pricing and online business model choice in the presence of retail competition. *Omega (United Kingdom)*, 94.
- Faehnle, A. e Guidolin, M. (2021). Dynamic Pricing Recognition on E-Commerce Platforms with VAR Processes. *Forecasting*, 3:166 – 180.
- Gerpott, T. J. e Berends, J. (2022). Competitive pricing on online markets: a literature review. *Journal of Revenue and Pricing Management*, 21:596 – 622.
- González, X. e Miles-Touya, D. (2018). Price dispersion, chain heterogeneity, and search in online grocery markets. *SERIEs*, 9:115 – 139.
- Hanspach, P., Sapi, G., e Wieting, M. (2024). Algorithms in the marketplace: An empirical analysis of automated pricing in e-commerce. *Information Economics and Policy*, 69.
- Hyndman, R. J. e Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.
- Lock, R. H., Lock, P. F., Morgan, K. L., Lock, E. F., e Lock, D. F. (2020). *Statistics: Unlocking the Power of Data*. John Wiley & Sons.
- Nalca, A., Ray, S., e Boyaci, T. (2018). Price Matching Strategy: Implications of Consumer Behaviour and Channel Structure.
- Ogasawara, E., Salles, R., Porto, F., e Pacitti, E. (2025). *Event Detection in Time Series*. Synthesis Lectures on Data Management. Springer Nature Switzerland, Cham, 1 edition.
- Rejón-Guardia, F. e Luna-Nevarez, C. (2017). “Showrooming” in Consumer Electronics Retailing: An Empirical Study. *Journal of Internet Commerce*, 16:174 – 201.
- Sarwate, D. V. e Pursley, M. B. (1980). Crosscorrelation Properties of Pseudorandom and Related Sequences. *Proceedings of the IEEE*, 68:593 – 619.
- Spann, M., Bertini, M., Koenigsberg, O., Zeithammer, R., Aparicio, D., Chen, Y., Fantini, F., Jin, G. Z., Morwitz, V. G., Popkowski Leszczyc, P., Vitorino, M. A., Williams, G. Y., e Yoo, H. (2025). Algorithmic pricing: Implications for marketing strategy and regulation. *International Journal of Research in Marketing*.
- Wang, W. e Li, F. (2020). What determines online transaction price dispersion? Evidence from the largest online platform in China. *Electronic Commerce Research and Applications*, 42:100968.