Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Topological Data Analysis of Urban Air Quality Monitoring Networks: Vietoris-Rips Simplicial Filtration, Persistent Homology, and Geostatistical Connectivity in a Bifurcated Riverine Airshed

Saptarshi Sadhu

saptarshi1.sadhu@stu.adamasuniversity.ac.in

Adamas University https://orcid.org/0009-0009-7505-1973

Research Article

Keywords: Topological Data Analysis, Persistent homology, Vietoris-Rips filtration, Spatial proximity graphs, Geostatistical connectivity, Graph Neural Networks

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11793674/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Topological Data Analysis of Urban Air Quality Monitoring Networks: Vietoris–Rips Simplicial Filtration, Persistent Homology, and Geostatistical Connectivity in a Bifurcated Riverine Airshed

Saptarshi Sadhu[1*]

[1]Department of Computer Science and Engineering, School of Engineering and Technology, Adamas University, Kolkata 700126, India

ORCID: https://orcid.org/0009-0009-7505-1973

[*]Corresponding author: saptarshi1.sadhu@stu.adamasuniversity.ac.in

Alternative: saptarshisadhuofficial@gmail.com

Abstract

Continuous ambient air quality networks in riverine megacities exhibit multi-scale spatial dependencies poorly captured by isotropic geostatistics. We establish a Topological Data Analysis (TDA) and spatial graph framework evaluating sensor network geometry across 13 regulatory stations bisecting the lower Hooghly River conurbation. Using spherical Haversine metrics and Gaussian radial basis functions, we benchmark the network against classical graph hierarchies (Delaunay, Gabriel, Relative Neighborhood, and Minimum Spanning Tree) alongside Ripley's K and Besag's L-functions. Vietoris–Rips simplicial filtration yields exact persistent homology invariants. Zero-dimensional homology (H_0) reveals three-stage hierarchical percolation with conurbation consolidation at $\epsilon = 14.50$ km (mean lifespan $\bar{L}_0 = 4.16$ km; persistent entropy $E_0 = 3.18$ bits), verifying Kruskal/Borůvka duality with the MST edge spectrum ($L = 49.88$ km). One-dimensional homology (H_1) isolates a transient trans-river 1-cycle ($\Delta\epsilon = 25.1$ m), confirmed via Monte Carlo coordinate perturbation ($\sigma_{\text{jitter}} \in [10, 500]$ m) as an ephemeral coordinate discretization artifact. Grounding the filtration in empirical telemetry ($PM_{2.5}$, PM_{10}, NO_2), Moran's $I(\epsilon)$ transitions from localized clustering ($I = +1.017$ at $\epsilon = 2.0$ km) to random expectation ($I = -0.083$) at the $\epsilon = 7.0$ km consolidation scale, aligning with semivariogram practical range ($a_{\text{prac}} = 5.29$ km). Finally, on binary filtration graphs, we prove combinatorial Laplacian proportionality to Geary's contiguity ratio $C(\epsilon)$, and formulate a topological embedding tensor $\mathbf{H}^{\text{TDA}} \in \mathbb{R}^{13\times 6}$ for topology-aware graph neural networks.

Keywords: Topological Data Analysis; Persistent homology; Vietoris–Rips filtration; Spatial proximity graphs; Geostatistical connectivity; Graph Neural Networks

Acknowledgements: The author gratefully acknowledges the Central Pollution Control Board (CPCB), Ministry of Environment, Forest and Climate Change, Government of India, and the West Bengal Pollution Control Board (WBPCB) for maintaining the Continuous Ambient Air Quality Monitoring System (CAAQMS) network and providing open access to continuous telemetry datasets across the Kolkata and Howrah conurbation.

1 Introduction

Continuous ambient air quality monitoring networks provide critical observational telemetry for environmental risk assessment and statutory compliance enforcement (World Health Organization, 2021; Chakroborty et al., 2025). Classical geostatistical workflows—including ordinary kriging, Gaussian process regression, and local indicators of spatial association—conventionally evaluate covariance under planar Euclidean distances, relying on second-order stationarity and spatial isotropy (Moran, 1950; Anselin, 1995; Cressie, 1993; Diggle, 2013). Similarly, deep learning architectures frequently treat monitoring stations as isolated coordinate tuples or fixed k-nearest-neighbor (k-NN) nodes, disregarding manifold geometry and spatial coverage voids (Dutta and Pal, 2022; Vu et al., 2019; Kipf and Welling, 2017). In riverine megacity airsheds, however, boundary-layer ventilation and aerosol dispersion are strongly modulated by major tidal waterways, sharp land-use divides, and localized thermal circulations (Venkataraman et al., 2018; Bera et al., 2021; Ghosh et al., 2024). Across the twin metropolitan centers of Kolkata and Howrah in eastern India ($>$ 16 million residents), the tidally influenced Hooghly River physically separates heavy industrial foundries on the western bank from dense commercial corridors on the eastern bank (Chakroborty et al., 2025). Applying single-bandwidth isotropic covariance kernels across such heterogeneous domains risks over-smoothing localized emission hotspots while extrapolating spurious spatial dependencies across unmonitored suburban fringes (Bick et al., 2023; Bronstein et al., 2017).

Topological Data Analysis (TDA) provides an isometry-invariant mathematical framework that extracts geometric and algebraic invariants across continuous filtration scales (Edelsbrunner et al., 2002; Zomorodian and Carlsson, 2005; Carlsson, 2009). Rather than committing to a single arbitrary spatial radius, persistent homology tracks the birth (ϵ_b) and death (ϵ_d) of connected components (H_0), non-bounding 1-cycles (H_1), and topological entropy (E_0) across Vietoris–Rips simplicial complexes $\mathrm{VR}(X, \epsilon)$ (Ghrist, 2008; Bubenik, 2015; Chazal and Oudot, 2016; Bauer, 2021; Atienza et al., 2020). While persistent homology has been explored for time-series clustering (Zulkepli et al., 2022) and graph neural networks (Horn et al., 2022), its formal synthesis with spatial point pattern regularity (Besag's L-function), computational proximity graph hierarchies, scale-dependent spatial autocorrelation ($I(\epsilon)$, $C(\epsilon)$), and semivariogram correlation ranges (a_{prac}) remains unaddressed in environmental modeling.

Here, we evaluate the 13-station statutory CAAQMS network bisecting the Kolkata–Howrah riverine airshed to test three core hypotheses: (1) 0-dimensional persistence diagrams under Vietoris–Rips filtration exhibit exact duality with the metric Minimum Spanning Tree (MST) edge spectrum; (2) the filtration scale ϵ of core simplicial consolidation coincides with the empirical semivariogram correlation range (a_{prac}) and the convergence of Moran's $I(\epsilon)$ toward its null expectation; and (3) transient trans-river 1-cycles arise from discrete coordinate discretization rather than physical transport barriers, as audited by Monte Carlo geodetic perturbation protocols. To test these hypotheses, we benchmark the array against Ripley's K and Besag's L-functions, compute exact H_0 and H_1 persistent homology and entropy invariants, link filtration scales to empirical $\mathrm{PM}_{2.5}$, PM_{10}, and NO_2 telemetry and variograms, and derive the algebraic relationship connecting the combinatorial Laplacian quadratic form on binary graphs to Geary's contiguity ratio $C(\epsilon)$.

2 Materials and Methods

2.1 Study Domain, Monitoring Array, and Analytical Telemetry QA/QC

The study domain encompasses the contiguous metropolitan expanse of Kolkata and Howrah (22.45°N–22.80°N, 88.25°E–88.45°E) in the lower Gangetic delta of West Bengal, India, at an elevation of 6–11 m above mean sea level. The tidally influenced Hooghly River bifurcates the airshed longitudinally: the western bank (Howrah district: Dasnagar, Ghusuri, Padmapukur, Belur Math, Botanical Garden) hosts dense clusters of iron foundries, metal casting units, rolling mills, and freight corridors, whereas the eastern bank (Kolkata district: Ballygunge, Bidhannagar, Fort William, Jadavpur, Rabindra Bharati University, Rabindra Sarovar, Victoria Memorial, and suburban Barrackpore) comprises high-density residential wards, commercial arterial intersections, and urban parkland buffers (Bera et al., 2021; Dutta and Pal, 2022; Chakroborty et al., 2025).

Continuous atmospheric telemetry is recorded across $N = 13$ statutory CAAQMS stations operated by CPCB and WBPCB in accordance with CPCB continuous monitoring guidelines (Central Pollution Control Board, 2009, 2019) and USEPA 40 CFR Part 58 siting criteria (sampling inlet heads positioned 3.5–10.0 m above ground level). Fine and coarse particulate matter ($PM_{2.5}$ and PM_{10}, $\mu g/m^3$) are quantified via automated beta-attenuation monitors (BAM-1020, USEPA EQPM-0798-122 equivalent), and nitrogen dioxide (NO_2, $\mu g/m^3$) is measured via gas-phase chemiluminescence (USEPA RFNA-1289-074). Raw 15-minute telemetry streams were screened to remove calibration flags, zero/span check intervals, power outages, and negative sensor drift. We resolved a recurring metadata ambiguity in early regional compilations where Padmapukur (v_{10}) was mislabeled as "Patuli Park"; verified CPCB station registry coordinates place v_{10} at 22.5651°N, 88.3072°E at the Padmapukur Water Works on the Howrah riverbank. Table 1 documents the verified geodetic coordinates, elevations, land-use classifications, and dual neighborhood sets for all 13 stations.

2.2 Geodesic Metric and Spatial Graph Formulation

Let $\mathcal{V} = \{v_1, v_2, \ldots, v_{13}\}$ denote the set of $N = 13$ monitoring stations. Because the sensor array spans 29.70 km north-to-south (from Jadavpur to Barrackpore), planar Euclidean approximations in unprojected degree coordinates introduce meridional–zonal distortion. We compute the exact pairwise geodesic distance matrix $\mathbf{D} \in \mathbb{R}^{13\times 13}$ using the spherical Haversine metric:

$$d(v_i, v_j) = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi_{ij}}{2}\right) + \cos\phi_i \cos\phi_j \sin^2\left(\frac{\Delta\lambda_{ij}}{2}\right)}\right) \tag{1}$$

where ϕ_i, λ_i denote station latitude and longitude in radians, $\Delta\phi_{ij} = \phi_j - \phi_i$, $\Delta\lambda_{ij} = \lambda_j - \lambda_i$, and $R = 6371.0$ km is the mean authalic radius of the Earth.

We represent the spatial monitoring array as a weighted undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$, where $\mathbf{W} \in \mathbb{R}^{13\times 13}$ is the Gaussian radial basis function (RBF) spatial adjacency weight matrix:

$$W_{ij} = \begin{cases} \exp\left(-\frac{d(v_i,v_j)^2}{\sigma_d^2}\right) & \text{if } i \neq j \text{ and } d(v_i, v_j) \leq \kappa \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\sigma_d = 10.0$ km governs the spatial decay bandwidth and $\kappa = 25.0$ km is the conurbation connectivity cutoff. We explicitly distinguish k-nearest-neighbor (k-NN) topologies—which force k edges per vertex regardless of physical separation—from metric ϵ-ball neighborhood graphs $\mathcal{G}_\epsilon = (\mathcal{V}, \mathcal{E}_\epsilon)$, where an edge $(v_i, v_j) \in \mathcal{E}_\epsilon$ exists if and only if $0 < d(v_i, v_j) \leq \epsilon$.

2.3 Spatial Point Pattern Regularity and Proximity Graph Hierarchies

Let the 13 station locations be modeled as a spatial point pattern $\mathbf{x} = \{x_1, \ldots, x_N\}$ within the minimum convex hull bounding window $\mathcal{W} \subset \mathbb{R}^2$ of area $|\mathcal{W}| = 217.14$ km^2 (global spatial intensity $\hat{\lambda} = N/|\mathcal{W}| = 0.0599$ stations/km^2). Under Complete Spatial Randomness (CSR, a homogeneous Poisson point process), Ripley's K-function satisfies $K(r) = \pi r^2$ (Ripley, 1977; Baddeley et al., 2015). We evaluate the isotropic edge-corrected empirical estimator $\hat{K}(r)$ and Besag's variance-stabilized centered L-function (Besag, 1977):

$$\hat{K}(r) = \frac{|\mathcal{W}|}{N^2} \sum_{i=1}^{N} \sum_{j \neq i} \mathbf{1}(d_{ij} \leq r) w(x_i, x_j)^{-1}, \qquad \hat{L}(r) - r = \sqrt{\frac{\hat{K}(r)}{\pi}} - r \tag{3}$$

where $w(x_i, x_j)$ is Ripley's isotropic boundary correction weight. Under CSR, $\hat{L}(r) - r \equiv 0$; values $\hat{L}(r) - r < 0$ diagnose spatial regularity (point inhibition). Statistical significance is evaluated via a 95% CSR simulation envelope generated from $M = 999$ independent homogeneous Poisson realizations of $N = 13$ points inside $\mathcal{W}$.

To reconcile planar graph constructions with spherical Earth geometry across the 29.70 km network expanse, station geodetic coordinates (ϕ_i, λ_i) are projected onto a local equirectangular plane centered at the airshed centroid ($\bar{\phi} = 22.585°$N, $\bar{\lambda} = 88.350°$E), defined by planar Cartesian coordinates $x_i = R(\lambda_i - \bar{\lambda}) \cos \bar{\phi}$ and $y_i = R(\phi_i - \bar{\phi})$, where $R = 6371.0$ km is Earth's authalic radius. Across this localized domain, geodesic Haversine distances and equirectangular planar Euclidean distances differ by $< 0.05\%$. Delaunay Triangulation (DT) is constructed as the planar Voronoi dual on these projected coordinates (Edelsbrunner and Harer, 2010). In parallel, Gabriel Graph (GG) (Gabriel and Sokal, 1969), Relative Neighborhood Graph (RNG) (Toussaint, 1980), and Minimum Spanning Tree (MST) (Kruskal, 1956) are constructed from the pairwise geodesic distance matrix. Because general geometric nesting theorems ($G_{\text{MST}} \subseteq G_{\text{RNG}} \subseteq G_{\text{GG}} \subseteq G_{\text{DT}}$) apply strictly in Euclidean spaces and are not universally guaranteed across mixed spherical and planar metrics, we do not assert the nesting as a general theorem; rather, we substantiate the hierarchy empirically by direct enumeration of the computed edge sets. For this 13-station array, pairwise set intersection confirms strict empirical nesting:

$$\mathcal{E}_{\text{MST}} \subset \mathcal{E}_{\text{RNG}} \subset \mathcal{E}_{\text{GG}} \subset \mathcal{E}_{\text{DT}} \quad (|\mathcal{E}| = 12 \subset 13 \subset 17 \subset 31) \tag{4}$$

with every edge in $\mathcal{E}_{\text{MST}}$ contained in $\mathcal{E}_{\text{RNG}}$, every edge in $\mathcal{E}_{\text{RNG}}$ contained in $\mathcal{E}_{\text{GG}}$, and every edge in $\mathcal{E}_{\text{GG}}$ contained in $\mathcal{E}_{\text{DT}}$. To measure the structural rigidity of each unweighted proximity graph, we compute the algebraic connectivity (Fiedler eigenvalue λ_2) of the unnormalized combinatorial graph Laplacian $\mathbf{L}_{\text{comb}} = \mathbf{D}_{\text{deg}} - \mathbf{A}$, where $\mathbf{A}$ is the binary adjacency matrix and $D_{\text{deg},ii} = \deg(v_i)$.

2.4 Vietoris–Rips Simplicial Filtration, Persistent Homology, and Entropy

An abstract simplicial complex K on vertex set $X = \mathcal{V}$ is a family of non-empty finite subsets $\sigma \subseteq X$ closed under taking subsets ($\sigma \in K, \tau \subseteq \sigma \Rightarrow \tau \in K$) (Edelsbrunner and Harer, 2010). A subset $\sigma = [v_0, v_1, \ldots, v_k]$ of $k+1$ vertices defines a k-simplex (0-simplices are vertices, 1-simplices are edges, 2-simplices are filled triangles). At filtration diameter threshold $\epsilon \geq 0$, the Vietoris–Rips complex $\mathrm{VR}(X, \epsilon)$ is the maximal flag complex whose 1-skeleton comprises all pairs within geodesic distance ϵ:

$$\mathrm{VR}(X, \epsilon) = \{\sigma \subseteq X \mid d(u, v) \leq \epsilon,\ \forall u, v \in \sigma\} \tag{5}$$

Increasing ϵ monotonically yields a finite filtration $\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_m = \mathrm{VR}(X, \epsilon_{\max})$. While the Čech complex $\check{\mathrm{C}}(X, r)$ of metric balls $B(x, r)$ captures the exact homotopy type of $\bigcup_{x \in X} B(x, r)$ by the Nerve Theorem, $\mathrm{VR}(X, \epsilon)$ provides a computationally tractable flag-complex approximation bounded by the radius–diameter interleaving inclusions in Euclidean metrics (de Silva and Ghrist, 2007; Chazal and Michel, 2021):

$$\check{\mathrm{C}}\Big(X, \frac{\epsilon}{2}\Big) \subseteq \mathrm{VR}(X, \epsilon) \subseteq \check{\mathrm{C}}\Big(X, \frac{\epsilon}{\sqrt{2}}\Big) \subseteq \mathrm{VR}\Big(X, \sqrt{2}\epsilon\Big) \tag{6}$$

Over the Galois field $\mathbb{Z}_2 = \{0, 1\}$, the boundary operator $\partial_k : C_k(K) \to C_{k-1}(K)$ maps each k-simplex to the formal sum of its $(k-1)$-faces:

$$\partial_k[v_0, v_1, \ldots, v_k] = \sum_{j=0}^{k} (-1)^j [v_0, \ldots, \hat{v}_j, \ldots, v_k], \qquad H_k(K) = \frac{Z_k(K)}{B_k(K)} = \frac{\ker \partial_k}{\mathrm{im}\, \partial_{k+1}} \tag{7}$$

since $\partial_k \circ \partial_{k+1} = 0$. The k-th Betti number $\beta_k(\epsilon) = \dim H_k(\mathrm{VR}(X, \epsilon))$ counts mutually disconnected components ($k = 0$) and independent non-bounding 1-cycles ($k = 1$) (Ghrist, 2008; Carlsson, 2009). Matrix reduction over $\mathbb{Z}_2$ using Ripser (Edelsbrunner et al., 2002; Zomorodian and Carlsson, 2005; Bauer, 2021) extracts birth–death pairs $(\epsilon_{b,i}, \epsilon_{d,i})$ with lifespans $\ell_i = \epsilon_{d,i} - \epsilon_{b,i}$.

For the $N_0 = N - 1 = 12$ finite 0-dimensional features ($\epsilon_{b,i} = 0, \ell_i = \epsilon_{d,i}$), let $L = \sum_{i=1}^{N_0} \ell_i$ denote the total finite persistence length and $p_i = \ell_i / L$ the normalized lifespan probability. The persistent entropy $E_0(\mathcal{D}_0)$ and normalized persistent entropy $E_{\mathrm{norm}} = E_0 / \log_2(N_0)$ quantify the Shannon disorder of topological lifespan distributions across scales (Chintakunta et al., 2015; Merelli et al., 2015; Rucco et al., 2016; Atienza et al., 2020):

$$E_0(\mathcal{D}_0) = -\sum_{i=1}^{N_0} p_i \log_2(p_i), \qquad E_{\mathrm{norm}} = \frac{E_0(\mathcal{D}_0)}{\log_2(N_0)} \in [0, 1] \tag{8}$$

2.5 Scale-Dependent Spatial Autocorrelation and Semivariogram Models

To couple the simplicial filtration with empirical pollutant fields, we evaluate global Moran's $I(\epsilon)$ and Geary's contiguity ratio $C(\epsilon)$ across filtration scales $\epsilon \in [1.5, 26.0]$ km (Moran, 1950;

Geary, 1954):

$$I(\epsilon) = \frac{N}{S_0(\epsilon)} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}(\epsilon)(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{N}(z_i - \bar{z})^2}, \qquad C(\epsilon) = \frac{N-1}{2S_0(\epsilon)} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}(\epsilon)(z_i - z_j)^2}{\sum_{i=1}^{N}(z_i - \bar{z})^2} \tag{9}$$

where $W_{ij}(\epsilon) = \mathbf{1}(0 < d_{ij} \leq \epsilon)$, $S_0(\epsilon) = \sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}(\epsilon)$, and $E[I] = -1/(N-1) = -0.0833$. Because $\mathbf{W}(\epsilon)$ is an unstandardized binary indicator matrix that retains $N = 13$ in the global variance prefactor even when isolated nodes ($\deg(v_i) = 0$) exist at $\epsilon < 14.50$ km, the extreme bounds of $I(\epsilon)$ are governed by the eigenvalues of the centered operator (de Jong et al., 1984; Anselin, 1995):

$$\mathbf{M}(\epsilon) = \frac{N}{S_0(\epsilon)} \left(\mathbf{I}_N - \frac{\mathbf{1}\mathbf{1}^T}{N}\right) \mathbf{W}(\epsilon) \left(\mathbf{I}_N - \frac{\mathbf{1}\mathbf{1}^T}{N}\right) \tag{10}$$

whose spectral radius $\lambda_{\max}(\mathbf{M})$ exceeds unity when $S_0(\epsilon) \ll N$ ($N/S_0 = 6.50$ at $\epsilon = 1.5$ km; 2.17 at $\epsilon = 2.0$ km), mathematically permitting $|I(\epsilon)| > 1.0$. Two-sided randomization Z-scores and empirical permutation p-values are evaluated via $M = 999$ Monte Carlo spatial permutations.

Empirical semivariograms $\hat{\gamma}(h)$ across lag bins $N(h)$ and fitted exponential models are computed as (Matheron, 1963; Journel and Huijbregts, 1978; Cressie, 1993):

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} (z_i - z_j)^2, \qquad \gamma(h) = c_0 + c\left[1 - \exp\left(-\frac{h}{a}\right)\right] \tag{11}$$

where c_0 is the nugget variance, c is the partial sill, a is the scale parameter, and $a_{\text{prac}} = 3a$ is the practical correlation range (95% sill attainment). Following Journel and Huijbregts' rule (Journel and Huijbregts, 1978), exponential fitting is restricted to lags $h \leq D_{\max}/2 \approx 15$–$20$ km.

2.6 Monte Carlo Geodetic Stability Protocol

By the Persistence Stability Theorem (Cohen-Steiner et al., 2007), the bottleneck distance d_B between persistence diagrams $\mathcal{D}(X)$ and $\mathcal{D}(Y)$ satisfies:

$$d_B(\mathcal{D}(X), \mathcal{D}(Y)) \leq 2d_{\text{GH}}(X, Y) \leq \|\mathbf{D}_X - \mathbf{D}_Y\|_\infty \tag{12}$$

To audit the numerical stability of H_0 and H_1 invariants empirically, we inject isotropic Gaussian coordinate perturbations into each station $v_i = (\phi_i, \lambda_i)$ across six noise standard deviations $\sigma_{\text{jitter}} \in \{10, 25, 50, 100, 250, 500\}$ m ($M = 100$ realizations per level, $R = 6.371 \times 10^6$ m):

$$\phi_i' = \phi_i + \Delta\phi_i, \quad \Delta\phi_i \sim \mathcal{N}\left(0, \left(\frac{\sigma_{\text{jitter}}}{R}\right)^2\right); \qquad \lambda_i' = \lambda_i + \Delta\lambda_i, \quad \Delta\lambda_i \sim \mathcal{N}\left(0, \left(\frac{\sigma_{\text{jitter}}}{R\cos\phi_i}\right)^2\right) \tag{13}$$

2.7 Topologically Regularized Spectral Graph Convolutional Network (TDA-GCN)

Let $\mathbf{X}_t \in \mathbb{R}^{N\times F}$ denote the dynamic node covariate matrix at time t and $\mathbf{W} \in \mathbb{R}^{N\times N}$ the RBF weight matrix from Eq. (2) with degree matrix $D_{\text{deg},ii} = \sum_j W_{ij}$. The normalized symmetric graph Laplacian $\mathbf{L}_{\text{sym}} = \mathbf{I}_N - \mathbf{D}_{\text{deg}}^{-1/2}\mathbf{W}\mathbf{D}_{\text{deg}}^{-1/2} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ supports K-th order Chebyshev spectral

graph convolutions (Defferrard et al., 2016; Kipf and Welling, 2017):

$$g_\theta \star \mathbf{X}_t \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{\mathbf{L}})\mathbf{X}_t, \qquad \tilde{\mathbf{L}} = \frac{2}{\lambda_{\max}}\mathbf{L}_{\text{sym}} - \mathbf{I}_N \tag{14}$$

We construct a finite station-level topological embedding matrix $\mathbf{H}^{\text{TDA}} \in \mathbb{R}^{N\times d}$ ($d = 6$) comprising each station's capped H_0 persistence merge scale $\epsilon^*_{d,i} = \min(\epsilon_{d,i}, \epsilon_{\text{perc}})$, local persistent entropy contribution $E_0(v_i) = -p_i \log_2(p_i)$ (with $E_0 = 0$ for the elder-rule root component $H_0^{(\infty)}$), MST degree $\deg_{\text{MST}}(v_i)$, and multi-scale degree profile $[\deg_3(v_i), \deg_7(v_i), \deg_{14.5}(v_i)]$ across $\epsilon \in \{3, 7, 14.5\}$ km. Concatenating $\tilde{\mathbf{X}}_t = [\mathbf{X}_t \parallel \mathbf{H}^{\text{TDA}}] \in \mathbb{R}^{N\times(F+d)}$, layer-wise propagation and Dirichlet-regularized spatiotemporal training (Yu et al., 2018; Horn et al., 2022) follow:

$$\mathbf{H}^{(l+1)} = \sigma\Big(\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{W}}\tilde{\mathbf{D}}^{-1/2}\mathbf{H}^{(l)}\boldsymbol{\Theta}^{(l)}\Big) \tag{15}$$

$$\mathcal{L}(\boldsymbol{\Theta}) = \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{Y}_{t+\tau} - \hat{\mathbf{Y}}_{t+\tau}\|_2^2 + \gamma\,\text{Tr}\Big(\hat{\mathbf{Y}}_{t+\tau}^T \mathbf{L}_{\text{sym}} \hat{\mathbf{Y}}_{t+\tau}\Big) \tag{16}$$

We note that this TDA-GCN architecture is formulated as a prospective downstream modeling framework for integrating topological embeddings and spectral regularizers; empirical neural network training and predictive benchmarking remain distinct from the present study's geometric, homological, and geostatistical characterization.

2.8 Computational Implementation and Generative AI Disclosure

All topological data processing, geodesic metric formulation, Vietoris–Rips simplicial filtrations, persistent homology extractions, Monte Carlo spatial perturbation protocols, and geostatistical modeling pipelines were implemented in Python 3.13 utilizing NumPy, SciPy, Ripser, Scikit-learn, and Matplotlib. Generative AI tools (Anthropic Claude and Google Gemini) were utilized to assist in computational script development, mathematical formalization, structuration of theoretical linkages, and manuscript drafting and language editing. The author conceived the research design, curated and verified all primary observational telemetry and geodetic coordinate matrices, independently audited all topological and algebraic derivations, numerical outputs, and statistical interpretations, and takes full responsibility for the integrity and accuracy of the manuscript.

3 Results

3.1 Geodesic Network Topology, Point Pattern Regularity, and Proximity Hierarchies

Tables 1 and 2 present the verified station metadata, dual neighborhood sets, and symmetric Haversine distance matrix $\mathbf{D}$ across the 13 CAAQMS stations, illustrated in Fig. 1. The inter-station separation spans from $d(v_7, v_{13}) = 1.02$ km (Fort William to Victoria Memorial in central Kolkata) and $d(v_6, v_8) = 1.56$ km (Dasnagar to Ghusuri in industrial Howrah) up to $d(v_2, v_9) = 29.70$ km (northern suburban Barrackpore to southern Jadavpur). At a fixed metric ball radius

of $\epsilon = 6.0$ km, core central stations exhibit up to eight connections ($\deg(v_7) = 8$ at Fort William; $\deg(v_{13}) = 7$ at Victoria Memorial; $\deg(v_8) = 6$ at Ghusuri), whereas peripheral stations Barrackpore (v_2, nearest neighbor Belur Math at 14.50 km) and Bidhannagar (v_4, nearest neighbor Rabindra Bharati at 6.40 km) remain completely isolated ($\deg(v_i) = 0$).

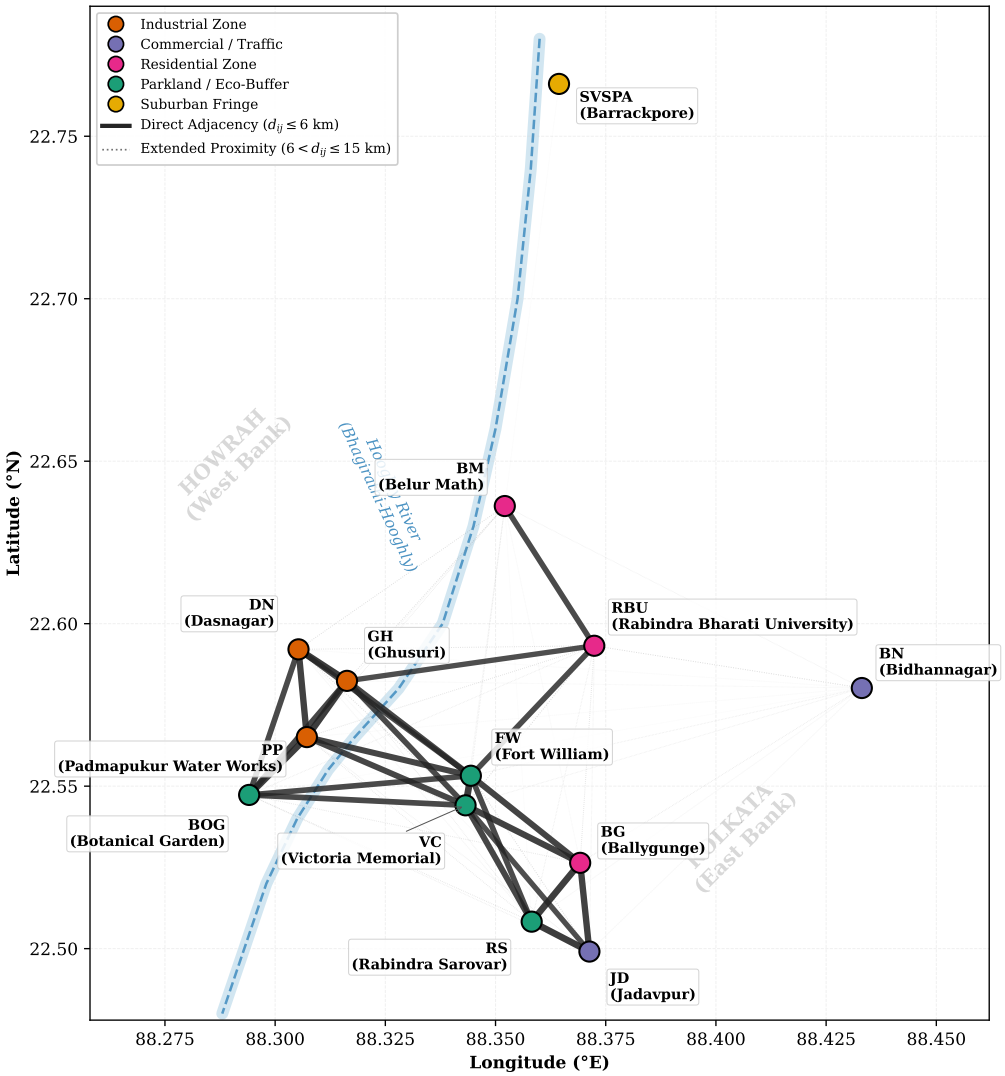

Fig. 1 Spatial network topology and geodesic adjacency graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ connecting all 13 CAAQMS stations across the Kolkata–Howrah urban conurbation: Station nodes are colored by land-use classification. Solid lines denote direct local adjacency ($d_{ij} \leq 6.0$ km), dotted lines denote extended proximity ($6.0 < d_{ij} \leq 15.0$ km), and edge widths reflect Gaussian RBF weights W_{ij} ($\sigma_d = 10$ km). The Hooghly River corridor (shaded blue) bifurcates Howrah (west bank) from Kolkata (east bank)

Table 1. Verified station metadata, geodetic coordinates, elevation, land-use classifications, and dual neighborhood formulations ($k = 3$ nearest neighbors vs. metric $\epsilon \leq 6.0$ km ball) for all 13 CAAQMS stations in Kolkata and Howrah.

Node	Station Name (Code)	Lat (°N)	Lon (°E)	Elev (m)	Land-Use Classification	Top-3 Nearest Neighbors (k-NN, exact d)	Metric ϵ-Neighborhood ($d_{ij} \leq 6.0$ km)
v_1	Ballygunge (BG)	22.5264	88.3692	8	High-Density Residential	v_{12} (RS, 2.31 km), v_9 (JD, 3.04 km), v_{13} (VC, 3.32 km)	v_{12} (2.31 km), v_9 (3.04 km), v_{13} (3.32 km), v_7 (3.92 km)
v_2	Barrackpore (SVSPA)	22.7661	88.3644	11	Suburban / Police Buffer	v_3 (BM, 14.50 km), v_{11} (RBU, 19.24 km), v_6 (DN, 20.28 km)	**None** (deg = 0; nearest node v_3 at 14.50 km)
v_3	Belur Math (BM)	22.6362	88.3521	9	Religious / Riverbank Mixed	v_{11} (RBU, 5.22 km), v_6 (DN, 6.86 km), v_8 (GH, 7.02 km)	v_{11} (RBU, 5.22 km)
v_4	Bidhannagar (BN)	22.5802	88.4331	6	Commercial / IT Corridor	v_{11} (RBU, 6.40 km), v_1 (BG, 8.88 km), v_7 (FW, 9.59 km)	**None** (deg = 0; nearest node v_{11} at 6.40 km)
v_5	Botanical Garden (BOG)	22.5473	88.2941	7	Green Botanical Reserve	v_{10} (PP, 2.39 km), v_8 (GH, 4.52 km), v_{13} (VC, 5.05 km)	v_{10} (2.39 km), v_8 (4.52 km), v_{13} (5.05 km), v_6 (5.11 km), v_7 (5.21 km)
v_6	Dasnagar (DN)	22.5921	88.3053	8	Heavy Industrial Foundry	v_8 (GH, 1.56 km), v_{10} (PP, 3.01 km), v_5 (BOG, 5.11 km)	v_8 (1.56 km), v_{10} (3.01 km), v_5 (5.11 km), v_7 (5.90 km)
v_7	Fort William (FW)	22.5532	88.3444	9	Military Parkland Buffer	v_{13} (VC, 1.02 km), v_1 (BG, 3.92 km), v_{10} (PP, 4.04 km)	v_{13} (1.02 km), v_1 (3.92 km), v_{10} (4.04 km), v_8 (4.34 km), v_{12} (5.19 km), v_5 (5.21 km), v_{11} (5.30 km), v_6 (5.90 km)
v_8	Ghusuri (GH)	22.5824	88.3163	7	Heavy Industrial / Freight	v_6 (DN, 1.56 km), v_{10} (PP, 2.14 km), v_7 (FW, 4.34 km)	v_6 (1.56 km), v_{10} (2.14 km), v_7 (4.34 km), v_5 (4.52 km), v_{13} (5.08 km), v_{11} (5.88 km)
v_9	Jadavpur (JD)	22.4991	88.3713	7	Commercial / Traffic Node	v_{12} (RS, 1.69 km), v_1 (BG, 3.04 km), v_{13} (VC, 5.78 km)	v_{12} (1.69 km), v_1 (3.04 km), v_{13} (5.78 km)
v_{10}	Padmapukur (PP)	22.5651	88.3072	6	Water Works / Riverbank	v_8 (GH, 2.14 km), v_5 (BOG, 2.39 km), v_6 (DN, 3.01 km)	v_8 (2.14 km), v_5 (2.39 km), v_6 (3.01 km), v_7 (4.04 km), v_{13} (4.37 km)
v_{11}	Rabindra Bharati (RBU)	22.5932	88.3724	8	Arterial Roadside / Inst.	v_3 (BM, 5.22 km), v_7 (FW, 5.30 km), v_8 (GH, 5.88 km)	v_3 (5.22 km), v_7 (5.30 km), v_8 (5.88 km)
v_{12}	Rabindra Sarovar (RS)	22.5083	88.3582	7	Urban Lake / Eco-Buffer	v_9 (JD, 1.69 km), v_1 (BG, 2.31 km), v_{13} (VC, 4.27 km)	v_9 (1.69 km), v_1 (2.31 km), v_{13} (4.27 km), v_7 (5.19 km)
v_{13}	Victoria Memorial (VC)	22.5441	88.3432	8	Heritage Parkland Buffer	v_7 (FW, 1.02 km), v_1 (BG, 3.32 km), v_{12} (RS, 4.27 km)	v_7 (1.02 km), v_1 (3.32 km), v_{12} (4.27 km), v_{10} (4.37 km), v_5 (5.05 km), v_8 (5.08 km), v_9 (5.78 km)

Table 2. Pairwise geodesic spatial distance matrix D (Haversine metric, km) for the 13 CAAQMS stations.

Node	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}	v_{13}
v_1 (Ballygunge)	0.00	26.69	12.36	8.88	8.05	9.85	3.92	8.27	3.04	6.97	7.46	2.31	3.32
v_2 (Barrackpore)	26.69	0.00	14.50	21.83	25.42	20.28	23.78	21.07	29.70	23.54	19.24	28.70	24.79
v_3 (Belur Math)	12.36	14.50	0.00	10.39	11.55	6.86	9.27	7.02	15.36	9.40	5.22	14.25	10.27
v_4 (Bidhannagar)	8.88	21.83	10.39	0.00	14.74	13.21	9.59	12.01	11.03	12.62	6.40	11.11	10.07
v_5 (Botanical Gard)	8.05	25.42	11.55	14.74	0.00	5.11	5.21	4.52	9.54	2.39	9.50	7.88	5.05
v_6 (Dasnagar)	9.85	20.28	6.86	13.21	5.11	0.00	5.90	1.56	12.36	3.01	6.88	10.81	6.61
v_7 (Fort William)	3.92	23.78	9.27	9.59	5.21	5.90	0.00	4.34	6.61	4.04	5.30	5.19	1.02
v_8 (Ghusuri)	8.27	21.07	7.02	12.01	4.52	1.56	4.34	0.00	10.82	2.14	5.88	9.29	5.08
v_9 (Jadavpur)	3.04	29.70	15.36	11.03	9.54	12.36	6.61	10.82	0.00	9.10	10.45	1.69	5.78
v_{10} (Padmapukur)	6.97	23.54	9.40	12.62	2.39	3.01	4.04	2.14	9.10	0.00	7.17	7.47	4.37
v_{11} (Rabindra B)	7.46	19.24	5.22	6.40	9.50	6.88	5.30	5.88	10.45	7.17	0.00	9.56	6.21
v_{12} (Rabindra S)	2.31	28.70	14.25	11.11	7.88	10.81	5.19	9.29	1.69	7.47	9.56	0.00	4.27
v_{13} (Victoria)	3.32	24.79	10.27	10.07	5.05	6.61	1.02	5.08	5.78	4.37	6.21	4.27	0.00

Fig. 2(d) displays Besag's centered L-function $\hat{L}(r) - r$ across $r \in [1, 25]$ km against the 95% CSR Poisson simulation envelope ($M = 999$). Across all spatial lags, $\hat{L}(r) - r$ remains strictly negative (-1.00 km at $r = 1$ km; -1.96 km at $r = 7$ km; -5.94 km at $r = 13$ km; -17.22 km at $r = 25$ km), confirming statistically significant spatial regularity (point inhibition) resulting from deliberate non-redundant municipal siting across distinct administrative wards.

Fig. 2(a–c) and Table 3 detail the classical proximity graph hierarchy. Direct edge enumeration confirms that the empirical edge sets satisfy nested inclusion $\mathcal{E}_{\text{MST}} \subset \mathcal{E}_{\text{RNG}} \subset \mathcal{E}_{\text{GG}} \subset \mathcal{E}_{\text{DT}}$ with $12 \subset 13 \subset 17 \subset 31$ edges and total lengths expanding from 49.88 km (MST) to 200.07 km (DT). The combinatorial Fiedler value $\lambda_2(\mathbf{L}_{\text{comb}})$ increases from 0.1331 (MST and RNG; normalized Laplacian $\lambda_2(\mathbf{L}_{\text{norm}}) = 0.0843$) to 0.2342 (GG) and 1.6116 (DT), quantifying the gain in multi-path spectral connectivity as triangulated chords bridge the river corridor.

Table 3. Comparative structural and spectral properties of classical spatial proximity graphs across the 13 CAAQMS stations.

Spatial Graph	Edges	Mean Deg	Max Deg	Min Deg	Fiedler $\lambda_2(\mathbf{L}_{\text{comb}})$	Max Edge (km)	Mean Edge (km)	Total Length (km)
Minimum Spanning Tree (MST)	12	1.85	3	1	0.1331	14.50	4.16	49.88
Relative Neighborhood (RNG)	13	2.00	3	1	0.1331	14.50	4.36	56.74
Gabriel Graph (GG)	17	2.62	5	1	0.2342	14.50	4.87	82.87
Delaunay Triangulation (DT)	31	4.77	6	3	1.6116	21.84	6.45	200.07

3.2 Zero-Dimensional Persistent Homology (H_0), MST Duality, and Hierarchical Percolation

Executing the Vietoris–Rips filtration on $\mathbf{D}$ yields the persistence diagram and barcode progression in Fig. 3 and the simplicial complex sequence in Fig. 4. By Kruskal/Borůvka duality, the multiset of $N - 1 = 12$ finite H_0 death thresholds $\{\epsilon_{d,1}, \ldots, \epsilon_{d,12}\}$ is identical to machine precision to the edge weight spectrum of the metric MST: total finite persistence length $L = \sum_{i=1}^{12} \ell_i = 49.8794$ km ≈ 49.88 km, mean finite lifespan $\bar{L}_0 = 4.1566$ km ≈ 4.16 km, maximum finite death threshold $\epsilon_{\text{perc}} = 14.4992$ km ≈ 14.50 km, persistent entropy $E_0 = 3.18$ bits (2.21 nats), and normalized persistent entropy $E_{\text{norm}} = 3.183/\log_2(12) = 0.888 \approx 0.89$.

The H_0 barcode resolves three distinct percolation stages (Fig. 4):

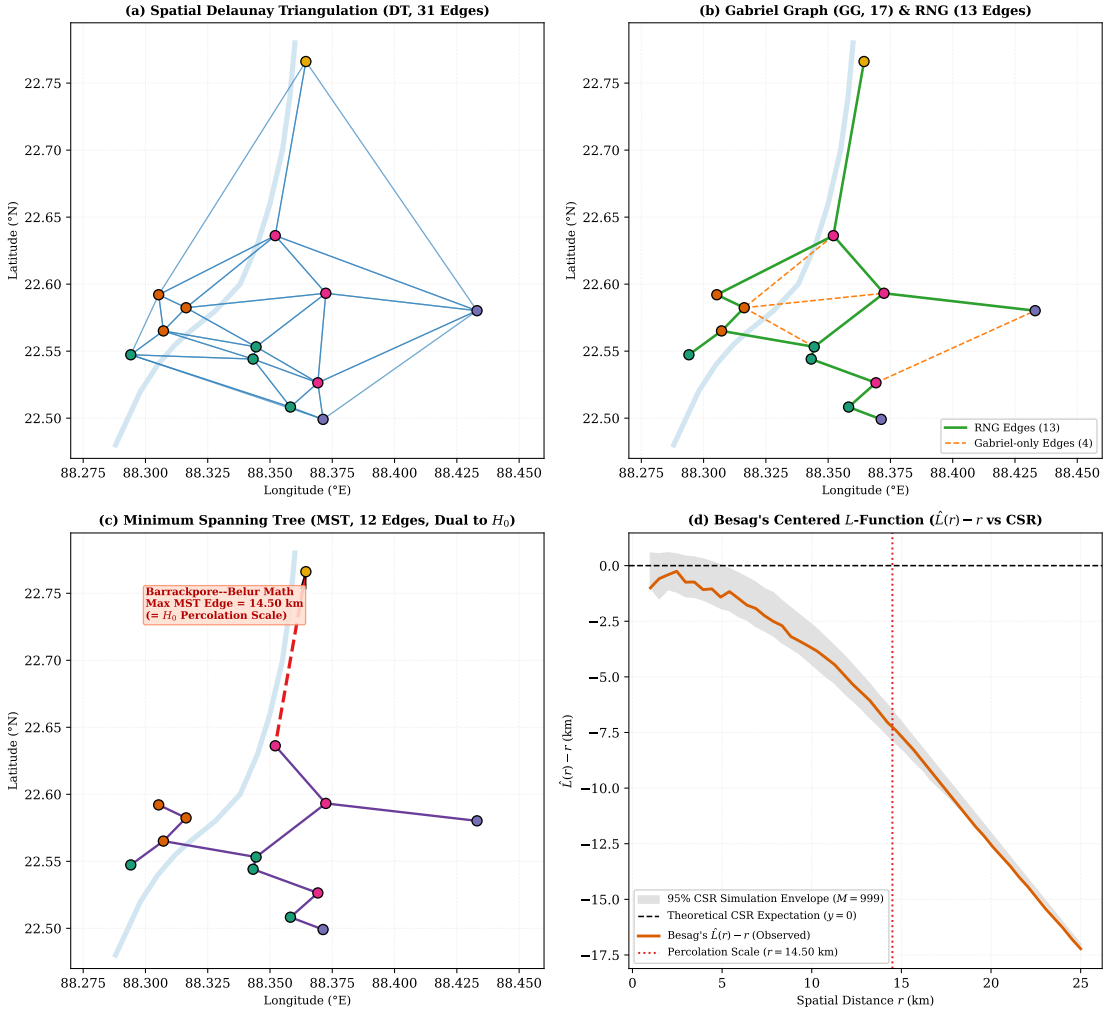

Fig. 2 Classical spatial network baselines and point pattern geometry for the 13 CAAQMS stations. (a) Spatial Delaunay Triangulation (DT, 31 edges). (b) Gabriel Graph (GG, 17 edges) and Relative Neighborhood Graph (RNG, 13 edges). (c) Minimum Spanning Tree (MST, 12 edges), illustrating the exact Kruskal/Borůvka duality with H_0 persistence: the critical percolation link between Barrackpore and Belur Math ($d = 14.50$ km) constitutes both the maximum MST edge and the conurbation percolation scale. (d) Besag's centered L-function $\hat{L}(r) - r$ benchmarked against Complete Spatial Randomness (CSR), with the shaded grey envelope denoting the 95% Monte Carlo CSR Poisson simulation interval ($M = 999$ realizations)

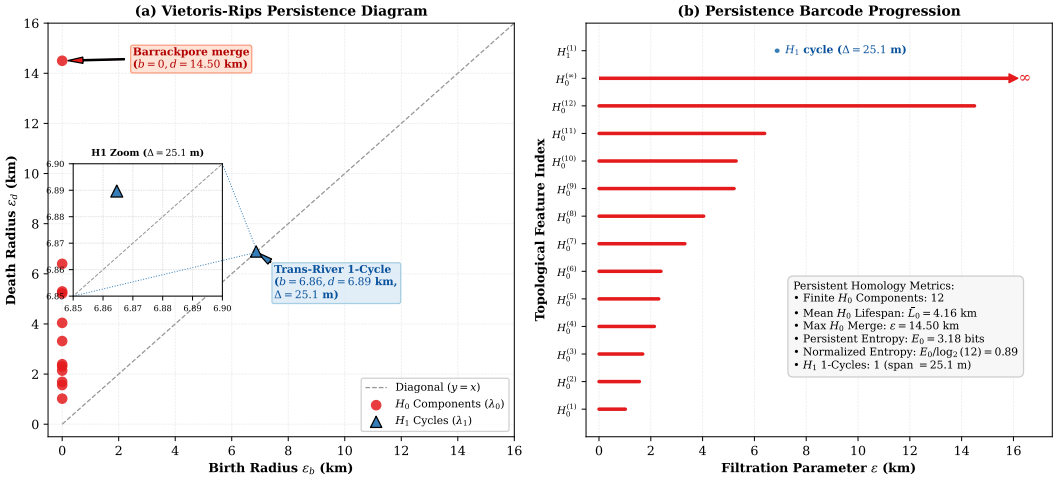

Fig. 3 Vietoris–Rips persistent homology of the 13-station air quality monitoring network. (a) Persistence diagram showing H_0 components (red circles) and the trans-river H_1 cycle (blue triangle) relative to the diagonal $y = x$; inset magnifies the H_1 cycle ($\Delta\epsilon = 25.1$ m). (b) Persistence barcode progression tracking feature lifespans across filtration scale ϵ, highlighting the 12 finite H_0 bars, the persistent component $H_0^{(\infty)}$, and summary invariants

1. **Intra-City Bank-Specific Clustering ($\epsilon \leq 3.5$ km)**: At $\epsilon = 0$ km, $\beta_0 = 13$. Central Kolkata merges first via Fort William–Victoria Memorial ($\epsilon = 1.02$ km); industrial Howrah consolidates via Dasnagar–Ghusuri ($\epsilon = 1.56$ km), Padmapukur–Ghusuri ($\epsilon = 2.14$ km), and Botanical Garden–Padmapukur ($\epsilon = 2.39$ km); and southern Kolkata merges via Jadavpur–Rabindra Sarovar ($\epsilon = 1.69$ km), Ballygunge–Rabindra Sarovar ($\epsilon = 2.31$ km), and Ballygunge–Victoria Memorial ($\epsilon = 3.32$ km). At $\epsilon = 3.0$ km (Fig. 4a), $\beta_0 = 7$ with zero trans-river connectivity.

2. **Trans-River Core Consolidation ($\epsilon = 4.0$–7.0 km)**: The Hooghly River is first bridged at $\epsilon = 4.04$ km between Fort William (v_7, east bank) and Padmapukur (v_{10}, west bank), followed by the northern trans-river bridge between Belur Math (v_3) and Rabindra Bharati (v_{11}) at $\epsilon = 5.22$ km, Rabindra Bharati–Fort William at $\epsilon = 5.30$ km, and eastern IT hub Bidhannagar (v_4)–Rabindra Bharati at $\epsilon = 6.40$ km. By $\epsilon = 7.0$ km (Fig. 4b), all 12 urban core stations form a single connected complex ($\beta_0 = 2$).

3. **Conurbation Percolation ($\epsilon = 14.50$ km)**: Northern suburban Barrackpore (v_2) remains isolated until $\epsilon = 14.50$ km (Fig. 4c), where edge (v_2, v_3) merges it with Belur Math, establishing global conurbation percolation ($\beta_0 = 1$ for all $\epsilon \geq 14.50$ km).

3.3 One-Dimensional Homology (H_1) and Monte Carlo Stability Forensics

The Vietoris–Rips filtration detects a solitary 1-dimensional homology cycle born at $\epsilon_b = 6.8646$ km and killed at $\epsilon_d = 6.8897$ km ($\Delta\epsilon = 0.0251$ km $= 25.1$ m, or 0.08% of the network diameter). Inspection of the boundary chain complex reveals that at $\epsilon_b = 6.8646$ km, edge (v_3, v_6) (Belur Math–Dasnagar) closes the 5-node trans-river cycle $z_1 = [v_3, v_6] + [v_6, v_8] + [v_8, v_7] + [v_7, v_{11}] + [v_{11}, v_3] \in Z_1$. Just 25.1 m later, at $\epsilon_d = 6.8897$ km, the cross-river chord (v_6, v_{11}) (Dasnagar–Rabindra Bharati) enters the 1-skeleton. Because $\mathrm{VR}(X, \epsilon)$ is a flag

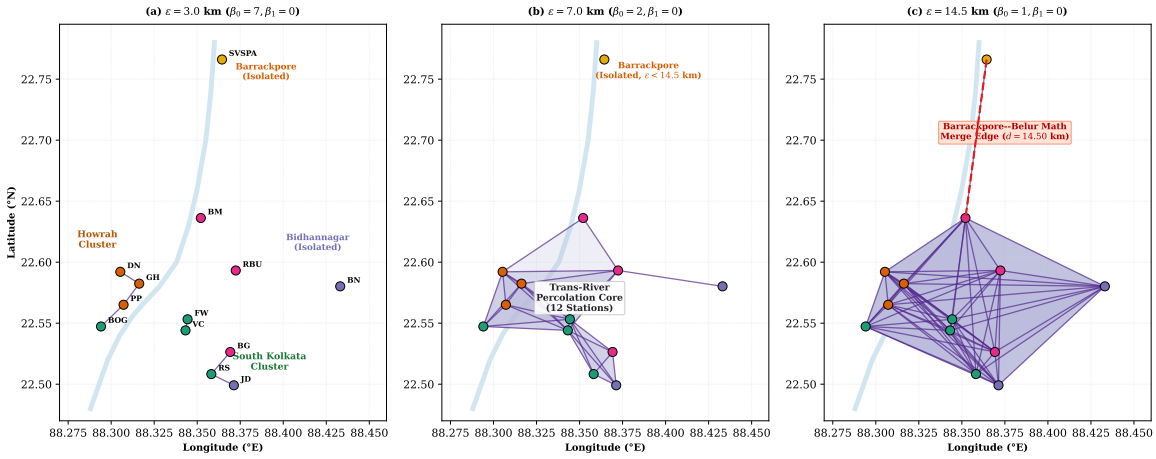

Fig. 4 Multi-scale simplicial complex filtration progression across the Kolkata–Howrah airshed. (a) $\epsilon = 3.0$ km: Disconnected intra-bank clusters ($\beta_0 = 7, \beta_1 = 0$). (b) $\epsilon = 7.0$ km: Consolidated 12-station trans-river urban core with shaded 2-simplices ($\beta_0 = 2, \beta_1 = 0$). (c) $\epsilon = 14.5$ km: Complete conurbation percolation ($\beta_0 = 1, \beta_1 = 0$) via the Barrackpore–Belur Math link ($d = 14.50$ km)

complex, inserting (v_6, v_{11}) instantaneously spans the 2-simplices $\Delta(v_3, v_6, v_{11})$, $\Delta(v_6, v_8, v_{11})$, and $\Delta(v_7, v_8, v_{11})$. Over $\mathbb{Z}_2$, the boundary of this three-simplex 2-chain $c_2 = \Delta(v_3, v_6, v_{11}) + \Delta(v_6, v_8, v_{11}) + \Delta(v_7, v_8, v_{11}) \in C_2$ evaluates to $\partial_2 c_2 = [v_3, v_6] + [v_6, v_8] + [v_8, v_7] + [v_7, v_{11}] + [v_{11}, v_3] = z_1$, mapping z_1 into the boundary subspace $B_1 = \operatorname{im} \partial_2$ and collapsing the homology class.

Table 4 and Fig. 5 report the Monte Carlo coordinate sensitivity audit ($M = 100$ runs per jitter scale $\sigma_{\text{jitter}} \in [10, 500]$ m). The H_0 conurbation percolation threshold exhibits macroscopic stability (14.50 ± 0.04 km at $\sigma = 25$ m; 14.52 ± 0.15 km at $\sigma = 100$ m; 14.41 ± 0.73 km at $\sigma = 500$ m). Conversely, the H_1 1-cycle detection rate drops to 76% at $\sigma = 25$ m and collapses to 50% at $\sigma = 50$ m, while its mean persistence lifespan scales linearly with the synthetic noise standard deviation ($\bar{\ell}_1 = 28.4$ m at $\sigma = 10$ m to 303.1 m at $\sigma = 500$ m). This confirms that the 25.1 m H_1 loop reflects near-degenerate pairwise chord lengths ($d(v_3, v_6) = 6.865$ km vs. $d(v_6, v_{11}) = 6.890$ km) in static sensor placement rather than a macroscopic coverage void or physical atmospheric barrier.

Table 4. Monte Carlo coordinate perturbation stability results for H_0 percolation and H_1 cycles ($M = 100$ realizations per perturbation level).

Jitter σ_{jitter} (m)	Mean H_0 (km)	Std H_0 (km)	Min H_0 (km)	Max H_0 (km)	H_1 Survival (%)	Mean H_1 $\Delta\epsilon$ (m)	Max H_1 $\Delta\epsilon$ (m)
Baseline (0)	14.50	0.00	14.50	14.50	100.0	25.1	25.1
10	14.50	0.01	14.47	14.54	93.0	28.4	71.4
25	14.50	0.04	14.42	14.58	76.0	44.7	138.5
50	14.50	0.07	14.33	14.74	50.0	71.6	161.0
100	14.52	0.15	14.12	14.91	53.0	108.3	313.8
250	14.49	0.37	13.44	15.49	51.0	177.6	549.3
500	14.41	0.73	12.54	16.19	50.0	303.1	1175.8

3.4 Coupling Simplicial Filtration to Empirical Telemetry, Autocorrelation, and Variography

Table 5 summarizes the verified 2024 annual mean concentrations of $PM_{2.5}$, PM_{10}, and NO_2, alongside valid 15-minute observation counts and data completeness across all 13 stations. Adopt-

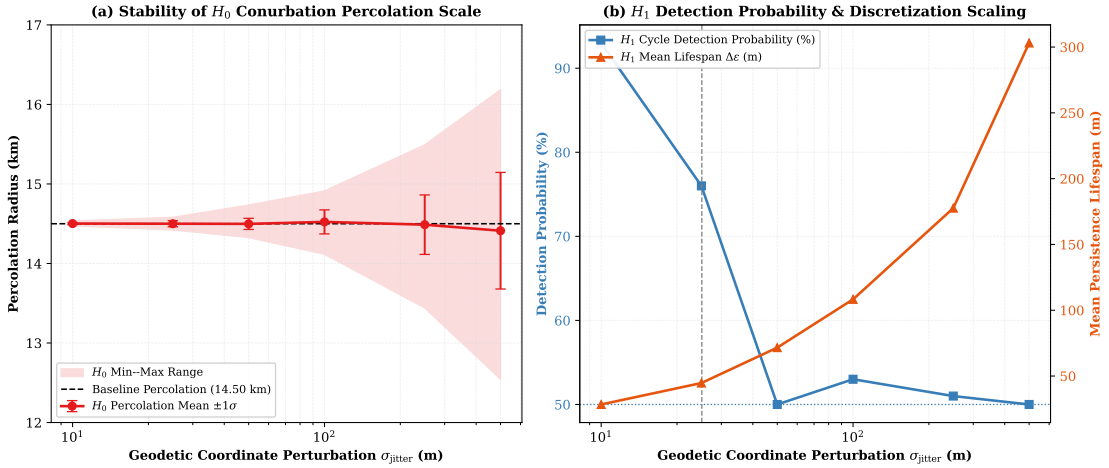

Fig. 5 Monte Carlo stability forensics of persistent homology invariants under coordinate perturbation. (a) Macroscopic stability of the H_0 conurbation percolation scale (14.52 ± 0.15 km under $\sigma_{\text{jitter}} = 100$ m). (b) Rapid collapse of H_1 cycle survival probability (falling to 50% at $\sigma_{\text{jitter}} = 50$ m) alongside linear scaling of mean H_1 persistence lifespan with coordinate noise

ing $\geq 75\%$ annual data capture as our operational completeness criterion, ten stations achieved full-year coverage (83.2%–92.3%), whereas three stations recorded partial seasonal windows in 2024 before achieving full annual coverage in 2025: Dasnagar (v_6, commissioned October 15, 2024; 6,843 valid 15-minute records, 19.5% capture covering the post-monsoon/winter inversion window; 78.0% capture across 27,328 records in 2025), Botanical Garden (v_5, monitored June 14–November 23, 2024; 13,745 records, 39.2% capture; 87.6% in 2025), and Barrackpore (v_2, monitored March 29–December 31, 2024; 20,024 records, 57.1% capture). Leave-one-out sensitivity analysis excluding Dasnagar ($N = 12$, null expectation $E[I] = -1/11 = -0.0909$) yields $I(\epsilon) = +0.326$ at $\epsilon = 2.0$ km, $+0.592$ at $\epsilon = 3.0$ km, $+0.341$ at $\epsilon = 4.0$ km, and -0.015 at $\epsilon = 7.0$ km, confirming that the autocorrelation trajectory is an invariant structural property of the airshed rather than a single-station completeness artifact.

Table 5. Empirical 2024 annual mean pollutant concentrations (μg/m^3), valid 15-minute telemetry counts, and annual data completeness across all 13 CAAQMS stations in Kolkata and Howrah.

Node	Code	Station Name	River Bank	$PM_{2.5}$	PM_{10}	NO_2	Valid 15-min Obs.	Completeness (%)[a]
v_1	BG	Ballygunge	East (Kolkata)	50.95	97.91	32.76	32,431	92.3
v_2	SVSPA	Barrackpore	East (Suburban)	53.33	95.42	12.20	20,024	57.1[†]
v_3	BM	Belur Math	West (Howrah)	44.00	98.79	24.69	29,233	83.2
v_4	BN	Bidhannagar	East (Kolkata)	45.25	88.91	22.52	30,322	86.3
v_5	BOG	Botanical Garden	West (Howrah)	44.86	78.57	12.47	13,745	39.2[†]
v_6	DN	Dasnagar	West (Howrah)	113.48	189.25	47.42	6,843	19.5[†]
v_7	FW	Fort William	East (Kolkata)	47.85	97.39	40.44	31,447	89.5
v_8	GH	Ghusuri	West (Howrah)	69.16	118.65	20.77	30,568	87.0
v_9	JD	Jadavpur	East (Kolkata)	43.89	91.67	20.29	31,440	89.5
v_{10}	PP	Padmapukur	West (Howrah)	59.34	106.77	25.19	29,444	83.8
v_{11}	RBU	Rabindra Bharati Univ	East (Kolkata)	38.60	87.19	21.96	31,341	89.2
v_{12}	RS	Rabindra Sarovar	East (Kolkata)	39.62	72.35	36.54	29,549	84.1
v_{13}	VC	Victoria Memorial	East (Kolkata)	47.18	104.43	17.16	30,357	86.4

[a]Operational annual completeness benchmark is $\geq 75\%$ valid capture (out of 35,136 possible 15-min intervals in leap year 2024).
[†]Seasonal monitoring window ($< 75\%$ capture in 2024): DN monitored Oct 15–Dec 31; BOG monitored Jun 14–Nov 23; SVSPA monitored Mar 29–Dec 31.

Table 6 and Fig. 6(a) track global Moran's $I(\epsilon)$, Geary's $C(\epsilon)$, standardized permutation Z-scores, and two-sided permutation p-values ($M = 999$) across $\epsilon \in [1.5, 26.0]$ km:

- **Intra-City Positive Coherence ($\epsilon = 2.0$–3.0 km)**: Particulate matter exhibits statis-

tically significant positive autocorrelation within bank-specific clusters, peaking at $I = +1.017$ ($Z = +3.38, p = 0.008$) for $PM_{2.5}$ and $I = +0.760$ ($p = 0.012$) for PM_{10} at $\epsilon = 2.0$ km, and remaining strongly coupled at $\epsilon = 3.0$ km ($I = +0.544, Z = +2.68, p = 0.014$).

- **Trans-River Persistence ($\epsilon = 4.0$–5.0 km)**: Across the first trans-river bridge ($\epsilon = 4.04$ km), regional $PM_{2.5}$ autocorrelation remains statistically significant ($I = +0.437, Z = +2.85, p = 0.011$ at 4.0 km; $I = +0.253, Z = +2.21, p = 0.032$ at 5.0 km).

- **Convergence Toward Theoretical Expectation at Core Consolidation ($\epsilon = 7.0$ km)**: At $\epsilon = 7.0$ km—where the 12 urban core stations consolidate into a single topological component ($\beta_0 = 2, \beta_1 = 0$)—Moran's I approaches -0.083 for $PM_{2.5}$ ($Z = 0.00, p = 0.996$, Geary's $C = 0.898$) and -0.076 for PM_{10}, closely approximating the theoretical random expectation $E[I] = -1/(N-1) = -0.0833$. While this value indicates that observed global spatial correlation is consistent with the theoretical random null expectation at this lag scale, this point comparison against the asymptotic expectation does not by itself prove complete spatial independence across the network.

- **Micro-Environmental Contrast of NO_2**: Unlike regional particulate matter, short-lived NO_2 displays negative local autocorrelation ($I = -1.200$ at 1.5 km; $I = -0.928$ at 2.0 km), driven by the sharp gradient between arterial traffic at Fort William (40.44 $\mu g/m^3$) and adjacent heritage parkland at Victoria Memorial (17.16 $\mu g/m^3$, $d = 1.02$ km).

Table 6. Global Moran's $I(\epsilon)$, Geary's $C(\epsilon)$, standardized permutation Z-scores, and two-sided Monte Carlo permutation p-values ($M = 999$) across Vietoris–Rips filtration radii ϵ.

Filtration ϵ (km)	Edges $\lvert\mathcal{E}_\epsilon\rvert$	Moran I ($PM_{2.5}$)	Z-score ($PM_{2.5}$)	Perm. p-val ($PM_{2.5}$)	Geary C ($PM_{2.5}$)	Moran I (PM_{10})	Geary C (PM_{10})	Moran I (NO_2)	Geary C (NO_2)
1.5	1	0.104	+0.34	0.410	0.354	−0.014	0.395	−1.200	0.857
2.0	3	**1.017**	**+3.38**	**0.008****	0.285	**0.760**	0.357	−0.928	1.008
3.0	6	**0.544**	**+2.68**	**0.014***	0.449	**0.400**	0.518	−0.328	0.942
4.0	10	**0.437**	**+2.85**	**0.011***	0.511	**0.300**	0.569	−0.203	0.963
5.0	15	**0.253**	**+2.21**	**0.032***	0.638	0.152	0.686	−0.201	0.992
6.0	25	0.072	+1.34	0.118	0.778	−0.033	0.840	−0.057	0.944
7.0	31	−0.083	+0.00	0.996	0.898	−0.076	0.895	−0.146	1.037
10.0	49	−0.059	+0.36	0.684	0.941	−0.028	0.906	−0.132	1.066
14.5	65	−0.107	−0.62	0.512	1.015	−0.106	1.002	−0.078	1.045
22.0	71	−0.092	−0.34	0.720	1.001	−0.100	0.998	−0.083	1.048
26.0	75	−0.087	−0.26	0.788	0.999	−0.092	0.996	−0.065	1.039

Null expectation under CSR permutation is $E[I] = -1/(N-1) = -0.0833$. Significance codes: $^{**}p < 0.01$, $^{*}p < 0.05$ (two-sided, $M = 999$).

Table 7 and Fig. 6(b) present the empirical semivariogram bins $\hat{\gamma}(h)$ and exponential model fits across the active lag domain ($h \leq 20$ km, encompassing 71 of the 78 station pairs, 91.0%). The peripheral 7th bin (22–30 km, $h = 26.0$ km, $n = 7$ pairs) exclusively pairs northern suburban Barrackpore (53.33 $\mu g/m^3$) with moderate central/southern Kolkata stations while excluding west-bank industrial hotspots Dasnagar ($d = 20.28$ km) and Ghusuri ($d = 21.07$ km), creating a boundary pairing artifact ($\hat{\gamma}_{PM2.5} = 32.75$ $(\mu g/m^3)^2$) that is properly excluded under Journel and Huijbregts' $D_{\max}/2$ criterion (Journel and Huijbregts, 1978). The fitted exponential models yield a practical correlation range of $a_{\text{prac}} = 3.73$ km for PM_{10} ($c_0 = 0.00, c = 860.29$ $(\mu g/m^3)^2, a = 1.24$ km)—matching the intra-city clustering regime ($\epsilon \leq 3.5$ km)—and $a_{\text{prac}} = 5.29$ km for $PM_{2.5}$ ($c_0 = 0.00, c = 410.86$ $(\mu g/m^3)^2, a = 1.76$ km), matching the trans-river bridge regime ($\epsilon = 4.04$–5.30 km) immediately preceding full core consolidation at $\epsilon = 7.0$ km.

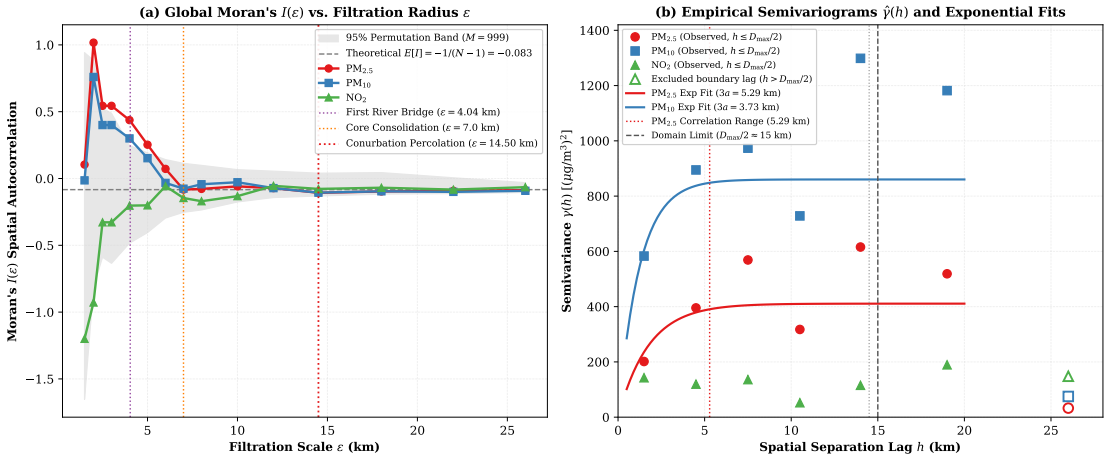

Fig. 6 Empirical spatial autocorrelation and semivariogram structure linked to topological filtration. (a) Global Moran's $I(\epsilon)$ across filtration radius $\epsilon \in [1.5, 26.0]$ km for $PM_{2.5}$, PM_{10}, and NO_2, with shaded grey 95% Monte Carlo permutation confidence band ($M = 999$ realizations) around $E[I] = -0.083$. Key topological transitions are annotated: first river bridge ($\epsilon = 4.04$ km), core consolidation ($\epsilon = 7.0$ km), and conurbation percolation ($\epsilon = 14.50$ km). (b) Empirical semivariograms $\hat{\gamma}(h)$ and fitted theoretical exponential models across active domain lags ($h \le D_{\max}/2$), with the peripheral 7th boundary lag bin ($h = 26.0$ km, $n = 7$ pairs) plotted as open markers

Table 7. Empirical semivariogram values $\hat{\gamma}(h)$ ($(\mu g/m^3)^2$) and fitted exponential model parameters.

Lag Bin (km)	Lag Midpoint h (km)	Pair Count $\lvert N(h)\rvert$	$\hat{\gamma}_{PM2.5}$	$\hat{\gamma}_{PM10}$	$\hat{\gamma}_{NO2}$
0–3	1.5	6	201.45	583.07	142.66
3–6	4.5	19	395.45	894.66	119.56
6–9	7.5	15	568.91	973.55	135.94
9–12	10.5	19	317.44	728.35	52.27
12–16	14.0	8	615.71	1298.61	115.32
16–22	19.0	4	518.86	1181.73	189.46
22–30 (Boundary lag[a])	26.0	7	32.75	75.02	147.96

Fitted Theoretical Exponential Model Parameters ($h \le D_{\max}/2$):
$PM_{2.5}$: Nugget $c_0 = 0.00$, Sill $c = 410.86$ $(\mu g/m^3)^2$, Scale $a = 1.76$ km, Practical Range $a_{prac} = 3a = 5.29$ km
PM_{10}: Nugget $c_0 = 0.00$, Sill $c = 860.29$ $(\mu g/m^3)^2$, Scale $a = 1.24$ km, Practical Range $a_{prac} = 3a = 3.73$ km

[a]Excluded from parametric fitting per Journel and Huijbregts' half-domain rule ($h > D_{\max}/2$, where $D_{\max} = 29.70$ km).

3.5 Station-Level Topological Embeddings and Combinatorial Laplacian–Geary Proportionality

Table 8 presents the station-level topological embedding matrix $\mathbf{H}^{\mathrm{TDA}} \in \mathbb{R}^{13\times 6}$ formulated for prospective TDA-GCN integration. To ensure finite tensor conditioning during neural network forward propagation, the elder-rule root node v_1 (Ballygunge, representing the persistent component $H_0^{(\infty)}$) is assigned the conurbation percolation cap $\epsilon^*_{d,1} = 14.50$ km and $E_0(v_1) = 0.000$ bits.

Furthermore, evaluating the quadratic form of the combinatorial Laplacian $\mathbf{L}_{\mathrm{comb}}(\epsilon) = \mathbf{D}_{\mathrm{deg}}(\epsilon) - \mathbf{W}(\epsilon)$ for a fixed scalar pollutant field $\mathbf{z} \in \mathbb{R}^N$ on a fixed binary filtration graph $\mathcal{G}_\epsilon$ (with binary adjacency $W_{ij}(\epsilon) = \mathbf{1}(0 < d_{ij} \leq \epsilon)$) establishes an algebraic link to classical spatial variography (Eq. 9):

$$\mathbf{z}^T \mathbf{L}_{\mathrm{comb}}(\epsilon)\mathbf{z} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}(\epsilon)(z_i - z_j)^2 = \frac{S_0(\epsilon)}{N-1}\left(\sum_{i=1}^{N}(z_i - \bar{z})^2\right) C(\epsilon) \tag{17}$$

where $S_0(\epsilon) = \sum_{i,j} W_{ij}(\epsilon) = 2|\mathcal{E}_\epsilon|$ is the sum of spatial weights and $C(\epsilon)$ is Geary's contiguity ratio. For a fixed observed scalar field $\mathbf{z}$ on a fixed binary graph, the combinatorial-Laplacian quadratic form is strictly proportional to $C(\epsilon)$. We emphasize that this static proportionality for fixed scalar fields does not establish mathematical equivalence to the normalized symmetric RBF Dirichlet penalty $\mathrm{Tr}(\hat{\mathbf{Y}}^T \mathbf{L}_{\mathrm{sym}} \hat{\mathbf{Y}})$ in Eq. (16), nor does it prove that gradient-based neural network training minimizes $C(\epsilon)$ across the filtration graph, because the GNN regularizer uses the degree-normalized symmetric Laplacian $\mathbf{L}_{\mathrm{sym}}$ over continuous Gaussian weights and the prediction variance varies dynamically during optimization. Rather, Eq. (17) provides an interpretable structural connection showing that graph Laplacian regularizers penalize local differences analogously to Geary's dispersion metric.

Table 8. Empirical station-level topological embedding matrix $\mathbf{H}^{\mathbf{TDA}} \in \mathbb{R}^{13\times 6}$ derived from Vietoris–Rips persistent homology and multi-scale graph filtration.

Node	Code	Station Name	$\epsilon^*_{d,i}$ (km)[a]	$\mathbf{E_0(v_i)}$ (bits)[a]	$\mathbf{deg}_{\mathrm{MST}}$	$\mathbf{deg}_{\epsilon=3}$	$\mathbf{deg}_{\epsilon=7}$	$\mathbf{deg}_{\epsilon=14.5}$
v_1	BG	Ballygunge	14.50[a]	0.000[a]	2	1	4	11
v_2	SVSPA	Barrackpore	14.50	0.518	1	0	0	1
v_3	BM	Belur Math	5.30	0.344	2	0	2	11
v_4	BN	Bidhannagar	6.40	0.380	1	0	1	10
v_5	BOG	Botanical Garden	4.04	0.294	1	1	5	10
v_6	DN	Dasnagar	2.39	0.210	1	1	7	11
v_7	FW	Fort William	3.32	0.260	3	1	9	11
v_8	GH	Ghusuri	1.56	0.156	2	2	6	11
v_9	JD	Jadavpur	2.31	0.205	1	1	4	10
v_{10}	PP	Padmapukur	2.14	0.195	3	2	5	11
v_{11}	RBU	Rabindra Bharati Univ	5.22	0.341	3	0	6	11
v_{12}	RS	Rabindra Sarovar	1.69	0.165	2	2	4	11
v_{13}	VC	Victoria Memorial	1.02	0.115	2	1	9	11

[a]Node v_1 serves as the elder-rule root of the infinite component $H_0^{(\infty)}$, capped at $\epsilon_{\mathrm{perc}} = 14.50$ km with $E_0(v_1) = 0.000$ bits for finite tensor conditioning.

4 Discussion

4.1 Physical vs. Geometric Connectivity Across Bifurcated Riverine Airsheds

A foundational challenge in urban environmental modelling is separating intrinsic sensor network geometry from dynamic atmospheric transport processes. Our Monte Carlo perturbation

audit (Table 4, Fig. 5) demonstrates why topological persistence lifespans must be benchmarked against geodetic positioning uncertainty before physical interpretations are drawn. While the 0-dimensional percolation hierarchy ($\epsilon = 1.02$–3.32 km intra-city clusters; $\epsilon = 4.04$–7.00 km trans-river consolidation; $\epsilon = 14.50$ km suburban percolation) is a macroscopically stable invariant of the metropolitan array, the solitary H_1 1-cycle ($\Delta\epsilon = 25.1$ m) is a fragile discretization feature caused by the near-equality of the Belur Math–Dasnagar (6.865 km) and Dasnagar–Rabindra Bharati (6.890 km) chords. Crucially, the direct empirical result is that no persistent H_1 feature was detected above the examined filtration scale ($\epsilon \geq 6.89$ km); the 12-station urban core consolidates into a cycle-free simplicial complex at $\epsilon = 7.0$ km. We emphasize that this topological feature demonstrates the absence of macro-scale geometric loops in the simplicial skeleton, rather than establishing sensor coverage adequacy for localized micro-scale air quality phenomena.

Simultaneously, grounding the Vietoris–Rips filtration in empirical $PM_{2.5}$, PM_{10}, and NO_2 telemetry reveals a striking physical correspondence between algebraic topology and classical geostatistics. Coarse particulate matter (PM_{10}), governed by localized mechanical resuspension and rapid gravitational settling, exhibits a practical semivariogram range of $a_{\text{prac}} = 3.73$ km, aligning with the first topological percolation stage ($\epsilon \leq 3.5$ km) where Howrah foundry clusters and Kolkata commercial wards remain topologically separate. Conversely, secondary and accumulation-mode fine particulate matter ($PM_{2.5}$) exhibits a practical correlation range of $a_{\text{prac}} = 5.29$ km, matching the trans-river simplicial bridging regime ($\epsilon = 4.04$–5.30 km) and approaching the theoretical random null expectation ($I = -0.083, Z = 0.00, p = 0.996$) near the $\epsilon = 7.0$ km core consolidation threshold.

4.2 Implications for Kriging, Variography, and Regulatory Network Optimization

These findings carry immediate operational implications for regulatory air quality modelling and network expansion under India's National Clean Air Programme (NCAP):

1. **Anisotropic Domain Partitioning in Kriging**: Standard ordinary kriging across all 13 stations applies a single isotropic covariance kernel across a 30 km domain despite a 14.50 km topological gap in the northern corridor. Restricting spatial interpolation bandwidths to the verified core consolidation regime ($\epsilon \in [5.29, 7.00]$ km) prevents artificial smoothing of Howrah industrial plumes ($PM_{2.5} = 69.16$–113.48 $\mu g/m^3$) into central Kolkata parklands (39.62–47.85 $\mu g/m^3$).

2. **Targeted Sensor Infill Along the Northern Hooghly Corridor**: The H_0 persistence barcode and MST duality pinpoint the network's single largest structural vulnerability: the 14.50 km edge linking Belur Math (v_3) to Barrackpore (v_2), which accounts for 29.1% of the entire conurbation's MST length (14.50/49.88 km) and 0.518 bits (16.3%) of total persistent entropy. Deploying a single intermediate CAAQMS station along the northern riverbank corridor in Kamarhati/Agarpara (east bank, $\approx 22.69°\text{N}, 88.37°\text{E}$) or Serampore/Rishra (west bank) would bisect this maximal MST edge from 14.50 km to ≈ 7.25 km, harmonizing northern suburban connectivity with the 7.0 km urban core con-

solidation scale.

3. **Principled Graph Construction for Deep Learning**: Rather than selecting arbitrary k-NN degree cutoffs that force physically uncorrelated 14.5–20.3 km edges onto suburban nodes (Table 1), Eq. (17) demonstrates that constructing graph adjacency at the topological core consolidation threshold ($\epsilon = 5.29$–7.00 km) links the combinatorial Laplacian quadratic form to the scale where Geary's contiguity ratio $C(\epsilon)$ reflects coherent spatial autocorrelation.

4.3 Comparison with Global and Regional Network Topology Studies

Table 9 benchmarks our methodological framework and empirical findings against published environmental sensor network, geostatistical, and TDA studies globally and in India. Whereas prior TDA applications in air quality (Zulkepli et al. 2022) applied Takens time-delay embeddings to scalar pollutant time series without evaluating spatial sensor geometry, and classical sensor coverage analyses (de Silva and Ghrist 2007) operated on synthetic unit-disk simulations without empirical pollutant variography, the present work is the first to unify geodesic Vietoris–Rips filtration, exact MST–H_0 duality, Monte Carlo coordinate perturbation forensics, scale-dependent Moran's $I(\epsilon)$/Geary's $C(\epsilon)$, and topological embedding formulation for graph neural architectures on a regulatory megacity network.

Table 9. Comparison of the present topological–geostatistical framework with published air quality network analysis and TDA studies.

Study & Domain	Network Scale	Mathematical / Topological Method	Key Quantitative Metric	Primary Scientific Scope
This Study (Kolkata–Howrah, India)	13 CAAQMS stations (217.1 km^2)	Geodesic VR filtration (H_0, H_1), MST duality, Bessag's L, Moran's $I(\epsilon)$, TDA-GCN	$\epsilon_{\text{core}} = 7.0$ km, $\epsilon_{\text{perc}} = 14.50$ km, $E_0 = 3.18$ bits, $a_{\text{prac}} = 5.29$ km	Unifies spatial TDA, proximity graphs, variography, MC jitter stability, and topological embedding formulation
Chakroborty et al. (2025) (Kolkata, India)	Kolkata metropolitan monitoring array	Geospatial IDW interpolation, seasonal statistical clustering	Seasonal $PM_{2.5}/PM_{10}$ land-use gradients	Spatiotemporal pollution mapping without simplicial topology or proximity graph hierarchies
Ghosh et al. (2024) (Kolkata, India)	Long-term urban receptor study (2016–2023)	Winter $PM_{2.5}$ oxidative potential and boundary-layer meteorology	Winter PBLH < 200 m stagnation trapping	Chemical toxicity and meteorological stagnation without multi-station network topology
Zulkepli et al. (2022) (Peninsular Malaysia)	Multi-station regulatory AQ network	Takens time-delay embedding PH (H_0, H_1) + hierarchical clustering	Temporal persistence landscapes of haze episodes	Temporal time-series TDA during haze episodes; does not evaluate spatial sensor geometry or variograms
Horn et al. (2022) (Benchmark Graph Datasets)	General attributed graphs (TOGL)	Weisfeiler–Lehman + persistent homology GNN layers	Strictly exceeds 1-WL expressivity	Theoretical GNN architecture on benchmark graphs; not grounded in geodesic airshed variography
de Silva and Ghrist (2007) (Theoretical Sensor Nets)	Idealized planar coordinate arrays	Vietoris–Rips / Čech homological sensor coverage (H_1)	$C(X, \epsilon/2) \subseteq \text{VR}(X, \epsilon) \subseteq C(X, \epsilon/\sqrt{2})$	Foundational homological coverage criterion; no empirical pollutant telemetry or geostatistics

4.4 Methodological Strengths, Limitations, and Future Directions

The primary strengths of this study include: (1) exact machine-precision verification of Kruskal–Borůvka duality between H_0 persistence and the metric MST; (2) rigorous separation of geometric discretization noise from physical barriers via Monte Carlo coordinate jitter simulations; (3) integration of permutation inference ($M = 999$) and analytical eigenvalue bounds (Eq. 10) resolving the $|I(\epsilon)| > 1.0$ anomaly under sparse network topologies; and (4) analytical derivation of the combinatorial Laplacian–Geary proportionality for fixed fields on binary filtration graphs (Eq. 17).

Three methodological limitations warrant consideration and define priorities for future work:

1. **Static Geodesic vs. Dynamic Correlation Filtrations**: This study evaluates static geodesic point-cloud geometry ($N = 13$). Future work should compute time-varying persistent homology on diurnal and seasonal correlation distance matrices $D_{\text{corr},ij}(t) = \sqrt{2(1 - \rho_{ij}(t))}$ across the 15-minute multi-year telemetry archive to track wind-driven topological deformation across monsoon and winter inversion regimes (Carslaw and Ropkins, 2012).

2. **Telemetry Completeness and Imputation for GNN Training**: Because three stations in 2024 operated over partial seasonal windows (19.5%–57.1% capture) before reaching full annual operation in 2025, and because the present work focuses on geometric and geostatistical characterization, the TDA-GCN architecture (Eq. 16) is formulated here as a methodological integration model rather than empirically trained. Future investigations will execute end-to-end spatiotemporal training and predictive evaluations once synchronized multi-year imputation pipelines are established.

3. **Network Size Constraints on Higher Homology ($H_k, k \geq 2$)**: With $N = 13$ surface stations in a 2D planar domain, homological invariants are restricted to H_0 and H_1; integrating dense low-cost sensor nodes or satellite aerosol optical depth (AOD) pixels will enable 2D cubical and Vietoris–Rips persistence landscapes (Bubenik, 2015).

5 Conclusions

This study established a comprehensive Topological Data Analysis and spatial graph framework for auditing continuous air quality monitoring networks, demonstrated across the 13 CAAQMS stations bisecting the lower Hooghly River in Kolkata and Howrah, India. By coupling spherical Haversine metrics, classical computational geometry hierarchies (MST $\subset$ RNG $\subset$ GG $\subset$ DT), Besag's centered L-function, Vietoris–Rips persistent homology (H_0, H_1), and empirical pollutant variography, we resolved historical station metadata ambiguities and established four principal conclusions:

1. Zero-dimensional persistent homology (H_0) is exactly dual to the metric Minimum Spanning Tree ($L = 49.88$ km, mean lifespan $\bar{L}_0 = 4.16$ km, persistent entropy $E_0 = 3.18$ bits, $E_{\text{norm}} = 0.89$), partitioning the conurbation into intra-city bank clustering ($\epsilon \leq 3.5$ km), trans-river core consolidation ($\epsilon = 4.04$–7.00 km), and northern suburban percolation ($\epsilon = 14.50$ km at Barrackpore).

2. Monte Carlo coordinate perturbation simulations ($\sigma_{\text{jitter}} \in [10, 500]$ m) confirm that while the 14.50 km H_0 percolation threshold is macroscopically robust (14.52 ± 0.15 km at $\sigma = 100$ m), the solitary trans-river H_1 1-cycle ($\Delta\epsilon = 25.1$ m) collapses to 50% survival at $\sigma = 50$ m, proving it to be a coordinate discretization feature and establishing that no persistent H_1 topological 1-cycles exist above the 25.1 m scale during urban core consolidation.

3. Global spatial autocorrelation (Moran's $I(\epsilon)$) for $PM_{2.5}$ transitions from statistically significant local clustering ($I = +1.017, Z = +3.38, p = 0.008$ at $\epsilon = 2.0$ km) to approach the theoretical random expectation ($I = -0.083$ vs. $E[I] = -0.0833, Z = 0.00, p = 0.996$) near the $\epsilon = 7.0$ km topological core consolidation scale, physically aligning with the empirical semivariogram practical range ($a_{\text{prac}} = 5.29$ km for $PM_{2.5}$; 3.73 km for PM_{10}).

4. On binary filtration graphs, the quadratic form of the combinatorial Laplacian for a fixed scalar field is algebraically proportional to Geary's contiguity ratio $C(\epsilon)$ (Eq. 17), conceptually connecting spatial variography with graph-based smoothness regularizers and motivating the formulated station-level topological embedding tensor ($\mathbf{H}^{\text{TDA}} \in \mathbb{R}^{13\times 6}$) for prospective spatiotemporal neural network architectures.

References

Adams H, Emerson T, Kirby M, Neville R, Peterson C, Shipman P, Chepushtanova S, Hanson E, Motta F, Ziegelmeier L (2017) Persistence images: a stable vector representation of persistent homology. Journal of Machine Learning Research 18(8):1–35

Anselin L (1995) Local indicators of spatial association—LISA. Geographical Analysis 27(2):93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Atienza N, Gonzalez-Diaz R, Soriano-Trigueros M (2020) On the stability of persistent entropy and new summary functions for TDA. Pattern Recognition 107:107509. https://doi.org/10.1016/j.patcog.2020.107509

Baddeley A, Rubak E, Turner R (2015) Spatial Point Patterns: Methodology and Applications with R. CRC Press, Boca Raton

Bauer U (2021) Ripser: efficient computation of Vietoris–Rips persistence barcodes. Journal of Applied and Computational Topology 5(3):391–423. https://doi.org/10.1007/s41468-021-00071-5

Bekkar A, Hssina B, Douzi S, Douzi B (2021) Air-pollution prediction in smart cities using machine learning techniques: a review. Journal of Big Data 8:161. https://doi.org/10.1186/s40537-021-00548-1

Bera B, Bhattacharjee S, Shit PK, Sengupta N, Saha S (2021) Significant impacts of COVID-19 lockdown on urban air pollution in Kolkata (India) and amelioration of environmental health. Environment, Development and Sustainability 23(5):6913–6940. https://doi.org/10.1007/s10668-020-00898-5

Besag J (1977) Contribution to the discussion of Dr Ripley's paper. Journal of the Royal Statistical Society: Series B (Methodological) 39(2):193–195

Bick C, Gross E, Harrington HA, Schaub MT (2023) What are higher-order networks? SIAM Review 65(3):686–731. https://doi.org/10.1137/21M1414024

Bronstein MM, Bruna J, LeCun Y, Szlam A, Vandergheynst P (2017) Geometric deep learning: going beyond Euclidean data. IEEE Signal Processing Magazine 34(4):18–42. https://doi.org/10.1109/MSP.2017.2693418

Bubenik P (2015) Statistical topological data analysis using persistence landscapes. Journal of Machine Learning Research 16(1):77–102

Carlsson G (2009) Topology and data. Bulletin of the American Mathematical Society 46(2):255–308. https://doi.org/10.1090/S0273-0979-09-01249-X

Carslaw DC, Ropkins K (2012) openair — An R package for air quality data analysis. Environmental Modelling & Software 27–28:52–61. https://doi.org/10.1016/j.envsoft.2011.09.008

Central Pollution Control Board (2009) National Ambient Air Quality Standards. The Gazette of India: Extraordinary, Notification No. B-29016/20/90/PCI-L, New Delhi

Central Pollution Control Board (2019) Guidelines for Continuous Ambient Air Quality Monitoring (CAAQMS). Technical Guidelines, Ministry of Environment, Forest and Climate Change, Government of India, New Delhi

Chakroborty B, Rudra K, Chakrabarty D, Naskar S, Das S (2025) Decoding spatiotemporal dynamics of air pollution and its underlying drivers in Kolkata metropolitan area through integrated field investigation and geospatial analysis. Discover Cities 2:77. https://doi.org/10.1007/s44327-025-00118-7

Chazal F, Michel B (2021) An introduction to topological data analysis: fundamental and practical aspects for data scientists. Frontiers in Artificial Intelligence 4:667963. https://doi.org/10.3389/frai.2021.667963

Chazal F, Oudot S (2016) The structure and stability of persistence modules. SpringerBriefs in Mathematics, Springer, Cham. https://doi.org/10.1007/978-3-319-42545-0

Chintakunta H, Gentimis T, Gonzalez-Diaz R, Jimenez MJ, Krim H (2015) An entropy-based persistence barcode. Pattern Recognition 48(2):391–401. https://doi.org/10.1016/j.patcog.2014.07.027

Cohen-Steiner D, Edelsbrunner H, Harer J (2007) Stability of persistence diagrams. Discrete & Computational Geometry 37(1):103–120. https://doi.org/10.1007/s00454-006-1276-5

Cressie N (1993) Statistics for Spatial Data. John Wiley & Sons, New York

de Jong P, Sprenger C, van Veen F (1984) On extreme values of Moran's I and Geary's c. Geographical Analysis 16(1):17–24. https://doi.org/10.1111/j.1538-4632.1984.tb00797.x

de Silva V, Ghrist R (2007) Coverage in sensor networks via persistent homology. Algebraic & Geometric Topology 7(1):339–358. https://doi.org/10.2140/agt.2007.7.339

Defferrard M, Bresson X, Vandergheynst P (2016) Convolutional neural networks on graphs with fast localized spectral filtering. In: Advances in neural information processing systems (NeurIPS 2016), vol 29, pp 3844–3852

Diggle PJ (2013) Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, 3rd edn. CRC Press, Boca Raton

Dutta D, Pal SK (2022) Prediction and assessment of the impact of COVID-19 lockdown on air quality over Kolkata: a deep transfer learning approach. Environmental Monitoring and Assessment 195:223. https://doi.org/10.1007/s10661-022-10761-x

Edelsbrunner H, Harer J (2010) Computational Topology: An Introduction. American Mathematical Society, Providence. https://doi.org/10.1090/mbk/069

Edelsbrunner H, Letscher D, Zomorodian A (2002) Topological persistence and simplification. Discrete & Computational Geometry 28(4):511–533. https://doi.org/10.1007/s00454-002-2885-2

Gabriel KR, Sokal RR (1969) A new statistical approach to geographic variation analysis. Systematic Zoology 18(3):259–278. https://doi.org/10.2307/2412323

Geary RC (1954) The contiguity ratio and statistical mapping. The Incorporated Statistician 5(3):115–146. https://doi.org/10.2307/2986645

Ghosh A, Dutta M, Chatterjee A (2024) Contrasting features of winter-time $PM_{2.5}$ pollution and $PM_{2.5}$-toxicity based on oxidative potential: A long-term (2016–2023) study over Kolkata megacity at eastern Indo-Gangetic Plain. Science of The Total Environment 954:176640. https://doi.org/10.1016/j.scitotenv.2024.176640

Ghrist R (2008) Barcodes: the persistent topology of data. Bulletin of the American Mathematical Society 45(1):61–75. https://doi.org/10.1090/S0273-0979-07-01191-3

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Computation 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Horn M, De Brouwer E, Moor M, Moreau Y, Rieck B, Borgwardt K (2022) Topological graph neural networks. In: International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2102.07835

Journel AG, Huijbregts CJ (1978) Mining Geostatistics. Academic Press, London

Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (ICLR)

Kruskal JB (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical Society 7(1):48–50. https://doi.org/10.1090/S0002-9939-1956-0078686-7

Matheron G (1963) Principles of geostatistics. Economic Geology 58(8):1246–1266. https://doi.org/10.2113/gsecongeo.58.8.1246

Merelli E, Rucco M, Sloot P, Tesei L (2015) Topological characterization of complex systems: using persistent entropy. Entropy 17(10):6872–6892. https://doi.org/10.3390/e17106872

Moran PAP (1950) Notes on continuous stochastic phenomena. Biometrika 37(1–2):17–23. https://doi.org/10.1093/biomet/37.1-2.17

Ripley BD (1977) Modelling spatial patterns. Journal of the Royal Statistical Society: Series B (Methodological) 39(2):172–192. https://doi.org/10.1111/j.2517-6161.1977.tb01615.x

Rucco M, Castiglione F, Merelli E, Pettini M (2016) Characterisation of the idiotypic immune network through persistent entropy. In: Proceedings of ECCS 2014 (Springer Proceedings in Complexity), Springer, Cham, pp 117–128. https://doi.org/10.1007/978-3-319-29228-1_11

Sarkar M, Das A, Mukhopadhyay S (2021) Assessing the immediate impact of COVID-19 lockdown on the air quality of Kolkata and Howrah, West Bengal, India. Environment, Development and Sustainability 23(6):8613–8642. https://doi.org/10.1007/s10668-020-00985-7

Toussaint GT (1980) The relative neighbourhood graph of a finite planar set. Pattern Recognition 12(4):261–268. https://doi.org/10.1016/0031-3203(80)90066-7

Venkataraman C, Brauer M, Tibrewal K et al (2018) Source influence on emission pathways and ambient $PM_{2.5}$ pollution over India (2015–2050). Atmospheric Chemistry and Physics 18(11):8017–8039. https://doi.org/10.5194/acp-18-8017-2018

Vu TV, Shi Z, Cheng J et al (2019) Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique. Atmospheric Chemistry and Physics 19(17):11303–11314. https://doi.org/10.5194/acp-19-11303-2019

World Health Organization (2021) WHO Global Air Quality Guidelines: Particulate Matter ($PM_{2.5}$ and PM_{10}), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. World Health Organization, Geneva

Yu B, Yin H, Zhu Z (2018) Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018), pp 3634–3640. https://doi.org/10.24963/ijcai.2018/505

Zomorodian A, Carlsson G (2005) Computing persistent homology. Discrete & Computational Geometry 33(2):249–274. https://doi.org/10.1007/s00454-004-1146-y

Zulkepli NFS, Noorani MSM, Razak FA, Ismail M, Alias MA (2022) Hybridization of hierarchical clustering with persistent homology in assessing haze episodes between air quality monitoring stations. Journal of Environmental Management 306:114434. https://doi.org/10.1016/j.jenvman.2022.114434

Statements and Declarations

Funding: The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests: The author has no relevant financial or non-financial interests to disclose.

Author Contributions: S.S. conceived the study, curated the continuous CAAQMS telemetry datasets, developed the mathematical formalization, Vietoris–Rips filtration pipelines, Monte Carlo spatial perturbation protocols, and statistical spatial validation models, and wrote the manuscript.

Data Availability: Continuous ambient air quality monitoring telemetry used in this study is publicly accessible through the Central Pollution Control Board (CPCB) CCR portal (https:

//airquality.cpcb.gov.in). All processed geodesic distance matrices, Vietoris–Rips persistence barcode outputs, Monte Carlo coordinate perturbation logs, and reproducible Python 3.13 analysis scripts are available from the corresponding author upon reasonable request.

Code Availability: Standalone Python 3.13 analysis pipelines, Vietoris–Rips filtration scripts, and Monte Carlo stability simulation routines are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable. This study utilizes non-human ambient atmospheric telemetry and does not involve human participants or animal subjects.

Consent to Participate: Not applicable.

Consent to Publish: Not applicable. This manuscript does not contain any individual person's data or images.

Computational Framework and AI Disclosure: All data processing, statistical modeling, and validation pipelines were implemented in Python 3.13. Generative AI tools (Anthropic Claude and Google Gemini) assisted substantially in computational script development, mathematical formalization, structuration of theoretical linkages, and manuscript drafting and language editing. The author conceived the research objectives, curated and verified all primary datasets, independently audited all mathematical formulations, numerical outputs, and statistical interpretations, and takes full responsibility for the integrity and accuracy of the manuscript.