Universidad Nacional de San Agustín de Arequipa

Escuela Profesional de Ciencia de la Computación

Tópicos en Ciencia de Datos

Grupo: Semestre 2025 B

Análisis visual multifactor de la evolución espacio-temporal de los casos de COVID-19 y su relación con factores sociodemográficos

Alumno:
John Edson Sanchez Chilo

Docente :
Ana María Cuadros Valdivia

Índice

1. Introducción

La enfermedad por coronavirus 2019 (COVID-19), causada por el virus SARS-CoV-2, emergió a finales de 2019 en Wuhan, China, y rápidamente se convirtió en una pandemia global que afectó todos los aspectos de la vida humana. Su rápida propagación [18], y la falta de preparación para afrontar la pandemia por parte de los gobiernos y las autoridades [1], pusieron a prueba los sistemas de salud, economía y la sociedad en su conjunto, afectando especialmente a ciertos sectores de la población.

La pandemia, evidenció las profundas desigualdades estructurales presentes en muchas regiones. Las poblaciones que ya enfrentaban condiciones sociales y económicas precarias —como acceso limitado a servicios de salud, empleo, mayor densidad poblacional, o mayor número de personas de edad avanzada— fueron particularmente vulnerables a una mayor propagación del virus [19]. En muchos casos, la ausencia de información clara sobre qué factores sociodemográficos influían con mayor peso en la dinámica de contagio dificultó la toma de decisiones efectivas por parte de las autoridades. Esta falta de conocimiento impidió aplicar medidas focalizadas que protegieran a los sectores más expuestos, limitando así la capacidad de respuesta y aumentando el impacto del virus en comunidades ya desfavorecidas.

Por ello, la integración de datos sociodemográficos es fundamental para comprender mejor la dinámica de propagación del COVID-19, ya que factores como la densidad poblacional, la edad, el nivel educativo, el ingreso económico y el acceso a servicios de salud influyen significativamente en la vulnerabilidad y exposición de las personas al virus [5]. La relación entre estas variables y la incidencia de casos varía según el contexto local y las desigualdades estructurales presentes [16], por lo que es esencial desarrollar estrategias de salud pública basadas en una comprensión integral de estas dinámicas sociales complejas.

La disponibilidad de datos sobre casos de COVID-19 por fecha y localidad ha permitido una mayor comprensión del desarrollo de la pandemia en diferentes territorios [6]. Gracias a los sistemas de vigilancia epidemiológica y al registro continuo de casos, es posible realizar análisis temporales y espaciales que revelan cómo evoluciona el virus en función de distintos contextos sociales. Sin embargo, esta abundancia de información también presenta importantes desafíos, ya que el volumen y la variedad de los datos dificultan su organización, procesamiento e interpretación efectiva.

Además, él análisis de la evolución de los casos de COVID-19 se ve notablemente dificultado por la amplia variedad de variables sociodemográficas involucradas, las cuales interactúan de manera dinámica y no lineal. Esta alta dimensionalidad de datos complica la detección de tendencias claras y la interpretación visual de los resultados usando métodos convencionales [21], especialmente cuando se busca mantener la coherencia temporal y geográfica de los datos. A pesar de su importancia, son aún limitados los estudios que integran de forma efectiva datos sociodemográficos con registros epidemiológicos mediante herramientas computacionales y visuales robustas [20], lo que limita la capacidad de los expertos en la toma de decisiones para generar respuestas informadas y focalizadas.

Para abordar este problema, se propone desarrollar una herramienta visual que permita analizar la evolución de los casos de COVID-19 a lo largo del tiempo y el espacio, en relación con los factores sociodemográficos de cada localidad. Esta herramienta facilitará la exploración y visualización de patrones espaciotemporales y sociodemográficos por localidad, además de permitir una interacción dinámica y focalizada con el usuario.

2. Trabajos Relacionados

En los últimos años, el análisis visual de datos se ha consolidado como una herramienta fundamental para comprender fenómenos complejos en el ámbito de la salud pública, especialmente en escenarios de propagación de enfermedades infecciosas. Diversos estudios han explorado la integración de datos espaciotemporales, movilidad humana y variables socio-demográficas para identificar patrones de transmisión y vulnerabilidad en distintas comunidades. Esta sección revisa investigaciones previas que han propuesto métodos de visualización aplicados a epidemias, y mas especificamente a conjuntos de datos de COVID-19, algunas como herramientas interactivas y otras con asitencia inteligente, todo ello con el objetivo de identificar sus aportes, limitaciones y oportunidades de mejora que fundamentan la presente propuesta.

A inicios del año 2020, la aparición del COVID-19 y su rápida propagación a nivel mundial, como lo evidencia el estudio de Park et al. [18], impulsaron a la comunidad científica a enfocar con mayor intensidad sus esfuerzos en el análisis epidemiológico. Riccaboni y Verginer [20] concluyeron que los términos médicos relacionados con COVID-19 experimentaron un incremento promedio de 6.5 veces en la producción científica desde el inicio de la pandemia, lo cual refleja un notable aumento en el interés por investigar enfermedades infecciosas y sus dinámicas de propagación.

En este contexto, el estudio de Rydow et al. [21] demostró que el análisis visual desempeña un papel clave en el estudio epidemiológico, al facilitar la interpretación de modelos complejos y apoyar la toma de decisiones basada en datos.

En las primeras etapas del desarrollo de herramientas de visualización aplicadas a estudios epidemiológicos, era común el uso de representaciones gráficas simples, como mapas de calor [22], series temporales [2], o gráficos compuestos de múltiples vistas [3]. Si bien estas visualizaciones permitían comunicar ciertos aspectos generales de la evolución de la enfermedad, su capacidad para representar relaciones complejas entre variables espaciales, temporales y socio-demográficas era limitada.

Los estudios sobre enfermedades infecciones se apoyan principalmente en dos tipos de datos, los datos espacio-temporales y los datos multidimensionales. Los primeros permiten capturar la evolución y distribución geográfica de las infecciones a lo largo del tiempo, es decir la visualizacion geográfica [11, 15, 23] puede representar de forma efectiva, datos multivariables en un espacio y tiempo específico.

Los segundos integran múltiples variables clínicas, epidemiológicas y ambientales para ofrecer una perspectiva más compleja y completa del fenómeno. Podemos incluir

datos sobre movilidad urbana [8, 14], factores económicos [26], sociodemográficos [10, 25] y otros más.

En el contexto específico de la visualización de datos relacionados con la pandemia de COVID-19, debido a la rápida y extensa propagación del virus a nivel global, surgió la necesidad de desarrollar múltiples tipos de gráficos y representaciones visuales que permitan analizar y comunicar eficazmente su impacto en distintas naciones [17] [4] [7].

Además de los esfuerzos gubernamentales, instituciones académicas han aportado significativamente al desarrollo de herramientas visuales para el seguimiento de la pandemia. Por ejemplo, la Universidad Johns Hopkins creó un dashboard interactivo ampliamente reconocido para visualizar en tiempo real la evolución global del COVID-19 [9]. Paralelamente, para facilitar el acceso y análisis de datos, también pusieron a disposición un repositorio público en GitHub que recopila y actualiza constantemente la información epidemiológica [13].

En la investigación relacionada con datos de COVID-19, comúnmente se manejan dos tipos principales de fuentes de datos, el primer tipo corresponde a datos directamente vinculados con la enfermedad, que incluyen información objetiva y cuantificable como el número de casos confirmados, las tasas de recuperación y las tasas de mortalidad [15, 24]. Estos datos son fundamentales para el seguimiento epidemiológico y permiten evaluar la evolución y gravedad del brote en diferentes regiones.

Por otro lado, existe un segundo tipo de datos, denominados indirectamente vinculados, que no reflejan factores objetivos directamente relacionados con la infección, pero que capturan aspectos comunitarios y sociales derivados de la pandemia. Entre estos se encuentran indicadores relacionados con los impactos sociales, como cambios en el comportamiento poblacional o en la movilidad [8, 14], así como los efectos económicos y financieros ocasionados por la crisis sanitaria [26] o también los datos sociodemográficos de la población [10]. Aunque estos datos no están directamente conectados con la transmisión o la incidencia del virus, resultan cruciales para comprender el alcance total de la pandemia y para diseñar estrategias integrales de respuesta que consideren tanto la salud pública como las dimensiones sociales y económicas.

3. Propuesta

En este trabajo, se propone un modelo integral para el análisis evolutivo espacio-temporal de la propagación de Covid-19, considerando tanto los casos de Covid-19 como los factores sociodemográficos que los influyen. El modelo busca abordar la complejidad de estas interacciones mediante un enfoque multifactorial que permite identificar, cuantificar y representar visualmente las relaciones dinámicas entre los casos de Covid-19 y variables sociodemográficas clave como densidad poblacional, promedio de ingresos, promedio de renta, promedio de ingreso familiar, población de diferentes edades, tamaño de los hogares, habitaciones por hogas, o medio de transporte utilizado.

3.1. Esquema de la propuesta

El esquema de la propuesta, representado en la Figura 4.1, se estructura en módulos interconectados diseñados para abordar de manera integral el análisis evolutivo espacio-temporal de la calidad del aire. Cada módulo incluye visualizaciones específicas que se adaptan a diferentes objetivos analíticos, permitiendo un enfoque personalizado según la tarea a realizar. Estas visualizaciones están respaldadas por procesos de pre-procesamiento diseñados para preparar los datos de manera óptima, garantizando que la información sea presentada de forma intuitiva, sintetizada y clara, facilitando la comprensión y la toma de decisiones informadas.

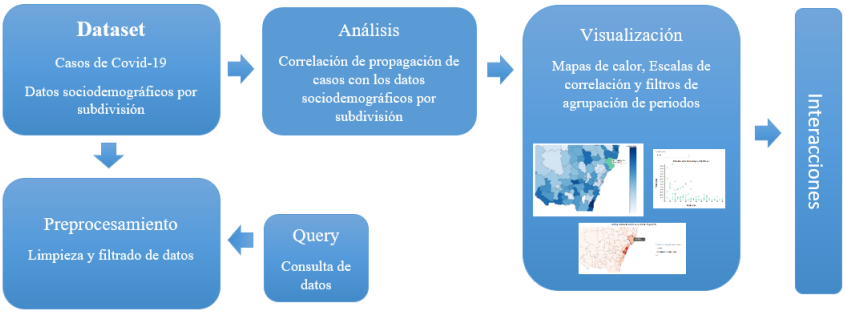

Figura 1: Pipeline del modelo propuesto

3.2. Tareas de Análisis

Después de revisar estudios previos para identificar los principales desafíos en el análisis de los datos. Los expertos destacaron la necesidad de herramientas que simplifiquen la representación, selección y análisis de la información, permitiendo estudiar las diferentes tendencias, estacionalidades y variaciones entre periodos. Asimismo, las tareas fueron complementadas con referencias a trabajos previos, resaltando la importancia de acelerar el estudio de la evolución multifactor de casos Covid-19 y su relación con factores sociodemográficos, de las cuales se seleccionaron cuatro tareas que se pueden realizar en la herramienta de análisis propuesta:

- T1: Comprender la visión general de los cambios en la distribución temporal de los casos de Covid-19 y los factores sociodemográficos. Al analizar una región específica, los usuarios deben identificar los cambios temporales clave, incluyendo cuando, dónde y que factores estan implicados en dichas variaciones.
- T2: Explorar patrones de tendencia y temporalidad específicos de casos Covid-19 en función de subdivisiones de casos de Covid-19. Los usuarios pueden identificar patrones filtrando y comparando datos de casos de Covid-19 y factores sociodemográficos según los diferentes periodos del año. Cuando se identifiquen patrones relevantes, como valores atípicos o aumentos bruscos de casos de Covid-19, los usuarios podrán profundizar en los detalles, analizando la regiones con mayor propagación, y sus

factores sociodemográficos que mas afectan a ello, con una visualización detallada para facilitar el análisis.

- T3: Realizar un análisis de correlación para explorar la relación entre los casos de Covid-19 y variables sociodemográficas. Este análisis ayudará a comprender cómo estos factores afectan la propagación de casos de Covid-19 en las diferentes subdivisiones.
- T4: Visualizar la evolución de patrones en los casos de Covid-19 y factores sociodemográficos. El sistema debe facilitar la representación gráfica de la evolución temporal de los patrones, permitiendo comparar los casos de Covid-19 y los factores sociodemográficos a lo largo de un periodo de tiempo o dentro de una agrupación seleccionada.

3.3. Dataset

El conjunto de datos utilizado en este estudio proviene de Nueva Gales del Sur, una región de Australia, y abarca el período desde el 25 de enero de 2020 hasta el 07 de febrero de 2022. Los datos fueron recopilados por el Ministerio de Salud de Nueva Gales del Sur y están disponibles en la pagina de Salud del Gobierno [12]. Este conjunto de datos incluye casos Covid-19 encontrados dentro de Nueva Gales del Sur, contienen la fecha, el código del distrito local de salud (LHD) y el código del Area de Gobierno Local (LGA). El segundo conjunto contiene datos sociodemográficos y económicos a nivel de Local Government Areas (LGAs), extraídos del censo de 2016, las cuales serán usadas con la intención de analizar características comunitarias y posiblemente correlacionarlas con los patrones de contagio.

3.4. Preprocesamiento

El preprocesamiento es una fase crítica en el análisis de datos, ya que prepara los datos brutos para su análisis posterior. Esta etapa asegura que los datos sean limpios, coherentes y adecuados para los métodos analíticos que se aplicarán. A continuación, se describen algunos aspectos clave del preprocesamiento:

3.4.1. Filtrado de datos

Los casos de Covid-19, tambien presentan una serie de problemas respecto a como se manejan. Lo primero es los datos faltantes de las ubicaciones de los casos encontrados, se tiene una serie de casos de los cuales aunque se sabe la fecha y el distrito local, no se conoce el LGA correspondiente, por tanto deben ser eliminados, no se realizará ningun proceso de imputación de datos debido que al tratarse de casos de Covid-19 los cuales van aumentando con el tiempo en las diferentes areas, no se puede estimar con ningún metodo a cual de todas pertenece.

Tampoco se trabajaran con diferentes areas de gobierno local las cuales presentan datos sociodemográficso incoherentes, o que no los tengan completos, esto con el fin evitar problemas con los modelos al usar data incompleta, además que en este caso tampoco se puede realizar ningún métodod e imputación al tratarse de areas de gobierno local totalmente independientes, por lo cual un acercamiento geográfico por los vecinos no sería de gran ayuda y se prefiere mantener solo los datos completos.

3.4.2. Agrupación de datos

Se tiene como datos casi un millon de casos de Covid-19 encontrados en las diferentes areas de gobierno local en Nueva Gales del Sur, esta cantidad inmensa de datos, no puede ser visualizada directamente, por lo cual se procederá agrupar mediante un conteo de casos por mes y area de gobierno local, de esta manera se busca reducir datos inncesarios y solo mantener un resumen de ellos, el cual podra ser usado a posterior por el modelo y la visualización.

3.5. Modelos usados

3.5.1. Modelo de Regresión de Datos Panel

Dado que los datos disponibles están estructurados por unidad geográfica (LGA) y por fecha, se optó por utilizar un modelo de regresión de datos panel. Este tipo de modelo permite capturar tanto la variación temporal como las diferencias estructurales entre regiones, controlando los efectos no observados específicos de cada LGA.

El modelo general adoptado se expresa como:

$$\text{Cases}_{it} = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i + \epsilon_{it} \tag{1}$$

donde:

- Cases_{it} representa el número de casos (o casos por cada 100,000 habitantes) en la LGA i durante el período t (semana o mes).
- $X_{1i}, X_{2i}, \ldots, X_{ki}$ son las variables sociodemográficas observadas para la LGA i, tales como edad mediana, ingreso familiar, densidad poblacional, porcentaje de adultos mayores, entre otras.
- u_i representa el efecto no observado y constante en el tiempo para cada LGA, que puede incluir factores estructurales como el acceso al sistema de salud, hábitos culturales o patrones de movilidad.
- ϵ_{it} es el término de error idiosincrático.

Este modelo permite estimar el efecto marginal de cada factor sociodemográfico sobre la incidencia de COVID-19, controlando por la heterogeneidad no observada entre regiones.

3.5.2. Análisis Espacial: Moran's I y LISA

Además del modelo panel, se realiza un análisis de autocorrelación espacial mediante el índice de Moran's I y el análisis LISA (*Local Indicators of Spatial Association*), con el fin de identificar patrones espaciales significativos en la distribución de los casos de COVID-19.

El índice de Moran's I global se expresa como:

$$I = \frac{N}{W} \cdot \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{2}$$

donde:

- x_i y x_j son los valores observados (por ejemplo, tasa de casos) en las LGA i y j respectivamente,
- $\bar{x}$ es el valor medio global,
- w_{ij} es el peso espacial entre las regiones i y j (basado, por ejemplo, en vecindad contigua o distancia),
- W es la suma total de los pesos espaciales,
- N es el número total de unidades espaciales.

Este análisis permite evaluar si los casos de COVID-19 tienden a concentrarse en ciertas áreas (aglomeración espacial) o si están distribuidos aleatoriamente en el espacio. El análisis LISA, por su parte, permite descomponer esta autocorrelación global en estadísticas locales para detectar clústeres significativos de alto o bajo contagio.

3.5.3. Justificación del uso combinado

El uso conjunto del modelo panel y de herramientas de análisis espacial permite abordar el problema desde una perspectiva integral: el modelo panel capta las relaciones causales entre variables explicativas y la evolución temporal de los casos, mientras que Moran's I y LISA aportan evidencia sobre la dependencia espacial y permiten identificar zonas críticas con comportamientos atípicos o agrupamientos inusuales.

Referencias

[1] Fahmi Y Al-Ashwal, Mohammed Kubas, Mohammed Zawiah, Ahmad Naoras Bitar, Ramzi Mukred Saeed, Syed Azhar Syed Sulaiman, Amer Hayat Khan, and Siti Maisharah Sheikh Ghadzi. Healthcare workers' knowledge, preparedness, counselling practices, and perceived barriers to confront covid-19: A cross-sectional study from a war-torn country, yemen. *PloS one*, 15(12):e0243962, 2020.

[2] Marco Angelini and Giorgio Cazzetta. Progressive visualization of epidemiological models for covid-19 visual analysis. In *AVI Workshop on Big Data Applications*, pages 163–173. Springer, 2020.

[3] Dario Antweiler, David Sessler, Maxim Rossknecht, Benjamin Abb, Sebastian Ginzel, and Jörn Kohlhammer. Uncovering chains of infections through spatio-temporal and visual analysis of covid-19 contact traces. *Computers & Graphics*, 106:1–8, 2022.

[4] Australian Government Department of Health and Aged Care. Monitoring and reporting on covid-19, 2025. Accedido: 23 de mayo de 2025.

[5] Jose Miguel Baena-Díez, María Barroso, Sara Isabel Cordeiro-Coelho, Jorge L Díaz, and María Grau. Impact of covid-19 outbreak by income: hitting hardest the most deprived. *Journal of Public Health*, 42(4):698–703, 08 2020.

[6] Md Arif Billah, Md Mamun Miah, and Md Nuruzzaman Khan. Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence. *PloS one*, 15(11):e0242128, 2020.

[7] California Department of Public Health. Weekly respiratory virus report, 2025. Accedido: 23 de mayo de 2025.

[8] Qi Cao, Renhe Jiang, Chuang Yang, Zipei Fan, Xuan Song, and Ryosuke Shibasaki. Metapopulation graph neural networks: Deep metapopulation epidemic modeling with human mobility. *arXiv preprint arXiv:2306.14857*, 2023.

[9] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 20(5):533–534, 2020.

[10] Yu Dong, Christy Jie Liang, Yi Chen, and Jie Hua. A visual modeling method for spatiotemporal and multidimensional features in epidemiological analysis: Applied covid-19 aggregated datasets. *Computational Visual Media*, 10(1):161–186, 2024.

[11] S. Goodwin, J. Dykes, A. Slingsby, and C. Turkay. Visualizing multiple variables across scale and geography. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):599–608, 2016.

[12] Government of New South Wales. Nsw government data portal, 2025. Accessed: 2025-06-12.

[13] Johns Hopkins University Center for Systems Science and Engineering (CSSE). Covid-19 data repository, 2020–2025. Accedido: 23 de mayo de 2025.

[14] Yuhao Kang, Song Gao, Yunlei Liang, Mingxiao Li, Jinmeng Rao, and Jake Kruse. Multiscale dynamic human mobility flow dataset in the us during the covid-19 epidemic. *Scientific data*, 7(1):390, 2020.

[15] Lee Mason, Blànaid Hicks, and Jonas S Almeida. Epivecs: exploring spatiotemporal epidemiological data using cluster embedding and interactive visualization. *Scientific Reports*, 13(1):21193, 2023.

[16] Gregorio A. Millett, Austin T. Jones, David Benkeser, Stefan Baral, Laina Mercer, Chris Beyrer, Brian Honermann, Elise Lankiewicz, Leandro Mena, Jeffrey S. Crowley, Jennifer Sherwood, and Patrick S. Sullivan. Assessing differential impacts of covid-19 on black communities. *Annals of Epidemiology*, 47:37–44, 2020.

[17] Ministerio de Salud del Perú. Sala situacional covid-19, 2025. Accedido: 23 de mayo de 2025.

[18] Mijeong Park, Alex R. Cook, Julian T. Lim, Yinxiao Sun, and Benjamin L. Dickens. A systematic review of covid-19 epidemiology based on current evidence. *Journal of Clinical Medicine*, 9(4):967, 2020.

[19] Brea L Perry, Brian Aronson, and Bernice A Pescosolido. Pandemic precarity: Covid-19 is exposing and exacerbating inequalities in the american heartland. *Proceedings of the National Academy of Sciences*, 118(8):e2020685118, 2021.

[20] Massimo Riccaboni and Luca Verginer. The impact of the covid-19 pandemic on scientific research in the life sciences. *PLOS ONE*, 17(2):e0263001, 2022.

[21] Emma Rydow, Rita Borgo, Haihan Fang, Thomas Torsney-Weir, Ben Swallow, Thibaud Porphyre, Cagatay Turkay, and Min Chen. Development and evaluation of two approaches of visual sensitivity analysis to support epidemiological modeling. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1255–1265, 2023.

[22] Paula Sanz-Leon, Louise H. W. Hamilton, Sarah J. Raison, Alexander J. X. Pan, Nicholas J. Stevenson, Robyn M. Stuart, Romesh G. Abeysuriya, Cliff C. Kerr, Stephen B. Lambert, and Jason A. Roberts. Modelling herd immunity requirements in queensland: Impact of vaccination effectiveness, hesitancy and variants of sars-cov-2. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 380(2233):20210311, 2022.

[23] M. Thöny, R. Schnürer, R. Sieber, L. Hurni, and R. Pajarola. Storytelling in interactive 3d geographic visualization systems. *ISPRS International Journal of Geo-Information*, 7(3):123, 2018.

[24] Guancheng Wan, Zewen Liu, Max SY Lau, B Aditya Prakash, and Wei Jin. Epidemiology-aware neural ode with continuous disease transmission graph. *arXiv preprint arXiv:2410.00049*, 2024.

[25] Yue Yu, Yifang Wang, Yongjun Zhang, Huamin Qu, and Dongyu Liu. Inclusiviz: Visual analytics of human mobility data for understanding and mitigating urban segregation. *arXiv preprint arXiv:2501.03594*, 2025.

[26] Haoran Zhang, Peiran Li, Zhiwen Zhang, Wenjing Li, Jinyu Chen, Xuan Song, Ryosuke Shibasaki, and Jinyue Yan. Epidemic versus economic performances of the covid-19 lockdown: A big data driven analysis. *Cities*, 120:103502, 2022.