Topología Persistente como Herramienta para Explorar la Conectividad Causal y la Predicción de Contaminantes

Santiago Peña, Alejandro Botero, Daniela Tellez Cobo

Noviembre 2025

Resumen

En este trabajo analizamos la estructura topológica de un conjunto de series de tiempo asociadas a la calidad del aire en Bogotá mediante técnicas de Análisis Topológico de Datos (ATD). El proyecto se centra en la construcción de un grafo dirigido donde cada nodo representa una serie estacionaria y las aristas corresponden a relaciones de causalidad estimadas entre las series. A partir de este grafo se construye un complejo de Dowker basado en relaciones de alcanzabilidad, y sobre él se simula una filtración definida por los niveles de significancia estadística asociados tanto a la causalidad como a la estacionariedad. Aplicamos un algoritmo de distancia tipo bottleneck y una variante filtrada del algoritmo de Dijkstra para determinar los nacimientos y muertes de componentes conexas en la homología persistente H_0. Los diagramas de persistencia y de barras resultantes permiten identificar componentes robustas, estructuras recurrentes y patrones de conectividad que reflejan dependencias dinámicas entre los contaminantes atmosféricos. Este enfoque ofrece una perspectiva complementaria a métodos clásicos, proporcionando una caracterización topológica de la interacción entre series temporales ambientales.

Palabras clave: Análisis Topológico de Datos, Homología Persistente, Complejo de Dowker, Series de Tiempo, Causalidad, Grafos Dirigidos.

1. Introducción

El análisis de series de tiempo ambientales es fundamental para comprender la dinámica de fenómenos complejos como la calidad del aire en grandes

centros urbanos. En ciudades como Bogotá, donde los niveles de contaminantes como PM2.5, PM10 y NO_2 representan un riesgo persistente para la salud pública, estudiar las relaciones temporales entre estas variables puede revelar patrones estructurales relevantes para la toma de decisiones y el diseño de políticas de mitigación. Tradicionalmente, estos estudios emplean métodos estadísticos clásicos o modelos de aprendizaje automático; sin embargo, estos enfoques suelen privilegiar relaciones locales o aproximaciones punto a punto, dejando de lado la estructura global del sistema.

El Análisis Topológico de Datos (ATD) ofrece una alternativa poderosa para caracterizar estructuras de alto nivel presentes en datos complejos. En particular, la homología persistente permite estudiar cómo emergen y desaparecen patrones de conectividad bajo una filtración, lo que proporciona información sobre la robustez de las relaciones subyacentes. Aunque este marco teórico ha sido aplicado con creciente frecuencia en dominios como biología, visión por computador y dinámica de sistemas, su uso para estudiar interacciones entre series de tiempo ambientales sigue siendo limitado.

En este trabajo utilizamos técnicas de ATD para analizar un conjunto de series de tiempo de calidad del aire en Bogotá. Para ello construimos un grafo dirigido en el que cada nodo corresponde a una serie estacionaria y las aristas representan relaciones de causalidad estadística. A partir de este grafo definimos un complejo de Dowker basado en relaciones de alcanzabilidad y simulamos una filtración determinada por niveles de significancia tanto en causalidad como en estacionariedad. Sobre este complejo calculamos la homología persistente en dimensión cero (H_0), lo que nos permite identificar componentes conexas persistentes, regiones de alta conectividad y patrones topológicos estables.

Este enfoque, inspirado en trabajos recientes sobre homología persistente en complejos de Dowker para grafos dirigidos, proporciona una descripción global de las dependencias entre las series temporales. Además, permite detectar estructuras que no son capturadas por métodos tradicionales basados únicamente en correlación, regresión o modelos VAR. Los resultados muestran que la topología subyacente de la red de contaminantes atmosféricos contiene información relevante sobre la interacción entre variables y su evolución temporal, lo que abre nuevas posibilidades para el estudio de sistemas ambientales mediante herramientas topológicas.

2. Marco Teórico

El presente marco teórico aborda la aplicación de técnicas de Análisis Topológico de Datos (TDA) al estudio y caracterización de las series de

tiempo [5] de las concentraciones de partículas finas $PM_{2,5}$, un contaminante atmosférico crítico [6], cuya dinámica exhibe a menudo comportamientos no lineales, complejos e incluso caóticos [1, 2].

¿Qué es la partícula $PM_{2,5}$?

El Material Particulado 2.5 ($PM_{2,5}$) se define como el conjunto de partículas en suspensión con un diámetro aerodinámico igual o inferior a 2,5 μm (oscilando generalmente entre 0,1 y 2,5 μm) [2, 6]. Debido a su tamaño microscópico, se considera un indicador crítico para la determinación de la calidad del aire [6].

Aspectos Clave

- **Clasificación Regulatoria:** Es catalogado por la Agencia de Protección Ambiental de Estados Unidos (EPA) como uno de los seis contaminantes criterio principales [6, 8].
- **Fuentes y Origen:** Su presencia es prevalente en zonas urbanas [6] y posee un origen complejo que combina fuentes naturales y antropogénicas. Su principal generador se define como la quema de combustibles fósiles.
- **Fuentes específicas:** Emisiones de vehículos (especialmente diésel), tráfico urbano, combustión industrial y doméstica de carbón/biomasa, quemas agrícolas y plantas de energía [6].

Historia del reconocimiento de $PM_{2,5}$

El reconocimiento formal del $PM_{2,5}$ es el resultado de un proceso evolutivo impulsado por la evidencia de sus graves impactos en la salud pública [15] y antecedentes históricos críticos, como la catástrofe de Donora (1948) y el esmog en Los Ángeles [8].

Hitos Regulatorios y Científicos

- **Legislación Temprana (EE. UU.):** El control comenzó con la *Air Pollution Control Act* (1955) y se consolidó con la *Clean Air Act* (1963) y sus enmiendas de 1970, las cuales establecieron la Agencia de Protección Ambiental (EPA) y los Estándares Nacionales de Calidad del Aire Ambiental (NAAQS) [8].

- **Transición de PM_{10} a $PM_{2,5}$:** Inicialmente, las regulaciones (como las enmiendas de 1990) se centraban en partículas más grandes (PM_{10}) [8]. Sin embargo, a finales de la década de 1990 y principios de los 2000, se produjo un cambio de paradigma regulatorio [6]. La evidencia científica demostró que el $PM_{2,5}$ es un indicador superior de contaminación urbana debido a su mayor toxicidad y capacidad de penetración en el torrente sanguíneo en comparación con el PM_{10} [6, 15].
- **Consenso Global:** La Organización Mundial de la Salud (OMS) formalizó este riesgo estableciendo guías de calidad del aire para $PM_{2,5}$ en 2005 [12], actualizando sus recomendaciones en 2021.

Riesgos de salud de la partícula $PM_{2,5}$

El $PM_{2,5}$ se consolida como el principal riesgo ambiental para la salud global [6], catalogado como un "Asesino Invisible" debido a su capacidad para inducir efectos sistémicos graves y mortalidad prematura [3, 6]. Según estimaciones de la Organización Mundial de la Salud (OMS), la contaminación atmosférica es responsable de 1 de cada 9 muertes a nivel mundial [6].

La peligrosidad del $PM_{2,5}$ radica en su tamaño microscópico ($< 2{,}5\,\mu$m) y composición. A diferencia de partículas más grandes (PM_{10}), el $PM_{2,5}$ actúa como portador de virus, metales pesados y compuestos orgánicos tóxicos [6], logrando una penetración profunda en el organismo [6].

La evidencia científica asocia la exposición (incluso a niveles bajos) con patologías graves en múltiples sistemas:

- **Sistema Respiratorio:** La exposición prolongada aumenta la incidencia y mortalidad por cáncer de pulmón [6, 3] (asociada al 5 % de muertes globales por esta causa) [6]. Además, compromete las defensas inmunológicas, incrementando la susceptibilidad a neumonía e infecciones respiratorias agudas (IRA) [15].
- **Sistema Cardiovascular:** Existe un vínculo directo con el aumento de la mortalidad cardiovascular [6, 15], incluyendo infartos de miocardio, accidentes cerebrovasculares (embolias), disfunción endotelial y aterosclerosis. Se estima que causa el 3 % de las muertes cardiopulmonares globales [6].
- **Sistema Neurológico:** Estudios recientes asocian la exposición a largo plazo con un mayor riesgo de demencia, deterioro cognitivo, ansiedad, y enfermedades neurodegenerativas como el Alzheimer y el Parkinson [6].

Justificación del Análisis de Series de Tiempo para $PM_{2,5}$

El uso de series de tiempo en el monitoreo del $PM_{2,5}$ trasciende el simple registro de datos; constituye el sustento técnico indispensable para la gestión de la calidad del aire y la protección de la salud pública [12].

1. **Cumplimiento Normativo y Gestión de Crisis:**
   El monitoreo continuo es esencial para verificar el cumplimiento de los límites establecidos por los Estándares Nacionales de Calidad del Aire Ambiental (NAAQS) de la EPA [8, 6] y las guías de la OMS (ej. promedio anual $< 10\,\mu g/m^3$) [12]. Además, proporciona la base empírica para que las autoridades declaren estados excepcionales (prevención, alerta o emergencia) [12] basándose en la relación crítica entre concentración y tiempo de exposición [12].

2. **Dinámica No Lineal y Caos:**
   El comportamiento del $PM_{2,5}$ se define como un sistema dinámico altamente no lineal, influenciado por variables meteorológicas (precipitación, temperatura) y otros contaminantes [12]. Esta complejidad introduce un comportamiento caótico que impone restricciones severas a la predictibilidad:

   - Modelos no lineales sugieren un horizonte de predicción confiable limitado a aproximadamente 3 horas.
   - A pesar de esta limitación temporal, el análisis a corto plazo es vital para la activación oportuna de protocolos de protección [12].

3. **Utilidad Predictiva:**
   El análisis temporal permite modelar cómo los valores históricos influyen en los actuales [11, 13] y corrobora la correlación espacial entre estaciones de monitoreo [12], facilitando la comprensión de la dinámica espacio-temporal del contaminante en zonas urbanas [12, 6].

Revisión de Literatura: Calidad del Aire

Cuadro 1: Análisis de series de tiempo en la Calidad de Aire

Método	**Aplicación**	**Objetivo/Hallazgos**	**Ref.**
ARIMA	Bogotá (PM_{10}, $PM_{2,5}$, O_3)	Análisis de variabilidad y pronóstico de alertas. Ajuste óptimo mensual.	Pinzón Hassan et al. (2010), Zafra et al. (2017).
ARIMA	Nilai, Malasia (PM_{10})	Descripción adecuada del componente estocástico y pronóstico corto plazo.	Hamid et al. (2016).
ARIMA	Fuzhou, China ($PM_{2,5}$)	Análisis de tendencias y estacionalidad equivalente a valores reales.	Zhang et al. (2018).
ARIMA	Londres (Transp.)	Simulación relación transporte-contaminación para predecir excedencias.	Catalano, M. G. (2016).
ARIMA	Oviedo, España (PM_{10})	Comparación vs. SVM, MLP, VARMA. Pronóstico mensual.	García et al. (2018).
ARIMA	Brunei (O_3)	Pronóstico de concentraciones diarias máximas de ozono.	Kumar et al. (2004).
Regresión Lineal	Escuelas Malasia (Indoor)	Modelado multivariado para PM_{10} y $PM_{2,5}$ en interiores.	Elbayoumi et al. (2014).
Regresión Lineal	Houston, EE.UU (O_3)	Comparación con redes neuronales. $R^2 = 0{,}79$ en un caso.	Prybutok et al. (2000).
Regresión Múltiple	Atenas/Helsinki	Precisión en pronóstico de NO_x y PM_{10}.	Vlachogianni et al. (2011).

Limitaciones de la Modelación Estadística Tradicional en PM2.5

La aplicación de modelos clásicos enfrenta restricciones fundamentales derivadas de la naturaleza física de los datos atmosféricos. La premisa central es la incompatibilidad estructural: se intenta medir un sistema dinámico no lineal y caótico utilizando herramientas lineales y deterministas [14].

Principales Deficiencias Identificadas

- **Incapacidad ante la No Linealidad y el Caos:** Los modelos de regresión lineal (ARIMA, MLR, ARMA) fallan al intentar capturar las leyes de cambio no lineales inherentes al PM2.5 [12, 18]. Dado que el sistema exhibe caos determinista (donde pequeñas variaciones inicia-

les generan grandes divergencias futuras), las aproximaciones lineales resultan insuficientes para describir la complejidad de la dinámica atmosférica [14, 1].

- **Métricas de Error y Precisión:** Los modelos existentes presentan márgenes de error significativos (frecuentemente superiores al 20 % en pronósticos a 24 horas) y pierden robustez ante valores extremos (picos de alta contaminación), que son precisamente los eventos más críticos para la salud pública [12, 18].
- **Subestimación de la Incertidumbre:** En marcos como los Modelos Aditivos Generalizados (GAM), la aproximación de términos suaves mediante funciones lineales tiende a sobreestimar los efectos de la contaminación mientras subestima la incertidumbre estadística real [18].
- **Limitación Espacial (Univariada):** Los modelos basados en series de tiempo univariadas restringen el pronóstico a la ubicación de la estación de monitoreo, impidiendo generar una distribución espacial (mapa de contaminación) de la ciudad y perdiendo la riqueza informativa que ofrecería un enfoque multivariable [12, 17].

Fundamentos y Herramientas del Análisis Topológico de Datos (TDA)

El Análisis Topológico de Datos (TDA) proporciona un marco matemático robusto para inferir la estructura cualitativa o "forma"de conjuntos de datos complejos, superando las limitaciones de la estadística clásica al enfocarse en la conectividad y la estructura global en lugar de métricas locales [19, 5].

Características Principales

- **Complejos Simpliciales:** Son la base estructural del TDA. Generalizan los grafos (redes) mediante el uso de "símplices"(bloques constructivos de dimensiones superiores como triángulos y tetraedros) [5]. Esto permite modelar relaciones de orden superior entre los puntos de datos que se perderían en una representación unidimensional [19].
- **Homología y Números de Betti (β_k):** Proviene de la topología algebraica clásica y cuantifica la presencia de agujeros o cavidades en el complejo simplicial [7].

- β_0: Número de componentes conectadas (agrupaciones).
- β_1: Número de bucles o ciclos (asociado a periodicidad en series de tiempo).
- β_k: Cavidades en dimensiones superiores.

- **Homología Persistente (PH):** Es el motor computacional del TDA. En lugar de analizar los datos a una escala fija, la PH observa cómo evolucionan las características topológicas (homología) a medida que cambian los parámetros de escala (filtración) [5]. Esto permite distinguir entre ruido"(características efímeras) y "señal"(estructuras que persisten a través de múltiples escalas) [7].

- **Diagramas de Persistencia:** Es la representación visual y cuantitativa de la Homología Persistente. En un plano 2D, cada punto representa una característica topológica con coordenadas de Nacimiento (*birth*) y Muerte (*death*) [19, 20].

  - La persistencia (vida útil = muerte - nacimiento) indica la robustez de la característica [7].
  - En series de tiempo, la homología de dimensión 1 (β_1) en estos diagramas es crucial para detectar patrones recurrentes o ciclos del sistema dinámico [1, 20].

Literatura de TDA en Series de Tiempo

Cuadro 2: Aplicaciones de TDA en Series de Tiempo

Dominio	**Autor (Año)**	**Técnica TDA**	**Objetivo**
Sist. Dinámicos y Caos	Skraba et al. (2012)	Homología Persistente (PH)	Análisis topológico de sistemas recurrentes.
Sist. Dinámicos y Caos	Khasawneh & Munch (2018); Tempelman & Khasawneh (2020)	PH en grafos (k-NN o redes de partición)	Distinción entre comportamiento caótico y periódico en series de tiempo.
Finanzas y Economía	Gidea & Katz (2018); Ismail et al. (2022)	PH y Persistence Landscapes	Análisis de crisis financieras y detección de señales de advertencia temprana (caídas de mercado).

Continúa...

Cuadro 2 – continúa...

Dominio	Autor (Año)	Técnica TDA	Objetivo
Finanzas y Economía	Majumdar & Laha (2020); Umeda (2017)	PH, Time Delay Embedding, RF-TDA	Clasificación/Clustering de acciones para predecir sector y similitudes entre modelos (AR, MA, GARCH).
Biomedicina	Perea et al. (2015)	SW1PerS (Sliding Windows Scoring)	Descubrimiento de periodicidad en series de tiempo de expresión genética.
Biomedicina	Emrani et al. (2014)	PH de Delay Embeddings	Detección de sibilancias (wheeze detection) en señales respiratorias.
Biomedicina	Karan & Kaygun (2021)	PH y Subwindowing	Clasificación de señales fisiológicas bajo condiciones de estrés y no estrés.
Biomedicina	Itzá-Ortíz et al. (2021)	Homología y Números de Betti	Método para analizar colecciones de series simultáneas (ej. polisomnografía).

Literatura de TDA en Grafos

Cuadro 3: Aplicaciones de TDA en Grafos

Dominio	**Autor (Año)**	**Técnica TDA**	**Objetivo**
Sist. Dinámicos	Myers, Munch, & Khasawneh (2019)	k-NN Graph, Takens, OPN	Pipeline TDA para distinguir eficazmente entre sistemas periódicos y caóticos.
Sist. Dinámicos	Tempelman & Khasawneh (2020)	Análisis topológico (Atractor)	Detección de caos en sistemas dinámicos aplicando TDA a la geometría del atractor.
Finanzas	Majumdar & Laha (2020)	RF-TDA, SOM-TDA	Demostración de que las características topológicas distinguen sectores en precios de acciones.
Finanzas	Gidea (2017)	Redes de Correlación	Detección de advertencia temprana de transiciones críticas en redes financieras.
Neurociencia	Giusti, Ghrist, & Bassett (2016)	Complejos Simpliciales	Modelado de relaciones neuronales de orden superior (más allá de aristas simples).
Correlación Series	Itzá-Ortíz et al. (2021)	Filtración de complejos ($\Delta(G_p)$)	Análisis de complejidad de colecciones simultáneas; distinción de matrices aleatorias.
Redes	Singh et al. (2023)	PH + Visibility Graphs	Extracción de vectores de características robustos para clasificación de series.
Redes	Chowdhury & Mémoli (2017)	Homología Trayectoria Persistente (PPH)	Manejo de asimetría y direccionalidad en redes (identificación de ciclos dirigidos).

Literatura de TDA en Análisis de Calidad del Aire

Cuadro 4: TDA aplicado a Contaminación Atmosférica

Contaminante	**Autor (Año)**	**Técnica**	**Objetivo**
Haze y PM_{10} (Malasia)	Zulkepli et al. (2022)	Hybrid HACA + PH + Wasserstein	Mejora en la agrupación de estaciones basada en similitud topológica; superó al método tradicional.
6 Contaminantes (PM, O_3, etc.)	Madukpe et al. (2025)	Conventional Mapper & Ball Mapper	Análisis de comportamiento: Mapper agrupa por concentración, Ball Mapper detecta patrones de tendencia.
Predicción PM_{10} y $PM_{2,5}$	(Estudio Deep Learning)	PH + CNN (AB-CNN)	Mejora de la precisión predictiva capturando estructuras complejas (ciclos, vacíos).
Colecciones de Series	Itzá-Ortíz et al. (2021)	Topología Algebraica (Matriz Corr.)	Análisis de la complejidad en fenómenos naturales vs. datos aleatorios mediante números de Betti.

Inferencia de Causalidad y Construcción de la Red

Inicialmente para nuestra investigacion fue necesario considerar las relaciones que tendrian entre si las series de tiempo. Para modelar las interacciones entre estas series se tomo en cuenta el enfoque propuesto por el autor Fulmyk, W. (2023) [24]. En este se utiliza el concepto de *Causalidad de Granger*, que establece que una serie X causa a Y si la historia de ambas predice mejor a Y que la historia de Y por sí sola. Clásicamente, esto se evalúa comparando la varianza del error de predicción entre un modelo restringido (sin X) y uno no restringido (con X) [24].

Sin embargo, dado que las interacciones atmosféricas suelen ser complejas, los modelos lineales clásicos presentan limitaciones intrínsecas para detectar relaciones no lineales sin transformaciones explícitas de los datos. Para superar esto, se justifica el uso del algoritmo *mlcausality* el cual fue propuesto e implementado por el autor del articulo. Este enfoque utiliza el "truco del kernel"(específicamente el kernel de base radial, RBF) para capturar no li-

nealidades y dependencias dinámicas complejas de manera eficiente, logrando puntuaciones AUC competitivas y tiempos de cómputo reducidos [24].

En este mismo articulo tambien se establece que en series de tiempo un requisito teórico fundamental para el análisis es la **estacionariedad** de las series temporales, asegurando que las propiedades estadísticas se mantengan constantes para que la inferencia predictiva sea válida.

Representación Topológica: Complejos de Dowker

Dado que la causalidad es una relación intrínsecamente asimétrica (dirigida), para la eleccion del complejo fue importante tener en cuenta la direccion de las aristas del grafo. Para tener esto en cuenta nos basamos en el estudio de Li, H., et al. (2024) [23], donde utilizan el **Complejo de Dowker** como estructura topologica principal representativa del grafo dirigido. A diferencia de los complejos de Vietoris-Rips que se centran en la vecindad directa y distancias simétricas, los complejos de Dowker capturan la estructura de **vecinos compartidos** y son sensibles a la dirección y el peso de las aristas [23].

El complejo de Dowker se define mediante la existencia de un nodo "sumidero"(*sink*) o "fuente"(*source*) común, lo que permite identificar relaciones de orden superior basadas en la alcanzabilidad dentro del grafo dinámico [23].

3. Metodología Algorítmica y Filtración

Finalmente, consideramos crucial el poder generar la filtracion con la menor complejidad computacional posible para esto tuvimos en cuenta el articulo de los autores Chowdhury, S., Mémoli, F. (2018) Este articulo simula una filtración basada en los niveles de significancia estadística asociados a la causalidad y la estacionariedad. Estos niveles actúan como el parámetro δ del filtro, permitiendo la evolución del complejo de Dowker a través de una secuencia anidada de subgrafos basada en los pesos de las aristas [23]. Esta simulacion es clave en nuestra investigacion ya que nos perimte realizar analisis para una amplia cantidad de series sin tener problemas de complejidad computacional

Para el cálculo de los nacimientos y muertes de las componentes conexas en la homología de dimensión cero (H_0), los autores aplican dos técnicas algorítmicas clave:

1. **Algoritmo de Distancia Tipo Bottleneck:** La distancia bottleneck (d_B) se utiliza para comparar diagramas de persistencia. Es fundamental en la Homología Persistente (HP), ya que el Teorema de Estabilidad garantiza que pequeñas perturbaciones en los datos de entrada resultan en pequeños cambios en los diagramas de persistencia, medidos por la distancia d_B [22].

2. **Variante Filtrada del Algoritmo de Dijkstra:** A su vez, nos apoyamos en el articulo "On the Bottleneck Shortest Path Problem" [21] el cual nos proporciona con una variamte del algoritmo de Dijkstra, para facilitarnos el calculo de la distancia Bottleneck. El algoritmo de Dijkstra es un método clásico utilizado para resolver el Problema de la Ruta más Corta (SP) desde una única fuente. En el contexto de grafos con capacidades o pesos, este se relaciona con el *Problema de la Ruta más Corta de Cuello de Botella*, que busca determinar la máxima capacidad de cualquier ruta entre dos vértices (o equivalentemente, el flujo máximo no divisible) [21]. Una ligera modificación del algoritmo de Dijkstra puede resolver el BSP de manera eficiente [21].

En el contexto de la filtración, la literatura demuestra que un algoritmo para el BSP que utiliza un ordenamiento conocido de los pesos de las aristas puede ser una reformulación del algoritmo de Dijkstra, resolviendo el problema en tiempo lineal $\mathcal{O}(m)$. El uso de una "variante filtrada"de Dijkstra sugiere que el cálculo de la persistencia se optimiza para encontrar las rutas críticas o los umbrales específicos (los niveles de significancia) que causan el nacimiento o la muerte de las componentes conexas (H_0).

4. Metodología

La metodología empleada combina técnicas de análisis de series de tiempo, teoría de grafos dirigidos y herramientas de Análisis Topológico de Datos (ATD). El proceso inicia con la obtención y consolidación de los datos, los cuales corresponden a registros históricos de calidad del aire (PM2.5) disponibles públicamente en la plataforma *AQICN* (Air Quality Open Data Platform)[1].

A partir de esta fuente, se extrajeron series temporales de 21 estaciones de monitoreo ubicadas en Bogotá y a su alrededor, ubicadas en sectores como *San Cristóbal*, *La Belleza*, *La Esperanza*, *General Santander*, entre otras. Los registros abarcan un período intermitente desde el 1 de febrero de 2024

[1]Datos obtenidos de: `https://aqicn.org/historical/es/#city:colombia/bogota/carvajal-sevillana`

hasta el 20 de septiembre de 2025, con mediciones diarias de la mediana de PM2.5 (en $\mu g/m^3$) y el conteo de observaciones por estación y fecha. Para garantizar la consistencia temporal, se realizó una homogenización de las fechas, asegurando que todas las estaciones compartieran el mismo vector de tiempos. Tras este proceso de depuración y alineación, se obtuvo un conjunto final de 122 registros diarios válidos y comunes para todas las estaciones.

A partir de estas series, se lleva a cabo un flujo de trabajo en cinco etapas principales: (i) preprocesamiento y verificación de estacionariedad, (ii) estimación de relaciones de causalidad mediante pruebas de Granger, (iii) construcción del grafo dirigido y del complejo de Dowker asociado, (iv) definición de una filtración basada en niveles de significancia estadística, y (v) cálculo de homología persistente en dimensión cero utilizando distancias tipo bottleneck obtenidas mediante un algoritmo de Dijkstra filtrado.

Algorithm 1 Pipeline general para el análisis topológico de series de tiempo

Require: Conjunto de series de tiempo $\{X_i(t)\}_{i=1}^{n}$

1: **Estacionariedad:** Calcular el p-valor ADF para cada serie y formar la matriz E
2: **Causalidad:** Calcular los p-valores de Granger para cada par ordenado (i, j) y formar la matriz C
3: **Construcción del grafo:** Crear un grafo dirigido G con pesos $w_{ij} = C_{ij}$
4: **Distancias bottleneck:** Para cada nodo ejecutar Dijkstra filtrado y construir la matriz D
5: **Testigos y Dowker:** Para cada par (u, v) calcular el peso de la arista W_{uv} mediante testigos t y la matriz D
6: **Filtración:** Para cada nivel de significancia τ:
7: — Incluir nodos con $E_{ii} \leq \tau$
8: — Incluir aristas con $W_{uv} \leq \tau$
9: **Homología**: Calcular $H_0(D_\tau)$, registrar nacimientos y muertes de componentes
10: **Resultado final:** Construir diagramas de barras y diagramas de persistencia

4.1. Preprocesamiento y Estacionariedad

Sea $\{X_i(t)\}_{t=1}^{T}$ el conjunto de series de tiempo para cada contaminante o estación. La estacionariedad se evalúa mediante la prueba aumentada de Dickey–Fuller (ADF):

$$\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \sum_{k=1}^{p} \phi_k \Delta X_{t-k} + \varepsilon_t,$$

cuyo valor p-valor se almacena en una matriz diagonal:

$$E_{ii} = p\text{-valor}(ADF(X_i)).$$

Cada nodo del grafo existirá en la filtración a partir del umbral τ para el cual:

$$E_{ii} < \tau.$$

4.2. Causalidad y Construcción del Grafo

Para cada par ordenado (X_i, X_j) se evalúa causalidad de Granger mediante el contraste:

$$X_j(t) = \sum_{k=1}^{p} a_k X_j(t-k) + \sum_{k=1}^{p} b_k X_i(t-k) + \varepsilon_t.$$

La hipótesis nula $H_0 : b_1 = \cdots = b_p = 0$ produce un p-valor que almacenamos en la matriz:

$$C_{ij} = p\text{-valor}(Granger(X_i \to X_j)).$$

La arista dirigida $i \to j$ existe con peso $w_{ij} = C_{ij}$.

4.3. Relación de Alcanzabilidad y Complejo de Dowker

El complejo de Dowker es adecuado para grafos dirigidos porque permite construir simplices a partir de relaciones bipartitas sin requerir simetría. Dado un grafo dirigido $G = (V, E)$ con pesos w_{ij}, definimos la relación de alcanzabilidad:

$$R = \{(u, t) : \text{existe un camino dirigido } u \rightsquigarrow t\}.$$

El complejo de Dowker asociado se define como:

$$D(R) = \{\sigma \subseteq V : \exists\, t \in V \text{ tal que } (u, t) \in R\ \forall u \in \sigma\}.$$

En este proyecto solo se requiere la estructura 1-dimensional del complejo, pues el cálculo de H_0 depende únicamente de nodos y aristas.

4.4. Distancia tipo Bottleneck entre Caminos

Para un camino dirigido $P = (v_0, v_1, \ldots, v_k)$, se define la distancia tipo bottleneck como:

$$d_{\text{bottleneck}}(P) = \max_{i=0,\ldots,k-1} w_{v_i v_{i+1}}.$$

Entre dos nodos u y v, la distancia es:

$$d(u, v) = \min_{P: u \rightsquigarrow v} d_{\text{bottleneck}}(P).$$

Esto corresponde a minimizar el cuello de botella del camino.

4.4.1. Dijkstra Filtrado

Implementamos una variante del algoritmo de Dijkstra donde la operación de suma se reemplaza por el operador máx:

$$\text{dist}(v) = \min_{u \in \text{pred}(v)} \max\big(\text{dist}(u), w_{uv}\big).$$

Algorithm 2 Dijkstra Filtrado

```
Inicializar distancias: dist(v) ← ∞ para todo v
Seleccionar nodo origen s y poner dist(s) ← 0
while existe un nodo no visitado do
    u ← nodo no visitado con menor dist(u)
    Marcar u como visitado
    for cada vecino v tal que (u, v) ∈ E do
        c ← máx(dist(u), w_uv)
        if c < dist(v) then
            dist(v) ← c
        end if
    end for
end while
```

Este procedimiento se ejecuta para cada nodo origen, obteniendo así la matriz completa de distancias bottleneck D.

4.4.2. Matriz de distancias bottleneck

La ejecución del algoritmo de Dijkstra filtrado para cada nodo del grafo produce una matriz completa de distancias tipo bottleneck:

$$D = (d(u,v))_{u,v \in V},$$

donde cada entrada $d(u,v)$ representa el mínimo valor de cuello de botella entre todos los caminos dirigidos que conectan u con v. Esta matriz constituye la base para determinar los valores de nacimiento de las aristas en el complejo de Dowker y, por tanto, la evolución de la filtración.

4.5. Nacimiento de Aristas en el Complejo

Para cada par de nodos u, v, consideramos todos los posibles vértices testigo t tales que:

$$u \rightsquigarrow t \quad \text{y} \quad v \rightsquigarrow t.$$

El peso asignado al complejo σ es:

$$W_\sigma = \min_{t \in V(G)} \max_{x \in \sigma} \big(d(x,t)\big).$$

Esto representa el nivel mínimo de significancia necesario para que ambos nodos alcancen un testigo común.

El simplice σ *nace* en la filtración cuando el umbral τ satisface:

$$\tau \geq W_\sigma.$$

4.6. Filtración e Ingreso de Simplices

En nuestro caso, la filtración queda determinada exclusivamente por nodos y aristas, ya que el cálculo de H_0 solo requiere información 1-dimensional. El complejo filtrado en un nivel de significancia τ está dado por

$$D_\tau \;=\; \{\{u\} : E_{uu} \leq \tau\} \;\cup\; \{\{u,v\} : W_{uv} \leq \tau\}\,.$$

Esto implica que la filtración se construye de manera incremental según las reglas:

- **Nacimiento de nodos:** un nodo u ingresa al complejo cuando satisface $E_{uu} \leq \tau$, es decir, cuando su nivel de significancia de estacionariedad es menor o igual que el umbral.
- **Nacimiento de aristas:** una arista no dirigida $\{u,v\}$ ingresa cuando $W_{uv} \leq \tau$, donde W_{uv} es el nivel mínimo de significancia necesario para que ambos nodos alcancen un testigo común según la matriz de distancias bottleneck.

Así, la familia $\{D_\tau\}_{\tau\in[0,1]}$ constituye una filtración creciente que describe cómo evolucionan las conexiones entre los nodos a medida que aumenta el nivel de significancia considerado.

4.7. Cálculo de Homología Persistente en Dimensión 0

Para cada nivel de la filtración, el complejo D_τ define una partición en componentes conexas. Medimos:

$$H_0(D_\tau) = \text{número de componentes conexas en } D_\tau.$$

Una componente nace cuando aparece un nodo aislado. Una componente muere cuando una arista conecta dos componentes previamente separadas.

Finalmente se construyen los:

- Diagramas de barras
- Diagramas de persistencia

que muestran la robustez de las estructuras topológicas presentes en la red.

4.8. Complejidad computacional

El proceso completo presenta la siguiente complejidad temporal aproximada:

- **Cálculo de causalidad de Granger:** Para n series de tiempo de longitud T, la prueba entre cada par ordenado (X_i, X_j) requiere $O(T)$, por lo que la matriz completa tiene costo $O(n^2T)$.
- **Cálculo de estacionariedad (ADF):** Evaluado de manera independiente en cada serie, con costo total $O(nT)$.
- **Distancias bottleneck (Dijkstra filtrado):** Cada ejecución del algoritmo para un nodo tiene complejidad $O(n^2)$ debido a la operación máx en lugar de sumas acumulativas. Al ejecutarse para todos los nodos, el costo total es $O(n^3)$.
- **Cálculo de los valores de nacimiento de aristas:** Para cada par (u, v) se consideran todos los vértices testigo t, lo que implica un costo $O(n)$ por par, para un total de $O(n^3)$.

Por tanto, el costo dominante del procedimiento es $O(n^3)$, proveniente tanto del cálculo completo de distancias bottleneck como del proceso de selección de testigos. Este orden de complejidad es manejable para los tamaños de grafo típicos en aplicaciones ambientales como la estudiada en este trabajo.

5. Resultados y análisis

En el presente estudio, tras el análisis del dígrafo mediante homología persistente, se obtuvo el siguiente diagrama de persistencia:

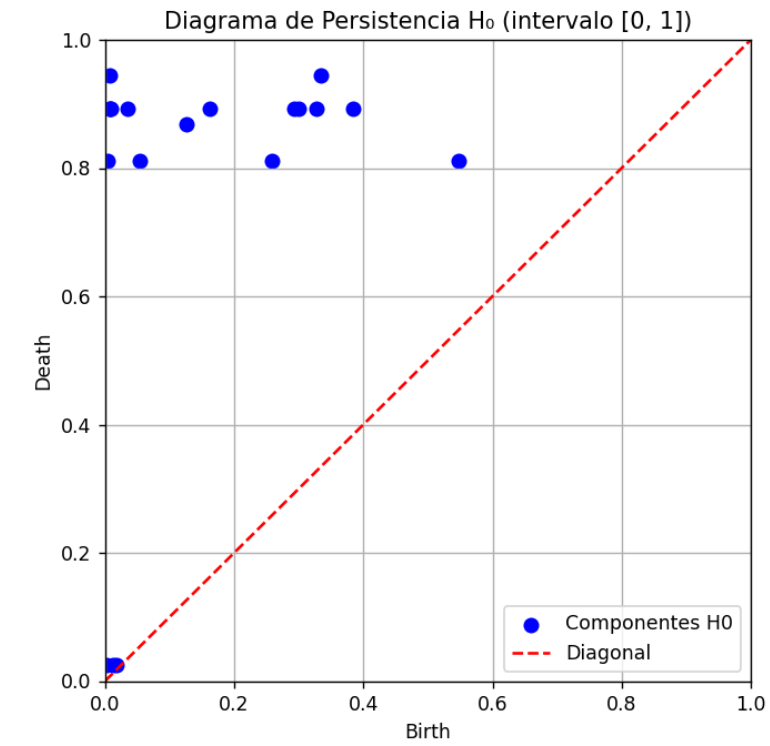

Figura 1: Diagrama de Persistencia H_0 del grafo estudiado.

En dicho diagrama se observa que, para valores altos del parámetro de significancia, se registran numerosas muertes de componentes conexas. Esto sugiere que a esas escalas —es decir, bajo filtrados menos estrictos— distintas regiones del grafo se fusionan, dando lugar a una estructura globalmente más conectada. En contraste, también se aprecia que aproximadamente tres

componentes mueren muy cerca del origen, lo que indica que bajo umbrales de significancia muy restrictivos la estructura del grafo se fragmenta de forma temprana, perdiendo conectividad rápidamente.

Esta dualidad revela un punto interesante: la topología y la estadística parecen, en principio, tensionarse entre sí. Desde la perspectiva topológica, imponer un umbral de significancia muy alto conduce a una fragmentación del dígrafo, lo cual puede ser útil para identificar clústeres o grupos de series temporalmente afines. Sin embargo, desde la perspectiva estadística, aumentar excesivamente la exigencia puede eliminar relaciones relevantes, preservando únicamente las más robustas pero sacrificando información potencialmente útil para el modelado. Por el contrario, adoptar umbrales más flexibles produce un grafo más unificado, pero también puede introducir relaciones espurias que incrementen ruido en las predicciones.

Precisamente por esta tensión, inicialmente consideramos que la topología podría aportar un criterio para determinar un valor "óptimo" de significancia. No obstante, al entrenar distintos modelos mediante Kernel Ridge Regression se encontró que —contrario a lo esperado— el análisis de homología persistente no determina por sí mismo un umbral que minimice el error. El desempeño del modelo resulta influenciado no solo por la estructura topológica inferida, sino también por propiedades estadísticas intrínsecas de las series. Esto sugiere que la topología, sin incorporar información estadística adicional, es insuficiente para identificar directamente un nivel de significancia óptimo.

Esta observación abre una vía interesante para trabajos futuros: integrar información estadística dentro del proceso de filtración topológica, de modo que el parámetro de significancia no solo controle la conectividad del grafo, sino también la estabilidad estadística de las relaciones entre series. Tal integración podría permitir modelar de manera más precisa el comportamiento del contaminante y mejorar la capacidad predictiva del sistema.

Adicionalmente, se construyó el siguiente diagrama de barras:

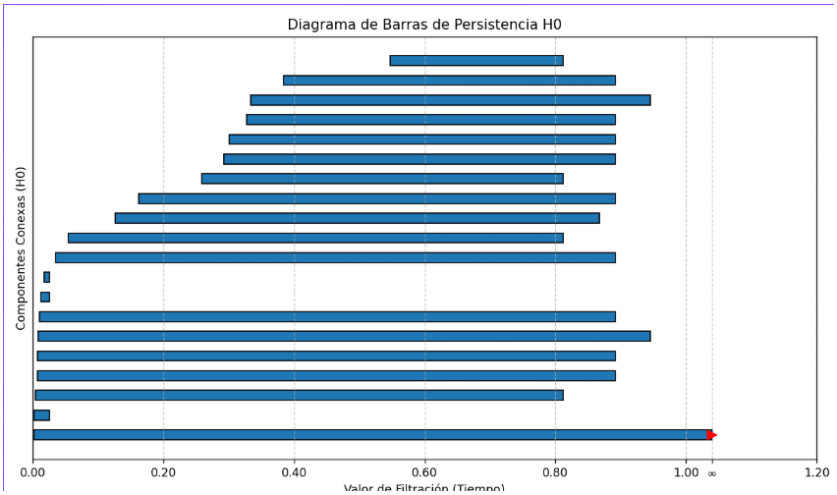

Figura 2: Diagrama de Barras de H_0 del grafo estudiado.

Este diagrama muestra que varias componentes persisten a lo largo del intervalo de filtración, y que existe una en particular cuya persistencia se mantiene hasta el final. Ello indica que, independientemente del nivel de significancia, al menos una componente conexa permanece invariante, actuando como un núcleo estructural robusto en el grafo. Asimismo, con el propósito de identificar posibles cavidades en la estructura, se calculó la homología H_1. En este caso, H_1 resultó vacío, lo cual implica que el complejo de Dowker construido a partir del grafo carece de ciclos persistentes: topológicamente, la estructura es enteramente de tipo árbol o unión de árboles bajo la filtración considerada.

En conjunto, estos resultados resaltan la importancia de considerar tanto la información topológica como la estadística al analizar relaciones entre series de contaminantes. Comprender esta interacción es crucial para realizar predicciones correctas y, más adelante, para estudiar de manera más profunda el comportamiento temporal del contaminante.

6. Conclusiones

En conclusión, este estudio desarrolla una perspectiva novedosa para el análisis de series de tiempo de calidad del aire en Bogotá, integrando la inferencia estadística clásica con el Análisis Topológico de Datos (ATD). La metodología propuesta trasciende el análisis univariado al incorporar la causalidad de Granger para modelar las interdependencias entre estaciones. Un aporte fundamental de este enfoque es la construcción de un complejo de Dowker, lo que permitió preservar la direccionalidad inherente del grafo de

causalidad, superando las limitaciones de los enfoques simétricos tradicionales. Sobre esta estructura, la aplicación de una filtración basada en niveles de significancia y el cálculo de la homología persistente mediante un algoritmo de Dijkstra filtrado demostró ser una estrategia computacionalmente eficiente, con una complejidad de $O(n^3)$.

Los resultados revelan una divergencia notable entre la significancia estadística y la topológica. Se observó que, aunque existen múltiples relaciones de causalidad estadísticamente significativas (bajos p-valores), muchas de ellas carecen de robustez topológica, manifestándose como componentes de vida corta en los diagramas de persistencia. Esto sugiere que el ATD actúa como un filtro eficaz, discriminando entre dependencias espurias o efímeras y la estructura estructural "dura"de la contaminación. Adicionalmente, el análisis de homología evidenció que $H_1 = \emptyset$, lo que indica que el sistema de contaminación en Bogotá es acíclico: los contaminantes se dispersan o acumulan siguiendo jerarquías de fuentes a sumideros sin formar bucles de retroalimentación detectables. Finalmente, se concluye que la información topológica es opuesto a la capacidad predictiva; si bien el ATD describe la forma y conectividad del sistema, no se traduce directamente en una minimización del error en modelos de regresión (como Kernel Ridge), lo que implica que la topología y la predicción estadística capturan dimensiones distintas pero complementarias de la dinámica ambiental.

Desde el contexto de los contaminantes, el análisis topológico sugiere que persistencia de una componente conexa robusta evidencia que, más allá de los focos locales de emisión, la ciudad enfrenta una dinámica de contaminación sistémica gobernada por factores globales. Sin embargo, la dificultad para predecir concentraciones basándose en esta estructura topológica indica que la magnitud de la contaminación está fuertemente modulada por variables meteorológicas externas y estocásticas, más que por la forma de los contaminantes.

Referencias

[1] Benmebarek, S., & Chettih, M. (2024). *Chaotic analysis of daily runoff time series using dynamic, metric, and topological approaches.* Acta Geophysica, 72, 2633–2651.

[2] Bui, Q. T., Jani, R., Mohajeri, F., Shabani, E., & Mehr, A. D. (2025). *Investigation of nonlinear dynamics and stochastic characteristics of fine particulate matter in urban environments.* Acta Geophysica (2025), 73, 1989–2004.

[3] Burnett, R., et al. (2018). *Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter*. Proc. Natl. Acad. Sci., 115, 9592–9597.

[4] Huang, X., Steinmetz, J., Marsh, E. K., Aravkin, A. Y., Ashbaugh, C., Murray, C. J. L., Yang, F., Ji, J. S., Zheng, P., Sorensen, R. J. D., Wozniak, S., Hay, S. I., McLaughlin, S. A., Garcia, V., Brauer, M., & Burkart, K. (2025). *A systematic review with a Burden of Proof meta-analysis of health effects of long-term ambient fine particulate matter (PM2.5) exposure on dementia*. Nature Aging. DOI: 10.1038/s43587-025-00844-y.

[5] Chazal, F., & Michel, B. (2021). *An introduction to topological data analysis: Fundamental and practical aspects for data scientists*. Front. Artif. Intell. 4.

[6] Hinojosa Rodríguez, M. R. (2020). *Contaminación por Partículas $PM_{2,5}$: Impacto en la Salud y Fuentes de Emisión*. Inteligencia Epidemiológica, 1: 23–28.

[7] Seversky, L. M., Davis, S., & Berger, M. (2016). *On Time-series Topological Data Analysis: New Data and Opportunities*. En Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 59–67).

[8] Esparza Puente, G. S. (2024). *La regulación de la calidad del aire en Estados Unidos de América (EUA)*.

[9] Hammer, M. S., et al. (2020). *Global Estimates and Long-Term Trends of Fine Particulate Matter (PM2.5) Concentrations (1998–2018)*. Environmental Science & Technology, 54(13), 7879–7890.

[10] Itzá-Ortíz, B. A., et al. (2021). *Un método topológico para el análisis de complejidad de series de tiempo*. Pädi Boletín Científico de Ciencias Básicas e Ingenierías del ICBI, 9(17), 103–107.

[11] Liu, M., Zhu, F., & Zhu, K. (2022). *Normalcy-dominant ordinal time series and its application to air quality data*. J. Time Ser. Anal., 43(3), 460–478.

[12] Pinzón Hassan, A. D., & Tique Ortiz, V. (2019). *Análisis de series de tiempo de contaminantes atmosféricos en la ciudad de Bogotá a partir del desarrollo de modelos estadísticos ARIMA*. Universidad Distrital Francisco José de Caldas, Bogotá D.C.

[13] Salcedo, R. L. R., Alvim Ferraz, M. C. M., Alves, C. A., & Martins, F. G. (1999). *Time-series analysis of air pollution data*. Atmospheric Environment, 33(15), 2361–2372.

[14] Salini, G. A. (2018). *Chaotic behaviour of PM2.5 concentration in an urban center in the south of Chile*. WIT Transactions on Ecology and the Environment, Vol 230, 131.

[15] Tan, X., et al. (2018). *The regulatory roles of PM2.5 in respiratory system diseases*. Biomed Res Int, 2018.

[16] Yan, X., et al. (2020). *A multi-model ensemble of EntityDenseNet for PM2.5 concentration estimation in China*. Environment International, 144, 106060.

[17] Shrikar, J., Nathezhtha, T., Abirami, S. & Sakthivel, G. (2025). *Enhancing urban air quality prediction using time-based-spatial forecasting framework*.Scientific Reports, vol. 15, art. no. 4139.

[18] B. S. Freeman, G. Taylor, B. Gharabaghi, y J. Thé, (2018) *Forecasting air quality time series using deep learning* Journal of the Air & Waste Management Association, vol. 68, no. 8, pp. 866-886.

[19] E.Munch, (2017) *A user's guide to topological data analysis* Journal of Learning Analytics, vol. 4, no. 2, pp. 47–61, 2017, doi: 10.18608/jla, 42.6.

[20] T.Ichinomiya, (2025) *Machine learning of time series data using persistent homology* Scientific Reports, vol. 15, no. 20508, pp. 1–11, doi: 10.1038/s41598-025-06551-3.

[21] Kaibel, V., & Peinhardt, M. A. F. (2006). *On the Bottleneck Shortest Path Problem*. Zuse Institute Berlin (ZIB), Report 06-22.

[22] Chowdhury, S., & Mémoli, F. (2017). *Persistent Path Homology of Directed Networks*. The Ohio State University. arXiv preprint.

[23] Li, H., Jiang, H., Fan, J., Ye, D., & Du, L. (2024). *Dynamic Neural Dowker Network: Approximating Persistent Homology in Dynamic Directed Graphs*. En Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM.

[24] Fulmyk, W. (2023). *Nonlinear Granger Causality using Kernel Ridge Regression*. arXiv preprint arXiv:2309.05107.