

Análise da sazonalidade do Vírus Sincicial Respiratório no Brasil usando *Clustering*

Alejandro Martins de Freitas¹, Fernanda Sumika Hojo de Souza¹

¹Departamento de Computação – Universidade Federal de Ouro Preto (UFOP)
Ouro Preto – MG – Brasil

alejandro.freitas@aluno.ufop.edu.br, fsumika@ufop.edu.br

Abstract. *Respiratory Syncytial Virus (RSV) is one of the main agents of lower respiratory tract infection in vulnerable populations. The COVID-19 pandemic and non-pharmacological interventions have temporarily altered its known seasonality. This study investigates whether RSV hospitalizations in Brazil present distinct regional patterns and evaluates their organization before, during, and after the COVID-19 pandemic. Using data from SIVEP-Gripe (2015-2025), STL decomposition and unsupervised learning were applied. Results indicate changes in the seasonality, duration, and amplitude of epidemic outbreaks in the analyzed periods.*

Resumo. *O Vírus Sincicial Respiratório (VSR) é um dos principais agentes de infecção do trato respiratório inferior em populações vulneráveis. A pandemia de COVID-19 e as intervenções não farmacológicas alteraram temporariamente sua sazonalidade conhecida. Este trabalho investiga se as hospitalizações por VSR no Brasil apresentam padrões regionais distintos e avalia sua organização antes, durante e após a pandemia da COVID-19. Utilizando dados do SIVEP-Gripe (2015-2025), aplicou-se decomposição STL e aprendizado não supervisionado. Resultados indicam mudanças na sazonalidade, duração e amplitude dos surtos epidêmicos nos períodos analisados.*

1. Introdução

O Vírus Sincicial Respiratório (VSR) é um dos principais agentes causadores de infecções agudas do trato respiratório inferior e acomete principalmente lactentes e crianças até os 5 anos, assim como idosos acima de 65 anos [Falsey and Walsh 2005]. Estimativas globais indicam aproximadamente 3,6 milhões de hospitalizações anuais e cerca de 100 mil mortes associadas ao VSR em crianças menores de cinco anos, com a maioria ocorrendo em países de baixa e média renda [Rago et al. 2024, Shi et al. 2022].

O VSR é de distribuição mundial, causando surtos sazonais que foram previsíveis até 2020. Com as intervenções não-farmacêuticas (NPIs) adotadas em 2020 devido à pandemia da COVID-19, tais como distanciamento social, uso de máscaras e higienização das mãos, houve uma significativa redução de casos e de hospitalização por VSR [Sanz-Muñoz et al. 2021]. Por outro lado, o relaxamento em relação às NPIs levou a surtos de casos por VSR fora dos períodos sazonais em 2021 [Sah et al. 2023], levantando preocupação sobre o fenômeno do “débito de imunidade”.

Apesar da relevância, estudos sobre as mudanças na sazonalidade do VSR pós-pandemia em diferentes países, incluindo o Brasil, ainda são escassos na literatura. Paralelamente, técnicas de aprendizado de máquina não supervisionado vem sendo utilizadas

para analisar padrões na dinâmica de epidemias. Por exemplo, [Silva et al. 2020] utilizaram métodos de agrupamento para identificar padrões na disseminação da COVID-19, demonstrando o potencial dessas técnicas na compreensão de dinâmicas epidemiológicas.

Neste contexto, este trabalho tem como objetivo analisar as hospitalizações associadas ao VSR no Brasil e identificar padrões de sazonalidade nas diferentes unidades federativas (UFs) entre 2015 e 2025, utilizando técnicas de análise de séries temporais e *clustering*. A compreensão desses padrões pode contribuir para o aprimoramento da vigilância epidemiológica e para o planejamento de ações de saúde pública.

2. Materiais e Métodos

A análise seguiu um *pipeline* de ciência de dados composto pelas etapas de preparação dos dados, decomposição temporal e agrupamento das séries. Foram utilizadas as seguintes bibliotecas para Python: Pandas, Scikit-learn, TSlearn, Matplotlib, NumPy, e Seaborn.

A base de dados utilizada é disponibilizada pelo Ministério da Saúde através do Sistema de Informação de Vigilância Epidemiológica da Gripe (SIVEP-Gripe). Foram extraídos registros de hospitalização por Síndrome Respiratória Aguda Grave (SRAG) confirmados para VSR entre os anos de 2015 e 2025¹. Após o tratamento dos dados obteve-se um conjunto contendo 167.399 registros. Os dados foram agregados espacialmente por UF e temporalmente por semana epidemiológica (1 a 52), produzindo séries temporais de hospitalizações associadas ao VSR para cada UF no período analisado.

Para analisar a estrutura das séries temporais de hospitalizações, foi aplicado o método de decomposição STL (*Seasonal-Trend decomposition using Loess*) [Cleveland et al. 1990]. A partir do método STL, são geradas três componentes de uma série, a tendência, a sazonalidade e o ruído, que auxiliam na compreensão de padrões subjacentes aos dados. A escolha do método é justificada por sua robustez a *outliers* e por sua flexibilidade na modelagem de componentes sazonais em séries temporais. A decomposição foi aplicada separadamente às séries semanais de hospitalizações de cada UF ao longo de todo o período analisado (2015-2025). A componente sazonal estimada foi extraída e segmentada em três períodos de análise: pré-pandemia (semanas 1 de 2015 a 10 de 2020), pandemia (semanas 11 de 2020 a 52 de 2022) e pós-pandemia (semanas 1 de 2023 a 40 de 2025). A decomposição foi realizada sobre a série completa para garantir uma estimativa mais estável da componente sazonal ao longo de múltiplos ciclos anuais.

Visando identificar UFs com comportamento sazonal semelhante, as componentes sazonais das séries temporais foram agrupadas por período utilizando aprendizado não supervisionado. Foi aplicada a normalização *Mean-Variance* pois o foco do agrupamento considerado eram as curvas (formato das ondas), desconsiderando a discrepância de magnitude populacional entre os estados. O algoritmo de agrupamento *K-Medoids* [Kaufman and Rousseeuw 1990], combinado com a métrica de distância *Dynamic Time Warping* (DTW) [Sakoe and Chiba 1978] pré-computada, foi escolhido por sua maior robustez a *outliers* e por permitir o uso direto de matrizes de distância não euclidianas. A métrica DTW foi utilizada para permitir alguma flexibilidade na comparação entre séries com pequenas defasagens temporais (comum na disseminação viral), sem deixar de capturar diferenças mais amplas nos padrões sazonais. Métodos tradicionais como distância

¹<https://dadosabertos.saude.gov.br>

Euclidiana podem falhar em capturar, justificando a escolha metodológica. A validação interna dos agrupamentos foi realizada avaliando-se o *Silhouette Score* e o somatório das distâncias intra-cluster. Foram analisados valores do hiperparâmetro $k \in \{4, 5, 6\}$.

Para auxiliar na interpretação dos agrupamentos, foram calculadas métricas descritivas das séries temporais originais de hospitalizações por UF: semana do pico epidêmico, duração da onda sazonal e amplitude da série. Essas medidas foram utilizadas para comparar diferenças no tempo e na intensidade das temporadas associadas ao VSR entre as UFs. Comparou-se a composição dos agrupamentos e possíveis mudanças de UFs entre grupos para observar a estabilidade temporal dos clusters ao longo dos períodos.

3. Resultados

A Figura 1 apresenta a evolução semanal das hospitalizações por VSR no Brasil entre 2015 e 2025. É possível observar um padrão sazonal no período pré-pandemia, com picos concentrados entre as semanas 15 e 22. Durante a pandemia da COVID-19 esse padrão sofre alterações, com antecipação das ondas epidêmicas e ocorrência de surtos em períodos menos usuais (após a semana 43). Nos anos pós-pandemia, observa-se uma retomada do comportamento sazonal, além de um aumento expressivo no número de hospitalizações.

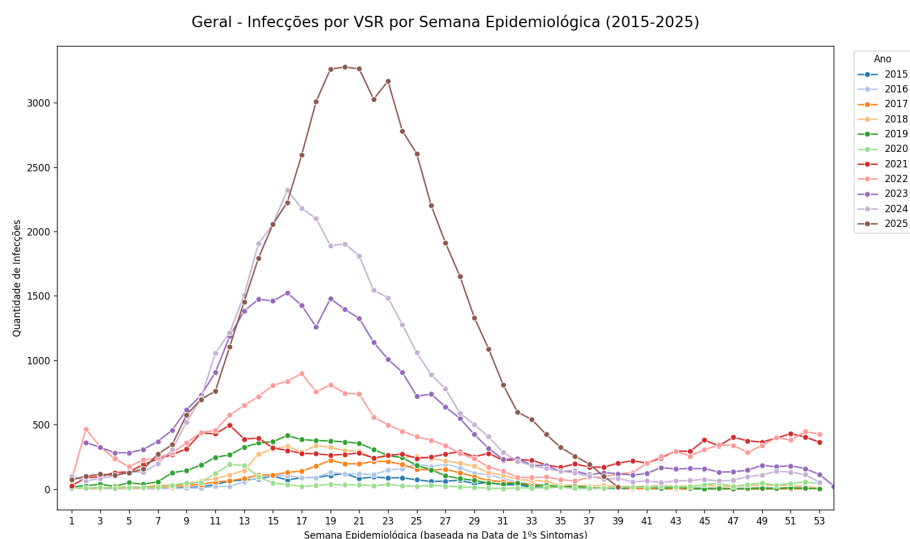


Figura 1. Séries temporais das hospitalizações por SRAG/VSR no Brasil entre 2015 e 2025.

A definição do número ideal de agrupamentos (k) foi baseada na métrica de *Silhouette Score* e no somatório das distâncias intra-cluster. A Tabela 1 sumariza os valores obtidos para os três períodos temporais. A configuração com $k = 5$ apresentou melhor equilíbrio entre separação dos agrupamentos, sendo adotado nas análises subsequentes. A Tabela 2 apresenta os padrões sazonais médios identificados para $k = 5$ nos períodos avaliados. No período pré-pandemia, é possível observar a formação de clusters com diferenças no momento do pico e na duração médias das ondas. Os picos concentram-se entre as semanas epidemiológicas 14 e 23. Durante o período da pandemia, observa-se uma reorganização dos clusters e mudanças na métricas sazonais. Os dados indicam uma antecipação dos picos (distribuídos entre as semanas 4 e 14) e com uma menor duração médias das ondas (entre 2 a 5 semanas na maioria dos clusters). No cenário pós-pandemia,

observa-se uma estrutura de clusters semelhante à observada anteriormente, porém, com mudanças nas métricas de sazonalidade. Os picos voltam a ocorrer entre as semanas 14 e 23, sugerindo uma retomada do padrão sazonal.

Tabela 1. Avaliação dos agrupamentos para diferentes valores de k

	$k = 4$		$k = 5$		$k = 6$	
Cenário	Silh.	$\sum$ Dist	Silh.	$\sum$ Dist	Silh.	$\sum$ Dist
Pré-Pandemia	0,17	161,38	0,18	148,74	0,19	136,76
Pandemia	0,21	96,11	0,21	88,72	0,19	82,23
Pós-Pandemia	0,21	97,15	0,22	89,33	0,20	82,66

Um comportamento distinto pode ser observado para o estado do Mato Grosso (MT), que aparece em um cluster próprio tanto no cenário da pandemia quanto na pós-pandemia, necessitando uma análise mais aprofundada. Outro aspecto notado é o aumento das amplitudes médias no cenário pós-pandemia em comparação aos demais períodos. Esse comportamento é consistente com os dados apresentados na Figura 1, que indica um maior número de hospitalizações nos anos posteriores à pandemia.

Tabela 2. Características dos clusters sazonais identificados para $k = 5$.

Cluster	UFs	Pico (sem.)	Duração (sem.)	Amplitude
Pré-pandemia				
CO	DF, ES, GO, MG	14,65	9,00	24,78
C1	AP, CE, PI, RN	22,50	1,45	34,17
C2	AC, MT, PB, RR	17,35	6,95	3,44
C3	AM, BA, MS, PA, PE, PR, RJ, RS, SC, SE, SP, TO	19,43	7,12	33,75
C4	AL, MA, RO	16,33	8,53	3,13
Pandemia				
CO	MT	4,00	3,00	2,21
C1	PI, RR	13,50	5,50	4,80
C2	AL, AM, CE, MG, PA, RN, SE, TO	10,00	4,19	29,92
C3	AC, AP, MA, PB, RO	12,40	2,80	12,71
C4	BA, DF, ES, GO, MS, PE, PR, RJ, RS, SC, SP	10,41	9,09	92,04
Pós-pandemia				
CO	BA, DF, ES, GO, MS, PE, PR, RJ, RS, SC, SP	17,91	8,67	156,60
C1	PI, RR	23,50	4,84	8,52
C2	AL, AM, CE, MG, PA, RN, SE, TO	19,83	4,38	46,44
C3	MT	14,00	3,00	2,21
C4	AC, AP, MA, PB, RO	22,20	2,73	21,72

A Figura 2 apresenta a sazonalidade das séries temporais agrupadas para o período da pandemia. As linhas cinzas representam as UFs pertencentes a cada cluster e a linha vermelha indica o medoide do agrupamento. Observa-se que os clusters formados capturaram diferentes padrões de forma e tempo dos surtos anuais, indicando uma reorganização das UFs nesse período. São apresentados apenas os resultados referentes à pandemia, uma vez que os períodos pré-pandemia e pós-pandemia apresentam comportamento consistente com os padrões descritos na Tabela 2.

4. Discussão e Conclusão

Os resultados obtidos indicam que houve alterações no padrão sazonal das hospitalizações por VSR no Brasil ao longo dos períodos analisados. Durante a pandemia, foi observada uma antecipação dos picos epidêmicos e uma duração mais curta dessas ondas. No

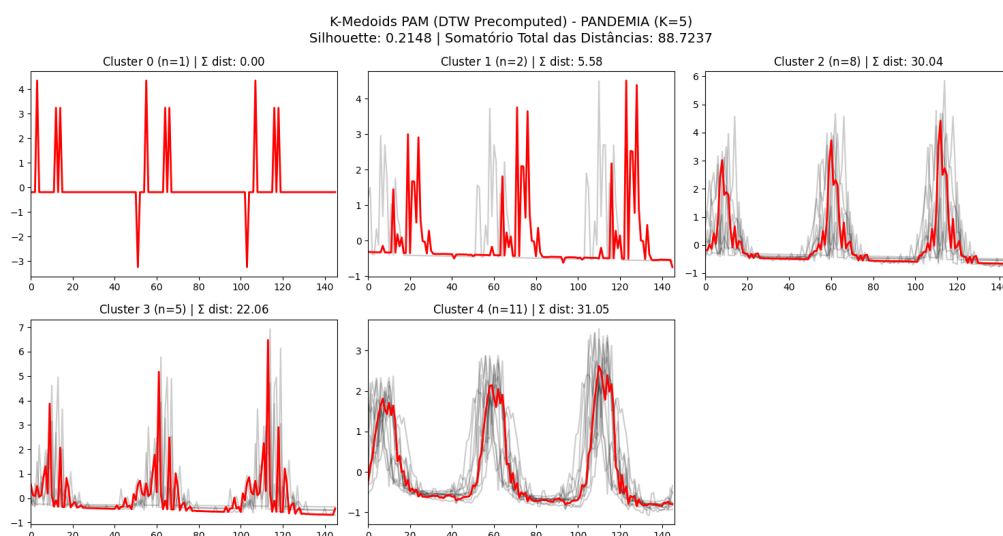


Figura 2. Séries temporais agrupadas pelo algoritmo K-Medoids com distância DTW para o período da pandemia.

período pós-pandemia, os picos voltaram a ocorrer em semanas próximas ao padrão observado antes de 2020, embora com maior amplitude, refletindo o aumento no número de casos de hospitalizações observado nas séries temporais agregadas.

A principal contribuição deste trabalho foi identificar padrões regionais distintos de temporalidade das epidemias do VSR nas diferentes UFs do Brasil por meio de técnicas de análise de séries temporais e aprendizado não supervisionado. O método proposto permitiu a identificação de clusters com padrões distintos mostrando que o VSR não circula de forma sincronizada no Brasil. Os resultados reforçam a heterogeneidade regional da sazonalidade do VSR no Brasil, sendo sujeita a sofrer reorganizações ao longo do tempo em função de fatores externos como a circulação de outros vírus, mudança climática, entre outros. Essa análise retrospectiva pode auxiliar no planejamento de ações de saúde pública como alertas epidemiológicos regionais em função dos momentos de pico observados, além de planejamento hospitalar em momentos de maior demanda.

Os resultados observados neste estudo são consistentes com estudos recentes. [Thindwa et al. 2024] analisaram séries temporais do VSR em diferentes países e identificaram mudanças na sazonalidade no período pós-pandemia e padrões distintos entre regiões. Segundo os autores, o início e o pico da primeira e segunda ondas epidêmicas pós-pandemia ocorreram mais cedo e posteriormente, os padrões da dinâmica do vírus tenderam a retornar gradualmente ao comportamento sazonal típico. O aumento na amplitude das ondas no período pós-pandêmico está alinhado com a hipótese do “débito de imunidade” discutida por [Sah et al. 2023, Zheng et al. 2021], sugerindo que a supressão temporária da circulação viral, devido às medidas de isolamento [Sanz-Muñoz et al. 2021], pode ter contribuído para o acúmulo de um contingente populacional suscetível.

Como limitações deste trabalho, alguns aspectos são listados a seguir. A base de dados é restrita a pacientes hospitalizados, cuja dinâmica pode diferir de casos ambulatoriais mais amplos. A ausência de variáveis climáticas não permite uma comparação com o atraso ou antecipação dos picos. Os valores da métrica silhouette indicam separação

moderada dos clusteres, embora seja esperado, dada a sazonalidade corresponde a um mesmo vírus, não havendo mudanças abruptas entre estados, e sim graduais. Como trabalhos futuros, será conduzida uma análise para caracterizar o perfil interno dos clusters utilizando variáveis clínicas e sociodemográficas presentes na base de dados. Pretende-se avaliar percentuais de lactentes e idosos, além da gravidade hospitalar (percentual de admissão em UTI, necessidade de suporte ventilatório e evolução para óbito).

5. Agradecimentos

Este trabalho foi realizado com apoio do CNPq, FAPEMIG (APQ-06570-24) e UFOP.

Referências

- Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. (1990). Stl: A seasonal-trend decomposition procedure based on loess (with discussion). *Journal of Official Statistics*, 6:3–73.
- Falsey, A. R. and Walsh, E. E. (2005). Respiratory syncytial virus infection in elderly adults. *Drugs & aging*, 22(7):577–587.
- Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley.
- Rago, A. R., D’Arrigo, S. F., Osmani, M., Espinosa, C. M., and Torres, C. M. (2024). Respiratory syncytial virus: Epidemiology, burden of disease, and clinical update. *Advances in Pediatrics*, 71(1):107–118. *Advances In Pediatrics*.
- Sah, R., Zaman, K., Mohanty, A., Al-Ahdal, T., Awad, H., Padhi, B. K., and Bhargava, A. (2023). Respiratory syncytial virus with ongoing covid-19: is it an emerging threat? *Annals of Medicine and Surgery*, 85(1):67–70.
- Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.
- Sanz-Muñoz, I., Tamames-Gómez, S., Castrodeza-Sanz, J., Eiros-Bouza, J. M., and de Lejarazu-Leonardo, R. O. (2021). Social distancing, lockdown and the wide use of mask; a magic solution or a double-edged sword for respiratory viruses epidemiology? *Vaccines*, 9(6):595.
- Shi, T., Vennard, S., Mahdy, S., and Nair, H. (2022). Risk factors for poor outcome or death in young children with respiratory syncytial virus-associated acute lower respiratory tract infection: A systematic review and meta-analysis. *The Journal of infectious diseases*, 226(Supplement_1):S10–S16.
- Silva, R., Xavier, F., Saraiva, A., and Cugnasca, C. (2020). Unsupervised machine learning and pandemics spread: the case of covid-19. In *Anais do XX Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 506–511, Porto Alegre, RS, Brasil. SBC.
- Thindwa, D., Li, K., Cooper-Wootton, D., Zheng, Z., Pitzer, V. E., and Weinberger, D. M. (2024). Global patterns of rebound to normal rsv dynamics following covid-19 suppression. *BMC Infectious Diseases*, 24:635.
- Zheng, Z., Pitzer, V. E., Shapiro, E. D., Bont, L. J., and Weinberger, D. M. (2021). Estimation of the timing and intensity of reemergence of respiratory syncytial virus following the covid-19 pandemic in the us. *JAMA Network Open*, 4(12):e2141779–e2141779.