Transactions in GIS

WILEY

Transactions in GIS

RESEARCH ARTICLE

Scale Snapshot Topology Distance: Quantifying the Spatial Scale Effect From a Topological Perspective

Jiawei Zhu[1] | Song Gao[2] | Chao Tao[3] | Yu Liu[4] | Haifeng Li[3]

[1]School of Architecture and Art, Central South University, Changsha, China | [2]Geospatial Data Science Lab, University of Wisconsin-Madison, Madison, USA | [3]School of Geosciences and Info-Physics, Central South University, Changsha, China | [4]Institute of Remote Sensing and Geographical Information Systems, Peking University, Beijing, China

Correspondence: Haifeng Li (lihaifeng@csu.edu.cn)

Received: 31 December 2024 | **Revised:** 14 May 2025 | **Accepted:** 4 June 2025

Funding: This work was supported by National Natural Science Foundation of China, 42301537, and the High-Performance Computing Platform of Central South University: HPC-CSU-1038.

Keywords: modifiable areal unit problem | persistent homology | scale effect | scale snapshot topology distance | topological data analysis

ABSTRACT

Recognizing the degree of scale effect is crucial for selecting appropriate scales in spatial analyses. However, few methods are available for quantifying this effect. Topology, which studies invariants preserved under continuous deformation, provides an effective means of characterizing data. Therefore, we propose quantifying the scale effect by calculating the distance between topological invariants of data aggregated at different scales, termed the scale snapshot topology distance (SSTD). We summarize data snapshots aggregated at different scales using persistence diagrams, capturing essential information about topological invariants. We then quantify SSTD by computing the Wasserstein distance between these diagrams and apply it to track data variation across scale changes. Experiments on origin–destination data from five cities validate the effectiveness of the SSTD metric. Results demonstrate that our method identifies critical spatial scales near the consistency boundary of the scale effect, providing guidance for scale selection and showcasing the feasibility of topology-based spatial data analysis.

1 | Introduction

The modifiable areal unit problem (MAUP) is a fundamental issue in geographical studies, which is acknowledged as the phenomenon that arbitrary choices of spatial aggregation configurations may cause different results and biases (Openshaw 1984; Fotheringham and Wong 1991; McMillen 2004; Martin and Schuurman 2020; Ye and Rogerson 2022). This problem can be divided into two different but related aspects, namely, the scale effect and the zoning effect (Fotheringham 1989; Miller 1999). Specifically, the former refers to the change in the results due to different sizes of units in which data are grouped; the latter signifies the effect brought by the manner in which the unit is divided with different spatial configurations with varying shapes and boundaries, such as regular grids or Thiessen polygons. In this paper, we focus on the scale effect. The term "scale" has several definitions, and we adopt the spatial scale defined as the measurement concept throughout the paper, which refers to the resolution of the spatial unit for analysis (Atkinson and Tate 2000).

Despite the long-standing recognition of the scale effect and extensive research efforts, there remains a lack of established rules or universal conventions to guide the spatial aggregation process (Openshaw 1977, 1984; Nelson 2001; Mu and Wang 2008; Stehman and Wickham 2011; Manley 2014). Choosing the "right" scale relies heavily on practitioners' empirical speculation due to the limited understanding of the

scale effect. One feasible solution is to find a quantitative way to measure the difference caused by different aggregation scales. Once the difference between the data aggregated at different scales can be precisely defined and measured, we can establish guiding principles for choosing an appropriate scale based on the degree of deviation between aggregations. We define data aggregated at different spatial scales as *scale snapshots* for ease of expression. Thus, the key to the problem can be expressed as follows: Given a sequence of ascending spatial scales, how can we quantify the differences between the corresponding scale snapshots?

To address this, we introduce a novel framework grounded in algebraic topology. Mathematically, topology involves the study of topological invariants preserved under continuous deformation, which can be used to characterize data (Carlsson 2009). Therefore, we propose a novel methodological framework that utilizes the summaries of topological invariants as surrogates to measure the distance between scale snapshots. In particular, we use persistent homology theory to extract summaries of topological invariants. Persistent homology studies the structure of complex and high-dimensional data through a multi-level analysis and has recently been used in exploratory data analysis and data mining (Hajij et al. 2018). Leveraging persistent homology, the proposed framework first extracts topological invariants across dimensions and then produces comprehensive topological summaries for scale snapshots. Then, we use a mathematically well-defined metric to assess the disparity among topological summaries of scale snapshots, which we name as the *scale snapshot topology distance (SSTD)*. Finally, we track how the scale snapshot changes across different spatial scales.

The contributions of our work are listed as follows. First, we propose the scale snapshot topology distance (SSTD) to measure the variation caused by the spatial scale effect by tracking the changes in the topology of scale snapshots. Second, in addition to clustering patterns to which traditional analysis on the scale effect is restricted, we measure the changes in higher-dimensional loop patterns to understand the scale effect more comprehensively. Third, we conduct case studies to validate the effectiveness of the proposed SSTD metric in the context of spatial interactions. According to our experimental results, the spatial clustering pattern evolves consistently across different aggregation settings, while abrupt changes in the high-dimensional patterns are observed. The critical scale, which is indicative of when the shift occurs in the topological evolution, can potentially provide a reference for choosing an appropriate spatial scale. To the best of our knowledge, this is the first time the scale effect has been explored with algebraic topology.

The remainder of the paper is organized as follows. Section 2 introduces the related works on studying scale effects. Section 3 introduces basic concepts in topology and persistent homology. The theoretical foundation of applying topology to analyze the scale effect and the pipeline of our work is given in Section 4. Next, we demonstrate the case studies in five cities and report our findings as well as those using traditional methods in Section 5. We further discuss the broader implications of the proposed framework in Section 6. In the last section, we conclude with a summary of the results and offer insights on future research directions.

1.1 | Related Work

Understanding the scale effect has long been a critical concern in geographic research, particularly due to the widespread practice of aggregating individual-level data into areal units (Atkinson et al. 2014). In the scope of spatial distribution, various studies on how different scales can affect the statistical results have been performed (Gehlke and Biehl 1934; Clark and Avery 1976; Fotheringham and Wong 1991; Amrhein and Flowerdew 1992; Amrhein 1995; Amrhein and Reynolds 1996; Green and Flowerdew 1996; Holt et al. 1996), including measurements such as mean, variance, covariance, and correlation coefficients, varying from a single variable to multiple variables, and from linear regression to the Poisson regression model. In order to understand the effect of changing spatial scales, researchers have also explored its impact on spatial autocorrelation statistics (Qi and Wu 1996; Jelinski and Wu 1996), spatial interpolation (Cressie 1996) and accessibility (Kotavaara et al. 2012; Stepniak and Rosik 2015). Xiao (2020) proposed a metric named spatial aggregation entropy (SAE) to quantify the changes in spatial heterogeneity and uncertainty caused by spatial aggregation, providing a new perspective and computable approach to spatial analysis. Zhou and Yeh (2021) utilized mobile positioning data to examine MAUP and determine unit size by calculating changes in the variation coefficient of indicators related to job-housing balance. Gao et al. (2022) explored the scale effects on commuting demand modeling in Shenzhen by analyzing nine spatial units from three partitioning schemes and applying geographically weighted regression to assess the effects on commuting demand. Briz-Redón (2022) proposed a Bayesian shared-effects modeling framework to quantify the global and local scale effects by analyzing the variation in covariate effects between different spatial scales and using posterior predictive checks to detect the contributing spatial units. Kosonen (2023) studied the scale effect by processing multi-scale migration and driving factor data, using linear regression to analyze the changes in variable correlations and explanatory power at different scales, and evaluating the impact of spatial autocorrelation with Moran's I test. Huang et al. (2024) studied the scale effect through kernel density analysis, spatial autocorrelation analysis, Pearson correlation analysis, and random forest model, exploring the relationship between urban vertical patterns and waterlogging at different scales.

For the scale effect on spatial interactions, the majority of studies focus on the distance decay effect, with the starting point being that spatial interaction is prone to attenuate with an increasing distance (Liu et al. 2014). Hence, researchers utilize the gravity model as a proxy to study the scale effect related to spatial interaction (Batty and Sikdar 1982; Ubøe 2004; Hagen-Zanker and Jin 2012; Arbia and Petrarca 2016; Stillwell et al. 2018). Apart from the distance decay mechanism, the structural pattern (in particular, the variation of the detected communities) of origin–destination (OD) matrices across scales was also studied (Zhang et al. 2018; Coscia et al. 2012). That is, how the clustering patterns of geographical units change from a spatial interaction

perspective. To summarize, existing analyses mostly rely on combinatorics and clustering analyses, which are limited to short-range interaction relationships and patterns of finite order.

1.2 | Preliminary

In this section, we provide a brief introduction to *topology* and *persistent homology*, as well as the necessary notations and concepts used throughout the rest of this paper.

In general, high-dimensional data can be represented in the form of a data cloud, that is, data points scattered around a manifold in a high-dimensional space (Chazal 2016). Data that are different in nature are assumed to be sampled from the manifolds that have disparate shapes that are invariant even with some deformations. Therefore, shape analysis is widely used in many domains to differentiate data distributions due to its great performance (Lum et al. 2013).

Similarly, if we use multiple scales to aggregate data, then each scale snapshot can form a point cloud, and the point clouds belonging to different scale snapshots may exhibit different shapes and diverse structures. We can then differentiate the scale snapshots based on their shapes and study the evolution of different structures with scale.

Topology is the branch of mathematics that analyzes and extracts insights from shapes (Ghrist 2008). This field probes topological invariants (e.g., connectedness and voidness) of shapes that are preserved under the coordinate variation, continuous deformation, and compression (Carlsson 2009). While traditional methods provide only partial information about the shape of the data (e.g., the connectedness provided by clustering methods), topology can provide a more holistic characterization of the data. Since data with disparate shapes are different in terms of connectedness and voidness, we can quantify the difference between data sets based on their topological characterizations.

1.3 | Homology

An algebraic topology method called *homology* plays a powerful role in providing information about invariants that characterize and describe shapes (Fugacci et al. 2016). In addition to the connected components (i.e., clusters), homology can extract other information about the given data. For example, describing more delicate structures, such as loops and voids that are invisible to other methods.

The first step in applying homology is constructing a meaningful abstraction to approximate the data structure and using the original data as the point set. Given a distance function d and a threshold $\epsilon \geq 0$, pairs of data points x_i, x_j are considered to be close if $d(x_i, x_j)$ is less than ϵ. Then, a widely adopted abstraction, the Vietoris-Rips (VR) complex (Fugacci et al. 2016), is defined as:

$$VR_\epsilon = \{\sigma = (x_0, \ldots, x_k) \mid d(x_i, x_j) < \epsilon, 0 \leq i, j \leq k\}, \quad (1)$$

where σ is a k-simplex composed of $k+1$ affinely independent points, and every point pair has a $d(x_i, x_j)$ at most ϵ. Two close points are connected to form a 1-simplex, and a k-simplex is generated once all its (k-1)-simplices are connected. An illustration of k-simplices and the Vietoris-Rips complex is shown in Figure 1. The VR complex can be utilized to approximate the shape of a scale snapshot, while the simplexes are the basic building blocks of the complex.

Given the VR complex, we can leverage homology to analyze the connected components and holes, thus deriving information about the shape of a dataset. The homology of different dimensionality derives information regarding the topological invariants (holes) of the corresponding dimensionality. Specifically, 0-dimensional homology consists of 0-dimensional holes (connected components), 1-dimensional homology consists of loops (non-bounding 1-cycles), and 2-dimensional homology corresponds to 2-dimensional holes (voids). For simplicity, we call them the H_0, H_1, and H_2 features, respectively. We can differentiate the data sampled from different shapes with the information on these features. As an example shown in Figure 2, we use the Betti number (Milnor 1964), a tool that counts the number of topological invariants of dimensions k, to demonstrate the discriminative power of homology. β_0, β_1, and β_2 count the number of H_0, H_1, H_2 features, respectively. All shapes have one connected component; therefore, they share $\beta_0 = 1$. However, they can be separated by the combination of β_1 and β_2, which counts H_1 and H_2 features.

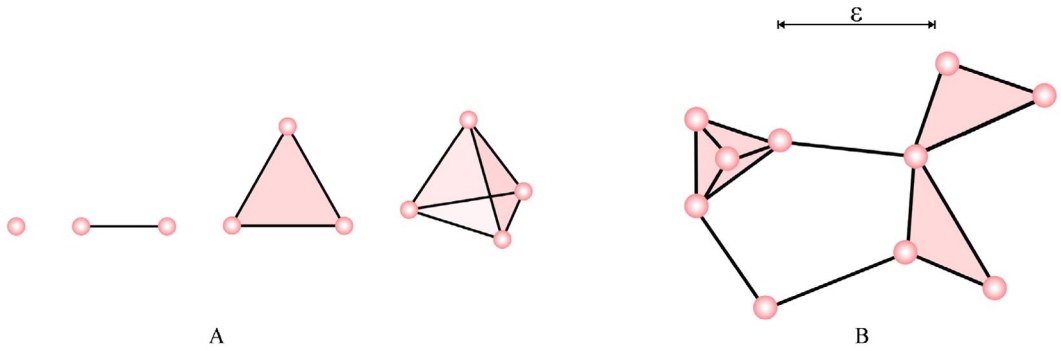

FIGURE 1 | Illustration of k-simplices and the Vietoris-Rips complex. (A) 0-, 1-, 2-, and 3-simplices (from left to right). An isolated point is a 0-simplex, a line connecting two points is a 1-simplex, a filled triangle is a 2-simplex, and so forth. (B) An example of the Vietoris-Rips complex.

1.4 | Persistent Homology

However, the Vietoris-Rips complex is determined by the value of ϵ. Given a distance function d, two units i, j are only connected when $d(i,j)$ is smaller than ϵ. How we define the proximity between spatial units determines the distance function we use and how we calculate it. A small value of ϵ makes the complex sparsely connected and scattered, thus providing little additional information than the original point cloud. In contrast, a sufficiently large value of ϵ ends in a fully connected complex, revealing no information about the proximity between spatial units. Different choices of ϵ generate complexes with different topological structures.

Therefore, instead of finding an optimal threshold value and studying a single complex to gain limited knowledge of the data, we compute the *persistent homology* related to these scale-dependent point clouds; that is, for each point cloud, we compute a sequence of Vietoris-Rips complexes by adapting a filtration process. The filtration is a finite sequence of complexes that starts from the empty complex:

$$0/ = VR_0 \subseteq VR_1 \subseteq VR_2 \subseteq \ \ldots \ \subseteq VR_n = VR, \tag{2}$$

where VR_i is the complex corresponding to ϵ_i whose nodes (i.e., spatial units) are connected only when the pair distance between the two nodes is below ϵ_i. Note that we adopt a series of ascending threshold values, so for $j < k$, $\epsilon_j < \epsilon_k$.

The evolution of filtration sheds light on the persistence of topological features. A topological feature born at the threshold value ϵ_i and dying at ϵ_j has persistence $\epsilon_j - \epsilon_i$. Persistence reveals the significance of the topological features. Features that persist across an extensive range of ϵ values are generally considered to be intrinsic data features.

As a toy example shown in Figure 3, we sparsely sample the data from a ring shape with a perturbation. We can see a large loop (H_1 feature, the uncolored part inside connected points) along with two small loops when $\epsilon = 0.24$ and the former continues for a large range while the latter ones disappear at $\epsilon = 0.36$. Although multiple loops (H_1 features) are detected

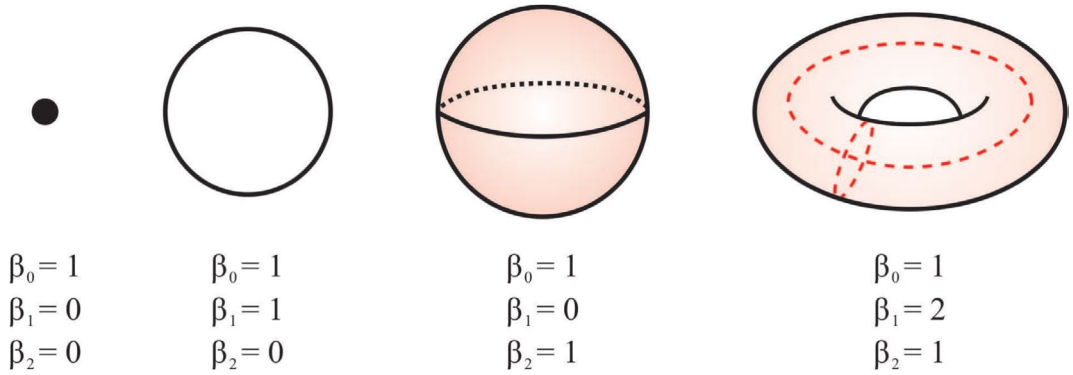

FIGURE 2 | Shapes and their Betti numbers. From left to right, a point, a loop, a hollow sphere, and a torus.

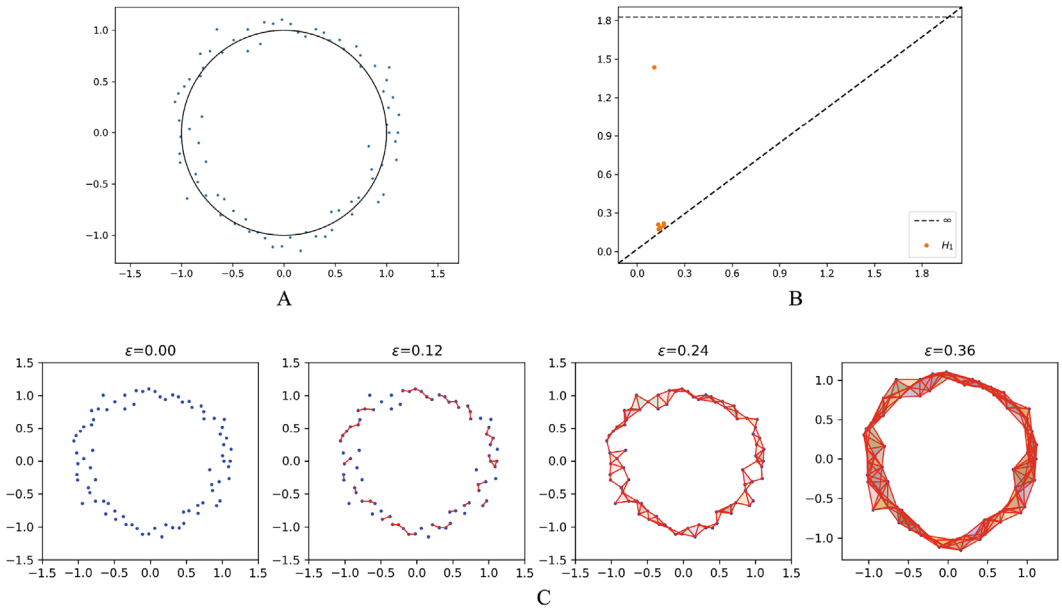

FIGURE 3 | Persistent homology was applied to data points sampled from a ring with perturbation. (A) The data cloud and the sampled ring; (B) A persistence diagram of the H_1 features, with each orange dot representing a loop feature that appears during the filtration process and the deviation from the diagonal line indicating the persistence of the feature. (C) A part of the filtration process.

during filtration, only one loop representing the ring lasts for a long time.

1.5 | Persistence Diagram and Wasserstein Distance

The persistence diagram (PD) (Kerber et al. 2017) can be used to summarize and visualize the persistence information of topological features that appear in the filtration process. PD has two-dimensional coordinates, with the x-coordinate representing the birth time and the y-coordinate representing the death time, as shown in Figure 3B. Topological features are represented by points distributed above the diagonal line because they "die after birth chronologically". Points that are distant from the diagonal persist longer and are considered to be intrinsic features of the data. To compare PDs, we can use the Wasserstein distance, a metric that measures the difference between distributions by solving Kantorovich's optimal transport problem[1]. It is mathematically proven to be stable (Cohen-Steiner et al. 2010), which means that a slight variation in the data will not result in an enormous difference in terms of the Wasserstein distance. However, in other methods, for example, mutual information-based Kullback–Leibler (KL) divergence, the divergence can be large even when the underlying data samples differ only slightly (Ozair et al. 2019).

The Wasserstein distance between two persistence diagrams $PD_a^{H_n}$ and $PD_b^{H_n}$ is defined as:

$$WD\left(PD_a^{H_n}, PD_b^{H_n}\right) = \left(\inf_{\eta: PD_a^{H_n} \to PD_b^{H_n}} \sum_{x \in PD_a^{H_n}} \| x - \eta(x) \|_\infty^p\right)^{\frac{1}{p}}, \quad (3)$$

where $\eta: PD_a^{H_n} \to PD_b^{H_n}$ ranges all bijections from $PD_a^{H_n}$ to $PD_b^{H_n}$ to match the points in $PD_a^{H_n}$ and $PD_b^{H_n}$, and p is commonly set to 2. This metric seeks a perfect match between points belonging to different PDs, while extra points are matched to artificial diagonal points if the cardinalities of the two PDs are inconsistent. Take the ring shape in Figure 3 as an example. If we densely sample the points from the ring without perturbation, we will derive a persistent diagram of H_1 with only one point distant from the diagonal line. When calculating the Wasserstein distance between PDs in Figure 3B and the new PD, the two points that are both away from the diagonal are matched. Since the new PD has no other points, those near the diagonal in Figure 3B are all matched with artificial points on the diagonal line. The final Wasserstein distance in this case will be small because the intrinsic features are matched, and the distance between short-lived parts and artificial points is negligible. For a more detailed description of the Wasserstein distance, please refer to Kerber et al. (2017).

Another commonly used metric for comparing persistence diagrams is the bottleneck distance, which also satisfies a stability theorem (Cohen-Steiner et al. 2005), but it focuses on the largest single feature difference. Since our goal is to capture overall topological variation rather than just the largest individual difference, the Wasserstein distance is more appropriate for our analysis.

2 | Methods

The main objective of this paper is to quantify the spatial scale effect from a topological perspective. The motivation that we choose topology as an analytic tool comes from the idea that data can be characterized by their intrinsic topological invariants.

The analytical pipeline of our topology-based framework is shown in Figure 4. First, we model the data aggregated under different scales as scale snapshots. Then, persistent homology is applied to extract the topological invariants in scale snapshots at multiple levels, and each scale snapshot is described by a persistence diagram that summarizes information on its topological invariants. Finally, we can calculate the scale snapshot topology distance between the scale snapshots by computing the difference between their persistence diagrams. The details are elucidated below.

2.1 | Data Preprocessing

This part introduces how to preprocess fine-grained data to be applied in the topological data analysis and investigation of the scale effect, which is shown in Figure 4A.

First, we adopt a series of ascending scales to partition the study area into non-overlapping spatial units. Then, we build scale snapshots to represent data aggregated at different scales in the matrix form. Specifically, the snapshot of scale s is of size $N_s \times N_s$, and the origin and destination of original origin–destination (OD) flows are redesignated to the spatial units they intersect according to their geographic coordinates. After this step, each origin and destination of an individual trip is assigned to the corresponding spatial unit, and we are no longer concerned with their precise geographic coordinates. With the rows as originating units and columns as destination units, each entry in the matrix refers to the absolute value of the frequency of interactions between the corresponding units; in the upscaling process (fine-to-coarse resolution aggregation), short-range OD flows may be converted to intra-unit interactions represented by diagonal entries. Finally, each row in the scale snapshot can serve as a vectorized representation of the corresponding spatial unit.

2.2 | Topological Representation of the Scale Snapshots

Now that we have the scale snapshots, which are the products of spatial aggregations under different scales. Then, each spatial unit can be represented by a point in a high-dimensional space, and the corresponding row vector in the snapshot matrix determines its coordinate. The distance between two vectors thus defines the distance between corresponding points. Each scale's spatial units form a point cloud (i.e., a point set with shape). Point clouds belonging to different scale snapshots can demonstrate diverse shapes and exhibit versatile topological structures. Therefore, it is feasible to study the spatial scale effect by monitoring the changes in the topological structure of snapshots.

After that, we apply persistent homology to each scale snapshot. During the filtration, we can extract both 0- and 1-dimensional

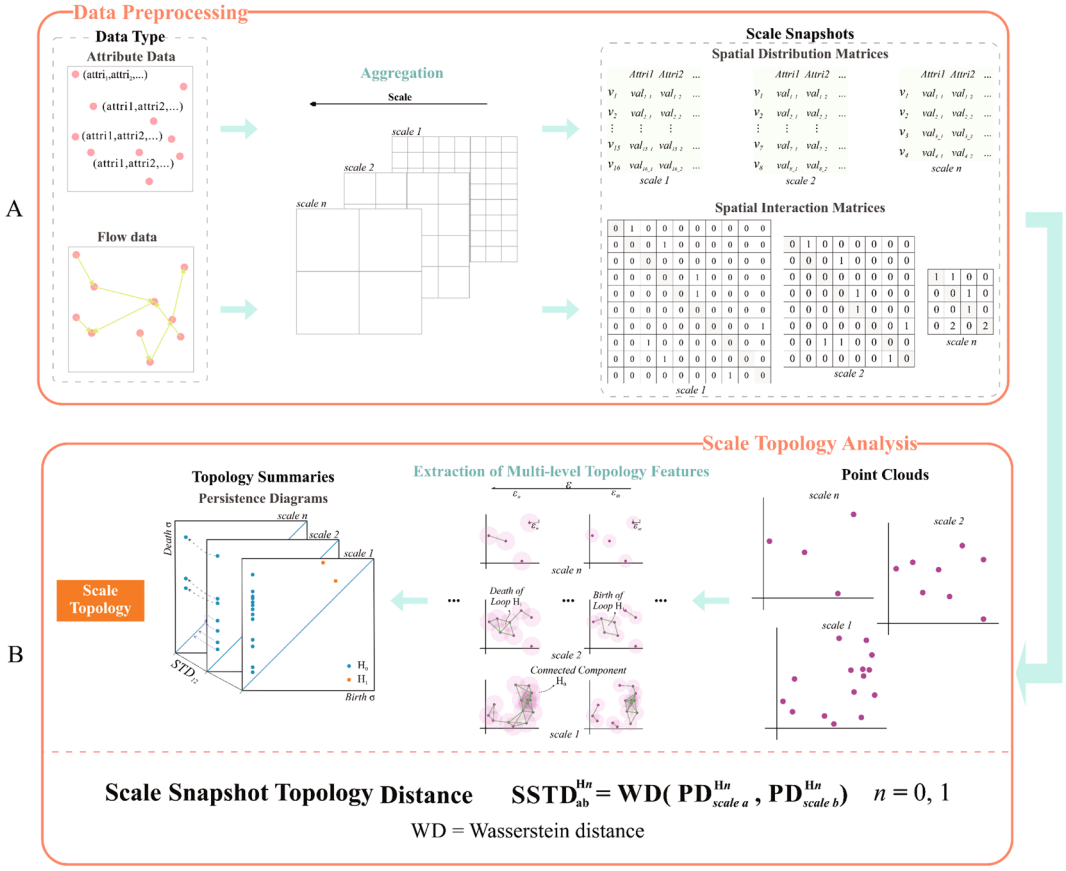

FIGURE 4 | The proposed framework for studying the spatial scale effect using topology.

features (H_0 and H_1 features, respectively) for each complex. H_0 features are clusters defined by their similarity, that is, the similarity in the interplay pattern for spatial interactions in our scenario. H_1 features, however, demonstrate more subtle connections. The presence of a loop structure indicates that the spatial units are not mutually similar, but instead exhibit similarity in a gradual or gradient-like manner. We later use the two topological features, H_0 and H_1, as proxies to capture changes in the scale snapshot.

Apart from information about the numbers and the existence of topological features, the PD also informs us of their significance. The topology of the snapshot changes as the scale evolves, as do the corresponding PDs. Therefore, the persistence diagram acts as a suitable topological representation for each scale snapshot. In order to track the variation of different patterns in the scale snapshots with scale, this study summarizes H_0 and H_1 separately, so that each scale snapshot has PD^{H_0} and PD^{H_1} as its representations.

2.3 | Scale Snapshot Topology Distance

To quantify the disparity in topological features between scale snapshots, we define the Scale Snapshot Topology Distance (SSTD) as the Wasserstein distance between the persistence diagrams of two snapshots. Since each scale snapshot has two topological representations that summarize the information of H_0 and H_1, respectively. Therefore, we compare the differences of H_0 and H_1 between the scale snapshots separately, as calculated in Equation 4:

$$\text{Given scale b and a, } \mathrm{SSTD}_{H_n}(a,b) = WD\left(\mathrm{PD}_a^{H_n}, \mathrm{PD}_b^{H_n}\right), n = 0, 1 \tag{4}$$

By definition, if a = b, then $\mathrm{SSTD}_{H_n}(a, b) = 0$. When topological features evolve stably across scales, the distance between consecutive PDs is small. Instead, if certain topological features of scale snapshots experience an abrupt transition, then the PD related to them will change abnormally. We can use such topological changes to understand the spatial scale effect on different aggregations. In the following, we introduce a case study of spatial interaction aggregations to verify the effectiveness of the proposed methodology.

3 | Case Study

3.1 | Data and Study Areas

With the proliferation of sensors and the development of location-awareness technologies, massive individual-level data are available for analyzing spatial interactions in both physical space and cyberspace (Yu and Shaw 2008; Gao et al. 2013; Li et al. 2021). In general, the selection of an analytic scale lacks standardized criteria and references when aggregating individual-level data to predefined spatial units (Liu et al. 2018). Therefore, the scale

effect needs to be explored via cross-scale quantitative empirical studies. Our study areas include five cities in China, namely, Beijing, Shanghai, Shenzhen, Wuhan, and Jiaxing, to examine the scale effect on spatial interaction patterns derived from taxi trajectories in these cities.

We collect the datasets of anonymous taxi GPS trajectories from major taxi companies in the five cities. The study area of each city is the area covered by recorded taxi trips, as shown in Figure 5. We delete dubious GPS records and organize valid GPS records as origin–destination (OD) trip pairs based on the pick-up and drop-off information (Liu, Wang, et al. 2012). A summary statistic of those records is shown in Table 1. Among the five cities, Beijing has the largest number of valid trip records, Shanghai has the longest time span of data, and Wuhan has the broadest coverage of records.

To better understand the datasets, we compute the trip-length distributions of the OD trip pairs in each city. From Figure 6, we can conclude that the trip-length distribution of Beijing is distinct from that of the other four cities. Fifty percent of the trips in Shanghai, Wuhan, and Shenzhen are shorter than 5 km, with even more than 80% for the city of Jiaxing. However, only 10% of the trips in Beijing are shorter than 5 km. Overall, most of the trips are limited to 40 km. The trip-length distribution informs us of an approximate upper bound of the size of the spatial units used to study the spatial interactions. Taking Shanghai as an example, if the unit size is greater than 5 km, more than half of the spatial interactions become intra-unit interactions, which may limit the discovery of fine-grained patterns.

3.2 | Configuration

Castro et al. (2013) proposed an overview of mechanisms for using taxi GPS data to analyze social and community dynamics and claimed it is more practical to decompose the city into separate areas and work with this decomposition. Liu, Kang, et al. (2012) explored intra-urban human mobility and land use variations based on taxi trajectory data, discretizing the study area into 1000 m × 1000 m cells. Liu et al. (2016) used grids sized 500 m × 500 km to divide the study area and used taxi data to classify and understand urban land use. Wei et al. (2020) identified and measured urban functional polycentricity using taxi GPS data aggregated across various scales, including grids sized 3000 m × 3000 m and 5000 m × 5000 m.

Following the literature review on OD trip studies, we map the pick-up and drop-off points onto non-overlapping regular grids (Liu, Wang, et al. 2012; Liu et al. 2015), and we adopt aggregation scales ranging from 250 m to 3000 m, with

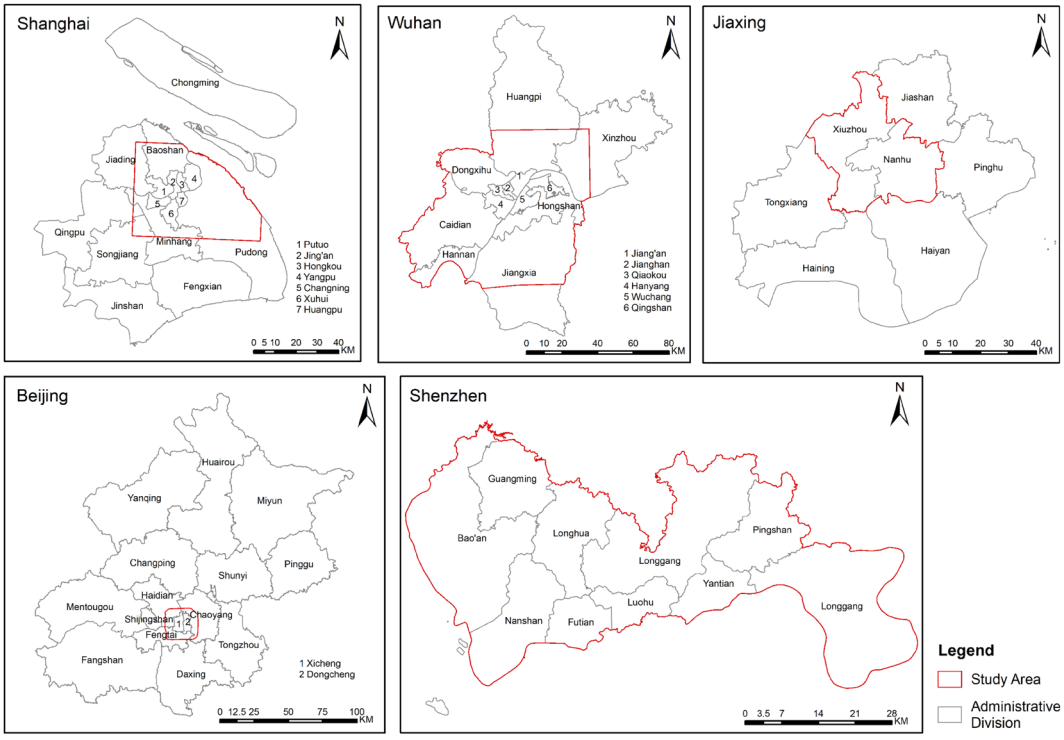

FIGURE 5 | Study area of Shanghai, Wuhan, Jiaxing, Beijing, and Shenzhen.

TABLE 1 | Details of the taxi trajectory records.

	Beijing	Shanghai	Shenzhen	Wuhan	Jiaxing
Time span (days)	25	30	6	10	7
Number of records	19,982,111	5,922,999	932,947	2,361,188	276,764
Coverage (km^2)	515	2184	2510	7060	1337

increments of 250 m. The size of grids at scale s is $s \times sm^2$. Note that the proposed pipeline can be extended to other spatial settings as well. The resulting numbers of spatial units after aggregation are presented in Table 2. Trips in Wuhan are more spatially dispersed than those of Beijing and Shenzhen, leading to more spatial units at coarser scales and fewer at the 250 m scale.

4 | Results

In this part, we present the topological variation analysis brought by the scale effect with the proposed scale snapshot topology distance.

As mentioned above, every scale snapshot (OD matrix) can form a point cloud; that is, 12 scale snapshots for each city. In each OD matrix, each row that stores the frequency of spatial interactions with all other units acts as the feature vector of the corresponding spatial unit. The feature vectors are the coordinates of the spatial units in high-dimensional space. We assume that the similarity of spatial units in their spatial interaction behavior determines the proximity, so we choose the Euclidean distance between their feature vectors to derive the proximity between all spatial unit pairs. The larger the Euclidean distance, the smaller the proximity.

We then apply the persistent homology to the point clouds to obtain persistence diagrams (PDs) on the chosen distance

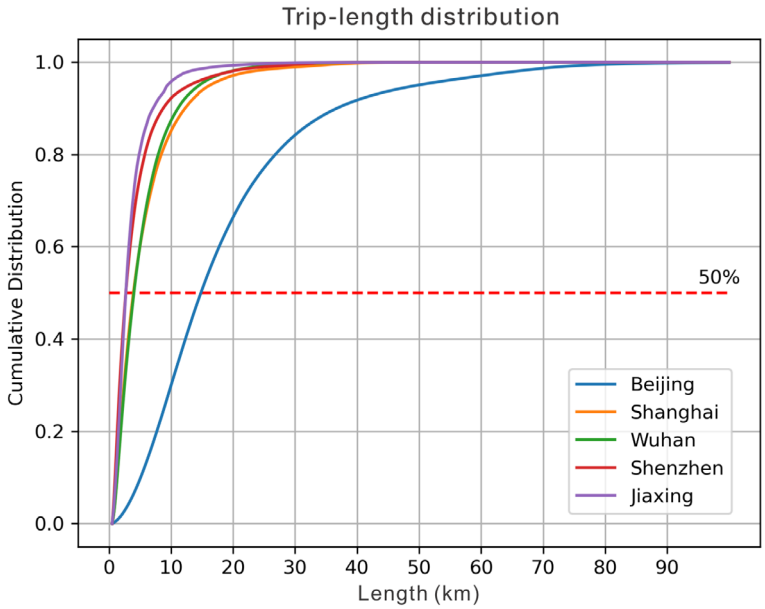

FIGURE 6 | The cumulative distributions of OD trip-length in five cities.

TABLE 2 | Number of spatial units w.r.t different scales.

Resolution	Beijing	Shanghai	Shenzhen	Wuhan	Jiaxing
250 m × 250 m	8377	6087	11,905	7175	2246
500 m × 500 m	2147	3396	4391	4710	1458
750 m × 750 m	982	2366	2346	3522	1105
1000 m × 1000 m	553	1525	1491	2875	910
1250 m × 1250 m	372	1008	1036	2091	662
1500 m × 1500 m	259	717	778	1619	504
1750 m × 1750 m	195	528	606	1315	396
2000 m × 2000 m	152	407	489	1096	326
2250 m × 2250 m	122	326	404	893	268
2500 m × 2500 m	99	271	339	778	224
2750 m × 2750 m	86	219	291	680	185
3000 m × 3000 m	76	190	255	608	164

function. The persistent homology computations are performed using an efficient Python package called Ripser (Tralie et al. 2018). We then calculate the $SSTD_{H_0}$ and $SSTD_{H_1}$ between every pair of scales based on PDs. To provide further visual context for the behavior of the SSTD measures across the five cities, we have included persistence diagrams and 2D projections of point clouds in the Supporting Information (see Appendix Figures S1–S5).

4.1 | Variation in the Clustering Pattern

Clustering is widely used to analyze data, and the way it is implemented is often ambiguous due to the various choices of thresholds and schemes (e.g., single linkage, complete linkage) and a lack of robustness (Carlsson 2009). Persistent homology addresses this problem by summarizing the behavior of clustering under a varying threshold. Therefore, in this part, we explore the variation in the clustering pattern of data clouds with $SSTD_{H_0}$.

The $SSTD_{H_0}$ analysis result is shown in Figure 7, where each line corresponds to a distinct fixed scale s and shows its $SSTD_{H_0}(s, x)$ with variable scales x (ranging 250–3000) with scale x on the horizontal axis. It can be seen that the line representing 250 monotonically increases in all five cities (except for Jiaxing's decline at 3000), while the line representing 3000 monotonically decreases (except for Jiaxing's rise at 2750). Lines representing other scales show the V-like shape. These results all indicate that the variation of H_0 features intuitively increases with increments in scale difference. For instance, $SSTD_{H_0}(250, 1250)$ (with a scale difference of 1000) is larger than $SSTD_{H_0}(250, 500)$ (with a scale difference of 250). It is worth noting that the more data we have, the smoother the observed H_0 variation brought by the scale effect is.

The results of $SSTD_{H_0}$ show a trend of an increasing distance with an increasing scale difference for all five cities. We quantify the relevance of the scale difference (with the starting scale 250 as the base) and the differences in H_0 features with quadratic polynomial fitting; the fitting results are shown in Figure 8. Each black square represents $SSTD_{H_0}$ between the scale determined by its x coordinate value and the scale 250; thus, the y coordinate value of the first square on each subfigure is 0. As we can see, $SSTD_{H_0}$ is polynomially related to the scale difference (with scale 250 as the base), and the fitting results of all five cities have high adjusted R-squared values.

4.1.1 | Variation in the Loop Pattern

In addition to clustering patterns, persistent homology can also provide information about higher-dimensional features, such as the H_1 features that reflect the data's loop pattern. Many applications have demonstrated the importance of loops and higher-dimensional features (Taylor et al. 2015; Stolz et al. 2017; Sizemore et al. 2018). A previous study of Wubie et al. (2018) found that H_1 appears to be sensitive to detecting small clusters that other clustering methods are incapable of; H_1 is also effective in detecting extreme patterns in data. Hence, we also leverage the $SSTD_{H_1}$ to measure the change in high-dimensional patterns across scales, aimed at exploring how this type of pattern evolves under the scale effect.

Different from the case of $SSTD_{H_0}$, the results of $SSTD_{H_1}$ are less regular. The variation from $SSTD_{H_1}(s, 250)$ to $SSTD_{H_1}(s, 3000)$ for all scales is cluttered together, so we only visualize the variation from $SSTD_{H_1}(250, 250)$ to $SSTD_{H_1}(250, 3000)$ in Figure 9. Although $SSTD_{H_1}$ starts with an increasing tendency, there are more deviations than $SSTD_{H_0}$ in larger spatial scales.

Intuitively, the topological distance becomes larger as the scale difference increases. However, the evolution of loop patterns (subtle structures in the data) exhibits unstable fluctuations, including an unexpected drop following a previously increasing trend. The critical scale at which this decline occurs varies across cities. Beijing has the most drastic drop at the scale of 1750 and a rebound at the scale of 2000. Shanghai and Wuhan

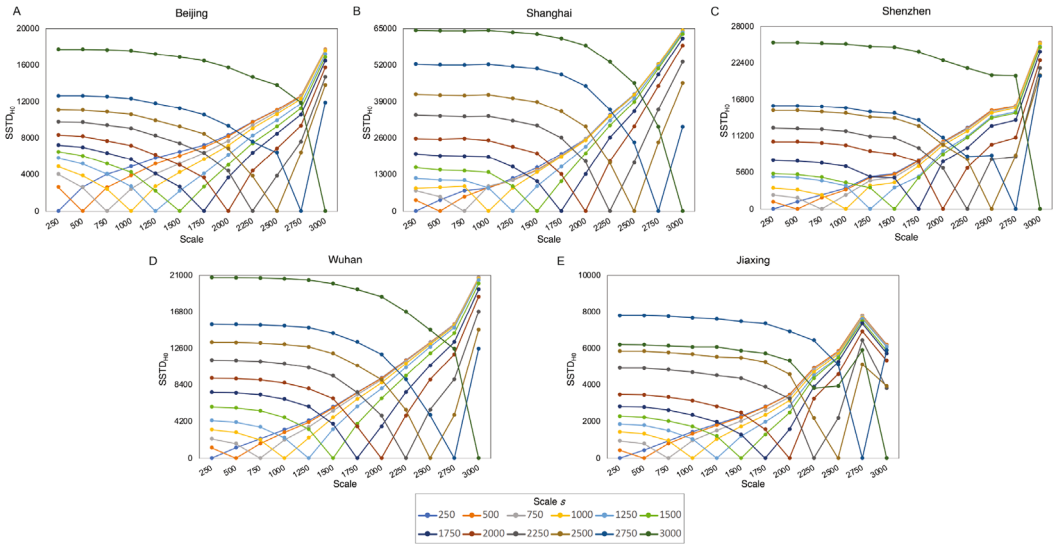

FIGURE 7 | The scale snapshot topology distance (SSTD) results w.r.t to H_0. The points in each subfigure represent the pairwise $SSTD_{H0}$ values between two spatial scales.

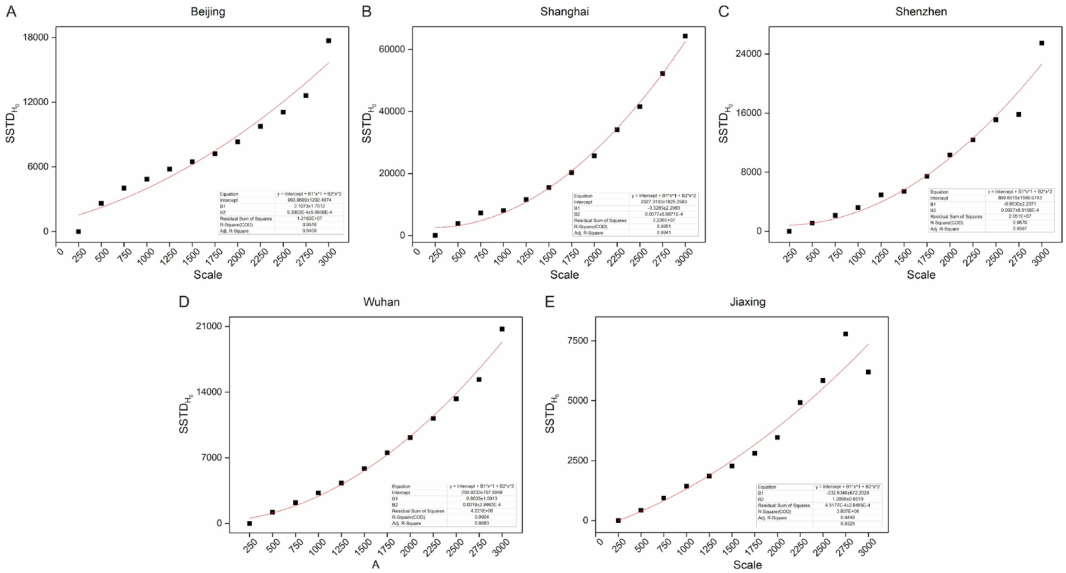

FIGURE 8 | Polynomial fitting of scale and $SSTD_{H_0}$ with scale 250 as the base.

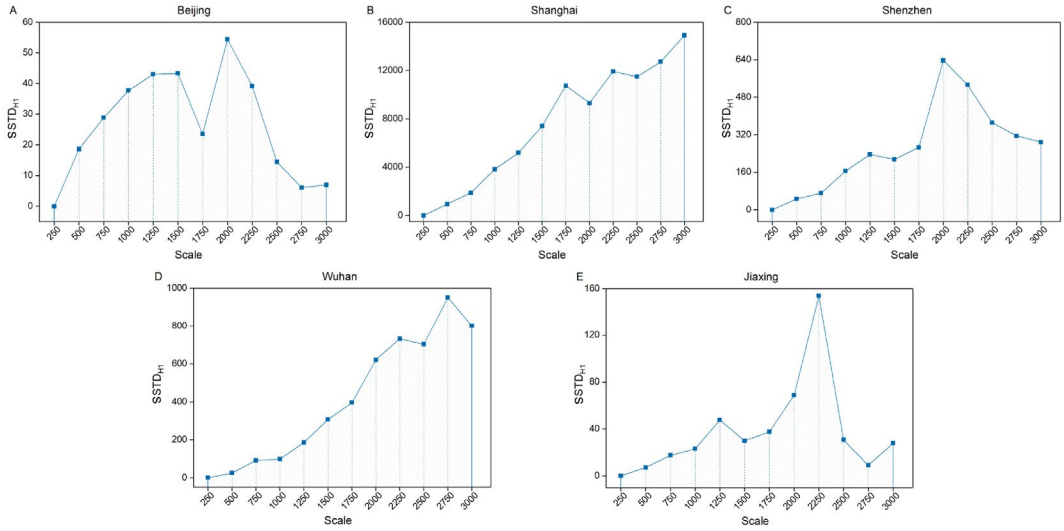

FIGURE 9 | $SSTD_{H_1}$ fluctuation with scale 250 as the base.

show similar behaviors, with a small rebound after a minor decline (2000 for Shanghai and 2500 for Wuhan). Shenzhen and Jiaxing both have a slight downhill (at the scale of 1500), followed by a drastic rise.

4.1.2 | Explanatory Analysis

To explain what happened when the $SSTD_{H_1}$ declines, we conduct an explanatory analysis. In this section, we want to examine how the clusters obtained using a conventional method change with scale, especially what happens at the critical spatial scales.

For our clustering analysis, we employed the K-means algorithm as implemented in the scikit-learn package (Pedregosa et al. 2011). To optimize the stability and reproducibility of the results, we adopted the k-means++ initialization method (Arthur and Vassilvitskii 2006) and performed 100 runs of independent initializations. This approach mitigates the sensitivity of K-means to initial centroid placement and enhances the robustness of the derived clusters. Given a scale, we feed its snapshot to the K-means algorithm to derive clusters. That is, we take the frequency of spatial interaction between a spatial unit and all other spatial units as its feature and then cluster the spatial units according to the spatial interaction pattern. If two spatial units interact with other units in a similar way, then these two spatial units will be grouped into the same category. The clusters extracted for each scale are projected back to the geographic map, and spatial units corresponding to different clusters are colored differently. The number of clusters is determined by the elbow method (Kodinariya and Makwana 2013).

Figure 10 shows the evolution of the clusters at different spatial scales, namely, from 250 to 3000 m. There are some empty grids in the maps, which are due to the lack of travel records. This phenomenon is more common at finer resolution, because small grids do not easily intersect with roads. We can

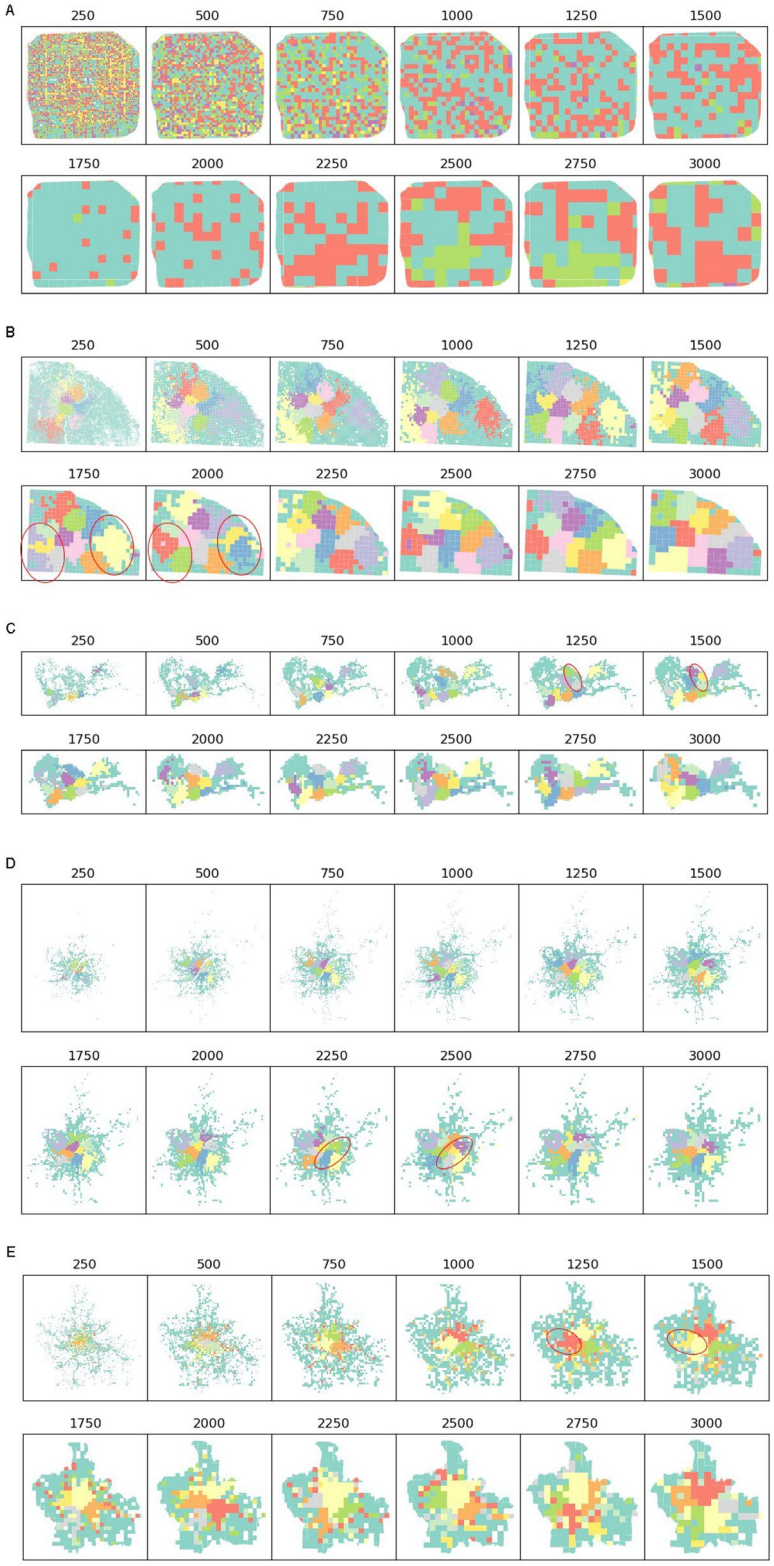

FIGURE 10 | The evolution of the clusters at different spatial scales for Beijing (A), Shanghai (B), Shenzhen (C), Wuhan (D), and Jiaxing (E).

observe that snapshots of the first few scales show a coherent clustering result. However, when the scale increases to the critical scale identified by SSTD $_{H_1}$, notable changes emerge in the clustering structure through significant reorganization or the subdivision of previously stable clusters. Ideally, as the spatial scale increases, clustering should evolve in an inclusive manner, such as the merging of multiple clusters or the gradual expansion of cluster boundaries, rather than the opposite way we observe.

We have marked in red the positions corresponding to the differences before and at the critical spatial scales. Beijing undergoes a dramatic change at the critical scale of 1750 m, so no particular location is circled. Moreover, any changes in the cluster results

after that scale could not be reasoned based on the previous results. In Shanghai, from the scale of 1750 to 2000 m, the clusters within the left marked region are reorganized, and the cluster within the right frame is split into two parts. Both Shenzhen and Wuhan have a cluster that splits into two at the critical scales (1500 for Shenzhen and 2500 for Wuhan), and the clusters immediately merge back into one after the critical scales. Likewise, Jiaxing shows the clusters splitting into two at the critical scale of 1500, only for that partition to merge back into one at the scale of 2000.

The fluctuations that break the consistency of H_1 evolution can potentially provide us with a reference for the choice of an appropriate scale for spatial analysis. These critical scales are indicative of when a disturbance occurs in the evolution of patterns.

4.2 | Results Using Traditional Methods

In this section, we analyze the scale effects using traditional analysis methods using gravity models and network statistics (Arbia and Petrarca 2016; Zhang et al. 2018; Coscia et al. 2012) and compare the findings with our proposed method, which shows the effectiveness of our method for detecting critical scales.

4.2.1 | Distance Decay Pattern

The gravity model is broadly used in the study of spatial interactions, and is named as such because it draws its main ideas from Newton's law of gravity (Haynes and Fotheringham 1984). The gravity model assumes that the spatial interaction between two locations is proportional to their mass and is impeded by geographic distance. The elementary formulation of the gravity model is shown below, where M_i and M_j represent the masses of the origin and the destination, respectively, d_{ij} is the spatial distance between two places, and β indicates to what extent the spatial interaction is harnessed by the distance decay effect (Kang et al. 2012).

$$G_{ij} = \alpha \frac{M_i M_j}{d_{ij}^{\beta}}$$

Investigating decay parameter β can help us explore the underlying distance decay mechanism of spatial interactions across scales. In this analysis, we utilize the PageRank (Page et al. 1999) score as a proxy for mass. PageRank is an algorithm originally used to rank web pages in Google Search, which measures the importance of a web page by considering its incoming links. The algorithm is suitable for any set of entities with reciprocal relationships (Thakkar et al. 2010), which makes the PageRank score a perfect candidate for representing the "mass". The fitting results of the distance-decay coefficient β at different scales are plotted in Figure 11. Among the cities analyzed, Shanghai consistently exhibits the highest β values across scales, followed by Wuhan, Shenzhen, and Jiaxing. Beijing has the lowest β values. The disparate β results suggest that the distance decay mechanism for spatial

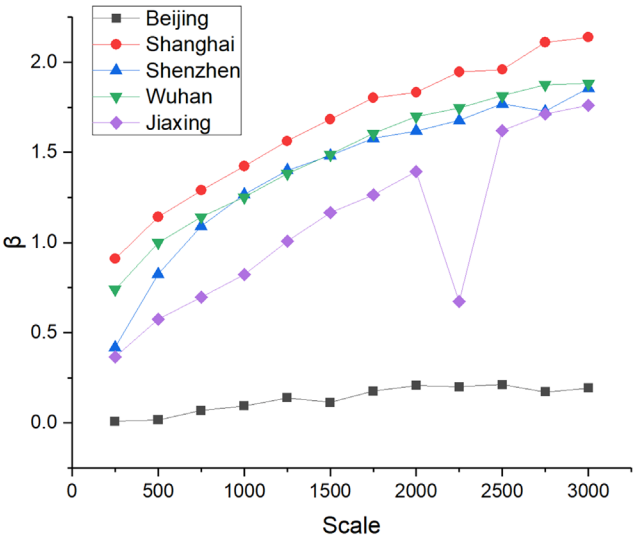

FIGURE 11 | The distributions of the distance decay coefficient *beta* across scales in different cities.

interaction is not uniform across cities, likely due to differences in urban size, structure, and travel behavior (Kang et al. 2012). Additionally, Shanghai, Shenzhen, and Wuhan display a steady upward trend in β with increasing spatial scale, suggesting that distance friction intensifies as spatial scale becomes coarser. Jiaxing deviates from this pattern, showing a sharp decline around the 2250 m scale, possibly due to the short observation period or limited sample size. Beijing has much lower β values than other cities, probably because its original distance distribution is much different from other cities, and its long trips are more frequent.

4.2.2 | Network Statistics

After we construct the OD matrix series as scale snapshots by varying the grid size, we can apply traditional network statistics to learn some basic features of the network at different scales, as depicted in Figure 12. This analysis was implemented using the Python package NetworkX (Hagberg et al. 2008). For most metrics, we calculated them directly on the original directed, weighted snapshots. For the clustering coefficient, average shortest path length, and small-world property, we first extracted the largest strongly connected component, converted it to an undirected graph, and then performed the computations.

Not surprisingly, the number of nodes and edges decreases with the increment in scales, but the number of edges drops more drastically than that of the nodes. For the statistics of the average weight, average degree, and average shortest path (ASP) in the maximum connected component, we can see that the average shortest path length and the average degree decrease with the drop in the number of nodes and edges. The last column describes the clustering coefficient and the small-worldness. Small-world networks have a high aggregation coefficient and a low average shortest path length; most of their nodes are not directly connected, but can be mutually accessed by a few edges. Small-worldness is evaluated by comparing the test network to a random network with the

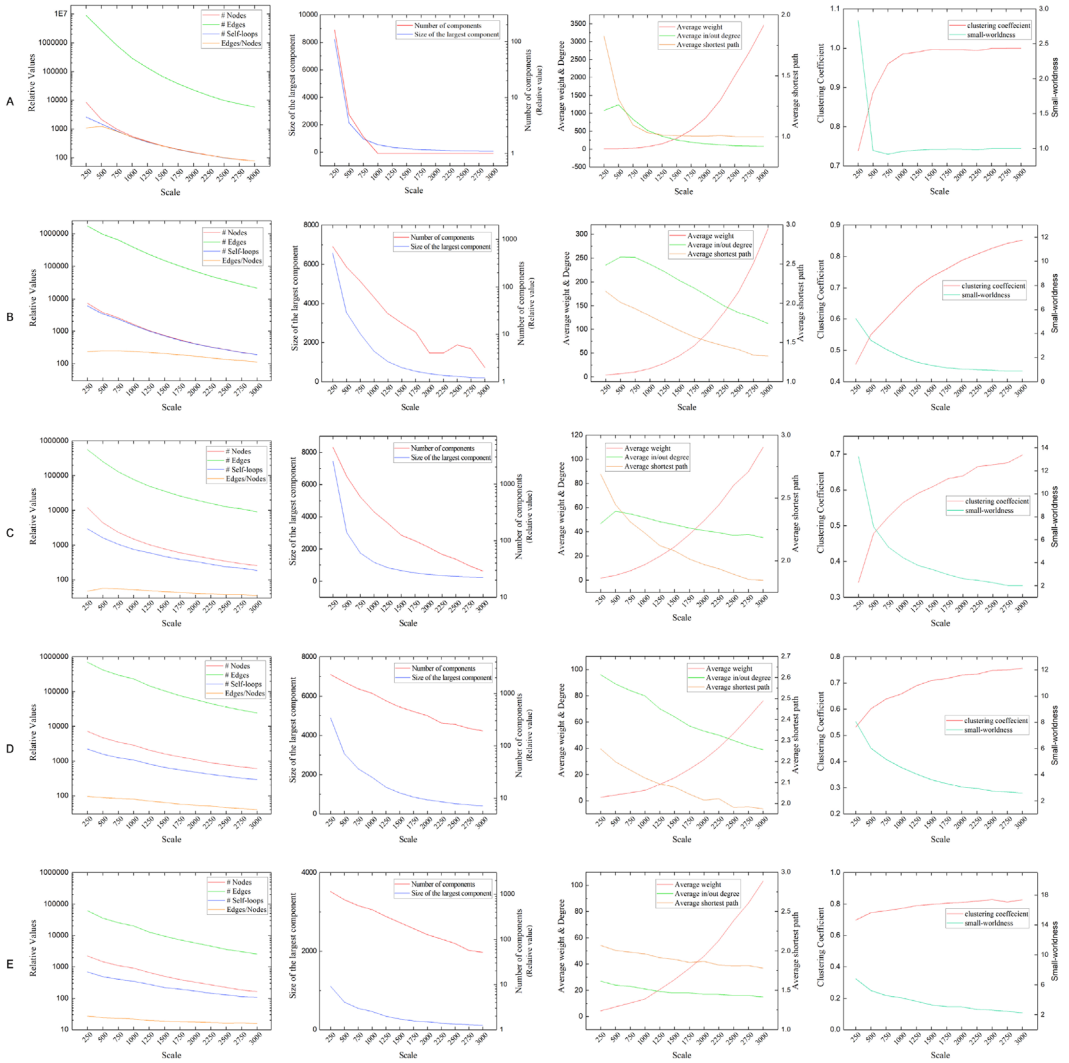

FIGURE 12 | Statistics of the OD matrices for Beijing (A), Shanghai (B), Shenzhen (C), Wuhan (D), and Jiaxing (E).

same number of nodes and edges. We can see that, although the clustering coefficient and the ASP length increase with an increasing scale, the small-worldness is suppressed, indicating that the high clustering coefficient and short ASP at large scales are the results of the small-size network. In summary, the OD network at large scales gradually loses some exclusive features of the complex network, which is in accordance with (Coscia et al. 2012).

While the results from network statistics and distance decay analysis exhibit relatively stable variation trends, these methods fall short in capturing abrupt transitions induced by the spatial scale effect, such as the critical scales identified by our approach, as shown in Figure 9.

5 | Discussion

5.1 | Guideline for Scale Selection

According to our experimental results, clustering patterns evolve consistently across spatial scales, and the scale effect on these patterns can be justified. However, the higher-dimensional H_1 pattern performs with less stability and more abnormalities with the scale variation. Concretely, SSTD $_{H_1}$ between scales smaller than the critical scale and the base scale presents a consistently increasing trend; Once the scale value reaches the critical scale, this consistency is violated and subsequently turned into chaos. The identified critical scale marks a fundamental shift in topological features compared to previous snapshots. While new consistency boundaries might emerge at coarser scales, we emphasize the significance of this first critical scale. This is because data aggregated at or before this critical scale retains greater fidelity to the finer-grained observations. Therefore, if researchers have to choose one single scale to aggregate data, scales preceding the critical threshold should be prioritized.

We propose a general framework that uses the topological signals inherent in the data to monitor changes in the data under scale effects. The critical scales detected by SSTD provide us with a reasonable range of scales for geographic studies. In a specific study, the researcher can safely choose scales within this range according to the research objective and the computational capability and should be more careful in selecting scales beyond this range.

5.2 | Flexibility of the Proposed Framework

In the case study, we investigated the scale effect on regular grid-based spatial interactions, following the study design of previous spatial interaction studies (Liu et al. 2015; Qi et al. 2011). However, the proposed framework is very flexible and compatible with other spatial aggregation configurations (i.e., irregular units); the only difference lies in the data preprocessing for generating scale snapshots. Without the loss of generality, our proposed framework can also be used to study the scale effects on spatial distributions. The scale snapshots generated from spatial distribution data can also be represented as point clouds, where the attributes of interest determine the coordinates of each observation. In this scenario, the snapshot of scale s is formed as $N_s \times M$, where N_s is the number of units, and M refers to the number of attributes. Each row of the scale snapshot is a statistical result of the original data points assigned to this unit. We can then apply the persistent homology and calculate the corresponding SSTD between different scale snapshots.

The workflow for applying our methodology to other research data is as follows: (1) Choose a series of increasing spatial scales $[s_1, s_2, \ldots, s_m]$. Note that the spatial unit is not restricted to a grid; it can be any spatial partitioning relevant to the research context. (2) Generate m scale snapshots under the spatial scales $[s_1, s_2, \ldots, s_m]$. (3) Apply persistent homology with a defined distance function and generate persistence diagrams for each snapshot (using Python packages such as Ripser (Tralie et al. 2018) and Dionysus (Morozov 2017)). If the computational conditions allow, persistence diagrams for features with higher dimensionality than H_1 can be derived for a more holistic characterization of the scale snapshot. (4) Compute SSTD $_{H_n}(s_1, s_i)$ for $i \in [2, m]$ w.r.t topological features with dimension n and plot the fluctuation as Figure 9. (5) Identify the critical scale by detecting the transition point in the fluctuation curves, which indicate significant changes in the underlying topological structure.

5.3 | Stability Brought by Using a Topological Perspective

The additional benefit of using persistent homology is that it provides assurances for extracting underlying patterns when facing uncertainties brought by noise or inadequate sampling (Carlsson 2009). Although tracking H_0 during filtration is similar to agglomerative hierarchical clustering, it is less sensitive to noise than the latter (Lee et al. 2012; Kim et al. 2015). Moreover, the stability theorem justifies the use of persistence diagrams (Cohensteiner et al. 2007). It claims that even given data perturbed by noise, the persistence diagram obtained from these data approximates the persistence diagram derived from the noise-free data. Remember that we sampled data from a ring shape with perturbations in the toy example, but we can still derive the point that represents the true ring (which is distant from the diagonal line). In contrast, uncertainties caused by noise or the results of inadequate sampling have short life spans and are distributed near the diagonal line, which makes them contribute little when calculating the Wasserstein distance.

6 | Conclusion and Future Work

The MAUP is of particular importance to quantitative geographical studies, of which the scale effect causes variations in results. When using individual-level data to perform spatial analysis, spatial aggregation is often used. Many researchers arbitrarily choose the size of the spatial unit and bypass the scale effect due to a lack of standard methods to guide their scale selection process. One solution to gain more knowledge of the scale effect is to quantify its influence on data.

In this paper, we propose the scale snapshot topology distance to quantify the scale effect from a topological perspective. We use scale snapshots to represent data aggregated under different scales and utilize the persistence diagram as a topological summary of each scale snapshot. The persistence diagram is obtained by introducing persistent homology, which has been proven to be an efficient method in various research areas. This technique encodes information on topological features of multiple dimensions, such as the H_0 feature (a connected component analogous to a cluster) and the H_1 feature (a loop represents a subtle structure in data). The informative and stable properties of persistence diagrams make them competent surrogates for measuring the difference between scale snapshots. Therefore, with mathematically well-defined metrics, we can eventually quantify the scale effect by computing the distance between persistence diagrams.

Topology opens a new venue for the study of the spatial scale effect. However, additional work is needed to better explain the disparities in the critical scale when the topology shift occurs for different cities. The relationships between the value of a critical scale and some characteristics of cities need to be further explored, which may shed light on the parameterization of critical scales. We will consider variables that describe urban morphology and land use. As fractality could also imply scale independence, the fractal dimension can be a potential variable. Furthermore, a key area of future research involves a deeper interpretation of topological features revealed through persistent homology, including the alignment of H_0 features (connected components) with traditional clustering methods and urban spatial partitions, as well as the meaning of H_1 features (loops) in spatial interaction patterns. Our future research will focus on these issues and also expand studies to a broader range of regions and different types of geospatial datasets.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Demo for scale snapshot topology distance at https://figshare.com, reference number https://doi.org/10.6084/m9.figshare.29066603.

Endnotes

[1] Named after Leonid Vaseršten, but first defined by Leonid Kantorovich; please refer to (Vershik 2013).

References

Amrhein, C., and H. Reynolds. 1996. "Using Spatial Statistics to Assess Aggregation Effects." *Journal of Geographical Systems* 2: 143.

Amrhein, C. G. 1995. "Searching for the Elusive Aggregation Effect: Evidence From Statistical Simulations." *Environment and Planning A* 27: 105–119.

Amrhein, C. G., and R. Flowerdew. 1992. "The Effect of Data Aggregation on a Poisson Regression Model of Canadian Migration." *Environment and Planning A* 24: 1381–1391.

Arbia, G., and F. Petrarca. 2016. "Effects of Scale in Spatial Interaction Models." In *Spatial Econometric Interaction Modelling*, 85–101. Springer.

Arthur, D., and S. Vassilvitskii. 2006. "k-Means++: The Advantages of Careful Seeding. Technical Report".

Atkinson, P., J. Zhang, and M. F. Goodchild. 2014. *Scale in Spatial Information and Analysis*. CRC Press.

Atkinson, P. M., and N. J. Tate. 2000. "Spatial Scale Problems and Geostatistical Solutions: A Review." *Professional Geographer* 52: 607–623.

Batty, M., and P. Sikdar. 1982. "Spatial Aggregation in Gravity Models: 3. Two-Dimensional Trip Distribution and Location Models." *Environment and Planning A* 14: 629–658.

Briz-Redón, Á. 2022. "A Bayesian Shared-Effects Modeling Framework to Quantify the Modifiable Areal Unit Problem." *Spatial Statistics* 51: 100689.

Carlsson, G. 2009. "Topology and Data." *Bulletin of the American Mathematical Society* 46: 255–308.

Castro, P. S., D. Zhang, C. Chen, S. Li, and G. Pan. 2013. "From Taxi Gps Traces to Social and Community Dynamics: A Survey." *ACM Computing Surveys* 46: 1–34.

Chazal, F. 2016. *High-Dimensional Topological Data Analysis*.

Clark, W. A., and K. L. Avery. 1976. "The Effects of Data Aggregation in Statistical Analysis." *Geographical Analysis* 8: 428–438.

Cohensteiner, D., H. Edelsbrunner, and J. Harer. 2007. "Stability of Persistence Diagrams." *Discrete & Computational Geometry* 37: 103–120.

Cohen-Steiner, D., H. Edelsbrunner, and J. Harer. 2005. "Stability of Persistence Diagrams." In *Proceedings of the Twenty-First Annual Symposium on Computational Geometry*, 263–271.

Cohen-Steiner, D., H. Edelsbrunner, J. Harer, and Y. Mileyko. 2010. "Lipschitz Functions Have l p-Stable Persistence." *Foundations of Computational Mathematics* 10: 127–139.

Coscia, M., S. Rinzivillo, F. Giannotti, and D. Pedreschi. 2012. "Optimal Spatial Resolution for the Analysis of Human Mobility." In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, IEEE*, 248–252.

Cressie, N. A. 1996. "Change of Support and the Modifiable Areal Unit Problem." *Geographical Systems* 3: 159–180.

Fotheringham, A. S. 1989. Scale-Independent Spatial Analysis. Accuracy of Spatial Databases.

Fotheringham, A. S., and D. W. Wong. 1991. "The Modifiable Areal Unit Problem in Multivariate Statistical Analysis." *Environment and Planning A: Economy and Space* 23: 1025–1044.

Fugacci, U., S. Scaramuccia, F. Iuricich, and L. D. F. Iuricich. 2016. "Persistent Homology: A Step-By-Step Introduction for Newcomers." In *Smart Tools and Apps for Graphics—Eurographics Italian Chapter Conference*, edited by G. Pintore and F. Stanco, 1–10. Eurographics Association.

Gao, F., J. Tang, and Z. Li. 2022. "Effects of Spatial Units and Travel Modes on Urban Commuting Demand Modeling." *Transportation* 49: 1549–1575.

Gao, S., Y. Liu, Y. Wang, and X. Ma. 2013. "Discovering Spatial Interaction Communities From Mobile Phone Data." *Transactions in GIS* 17: 463–481.

Gehlke, C. E., and K. Biehl. 1934. "Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material." *Journal of the American Statistical Association* 29: 169–170.

Ghrist, R. 2008. "Barcodes: The Persistent Topology of Data." *Bulletin of the American Mathematical Society* 45: 61–75.

Green, M., and R. Flowerdew. 1996. "New Evidence on the Modifiable Areal Unit Problem." *Spatial Analysis: Modelling in a GIS Environment*: 41–54.

Hagberg, A., P. J. Swart, and D. A. Schult. 2008. "Exploring Network Structure, Dynamics, and Function Using NetworkX. Technical Report. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States)".

Hagen-Zanker, A., and Y. Jin. 2012. "A New Method of Adaptive Zoning for Spatial Interaction Models." *Geographical Analysis* 44: 281–301.

Hajij, M., B. Wang, C. Scheidegger, and P. Rosen. 2018. "Visual Detection of Structural Changes in Time-Varying Graphs Using Persistent Homology." In *Pacific Visualization Symposium (PacificVis), 2018 IEEE, IEEE*, 125–134.

Haynes, K. E., and A. S. Fotheringham. 1984. *Gravity and Spatial Interaction Models*. Vol. 2. Sage.

Holt, D., D. Steel, and M. Tranmer. 1996. "Area Homogeneity and the Modifiable Areal Unit Problem." *Geographical Systems* 3: 181–200.

Huang, Y., J. Lin, X. He, Z. Lin, Z. Wu, and X. Zhang. 2024. "Assessing the Scale Effect of Urban Vertical Patterns on Urban Waterlogging: An Empirical Study in Shenzhen." *Environmental Impact Assessment Review* 106: 107486.

Jelinski, D. E., and J. Wu. 1996. "The Modifiable Areal Unit Problem and Implications for Landscape Ecology." *Landscape Ecology* 11: 129–140.

Kang, C., X. Ma, D. Tong, and Y. Liu. 2012. "Intra-Urban Human Mobility Patterns: An Urban Morphology Perspective." *Physica A: Statistical Mechanics and Its Applications* 391: 1702–1717.

Kerber, M., D. Morozov, and A. Nigmetov. 2017. "Geometry Helps to Compare Persistence Diagrams." *ACM Journal of Experimental Algorithmics* 22: 1–20.

Kim, H., J. Hahm, H. Lee, E. Kang, H. Kang, and D. S. Lee. 2015. "Brain Networks Engaged in Audiovisual Integration During Speech Perception Revealed by Persistent Homology-Based Network Filtration." *Brain Connectivity* 5: 245–258.

Kodinariya, T. M., and P. R. Makwana. 2013. "Review on Determining Number of Cluster in k-Means Clustering." *International Journal* 1: 90–95.

Kosonen, M. 2023. "The Scaling Effect of the Modifiable Areal Unit Problem in Assessing the Social and Environmental Drivers of Global Migration".

Kotavaara, O., H. Antikainen, M. Marmion, and J. Rusanen. 2012. "Scale in the Effect of Accessibility on Population Change: Gis and a Statistical Approach to Road, Air and Rail Accessibility in Finland, 1990–2008." *Geographical Journal* 178: 366–382.

Lee, H., H. Kang, M. K. Chung, B. N. Kim, and D. S. Lee. 2012. "Persistent Brain Network Homology From the Perspective of Dendrogram." *IEEE Transactions on Medical Imaging* 31: 2267–2277.

Li, M., S. Gao, F. Lu, K. Liu, H. Zhang, and W. Tu. 2021. "Prediction of Human Activity Intensity Using the Interactions in Physical and Social Spaces Through Graph Convolutional Networks." *International Journal of Geographical Information Science* 35: 2489–2516.

Liu, X., L. Gong, Y. Gong, and Y. Liu. 2015. "Revealing Travel Patterns and City Structure With Taxi Trip Data." *Journal of Transport Geography* 43: 78–90.

Liu, X., C. Kang, L. Gong, and Y. Liu. 2016. "Incorporating Spatial Interaction Patterns in Classifying and Understanding Urban Land Use." *International Journal of Geographical Information Science* 30: 334–350.

Liu, Y., C. Kang, S. Gao, Y. Xiao, and Y. Tian. 2012. "Understanding Intra-Urban Trip Patterns From Taxi Trajectory Data." *Journal of Geographical Systems* 14: 463–483.

Liu, Y., Z. Sui, C. Kang, and Y. Gao. 2014. "Uncovering Patterns of Inter-Urban Trip and Spatial Interaction From Social Media Check-In Data." *PLoS One* 9: e86026.

Liu, Y., F. Wang, Y. Xiao, and S. Gao. 2012. "Urban Land Uses and Traffic 'Source-Sink Areas': Evidence From Gps-Enabled Taxi Data in Shanghai." *Landscape and Urban Planning* 106: 73–87.

Liu, Y., Z. Zhan, D. Zhu, Y. Chai, X. Ma, and L. Wu. 2018. "Incorporating Multi-Source Big Geo-Data to Sense Spatial Heterogeneity Patterns in an Urban Space." *Geomatics and Information Science of Wuhan University* 43: 327–335.

Lum, P. Y., G. Singh, A. Lehman, et al. 2013. "Extracting Insights From the Shape of Complex Data Using Topology." *Scientific Reports* 3: 1236.

Manley, D. 2014. "Scale, Aggregation, and the Modifiable Areal Unit Problem." In *Handbook of Regional Science*, 1157–1171.

Martin, M. E., and N. Schuurman. 2020. "Social Media Big Data Acquisition and Analysis for Qualitative Giscience: Challenges and Opportunities." *Annals of the American Association of Geographers* 110.

McMillen, D. P. 2004. "Geographically Weighted Regression: The Analysis of Spatially Varying Relationships." *American Journal of Agricultural Economics* 86.

Miller, H. J. 1999. "Potential Contributions of Spatial Analysis to Geographic Information Systems for Transportation (Gis-t)." *Geographical Analysis* 31: 373–399.

Milnor, J. 1964. "On the Betti Numbers of Real Varieties." *Proceedings of the American Mathematical Society* 15: 275–280.

Morozov, D. 2017. "Dionysus Documentation." https://mrzv.org/software/dionysus/.

Mu, L., and F. Wang. 2008. "A Scale-Space Clustering Method: Mitigating the Effect of Scale in the Analysis of Zone-Based Data." *Annals of the Association of American Geographers* 98: 85–101.

Nelson, A. 2001. "Analysing Data Across Geographic Scales in Honduras: Detecting Levels of Organisation Within Systems." *Agriculture, Ecosystems & Environment* 85: 107–131.

Openshaw, S. 1977. "Optimal Zoning Systems for Spatial Interaction Models." *Environment and Planning A* 9: 169–184.

Openshaw, S. 1984. *The Modifiable Areal Unit Problem* Concepts and techniques in modern geography.

Ozair, S., C. Lynch, Y. Bengio, A. van den Oord, S. Levine, and P. Sermanet. 2019. "Wasserstein Dependency Measure for Representation Learning. Advances in Neural Information Processing Systems 32".

Page, L., S. Brin, R. Motwani, and T. Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web. Technical Report*. Stanford InfoLab.

Pedregosa, F., G. Varoquaux, A. Gramfort, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830.

Qi, G., X. Li, S. Li, G. Pan, Z. Wang, and D. Zhang. 2011. "Measuring Social Functions of City Regions From Large-Scale Taxi Behaviors, in: 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), IEEE."

Qi, Y., and J. Wu. 1996. "Effects of Changing Spatial Resolution on the Results of Landscape Pattern Analysis Using Spatial Autocorrelation Indices." *Landscape Ecology* 11: 39–49.

Sizemore, A. E., C. Giusti, A. Kahn, J. M. Vettel, R. F. Betzel, and D. S. Bassett. 2018. "Cliques and Cavities in the Human Connectome." *Journal of Computational Neuroscience* 44: 115–145.

Stehman, S. V., and J. D. Wickham. 2011. "Pixels, Blocks of Pixels, and Polygons: Choosing a Spatial Unit for Thematic Accuracy Assessment." *Remote Sensing of Environment* 115: 3044–3055.

Stepniak, M., and P. Rosik. 2015. "The Impact of Data Aggregation on Potential Accessibility Values." In *Geoinformatics for Intelligent Transportation*, 227–240. Springer.

Stillwell, J., K. Daras, and M. Bell. 2018. "Spatial Aggregation Methods for Investigating the Maup Effects in Migration Analysis." *Applied Spatial Analysis and Policy* 11: 693–711.

Stolz, B. J., H. A. Harrington, and M. A. Porter. 2017. "Persistent Homology of Time-Dependent Functional Networks Constructed From Coupled Time Series." *Chaos (Woodbury, N.Y.)* 27: 047410.

Taylor, D., F. Klimm, H. A. Harrington, et al. 2015. "Topological Data Analysis of Contagion Maps for Examining Spreading Processes on Networks." *Nature Communications* 6: 7723.

Thakkar, K. S., R. V. Dharaskar, and M. Chandak. 2010. *Graph-Based Algorithms for Text Summarization, in: 2010 3rd International Conference on Emerging Trends in Engineering and Technology, IEEE*, 516–519.

Tralie, C., N. Saul, and R. Bar-On. 2018. "Ripser.Py: A Lean Persistent Homology Library for Python." *Journal of Open Source Software* 3: 925. https://doi.org/10.21105/joss.00925.

Ubøe, J. 2004. "Aggregation of Gravity Models for Journeys to Work." *Environment and Planning A* 36: 715–729.

Vershik, A. M. 2013. "Long History of the Monge-Kantorovich Transportation Problem." *Mathematical Intelligencer* 35: 1–9.

Wei, L., Y. Luo, M. Wang, et al. 2020. "Multiscale Identification of Urban Functional Polycentricity for Planning Implications: An Integrated Approach Using Geo-Big Transport Data and Complex Network Modeling." *Habitat International* 97: 102134.

Wubie, B. A., A. Andres, R. Greiner, et al. 2018. "Cluster Identification via Persistent Homology and Other Clustering Techniques, With Application to Liver Transplant Data." In *Research in Computational Topology*, 145–177. Springer.

Xiao, J. 2020. "Spatial Aggregation Entropy: A Heterogeneity and Uncertainty Metric of Spatial Aggregation." *Annals of the American Association of Geographers* 111.

Ye, X., and P. Rogerson. 2022. "The Impacts of the Modifiable Areal Unit Problem (Maup) on Omission Error." *Geographical Analysis* 54: 32–57.

Yu, H., and S. L. Shaw. 2008. "Exploring Potential Human Activities in Physical and Virtual Spaces: A Spatio-Temporal Gis Approach." *International Journal of Geographical Information Science* 22: 409–430.

Zhang, S., D. Zhu, X. Yao, X. Cheng, H. He, and Y. Liu. 2018. "The Scale Effect on Spatial Interaction Patterns: An Empirical Study Using Taxi Od Data of Beijing and Shanghai." *IEEE Access* 6: 51994–52003.

Zhou, X., and A. G. Yeh. 2021. "Understanding the Modifiable Areal Unit Problem and Identifying Appropriate Spatial Unit in Jobs–Housing Balance and Employment Self-Containment Using Big Data." *Transportation* 48: 1267–1283.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Appendix

Visualizations of 2D Projections of Point Clouds and Persistence Diagrams for Five Cities

This appendix provides supplementary visualizations to illustrate the data structure and topological features of the five cities across spatial scales. The 2D scatter plots are obtained by using PCA to project the original high-dimensional spatial data onto the first two principal components (Figures S1A–S5A), with the first principal component (PC1) displayed on the x-axis and the second principal component (PC2) on the y-axis. The persistence diagrams are computed directly from the original point cloud data and summarize the evolution of topological features across scales. The generated PDs (Figures S1B–S5B) show H_0 features as blue dots and H_1 features as orange dots. Predominantly, the H_1 features are close to the diagonal, indicating noise; however, persistent H_1 features appear on some scales.

As spatial scale increases, 2D PCA projections across all five cities show a consistent trend of point clouds becoming more dispersed, reflecting a loss of fine-grained structural consistency. Correspondingly, persistence diagrams (PDs) reveal a broader spread of features around the critical scales, particularly in H_1, indicating the emergence of higher-dimensional topological structures. For example, in Shenzhen and Jiaxing (critical scale: 1500), point clouds transition from tight clusters to more diffuse forms, with PDs showing more persistent H_0 features and late-born H_1 features. In Beijing (1750) and Shanghai (2000), similar to Shenzhen and Jiaxing, the dispersion and emergence of later-born persistent H_1 features are noticeable around their critical scales, while in Wuhan (2500), the changes are more pronounced, with both H_0 and H_1 features exhibiting increased persistence and emergence at later birth times.

It should be noted that PCA-based 2D projections only retain a limited proportion of the variance in the original high-dimensional space. As a result, the scatter plots may fail to capture essential structural features of the original data, particularly when the principal components explain only a small portion of the total variance. Although persistence diagrams preserve essential information about topological invariants, interpreting them visually to identify precise critical scales is inherently challenging. The changes in diagram structure tend to be gradual and subtle, often requiring subjective judgment. In contrast, the proposed SSTD metric offers a quantitative and systematic approach to identify scale thresholds where topological features change most significantly.