Spatially Aware Adjusted Rand Index for Evaluating Spatial Transcriptomics Clustering

Yinqiao Yan[1], Xiangnan Feng[2], and Xiangyu Luo[*3]

[1]School of Mathematics, Statistics and Mechanics, Beijing University of Technology
[2]Department of Statistics and Data Science, Fudan University
[3]Institute of Statistics and Big Data, Renmin University of China

Abstract

The spatial transcriptomics (ST) clustering plays a crucial role in elucidating the tissue spatial heterogeneity. An accurate ST clustering result can greatly benefit downstream biological analyses. As various ST clustering approaches are proposed in recent years, comparing their clustering accuracy becomes important in benchmarking studies. However, the widely used metric, adjusted Rand index (ARI), totally ignores the spatial information in ST data, which prevents ARI from fully evaluating spatial ST clustering methods. We propose a spatially aware Rand index (spRI) as well as spARI that incorporate the spatial distance information. Specifically, when comparing two partitions, spRI provides a disagreement object pair with a weight relying on the distance of the two objects, whereas Rand index assigns a zero weight to it. This spatially aware feature of spRI adaptively differentiates disagreement object pairs based on their distinct distances, providing a useful evaluation metric that favors spatial coherence of clustering. The spARI is obtained by adjusting spRI for random chances such that its expectation takes zero under an appropriate null model. Statistical properties of spRI and spARI are discussed. The applications to simulation study and two ST datasets demonstrate the improved utilities of spARI compared to ARI in evaluating ST clustering methods. The R package to compute spRI and spARI is available at https://github.com/yinqiaoyan/spARI.

Key words: clustering evaluation; hypergeometric distribution; Rand index; spatial transcriptomics.

*Corresponding author, xiangyuluo@ruc.edu.cn

1 Introduction

The spatial transcriptomics (ST) provides unprecedented opportunity to elucidate the spatial context of gene expressions thanks to the advances in biological technologies such as in situ hybridization, in situ sequencing, and spatial barcoding (Williams et al., 2022). Depending on the spatial resolution, ST data can be classified into the spot level and the single-cell level, where in the previous case spots are arranged in a regular lattice format and in the latter case cells distribute irregularly in the tissue section. The ST clustering for spots or cells is an important step in the statistical analysis to understand and investigate the solid tissue spatial heterogeneity especially for cancer and brain. For spot level ST data, the ST clustering becomes region segmentation that aims to partition the tissue section into disjoint domains with different gene expression or morphological features. For single-cell ST data, the ST clustering comes in the form of spatial cell typing that assigns cells to different types. Therefore, downstream biological analyses that rely on the elucidation of spatial heterogeneity, such as marker gene detection (Song et al., 2023), domain/cell type annotation (Shi et al., 2023), imputation (Wan et al., 2023), and region-specific deconvolution (Ma and Zhou, 2022), would greatly benefit from an accurate ST clustering result.

Various ST clustering approaches have been developed in recent several years for region segmentation (Hu et al., 2021; Zhao et al., 2021; Dong and Zhang, 2022; Chidester et al., 2023; Yan and Luo, 2024; Ma and Zhou, 2024) as well as spatial cell typing (Teng et al., 2022; Dong and Zhang, 2022; Li and Zhou, 2022; Singhal et al., 2024; Guo et al., 2024; Yan and Luo, 2025). Different from clustering methods for single-cell RNA-sequencing data, these ST clustering approaches can incorporate spatial information (i.e., the neighbors or spatial coordinates of spots/cells) to improve accuracy, thus offering more insights into the complex disease molecular mechanisms (Park et al., 2023). Considering a large volume of ST clustering methods available, a natural and practical question is that how to compare their performances in a reasonable and quantitative way to provide guidance for an appropriate choice of these ST clustering methods in diverse situations (Yuan et al., 2024).

To evaluate the clustering accuracy, several statistical metrics have been come up with to quantify the comparison of the clustering result against a reference. The Rand index (Rand, 1971) is the first attempt to compare two partitions for a given number of objects, which are spots or cells in our situation, and quantify the "similarity" of the two partitions. Specifically, for two different objects, we first observe whether the two objects are in the same cluster for one partition but are in different groups for the other partition. If it is the case, such two objects are then called a disagreement pair; otherwise, they are named an agreement pair. Subsequently, the Rand index is defined as the ratio of the number of agreement pairs to the total number of pairs, so it takes values between zero and one, and Rand index one indicates that two partitions are identical. The commonly used adjusted Rand index (ARI) (Hubert and Arabie, 1985) is a variant of the Rand index that is corrected for random chance. By counting the common object numbers of two partitions in a contingency table and assuming a generalized hypergeometric distribution for the table entries with fixed column and row sums, ARI has an expectation zero and is bounded above by one. Other clustering evaluation measures include the normalized mutual information and the V-measure (Wu and Wu, 2020). Sometimes, Moran's I is used to measure spatial autocorrelations in a single partition, but it

cannot capture the similarity between two partitions. In the current ST clustering literature, ARI is the most widely used clustering evaluation metric.

However, ARI ignores an invaluable characteristic of the ST partition—the spatial locations of objects, so it cannot evaluate the spatial coherence ability of the clustering methods. For example, two different spatial clusterings can result in the same ARI (Sections 4.1 and 4.2). In the analysis of human dorsolateral prefrontal cortex data (Section 5.1), the omission of spatial information by ARI may lead to unsatisfactory evaluation results, where a method with comparable clustering accuracy but weaker spatial coherence receives a higher ARI value. Therefore, there is an urgent need to propose a subtle *spatially aware* clustering metric that can adequately assess a large number of ST clustering algorithms.

We remark that the focus of this paper is not on benchmarking ST clustering algorithms, but on developing a new clustering metric that can directly make use of spatial information to provide more biologically meaningful measurement. In the field of single-cell RNA-seq data clustering, Wu and Wu (2020) begin to improve the Rand index by accounting for the cell type hierarchy. For a disagreement cell pair, its contribution to the weighted Rand index proposed by Wu and Wu (2020) is not identical to zero but relies on the hierarchical relationship between two cells. However, it is not straightforward to incorporate the spatial information of these cells into wRI due to the biological distinction between the cell type hierarchy and the spatial pattern of spots/cells. Recently, Xu et al. (2024) propose a spatial grouping discrepancy metric for ST data, which enables the measurement of spatial distribution discrepancies between predicted and reference labels. Nevertheless, this metric is primarily for anomalous tissue region detection and is not entirely compatible with the general ST clustering accuracy quantification in our task.

In this paper, we present two novel spatially aware clustering metrics based on Rand index and ARI, which are named spRI and spARI, respectively, to quantitatively evaluate the ST clustering performance. Specifically, inspired by the framework of Wu and Wu (2020), when comparing two partitions, spRI provides a disagreement pair of objects with a weight relying on the spatial distance of the two objects, directly employing the available spatial coordinates. In contrast, the classical Rand index assigns an identical zero weight to all disagreement pairs. This spatially aware feature of spRI successfully differentiates disagreement object pairs based on their distinct distances, which cannot be achieved by the Rand index. We adjust spRI to correct for random chances so that its expectation takes on zero value under an appropriate random null model, which results in spARI. We further investigate the statistical properties of spRI and spARI, including the analytical forms between two similar clusterings, the sensitivity to subsampling, hypothesis testings, and computational efficiency. Importantly, in simulation studies, two spatial clustering results with the same ARI can have different spARI values, demonstrating that spARI prefers a more spatially coherent clustering. We also provide an R package for computing spRI and spARI which is user-friendly for statisticians and biologists. The applications to two ST datasets demonstrate the improved utilities of spARI compared to ARI in evaluating ST clustering methods.

2 Method

2.1 Review of Rand index and adjusted Rand index

We consider n objects (spots or cells) with each object i ($1 \leq i \leq n$) having a spatial coordinate z_i. We assume there is a reference partition of the n objects, denoted by $\mathbf{U} = \{U_1, \ldots, U_r, \ldots, U_R\}$. In the reference partition $\mathbf{U}$, the n objects are clustered into disjoint R groups, which represent regions for spots and cell types for cells, and U_r is the set of objects that belong to group r ($1 \leq r \leq R$). The reference partition can be obtained from data sources with highly confident manual region or cell type annotations (Wu and Wu, 2020; Yuan et al., 2024) and thus serves as a gold standard to compare ST clustering algorithms. In addition, some clustering method partitions the same n objects into C clusters, denoted by $\mathbf{V} = \{V_1, \ldots, V_c, \ldots, V_C\}$, where the cluster number C is not necessarily equal to the group number R in the reference.

Based on the two partitions $\mathbf{U}$ and $\mathbf{V}$, for any two different objects i and j, they are named an *agreement pair* if they are grouped in both the reference and the clustering, i.e., $i, j \in U_r$ and $i, j \in V_c$ hold for some r and c, or if they are separated in both the reference and the clustering, i.e., $i, j \notin U_r$ and $i, j \notin V_c$ hold for any r and c. On the other hand, if two objects i and j are grouped in one partition but separated in the other partition, this pair is referred to as a *disagreement* pair, following the notations in Hubert and Arabie (1985). For all $N = \binom{n}{2}$ object pairs, according to the "grouped" case or the "separated" case in the reference $\mathbf{U}$ and the clustering $\mathbf{V}$, we have four types for these pairs, and the corresponding contingency table is in Table 1.

Table 1: The numbers for the four object pair types. Type "g×g", the two objects are in the same group in $\mathbf{U}$ and in the same cluster in $\mathbf{V}$. Type "s×s", the two objects are in different groups in $\mathbf{U}$ and in different clusters in $\mathbf{V}$. Type "s×g", the two objects are in different groups in $\mathbf{U}$ but in the same cluster in $\mathbf{V}$. Type "g×s", the two objects are in the same group in $\mathbf{U}$ but in different clusters in $\mathbf{V}$.

Reference $\mathbf{U}$ \ Clustering $\mathbf{V}$	Grouped	Separated	Sums
Grouped	"g×g" N_{11}	"g×s" N_{12}	$N_{1\cdot}$
Separated	"s×g" N_{21}	"s×s" N_{22}	$N_{2\cdot}$
Sums	$N_{\cdot 1}$	$N_{\cdot 2}$	N

In Table 1, we denote the object pair numbers of types "g×g", "s×s", "s×g" and "g×s" by N_{11}, N_{22}, N_{21}, and N_{12}, respectively. We also use the dot index to represent the row sums or column sums in Table 1. For example, $N_{1\cdot} = N_{11} + N_{12}$. Subsequently, the number of agreement object pairs is equal to the type "g×g" number plus the type "s×s" number (i.e., $N_{11} + N_{22}$), while the number of disagreement object pairs becomes the type "s×g" number plus the type "g×s" number (i.e., $N_{21} + N_{12}$). The Rand index (RI) (Rand, 1971) is then defined as $\mathrm{RI} = (N_{11} + N_{22})/N$, which is exactly the proportion of agreement pairs among all the N object pairs.

Rand (1971) and Hubert and Arabie (1985) provide a convenient way to calculate RI using the common object numbers of the two partitions. In Table 2, n_{rc} is the number of common objects in group U_r and cluster V_c ($r = 1, \ldots, R;\ c = 1, \ldots, C$), $n_{\cdot c} = \sum_{r=1}^{R} n_{rc}$, and

Table 2: The common object number table by comparing the reference partition $\mathbf{U}$ and the clustering partition $\mathbf{V}$.

Reference \ Clustering		$\mathbf{V}$				Sums
		V_1	V_2	$\cdots$	V_C	
$\mathbf{U}$	U_1	n_{11}	n_{12}	$\cdots$	n_{1C}	$n_{1\cdot}$
	U_2	n_{21}	n_{22}	$\cdots$	n_{2C}	$n_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	U_R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	$n_{R\cdot}$
Sums		$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot C}$	$n_{\cdot\cdot} = n$

$n_{r\cdot} = \sum_{c=1}^{C} n_{rc}$. Using these notations, it is easy to derive that

$$N_{12} = \sum_{r=1}^{R} \sum_{1 \le c_1 < c_2 \le C} \binom{n_{rc_1}}{1} \binom{n_{rc_2}}{1} = \sum_{r=1}^{R} \sum_{1 \le c_1 < c_2 \le C} n_{rc_1} n_{rc_2} \tag{1}$$
$$= \frac{1}{2} \left[\sum_{r=1}^{R} \left(\sum_{c=1}^{C} n_{rc} \right)^2 - \sum_{r=1}^{R} \sum_{c=1}^{C} n_{rc}^2 \right] = \frac{1}{2} \left[\sum_{r=1}^{R} n_{r\cdot}^2 - \sum_{r=1}^{R} \sum_{c=1}^{C} n_{rc}^2 \right],$$

$$N_{21} = \sum_{c=1}^{C} \sum_{1 \le r_1 < r_2 \le R} \binom{n_{r_1 c}}{1} \binom{n_{r_2 c}}{1} = \sum_{c=1}^{C} \sum_{1 \le r_1 < r_2 \le R} n_{r_1 c} n_{r_2 c} \tag{2}$$
$$= \frac{1}{2} \left[\sum_{c=1}^{C} \left(\sum_{r=1}^{R} n_{rc} \right)^2 - \sum_{r=1}^{R} \sum_{c=1}^{C} n_{rc}^2 \right] = \frac{1}{2} \left[\sum_{c=1}^{C} n_{\cdot c}^2 - \sum_{r=1}^{R} \sum_{c=1}^{C} n_{rc}^2 \right].$$

In Equation (1), $\sum_{1 \le c_1 < c_2 \le C} n_{rc_1} n_{rc_2}$ represents the number of object pairs that are in group U_r of $\mathbf{U}$ but are separated in $\mathbf{V}$. Similarly, in Equation (2), $\sum_{1 \le r_1 < r_2 \le R} n_{r_1 c} n_{r_2 c}$ is the number of object pairs that are separated in $\mathbf{U}$ but are in cluster V_c of $\mathbf{V}$. Subsequently,

$$\mathrm{RI} = \frac{N_{11} + N_{22}}{N} = \frac{N - N_{12} - N_{21}}{N} = \frac{\binom{n}{2} + \sum_{r=1}^{R} \sum_{c=1}^{C} n_{rc}^2 - \frac{1}{2} \left[\sum_{r=1}^{R} n_{r\cdot}^2 + \sum_{c=1}^{C} n_{\cdot c}^2 \right]}{\binom{n}{2}}$$
$$= \frac{\binom{n}{2} + 2 \sum_{r=1}^{R} \sum_{c=1}^{C} \binom{n_{rc}}{2} - \left[\sum_{r=1}^{R} \binom{n_{r\cdot}}{2} + \sum_{c=1}^{C} \binom{n_{\cdot c}}{2} \right]}{\binom{n}{2}}.$$

To account for the randomness in the contingency table 2, Hubert and Arabie (1985) set a null model where the common object numbers $\{n_{rc} : 1 \le r \le R,\ 1 \le c \le C\}$ follow a generalized hypergeometric distribution with the cluster numbers R and C as well as the column and row sums ($n_{\cdot c}$ and $n_{r\cdot}$) fixed. Under this assumption, the marginal distribution of n_{rc} is actually a hypergeometric distribution, and it turns out that (Hubert and Arabie, 1985) $\mathbb{E}\left[\binom{n_{rc}}{2}\right] = \binom{n_{r\cdot}}{2}\binom{n_{\cdot c}}{2} \Big/ \binom{n}{2}$. Accordingly, the expectation of RI in the null random model

is obtained by

$$\mathbb{E}\left(\mathrm{RI}\right) = 1 + 2\sum_{r=1}^{R}\binom{n_{r\cdot}}{2}\sum_{c=1}^{C}\binom{n_{\cdot c}}{2}\Big/\binom{n}{2}^2 - \left[\sum_{r=1}^{R}\binom{n_{r\cdot}}{2} + \sum_{c=1}^{C}\binom{n_{\cdot c}}{2}\right]\Big/\binom{n}{2}. \tag{3}$$

We remark that $\mathbb{E}(\mathrm{RI}) = 1$ if and only if the reference $\mathbf{U}$ has only one group and all the objects are in one cluster of $\mathbf{V}$, i.e., $R = C = 1$ and $n_{1\cdot} = n_{\cdot 1} = n$; or each cluster has only one object in both $\mathbf{U}$ and $\mathbf{V}$, i.e., $n_{r\cdot} = n_{\cdot c} = 1$ for $\forall r$, $\forall c$. The proof is in Web Appendix A. When $\mathbb{E}(\mathrm{RI}) = 1$ (two partitions are identical), the adjusted Rand index (ARI) is defined as one; and when $\mathbb{E}(\mathrm{RI}) < 1$, ARI has the following form (Hubert and Arabie, 1985),

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}(\mathrm{RI})}{1 - \mathbb{E}(\mathrm{RI})} = \frac{\sum_{r=1}^{R}\sum_{c=1}^{C}\binom{n_{rc}}{2} - \sum_{r=1}^{R}\binom{n_{r\cdot}}{2}\sum_{c=1}^{C}\binom{n_{\cdot c}}{2}\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_{r=1}^{R}\binom{n_{r\cdot}}{2} + \sum_{c=1}^{C}\binom{n_{\cdot c}}{2}\right] - \sum_{r=1}^{R}\binom{n_{r\cdot}}{2}\sum_{c=1}^{C}\binom{n_{\cdot c}}{2}\Big/\binom{n}{2}}.$$

ARI is bounded above by one and has an expectation of zero in the null random model. It is the most widely used clustering evaluation metric in the current ST clustering literature.

2.2 The proposed spatially aware adjusted Rand index

In evaluating ST clustering, RI and ARI do not directly account for the spatial information of objects. Following Wu and Wu (2020), we first define a weight function $W(i,j)$ for each object pair (i,j) as follows

$$W(i,j) = \begin{cases} 1 & \text{if } (i,j) \text{ is an agreement pair} \\ 0 & \text{if } (i,j) \text{ is a disagreement pair} \end{cases}.$$

Using this notation, RI can be reformulated as $\mathrm{RI} = \sum_{1\le i<j\le n} W(i,j)/N$. We observe that for any disagreement pair, its weight in RI is identical to zero, so this weight function cannot make any difference for two disagreement pairs where their spatial distances are distinct. Therefore, we propose a new weight function $\widetilde{W}(i,j)$ that can incorporate the spatial distance information as follows,

$$\widetilde{W}(i,j) = \begin{cases} 1 & \text{if } (i,j) \text{ is an agreement pair} \\ f_{ij} := f(d(z_i,z_j)) \in [0,\alpha] & \text{if } (i,j) \text{ is a type “s×g” object pair} \\ h_{ij} := h(d(z_i,z_j)) \in [0,\alpha] & \text{if } (i,j) \text{ is a type “g×s” object pair} \end{cases},$$

where $d(z_i,z_j)$ is a distance between the spatial locations of object i (z_i) and object j (z_j), f is a decreasing function from $[0,+\infty)$ to $[0,\alpha]$, and h is an increasing function from $[0,+\infty)$ to $[0,\alpha]$. The coefficient α belongs to the open interval $(0,1)$ to keep a positive gap between the maximal weight of the disagreement pair and the weight one of the agreement pair. Throughout this paper, we adopt the Euclidean distance $d(z_i,z_j) = \|z_i - z_j\|_2$, $f(t) = \alpha\exp\{-t^2\}$, $h(t) = \alpha(1-\exp\{-t^2\})$, and α is fixed at 0.8; the rationales for these choices are provided in the next subsection. To remove the effect of spatial location unit, we normalize z_i by $(z_{i\ell} -$

$\min_i z_{i\ell})/(\max_i z_{i\ell} - \min_i z_{i\ell})$ for each dimension ℓ of z_i such that $z_{i\ell} \in [0, 1]$.

In this definition, even though two disagreement pairs (i_1, j_1) and (i_2, j_2) are of the same disagreement type ("s×g" or "g×s"), their weights are different as long as $d(z_{i_1}, z_{j_1}) \neq d(z_{i_2}, z_{j_2})$. More importantly, the weight function $\widetilde{W}$ has advantages over W in the following two situations which are often encountered in practice.

(i) In situation one where both (i_1, j_1) and (i_2, j_2) are of type "s×g", if $d_1 := d(z_{i_1}, z_{j_1})$ is larger than $d_2 := d(z_{i_2}, z_{j_2})$, then $\widetilde{W}(i_1, j_1)$ is less than $\widetilde{W}(i_2, j_2)$ due to the decreasing function f. In other words, if two objects are separated in the reference but grouped together in the ST clustering, $\widetilde{W}$ assigns a higher weight to the object pair with a smaller distance. In this "grouped incorrectly" clustering case, we are more willing to forgive an incorrect grouping with a smaller distance.

(ii) In situation two where both (i_1, j_1) and (i_2, j_2) are of type "g×s", if $d_1 > d_2$, then $\widetilde{W}(i_1, j_1) > \widetilde{W}(i_2, j_2)$ due to the increasing function h. In other words, if two objects are in one group in the reference but separated in the ST clustering, $\widetilde{W}$ assigns a higher weight to the object pair with a larger distance. In this "separated incorrectly" clustering case, we are more willing to forgive an incorrect separation with a larger distance.

Subsequently, similar to RI, we define the spatially aware Rand index (spRI) as

$$\text{spRI} = \frac{\sum_{1 \leq i < j \leq n} \widetilde{W}(i, j)}{N}.$$

Following the random model used in ARI where column/row sums, the reference group number R, and the cluster number C in Table 2 are fixed, we aim to calculate the expectation of spRI. We consider two objects (i, j) and $i \neq j$. The probability that the two object are in one group of the reference $\mathbf{U}$ is denoted by p_{ij}, and the probability that they are in one group of the clustering $\mathbf{V}$ is denoted by q_{ij}. Using this randomness, it turns out that

$$p_{ij} = \frac{\binom{n_{1\cdot}}{2} + \binom{n_{2\cdot}}{2} + \ldots + \binom{n_{R\cdot}}{2}}{\binom{n}{2}} = \frac{\sum_{r=1}^{R} \binom{n_{r\cdot}}{2}}{\binom{n}{2}},$$

$$q_{ij} = \frac{\binom{n_{\cdot 1}}{2} + \binom{n_{\cdot 2}}{2} + \ldots + \binom{n_{\cdot C}}{2}}{\binom{n}{2}} = \frac{\sum_{c=1}^{C} \binom{n_{\cdot c}}{2}}{\binom{n}{2}}.$$

Notice that the terms on the right hand side are not relevant to indices (i, j), so we simply let $p_{ij} = p$ and $q_{ij} = q$. Owing to the independence in assigning a type to an object pair for $\mathbf{U}$ and $\mathbf{V}$, for any object pair, the probability of being type "g×g", "s×s", "s×g", and "g×s" is pq, $(1-p)(1-q)$, $(1-p)q$, and $p(1-q)$, respectively. Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[\widetilde{W}(i, j)\right] = 1 \cdot \mathbb{P}(\text{object pair (i, j) is of type "g×g" or "s×s"}) \\
&+ f_{ij} \cdot \mathbb{P}(\text{object pair (i, j) is of type "s×g"}) + h_{ij} \cdot \mathbb{P}(\text{object pair (i, j) is of type "g×s"}) \\
&= pq + (1-p)(1-q) + f_{ij}(1-p)q + h_{ij}p(1-q).
\end{aligned}$$

The expectation of spRI is derived as follows,

$$\mathbb{E}(\text{spRI}) = \frac{\sum_{1\le i<j\le n} \mathbb{E}\widetilde{W}(i,j)}{\binom{n}{2}} = pq + (1-p)(1-q) + F(1-p)q\Big/\binom{n}{2} + Hp(1-q)\Big/\binom{n}{2}$$
$$= 1 + 2pq - (p+q) + \frac{Fq + Hp - (H+F)pq}{\binom{n}{2}} = \mathbb{E}(\text{RI}) + \frac{Fq + Hp - (H+F)pq}{\binom{n}{2}},$$

where $F = \sum_{1\le i<j\le n} f_{ij}$, $H = \sum_{1\le i<j\le n} h_{ij}$, and the last equality holds due to Equation (3). When $f_{ij} = 0$ and $h_{ij} = 0$ for any (i,j), $\mathbb{E}(\text{spRI})$ becomes $\mathbb{E}(\text{RI})$. We notice that for $p, q \in [0,1]$, $\mathbb{E}(\text{spRI}) = 1$ if and only if $p = q = 1$ or $p = q = 0$ (the proof is in Web Appendix B). When $\mathbb{E}(\text{spRI}) < 1$, we define the adjusted version of spRI, named spARI, as

$$\text{spARI} = \frac{\text{spRI} - \mathbb{E}(\text{spRI})}{1 - \mathbb{E}(\text{spRI})}$$
$$= \frac{\sum_{1\le i<j\le n} \widetilde{W}(i,j) - \binom{n}{2} + \binom{n}{2}(p+q-2pq) - (Fq + Hp - (H+F)pq)}{\binom{n}{2}(p+q-2pq) - (Fq + Hp - (H+F)pq)}.$$

In this way, the expectation of spARI is exactly zero under the null random model, and this property is the same as ARI (Hubert and Arabie, 1985). When $\mathbb{E}(\text{spRI}) = 1$ (two partitions are identical), we define spARI as one. Notice that if we denote the spARI value with $\mathbf{U}$ in rows and $\mathbf{V}$ in columns as in Table 2 by $\text{spARI}(\mathbf{U}, \mathbf{V})$, then $\text{spARI}(\mathbf{U}, \mathbf{V}) \neq \text{spARI}(\mathbf{V}, \mathbf{U})$ as "$s \times g$" pairs and "$g \times s$" pairs are not symmetric. Therefore, to ensure uniformity, we always put the reference partition in the first position of spARI.

In practice, when we compare two clustering methods (against the same reference), it is helpful to investigate whether one method with a higher spRI than the other method can still give a higher spARI. Web Appendix C provides a practical guideline about when spRI and spARI can reach the same evaluation conclusion for two ST clustering methods.

2.3 The choices for the parameters in spARI

Currently, most ST datasets are collected from tissue slices in a two-dimensional plane, so the Euclidean distance is a natural choice to compute the pairwise distances between objects when calculating spARI. In general, alternative distances can be employed. For example, the use of Euclidean distance implicitly assumes that all coordinate dimensions share equal weights. As a generalization, the Mahalanobis distance is able to assign different weights or even introduce correlation structures for dimensions through the covariance matrix Σ, $d_M(z_i, z_j) = \sqrt{(z_i - z_j)^\top \Sigma^{-1} (z_i - z_j)}$. When Σ is a diagonal matrix, the inverse of each diagonal entry serves as a specific weight for the corresponding dimension. In addition, for geospatial data with large spatial extent on the sphere, the geodesic distance may provide a more appropriate measure of spatial proximity than Euclidean distances.

The functions f and h are defined as $f(t) = \alpha \exp\{-t^2\}$ and $h(t) = \alpha(1 - \exp\{-t^2\})$. Their corresponding function curves are illustrated in Supplementary Figure S1. Both f and h take values in the interval $[0, \alpha]$, with f being decreasing and h increasing. The choice of the exponential function $\exp\{-t^2\}$ stems from its rapid decay as t increases, which allows spARI to better distinguish between nearby and distant object pairs. Generally, such exponential (or

Gaussian) functions are widely employed in spatial statistics to model the decay rate of the spatial correlation with distance (Williams and Rasmussen, 2006). To address the situation where other types of functions may be more appropriate, the R package of spARI now supports user-specified functions f and h with the following constraints—they should map from their domains to $[0, \alpha]$ with f decreasing and h increasing.

Regarding the parameter α, we conducted a sensitivity analysis in Web Appendix D to assess its impact on the spARI value. When α approaches zero, the contribution of disagreement pairs to spARI becomes negligible, so spARI becomes close to ARI. In contrast, as α increases toward one, the accumulated weights of disagreement pairs become larger. Thus, we set the default value of α to 0.8 to ensure a reasonable positive gap between the weights of disagreement and agreement pairs, also capturing meaningful spatial information.

In practice, we should be cautious about the alternative choices of the parameters in spARI unless users have strong and reliable domain expertise that can accurately specify the distance function or associated parameters. The introduction of potentially unjustified choices may mislead the evaluation. Since spARI is designed to serve as an objective evaluator of spatial clustering methods, we recommend that the need for manual or data-driven tuning should be minimized as much as possible if we do not have strong prior information.

3 Some results for spRI and spARI

3.1 Expressions for spRI between two similar clusterings

As all clusterings can be formed by applying a series of operations to an initial clustering with a simple structure, it is important to derive the form of spRI for cases where one clustering is constructed from another based on some basic operation. Following the evaluation framework in Rand (1971), in Web Appendix E, we provide the analytical forms of spRI (a) between the initial clustering and a new clustering formed by simple operations; (b) between two clusterings, each of which is created by applying the same operation to different parts of the same initial clustering; and (c) in some standard settings including all-objects-in-one-cluster, one-object-in-one-cluster, as well as the opposite clustering.

3.2 The sensitivity of spRI to subsampling

The value of spRI may change when we focus on a subset of all objects, and this subsampling technique is often adopted in a pilot study when we aim to cluster hundreds of thousands of spots or cells. Hence, it is useful to look into the sensitivity of spRI, which measures how the spRI varies with the subsampling proportion τ $(0 < \tau < 1)$. Let Ω represent the set of all n objects, and $\tau\Omega$ the subset of Ω with $n\tau$ objects. In Table 2, if we use the subsample set $\tau\Omega$, the expected object number falling into the rth row and cth column is τn_{rc}. For the ease of computation, we assume all these τn_{rc} $(1 \leq r \leq R, 1 \leq c \leq C)$ are integers. In Web Appendix F, we prove that under mild conditions the difference between spRI on the subset and spRI on the whole dataset can be small as long as n is large enough, and the subsampling approximation error is given by $|\text{spRI}_\tau - \text{spRI}| \leq 3\frac{1/\tau - 1}{n - 1/\tau}$.

3.3 Hypothesis testings for spARI

In most ST clustering literature, only a single clustering index is calculated and reported for method evaluations. It is useful to investigate whether the clustering partition is significantly associated with the reference partition. This corresponds to testing the null hypothesis that the spARI value is identical to zero (no association between reference and clustering). We denote the spARI value using the observed partition by $\text{spARI}_{\text{obs}}$, and we test the null hypothesis $H_0 : \text{spARI} = 0$ versus the alternative hypothesis $H_a : \text{spARI} > 0$. Without any assumption about the probability distribution of spARI, we adopt the permutation test (Pitman, 1937) to implement the nonparametric test. The detailed steps in the algorithm are presented in Web Appendix G. Notice that as far as we know, there still does not exist any statistical method that can test the difference between two ARI values (i.e., H_0: ARI_1=ARI_2 v.s. H_a: $\text{ARI}_1 > \text{ARI}_2$); developing a powerful hypothesis testing for comparing two spARI values is challenging and left for future work.

3.4 Computational efficiency of spARI

The R package "spARI" is developed to encapsulate the computational procedure of spRI and spARI, and is publicly available at `https://github.com/yinqiaoyan/spARI`. Users only need to input the two partition results along with spatial coordinates or a precomputed distance matrix, and then the main function efficiently computes the spRI and spARI values. The computational complexity of computing ARI is $\mathcal{O}(n^2)$. In Web Appendix H, we find that if we use a precomputed distance matrix as input, the runtime is under one second for the object number $n \leq 8,000$ on a MacBook Air powered by Apple M4 CPU with 16GB of RAM. When $n = 100,000$, if a sparse distance matrix is passed to the R function to compute the spARI value, the runtime is around 4.8 seconds. This shows that spARI is scalable to large-scale ST data if users are able to provide the distance matrix in a sparse format.

4 Simulation study

In this section, we generate three synthetic ST data to illustrate the distinction between RI/ARI and spRI/spARI when evaluating two clustering results. We demonstrate that the proposed spRI/spARI can effectively distinguish structural differences between two clustering partitions using the spatial information of disagreement object pairs, even when their RI/ARI values are identical. For presentation convenience, we denote the sum of weights contributed by type "s×g" object pairs, $\sum_{(i<j):\text{type "s×g"}} \widetilde{W}_{ij} = \sum_{(i<j):\text{type "s×g"}} f_{ij}$, by $F_{s\times g}$, and the sum of weights by type "g×s" object pairs, $\sum_{(i<j):\text{type "g×s"}} \widetilde{W}_{ij} = \sum_{(i<j):\text{type "g×s"}} h_{ij}$, by $H_{g\times s}$.

4.1 Two clustering results with different $F_{s\times g}$ and identical $H_{g\times s}$

The reference dataset consists of two groups, each with 6 objects in the left panel of Figure 1(a). Objects numbered 1 to 6 belong to group R1, and the other six objects numbered 7 to 12 belong to group R2. There exist two clustering partitions that are illustrated in the middle and right panels of Figure 1(a), respectively. In clustering result i, object number 5, which is in

reference group R1, is incorrectly clustered into C2, while in clustering result ii, object number 2 is incorrectly clustered into C2. Both the clustering results contain only one misclustered object, and they have the same object numbers in C1 and C2, respectively, so their RI and ARI values are identical.

However, the positions of the misclustered objects are different between the two results, leading to distinct spRI and spARI shown in Figure 1(a). Specifically, in clustering result i, the type "s×g" object pairs consist of (5,7), (5,8), (5,9), (5,10), (5,11), and (5,12), giving $F_{s\times g} = 3.143$; and the type "g×s" object pairs include (1,5), (2,5), (3,5), (4,5), and (5, 6), giving $H_{g\times s} = 0.923$. In contrast, in clustering result ii, the type "s×g" object pairs, which are (2,7), (2,8), (2,9), (2,10), (2,11), and (2,12), result in a smaller value of $F_{s\times g} = 2.065$ than that of clustering i because object number 2 is more far away from other objects in cluster C2 compared to object number 5. Type "g×s" object pairs in clustering ii are (1,2), (2,3), (2,4), (2,5), and (2,6), providing an identical $H_{g\times s} = 0.923$ to that in clustering i due to the location symmetry of object number 2 and object number 5 in reference group R1.

Based on the observations above, it is easy to show that the spRI value of clustering result i (0.895) is higher than that of clustering result ii (0.879). Since the expectations of spRI are the same for both the clustering results (i.e., $\eta = 0$ in Supplementary Proposition S1), clustering result i also achieves a higher spARI value.

4.2 Two clustering results with identical $F_{s\times g}$ and different $H_{g\times s}$

In this case, we consider a total of 10 objects. In the reference partition in Figure 1(b), objects numbered 2, 4, 5, 7, 8, and 10 belong to group R1, while objects numbered 1, 3, 6, and 9 in the middle part belong to R2. Similar to the previous case, each of the two clustering results has only one object that is clustered incorrectly in the middle and right panels of Figure 1(b), and they share the same object numbers in C1 and C2, respectively, resulting in their identical RI and ARI values.

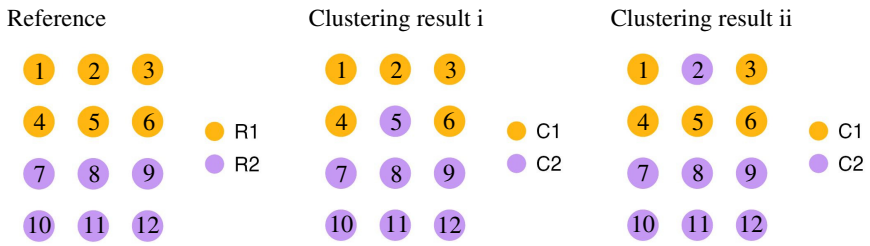
In clustering result i, the type "s×g" object pairs include (2,9), (4,9), (5,9), (7,9), (8,9), and (9,10), with a corresponding sum of weights $F_{s\times g} = 3.160$. The type "g×s" object pair comprises (1,9), (3,9), and (6,9), offering a sum of weights $H_{g\times s} = 0.877$. In the meanwhile, in clustering result ii, the type "s×g" object pairs (2,3), (3,4), (3,5), (3,7), (3,8), and (3,10) have a sum of weights, $F_{s\times g} = 3.160$, the same as that in clustering result i, and the type "g×s" object pairs (1,3), (3,6), and (3,9) obtain a sum of weights, $H_{g\times s} = 0.455$, smaller than the counterpart in clustering result i. Thus, clustering result i has a higher spRI value than clustering result ii. Due to the equal expectations of spRI in both clustering results, clustering result i also exhibits a higher spARI value, according to Supplementary Proposition S1.

These two cases demonstrate that even when ARI values are the same, spARI can still effectively evaluate and distinguish various clustering partitions, also highlighting that spARI favors partitions with compact and continuous region structures.

4.3 A general clustering setting

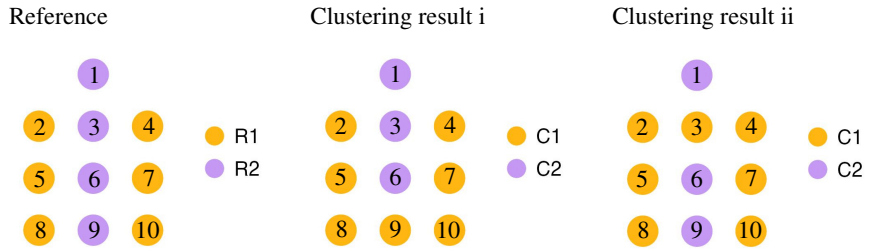
We expand the object number to 160 and consider a general setting where there are four reference groups, which mimics the layered structures observed in mouse brain tissue sections;

(a) Clustering results in the first simulation study



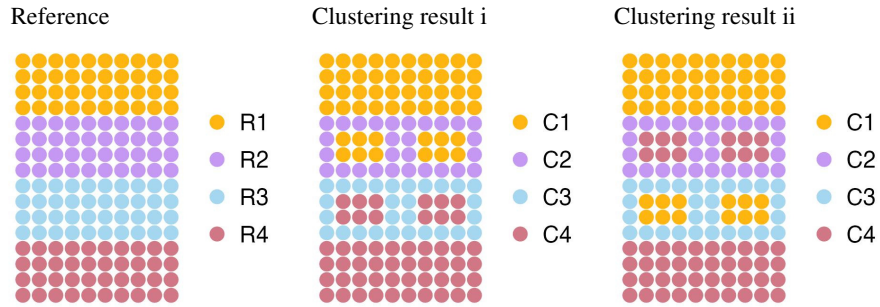
Metrics	Clustering result i	Clustering result ii
RI	0.833	0.833
ARI	0.665	0.665
spRI	**0.895**	0.879
spARI	**0.647**	0.592

(b) Clustering results in the second simulation study



Metrics	Clustering result i	Clustering result ii
RI	0.800	0.800
ARI	0.602	0.602
spRI	**0.890**	0.880
spARI	**0.627**	0.595

(c) Clustering results in the third simulation study



Metrics	Clustering result i	Clustering result ii
RI	0.872	0.872
ARI	0.664	0.664
spRI	**0.926**	0.917
spARI	**0.672**	0.632

Figure 1: The clustering results in (a) the first, (b) the second, and (c) the third simulation study. There are 12, 10, and 160 objects in the simulated data, respectively. The left, middle, and right panels correspond to the reference partitions of the objects, the clustering partitions by method i, and the clustering partitions by method ii, respectively. The tables at the bottom of panels demonstrate the specific values of RI, ARI, spRI, and spARI for the two clustering results.

see Figure 1(c). Each layer contains 40 objects. Two clustering results represent the partitions obtained by two ST clustering methods, as shown in the middle and right panels of Figures 1(c), respectively. Each of the two partitions contains 24 incorrectly clustered objects, and the number of objects in each cluster is identical for the two partitions, leading to equal values of RI and ARI. However, in clustering result i, these misclustered objects are spatially closer to objects in the same clusters than those in clustering result ii, so clustering result i has a larger $F_{s\times g}$ value. On the other hand, since the distances between misclustered objects and other objects that belong to their reference group remain the same in both clustering partitions, the two clustering results have the same $H_{g\times s}$ value. Therefore, the clustering result i achieves higher spRI and spARI values than clustering result ii, suggesting that clustering method i shows more efforts to keep regions integral than clustering method ii.

5 Real application

We apply six region segmentation methods—STAGATE (Dong and Zhang, 2022), SpaGCN (Hu et al., 2021), BayesSpace (Zhao et al., 2021), BANKSY (Singhal et al., 2024), BINRES (Yan and Luo, 2024), and IRIS (Ma and Zhou, 2024)—to a spot level ST dataset, and five cell typing methods—STAGATE, SpaGCN, BANKSY, BACT (Yan and Luo, 2025), and BASS (Li and Zhou, 2022)—to a single-cell level ST dataset, and their ARI and spARI values are then calculated and compared. We remark that this application does not perform a benchmarking study for existing ST clustering but demonstrates the advantages of spARI in capturing spatial information of objects over ARI.

5.1 Region segmentation evaluation in human dorsolateral prefrontal cortex data

The human dorsolateral prefrontal cortex (DLPFC) is a functional area of the human brain involved in cognitive control and attention regulation. The dataset contains twelve DLPFC tissue sections with manual annotations (reference) provided by Maynard et al. (2021). The DLPFC section 151509 is used for downstream analysis, which contains 4789 spots and 33538 genes. The upper panel of Figure 2(a) shows the bar plots of the ARI and spARI values for each method. Based on ARI, the method performance ranking in this dataset is IRIS, SpaGCN, STAGATE, BINRES, BANKSY, and BayesSpace, while the ranking becomes IRIS, STAGATE, SpaGCN, BINRES, BayesSpace, and BANKSY using spARI. Specifically, BANKSY has a higher ARI value (0.357) than BayesSpace (0.343), but its spARI (0.342) is lower than BayesSpace (0.359).

To clarify the discrepancy between ARI and spARI for BayesSpace and BANKSY, we visualize their spatial partitions in the lower middle and right panels of Figure 2(a) to provide further insights. We also show the number of agreement pairs denoted by #{s×s, g×g}, the sum of weights f_{ij} for type "s×g" object pairs ($F_{s\times g}$), the sum of weights h_{ij} for type "g×s" object pairs ($H_{g\times s}$), spRI, $\mathbb{E}$(spRI), and spARI in the first two rows of Table 3.

Based on the reference illustrated in the lower left panel of Figure 2(a), we observe that both BayesSpace and BANKSY identify the major part of Layer 1, but exhibit unsatisfactory

Table 3: Related quantities used to calculate the spARI for BayesSpace and BANKSY in the analysis of DLPFC section 151509 data and for STAGATE and SpaGCN in the analysis of STARmap* data.

Data	Methods	$\#\{s\times s, g\times g\}/\binom{n}{2}$	$H_{g\times s}/\binom{n}{2}$	$F_{s\times g}/\binom{n}{2}$	spRI	$\mathbb{E}$(spRI)	spARI
DLPFC	BayesSpace	0.7760	**0.0301**	**0.0550**	**0.8611**	0.7832	**0.3594**
	BANKSY	**0.7874**	0.0294	0.0420	0.8588	**0.7853**	0.3421
STARmap*	STAGATE	0.8684	**0.0103**	**0.0509**	0.9296	0.8849	0.3884
	SpaGCN	**0.9037**	0.0094	0.0251	**0.9382**	**0.8947**	**0.4135**

performances in detecting Layers 2 and 3. For Layers 4, 5 and 6, neither method captures the complete three-layer structure well. Quantitatively, the first two rows of Table 3 show that BANKSY has a larger number of agreement pairs than BayesSpace. Nevertheless, the dashed domains in the lower right panel of Figure 2(a) tell us that regions C3 and C7 in the BANKSY result are spatially fragmented. This fragmentation leads to a smaller value of $H_{g\times s}$ for the type "g×s" object pairs and also a smaller value of $F_{s\times g}$ for the type "s×g" object pairs than BayesSpace. In contrast, the region structures in BayesSpace's partitions look more integral. Consequently, although BayesSpace has a lower ARI, BayesSpace can attain a larger spARI score (0.3593) than BANKSY (0.3420) by effectively incorporating the spatial information. These findings are further supported by the biological interpretations provided in Web Appendix I. This indicates that spARI prefers clustering methods that can yield more compact spatial structures.

5.2 Cell typing evaluation in mouse visual cortex STARmap* data

STARmap* is an imaging-based ST technique that captures gene expression patterns of single cells (Wang et al., 2018). This dataset consists of 1207 cells and 1020 genes. The ARI/spARI barplots of cell typing methods are demonstrated in the upper panel of Figure 2(a). The ranking of these methods using ARI is the same as that using spARI, meaning that the evaluation results by ARI and spARI are consistent in this case. This also indicates that the incorporation of spatial information into ARI does not change the ranking of these cell typing methods. However, there are still some interesting observations. For example, when comparing SpaGCN and STAGATE, their ARI difference is 0.098, but the difference decreases to 0.026 using spARI.

To look into what makes this difference, we provide the reference and the clustering results by STAGATE and SpaGCN in the lower panels of Figure 2(b) as well as related quantities during the calculations for their spARI. We can observe that in the lower middle and right panels of Figure 2(b), the cell cluster boundaries in STAGATE (e.g., clusters C3 and C4) are more clear than those in SpaGCN. In other words, more spatially coherent regions are formed in STAGATE. Nevertheless, as spRI is a trade off between RI and spatial proximity, the spRI of STAGATE is still lower than that of SpaGCN (last two rows of Table 3). To put it another way, in the lower left panel of Figure 2(b), some reference cell types exhibit irregular and scattered distributions in the tissue section (e.g., Astro-2, Micro, and Endo) rather than forming well-partitioned layers. Consequently, the more compact clustering characteristic of STAGATE does not enable it to surpass SpaGCN in terms of spRI scores.

Moreover, according to Supplementary Proposition S1, SpaGCN has a larger spARI value

than STAGATE. Although the ranking of STAGATE and SpaGCN remains unchanged for ARI and spARI, the narrowed gap in spARI values compared to ARI indicates that spARI not only assesses the clustering accuracy but also captures the structural information of the resulting spatial clusters, demonstrating its utility for a thorough evaluation of the ST clustering.

6 Discussion

ARI has been widely used for comparing clustering methods, but it lacks spatial awareness, which prevents it from fully evaluating and comparing spatial clustering methods. In this paper, we propose the spatially aware adjusted Rand index, spARI, to evaluate the ST clustering methods. Built upon ARI that only uses the partitions to provide a clustering accuracy index, spARI can further incorporate the spatial information to assess the ST clustering accuracy. Specifically, the weights in spARI for disagreement object pairs depend on the the distances of objects instead of being identically zero in ARI. It is this feature that equips spARI with the ability to differentiate two ST clustering methods with the same ARI value and that makes spARI in favor of a more spatially coherent clustering. Related properties for spRI and spARI are provided and discussed. Simulation studies and real applications demonstrate that spARI not only considers the consistency of two partitions as reflected by ARI but also takes care of the ability to preserve continuous spatial structures. More importantly, the application of spARI is not restricted to the spatial transcriptomics field but actually can be adopted in any spatial clustering area, such as meteorological data clustering (Pan et al., 2024) and geographical data clustering (Kang et al., 2022) (see Web Appendix J). Therefore, we envision that spARI will be an effective and statistically sound index for any problems associated with spatial clustering evaluations.

Acknowledgements

We thank Editor, Associate Editor, and reviewers for their insightful suggestions that have significantly improved the quality of this paper. This research was supported by National Natural Science Foundation of China with grant numbers 12471275 and 72271060.

Supplementary Materials

Web Appendices, Figures, code and data referenced in the paper are available. The R package to compute spRI and spARI is available at https://github.com/yinqiaoyan/spARI.

Data Availability

All data used in this paper are available in the supplementary materials.

References

Chidester, B., T. Zhou, S. Alam, and J. Ma (2023). SPICEMIX enables integrative single-cell spatial modeling of cell identity. *Nature Genetics 55*(1), 78–88.

Dong, K. and S. Zhang (2022). Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications 13*(1), 1–12.

Guo, Y., B. Zhu, C. Tang, R. Rong, Y. Ma, G. Xiao, et al. (2024). BayeSMART: Bayesian clustering of multi-sample spatially resolved transcriptomics data. *Briefings in Bioinformatics 25*(6), bbae524.

Hu, J., X. Li, K. Coleman, A. Schroeder, N. Ma, D. J. Irwin, et al. (2021). SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods 18*(11), 1342–1351.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*, 193–218.

Kang, Y., K. Wu, S. Gao, I. Ng, J. Rao, S. Ye, et al. (2022). STICC: a multivariate spatial clustering method for repeated geographic pattern discovery with consideration of spatial contiguity. *International Journal of Geographical Information Science 36*(8), 1518–1549.

Li, Z. and X. Zhou (2022). BASS: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. *Genome Biology 23*(1), 168.

Ma, Y. and X. Zhou (2022). Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature Biotechnology 40*(9), 1349–1359.

Ma, Y. and X. Zhou (2024). Accurate and efficient integrative reference-informed spatial domain detection for spatial transcriptomics. *Nature Methods 21*(7), 1231–1244.

Maynard, K. R., L. Collado-Torres, L. M. Weber, C. Uytingco, B. K. Barry, S. R. Williams, et al. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience 24*(3), 425–436.

Pan, T., W. Shen, and G. Hu (2024). Clustering spatial functional data using a geographically weighted Dirichlet process. *Canadian Journal of Statistics 52*(3), 696–712.

Park, H.-E., S. H. Jo, R. H. Lee, C. P. Macks, T. Ku, J. Park, et al. (2023). Spatial transcriptomics: technical aspects of recent developments and their applications in neuroscience and cancer research. *Advanced Science 10*(16), 2206939.

Pitman, E. (1937). Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society Series B: Statistical Methodology 4*(1), 119–130.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association 66*(336), 846–850.

Shi, X., Y. Yang, X. Ma, Y. Zhou, Z. Guo, C. Wang, et al. (2023). Probabilistic cell/domain-type assignment of spatial transcriptomics data with SpatialAnno. *Nucleic Acids Research 51*(22), e115–e115.

Singhal, V., N. Chou, J. Lee, Y. Yue, J. Liu, W. K. Chock, et al. (2024). BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis. *Nature Genetics 56*(3), 431–441.

Song, D., K. Li, X. Ge, and J. J. Li (2023). ClusterDE: a post-clustering differential expression method robust to false-positive inflation caused by double dipping. *bioRxiv*.

Teng, H., Y. Yuan, and Z. Bar-Joseph (2022). Clustering spatial transcriptomics data. *Bioinformatics 38*(4), 997–1004.

Wan, X., J. Xiao, S. S. T. Tam, M. Cai, R. Sugimura, Y. Wang, et al. (2023). Integrating spatial and single-cell transcriptomics data using deep generative models with SpatialScope. *Nature Communications 14*(1), 7848.

Wang, X., W. E. Allen, M. A. Wright, E. L. Sylwestrak, N. Samusik, S. Vesuna, et al. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science 361*(6400), eaat5691.

Williams, C. G., H. J. Lee, T. Asatsuma, R. Vento-Tormo, and A. Haque (2022). An introduction to spatial transcriptomics for biomedical research. *Genome Medicine 14*(1), 68.

Williams, C. K. and C. E. Rasmussen (2006). *Gaussian processes for machine learning*, Volume 2. MIT press Cambridge, MA.

Wu, Z. and H. Wu (2020). Accounting for cell type hierarchy in evaluating single cell RNA-seq clustering. *Genome Biology 21*(1), 123.

Xu, K., Y. Lu, S. Hou, K. Liu, Y. Du, M. Huang, et al. (2024). Detecting anomalous anatomic regions in spatial transcriptomics with STANDS. *Nature Communications 15*(1), 8223.

Yan, Y. and X. Luo (2024). Bayesian integrative region segmentation in spatially resolved transcriptomic studies. *Journal of the American Statistical Association 119*(547), 1709–1721.

Yan, Y. and X. Luo (2025). BACT: nonparametric Bayesian cell typing for single-cell spatial transcriptomics data. *Briefings in Bioinformatics 26*(1), bbae689.

Yuan, Z., F. Zhao, S. Lin, Y. Zhao, J. Yao, Y. Cui, et al. (2024). Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nature Methods 21*(4), 712–722.

Zhao, E., M. R. Stone, X. Ren, J. Guenthoer, K. S. Smythe, T. Pulliam, et al. (2021). Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology 39*(11), 1375–1384.

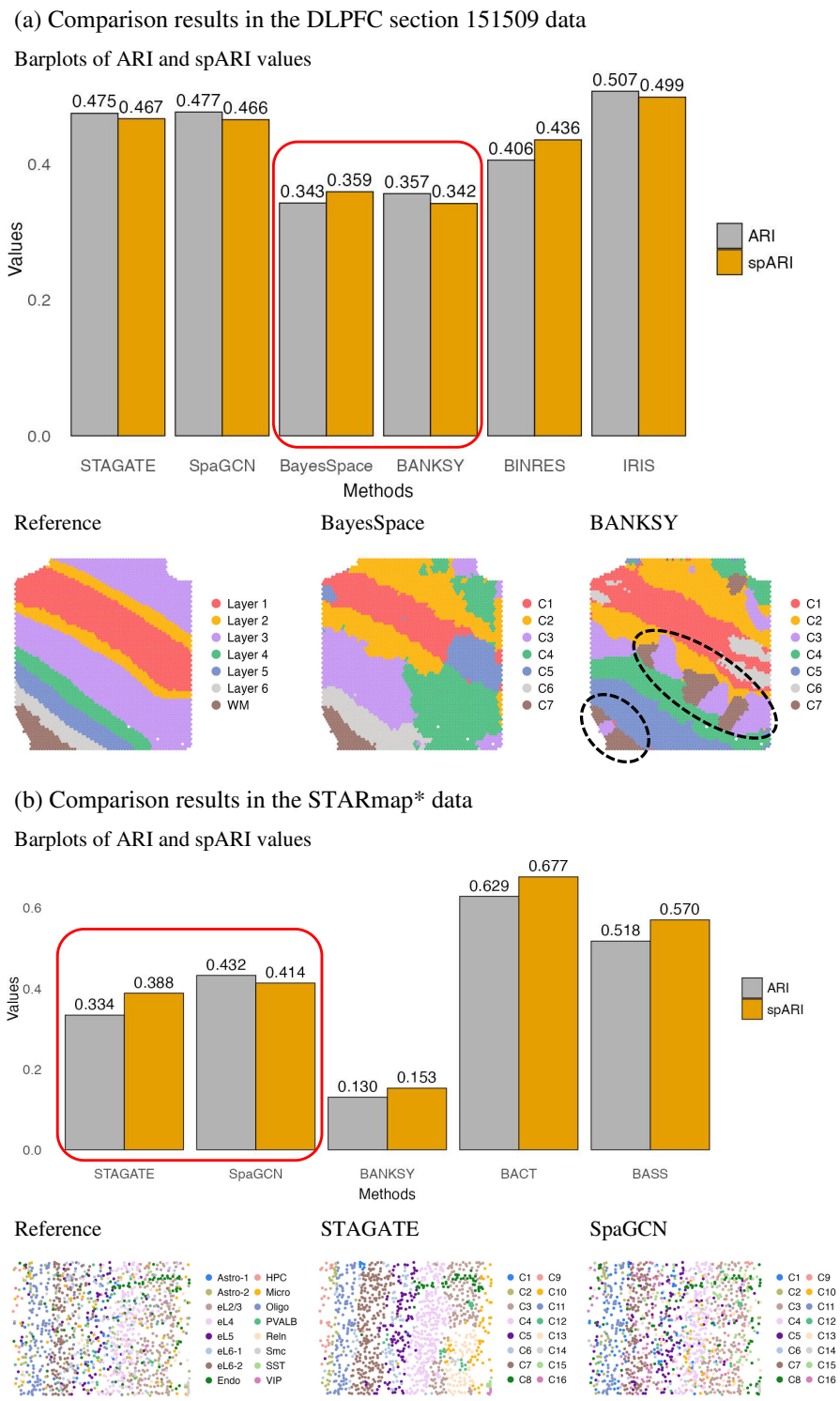


Figure 2: (a) The comparison results in the DLPFC section 151509 data. The upper panel displays the barplots of ARI and spARI values for the six region segmentation methods—STAGATE, SpaGCN, BayesSpace, BANKSY, BINRES, and IRIS. The ARI/spARI values are displayed at the top of bars. The lower panels from left to right show the reference annotations of the DLPFC 151509 section (including six DLPFC layers and the white matter (WM)), the region segmentation by BayesSpace, and the region segmentation by BANKSY. (b) The comparison results in the STARmap* data. The upper panel shows the barplots of ARI and spARI values for the five cell typing methods—STAGATE, SpaGCN, BANKSY, BACT, and BASS. The ARI/spARI values are displayed at the top of bars. The lower panels from left to right show the reference annotations of the STARmap* data (including sixteen cell types), the cell typing by STAGATE, and the cell typing by SpaGCN.

Supplementary Materials for "Spatially Aware Adjusted Rand Index for Evaluating Spatial Transcriptomics Clustering"

by

Yinqiao Yan, Xiangnan Feng and Xiangyu Luo

Web Appendix A: Proof of the sufficient and necessary conditions for $\mathbb{E}(\mathrm{RI}) = 1$

Proof. We first define that $\binom{1}{2} = 0$ and observe that $n_{r\cdot} \geq 1$ and $n_{\cdot c} \geq 1$. By definition, $\mathbb{E}(\mathrm{RI}) = 1$ if and only if

$$2\sum_{r=1}^{R}\binom{n_{r\cdot}}{2}\sum_{c=1}^{C}\binom{n_{\cdot c}}{2}\Big/\binom{n}{2}^2 = \left[\sum_{r=1}^{R}\binom{n_{r\cdot}}{2} + \sum_{c=1}^{C}\binom{n_{\cdot c}}{2}\right]\Big/\binom{n}{2}. \tag{1}$$

Multiplying both sides of Equation (1) by $\binom{n}{2}^2$ yields

$$2\sum_{r=1}^{R}\binom{n_{r\cdot}}{2}\sum_{c=1}^{C}\binom{n_{\cdot c}}{2} = \binom{n}{2}\left[\sum_{r=1}^{R}\binom{n_{r\cdot}}{2} + \sum_{c=1}^{C}\binom{n_{\cdot c}}{2}\right]. \tag{2}$$

Next, we examine whether this equation holds under the following four distinct cases.

Case 1. $R = 1$ and $C = 1$. In this case, $n_{r\cdot} = n_{\cdot c} = n$, and hence Equation (2) holds directly.

Case 2. $R = 1$ and $C > 1$. Here, $n_{r\cdot} = n$. Equation (2) then implies that $\sum_{c=1}^{C}\binom{n_{\cdot c}}{2} = \binom{n}{2}$. However, it is straightforward to verify that the inequality $\sum_{c=1}^{C}\binom{n_{\cdot c}}{2} < \binom{n}{2}$ holds whenever $C > 1$, which leads to a contradiction. Thus, Equation (2) does not hold in this case.

Case 3. $R > 1$ and $C = 1$. Similar to Case 2, we obtain that $n_{\cdot c} = n$ and Equation (2) implies that $\sum_{r=1}^{R}\binom{n_{r\cdot}}{2} = \binom{n}{2}$, which again contradicts the inequality $\sum_{r=1}^{R}\binom{n_{r\cdot}}{2} < \binom{n}{2}$ when $R > 1$. Therefore, Equation (2) does not hold in this case either.

Case 4. $R > 1$ and $C > 1$. Let $A = \sum_{r=1}^{R}\binom{n_{r\cdot}}{2}$ and $B = \sum_{c=1}^{C}\binom{n_{\cdot c}}{2}$, both nonnegative and strictly less than $\binom{n}{2}$. Then the following two inequalities hold: $2\min(A, B) \leq A + B$

and $\max(A, B) < \binom{n}{2}$, which yields

$$2AB = 2\min(A, B)\max(A, B) \leq (A + B)\binom{n}{2}.$$

The equality holds (i.e., Equation (2)) if and only if $A = 0$ and $B = 0$, which are satisfied only when $n_{r\cdot} = 1$ for all $1 \leq r \leq R$ and $n_{\cdot c} = 1$ for all $1 \leq c \leq C$.

Therefore, we conclude that $\mathbb{E}(\mathrm{RI}) = 1$ if and only if one of the following two conditions holds: (i) the reference $\mathbf{U}$ has only one group and all the objects are in one cluster of $\mathbf{V}$, i.e., $R = C = 1$ and $n_{1\cdot} = n_{\cdot 1} = n$; or (ii) each cluster has only one object in both $\mathbf{U}$ and $\mathbf{V}$, i.e., $R = C = n$ and $n_{r\cdot} = n_{\cdot c} = 1$ for all $1 \leq r \leq R$ and $1 \leq c \leq C$. □

Web Appendix B: Proof of the sufficient and necessary conditions for $\mathbb{E}(\mathrm{spRI}) = 1$

Proof. We recall that the expectation of spRI is formulated as:

$$\mathbb{E}(\mathrm{spRI}) = \mathbb{E}(\mathrm{RI}) + \frac{Fq + Hp + (H + F)pq}{\binom{n}{2}},$$

where $\mathbb{E}(\mathrm{RI}) = 1 + 2pq - (p + q)$, $F = \sum_{1 \leq i < j \leq n} f_{ij}$ and $H = \sum_{1 \leq i < j \leq n} h_{ij}$. Then $\mathbb{E}(\mathrm{spRI}) = 1$ holds if and only if

$$2pq - (p + q) + \frac{Fq + Hp - (H + F)pq}{\binom{n}{2}} = 0. \tag{3}$$

We consider the following four cases to identify whether this equality holds.

Case 1. $R = 1$ and $C = 1$. In this setting, $n_{r\cdot} = n_{\cdot c} = n$ and $p = q = 1$. We then obtain that Equation (3) holds directly.

Case 2. $R = 1$ and $C > 1$. We have $n_{r\cdot} = n$, $p = 1$ and $q \in [0, 1)$. Substituting these values into Equation (3) gives $q - 1 + H(1 - q)/\binom{n}{2} = 0$, which implies that $H = \binom{n}{2}$. This contradicts the fact that $H < \binom{n}{2}$ when $C > 1$, so Equation (3) does not hold in this case.

Case 3. $R > 1$ and $C = 1$. We similarly obtain that $n_{\cdot c} = n$, $q = 1$ and $p \in [0, 1)$, and Equation (3) implies that $p - 1 + F(1 - p)/\binom{n}{2} = 0$, which again leads to a contradiction, since $F < \binom{n}{2}$ when $R > 1$. Thus, Equation (3) does not hold in this case either.

Case 4. $R > 1$ and $C > 1$. In this case, $p, q \in [0, 1)$. Through Equation (3), q can be expressed as a function of p

$$q(p) = \frac{\left(1 - H/\binom{n}{2}\right) p}{\left(2 - (H + F)/\binom{n}{2}\right) p - \left(1 - F/\binom{n}{2}\right)}.$$

To ensure that the denominator is nonzero, p must not equal $\frac{1 - F/\binom{n}{2}}{2 - (H+F)/\binom{n}{2}}$, which is denoted by p^*. Given that $F < \binom{n}{2}$ and $H < \binom{n}{2}$, it follows that $0 < 1 - F/\binom{n}{2} \leq 1$ and

$0 < 1 - H/\binom{n}{2} \leq 1$, and it is easy to verify that $0 < p^* < 1$. The derivative of $q(p)$ is given by

$$q'(p) = \frac{-\left(1 - F/\binom{n}{2}\right)\left(1 - H/\binom{n}{2}\right)}{\left[\left(2 - (H+F)/\binom{n}{2}\right)p - \left(1 - F/\binom{n}{2}\right)\right]^2},$$

which is negative in the domain of $q(p)$, indicating that q is decreasing on each of the intervals $[0, p^*)$ and $(p^*, 1)$ with respective to p.

Moreover, from the form of $q(p)$, we observe that $\lim_{p\to 1-} q(p) = 1$ and $\lim_{p\to 0+} q(p) = 0$. Since both p and q are restricted to the interval $[0, 1)$, the only point in this domain that satisfies the expression of $q(p)$ is $p = q = 0$, under which Equation (3) is satisfied.

Thus, $\mathbb{E}(\text{spRI}) = 1$ holds if and only if one of the following two conditions holds: (i) the reference $\mathbf{U}$ consists of a single group and all objects in $\mathbf{V}$ belong to one cluster, i.e., $R = C = 1$ and $n_{1\cdot} = n_{\cdot 1} = n$ ($p = q = 1$); or (ii) every cluster in both $\mathbf{U}$ and $\mathbf{V}$ contains only one object, i.e., $R = C = n$ and $n_{r\cdot} = n_{\cdot c} = 1$ for all $1 \leq r \leq R$ and $1 \leq c \leq C$ ($p = q = 0$).

□

Web Appendix C: Consistency between spRI and spARI in method comparison

We aim to determine whether a method with a higher spRI can consistently yield a higher spARI compared to another method. The following proposition answers the question.

Proposition S1. *Consider two clustering partitions $\mathbf{V}_1$ and $\mathbf{V}_2$. We denote the spRI and spARI values of $\mathbf{V}_1$ by $spRI_1$ and $spARI_1$, respectively, and those of $\mathbf{V}_2$ by $spRI_2$ and $spARI_2$, respectively. Let $\xi = spRI_1 - spRI_2$ and $\eta = \mathbb{E}(spRI_1) - \mathbb{E}(spRI_2)$, then $spARI_1 > spARI_2$ if and only if $\xi > \frac{1-spRI_2}{1-\mathbb{E}(spRI_2)}\eta$.*

Proof. Using the notations ξ and η, we can express spARI_1 in the following form

$$\text{spARI}_1 = \frac{\text{spRI}_1 - \mathbb{E}(\text{spRI}_1)}{1 - \mathbb{E}(\text{spRI}_1)} = \frac{\text{spRI}_2 - \mathbb{E}(\text{spRI}_2) + (\xi - \eta)}{1 - \mathbb{E}(\text{spRI}_2) - \eta}.$$

Accordingly, $\text{spARI}_1 > \text{spARI}_2$ if and only if

$$\frac{\text{spRI}_2 - \mathbb{E}(\text{spRI}_2) + (\xi - \eta)}{1 - \mathbb{E}(\text{spRI}_2) - \eta} > \frac{\text{spRI}_2 - \mathbb{E}(\text{spRI}_2)}{1 - \mathbb{E}(\text{spRI}_2)}. \tag{4}$$

As mentioned in the main text, unless in the cases $R = C = 1$ or $R = C = n$, the expected value of spRI is strictly less than one, implying that $1 - \mathbb{E}(\text{spRI}_1)$ and $1 - \mathbb{E}(\text{spRI}_2)$ are strictly positive. Next, we multiply both sides in inequality (4) by $(1 - \mathbb{E}(\text{spRI}_1))(1 - \mathbb{E}(\text{spRI}_2))$, resulting in

$$[\text{spRI}_2 - \mathbb{E}(\text{spRI}_2) + (\xi - \eta)]\,[1 - \mathbb{E}(\text{spRI}_2)] > [\text{spRI}_2 - \mathbb{E}(\text{spRI}_2)]\,[1 - \mathbb{E}(\text{spRI}_2) - \eta]\,. \tag{5}$$

After simple algebra and using the positivity of $1 - \mathbb{E}(\text{spRI}_2)$, we can obtain that the inequality

(5) holds if and only if

$$\xi > \frac{1 - \mathrm{spRI}_2}{1 - \mathbb{E}(\mathrm{spRI}_2)}\eta.$$

□

Without loss of generality, we assume that the clustering partition $\mathbf{V}_1$ has a higher spRI value than $\mathbf{V}_2$, i.e., $\xi > 0$. Based on Proposition S1, when $\eta \leq 0$, it is evident that $\mathbf{V}_1$ also gives a higher spARI value. When $\eta > 0$, $\mathbf{V}_1$ has a higher spARI value if $\xi > \frac{1-\mathrm{spRI}_2}{1-\mathbb{E}(\mathrm{spRI}_2)}\eta$ is satisfied. This provides a practical guideline about when spRI and spARI can reach the same evaluation conclusion for two ST clustering methods.

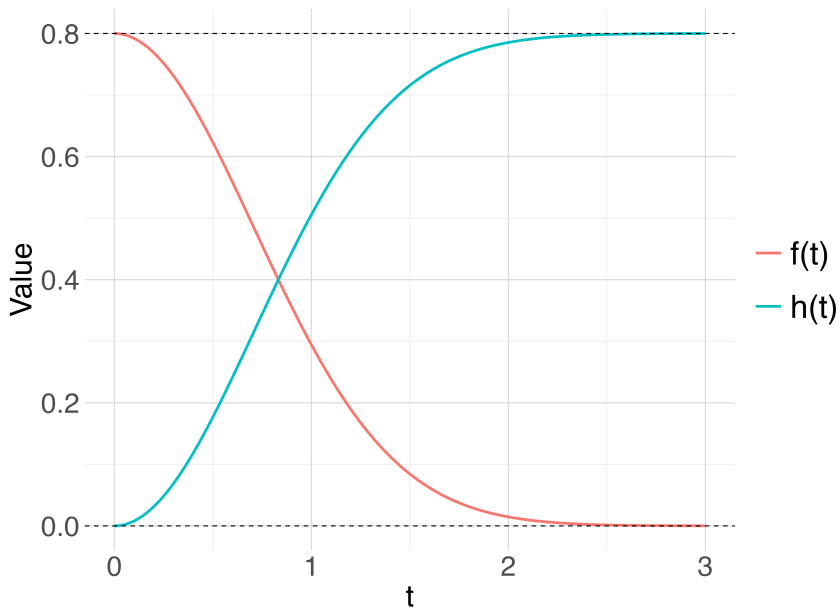


Figure S1: The function curves of $f(t) = \alpha \exp\{-t^2\}$ and $h(t) = \alpha(1 - \exp\{-t^2\})$.

Web Appendix D: Sensitivity analysis for α

Regarding the parameter α, we conducted a sensitivity analysis based on the third simulation study (Section 4.3 in the main text) to assess its impact on the spARI value. Specifically, we varied α over the set $\{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and computed the spARI value of the first clustering partition for each α. We also calculated the ARI value of the first clustering result, which does not change with α. Figure S2 illustrates the difference between spARI and ARI, and this difference becomes more noticeable as α increases. This phenomenon arises because α controls the gap between the maximal weight of disagreement object pairs and the weight one of agreement pairs. When α approaches zero, the contribution of disagreement pairs to spARI becomes negligible, so spARI becomes close to ARI. In contrast, as α increases toward one, the accumulated weights of disagreement pairs become larger, resulting in a relatively large difference between spARI and ARI in this example. In practice, we set the default value of α to 0.8 to ensure a reasonable positive gap between the weights of disagreement and agreement pairs, which can also capture meaningful spatial information.

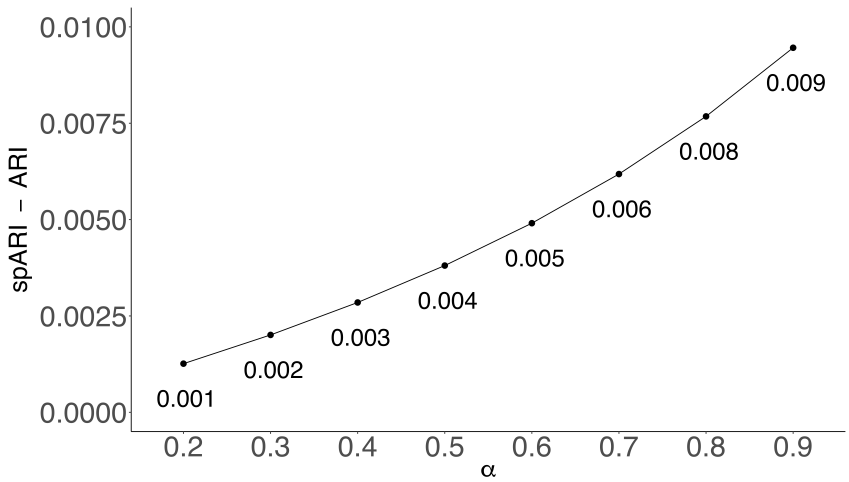


Figure S2: The differences between spARI and ARI of the first clustering result in the third simulation study in the main text, and the parameter α varies from 0.2 to 0.9 with gap 0.1.

Web Appendix E: Analytical forms of spRI between two similar clusterings

Specifically, for any two groups U_r and $U_{r'}$ $(1 \leq r, r' \leq k)$ in the reference, we further assume that the limit of the arithmetic mean of f_{ij} or h_{ij} on U_r and $U_{r'}$ exist when the object size m per group increases. In other words, $\frac{1}{\#\{i \in U_r, j \in U_{r'}\}} \sum_{i \in U_r, j \in U_{r'}} f_{ij}$ converges to $\mu_{f,U}(r, r')$, and $\frac{1}{\#\{i \in U_r, j \in U_{r'}\}} \sum_{i \in U_r, j \in U_{r'}} h_{ij}$ to $\mu_{h,U}(r, r')$, as m goes to positive infinity.

In the first circumstance, we consider the following four operations on the reference partition $\mathbf{U}$ to generate four different clustering partitions $\mathbf{V}$, respectively, as illustrated in Figure S3. For each operation, we demonstrate its specific spRI value between $\mathbf{U}$ and $\mathbf{V}$, as well as its limits when either m or k tends to infinity in the rows (A1)-(A4) in Table S1.

(A1) Merge two groups U_r and $U_{r'}$ into a single cluster.

(A2) Split one group in $\mathbf{U}$ into two clusters V_c and $V_{c'}$ with each containing an equal number of objects (assuming m is even).

(A3) Split one group U_r into m clusters, each containing exactly one object.

(A4) Select one object from each group in $\mathbf{U}$, and assign them together to a new cluster of size k. Let $U(i)$ denote the original group in $\mathbf{U}$ to which object i belongs. Let ϕ_V be $\lim_k 2/k^2 \sum_{i \neq j \in V_{k+1}} f_{ij}$, which is assumed to exist.

In the second circumstance, we apply the three following types of operations to the reference partition $\mathbf{U}$, and each operation generates two different clustering results $\mathbf{V}_1$ and $\mathbf{V}_2$ on different parts of $\mathbf{U}$, as illustrated in Figure S4. We then compute the spRI values between $\mathbf{V}_1$ and $\mathbf{V}_2$ as well as its limits in the rows (B1)-(B3) in Table S1.

(B1) Move an object a into two different reference groups, U_s and U_t, respectively. Denote the clusters containing object a in $\mathbf{V}_1$ and $\mathbf{V}_2$ by V_{1c} and $V_{2c'}$, respectively.

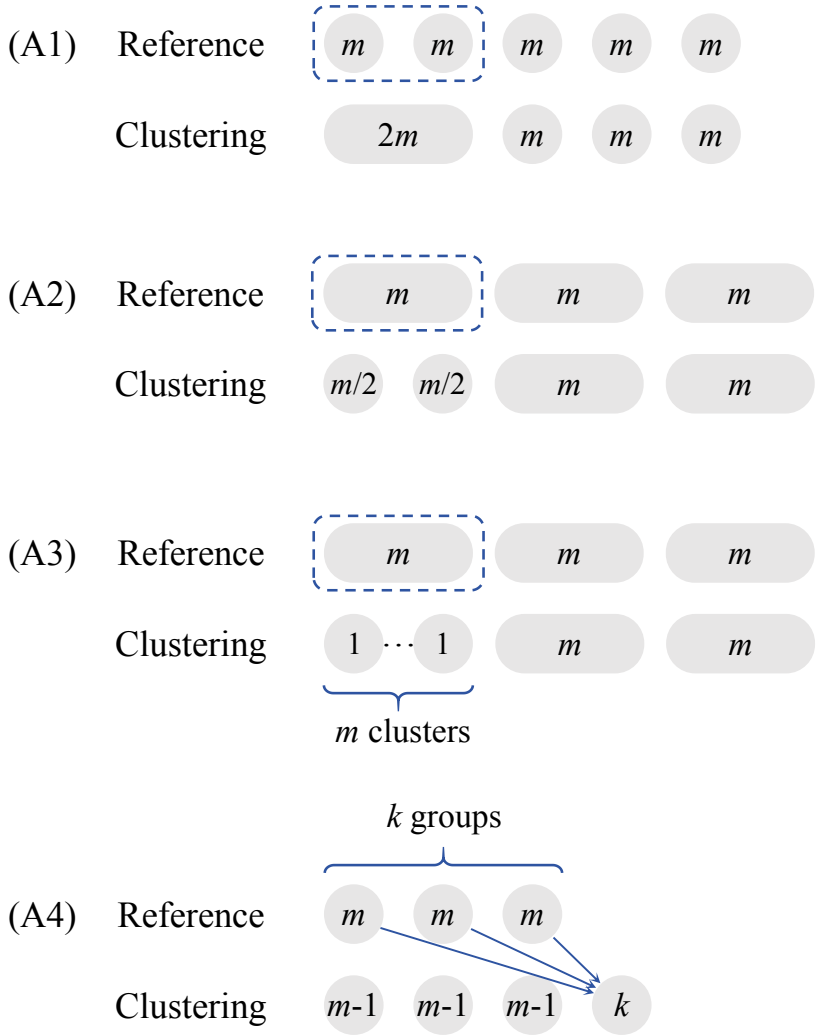


Figure S3: Illustrations of the four operations (A1)–(A4) applied to the reference partition, each resulting in a distinct clustering partition.

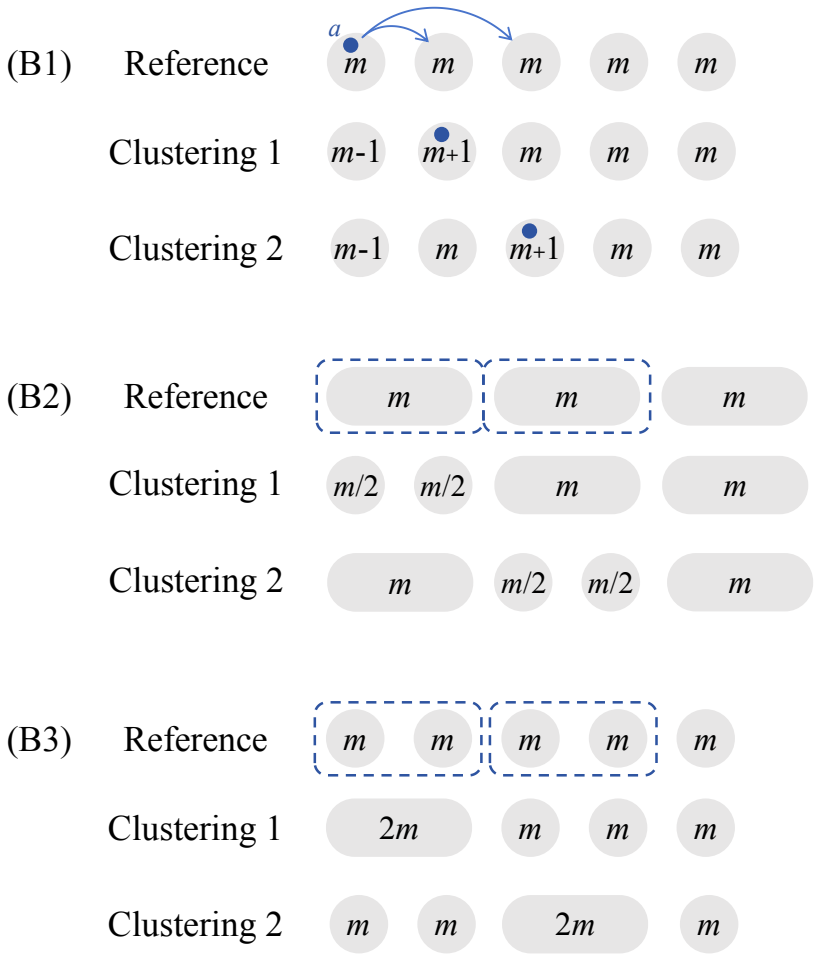


Figure S4: Illustrations of the three operations (B1)–(B3) applied to the reference partition, each resulting in two distinct clustering partitions.

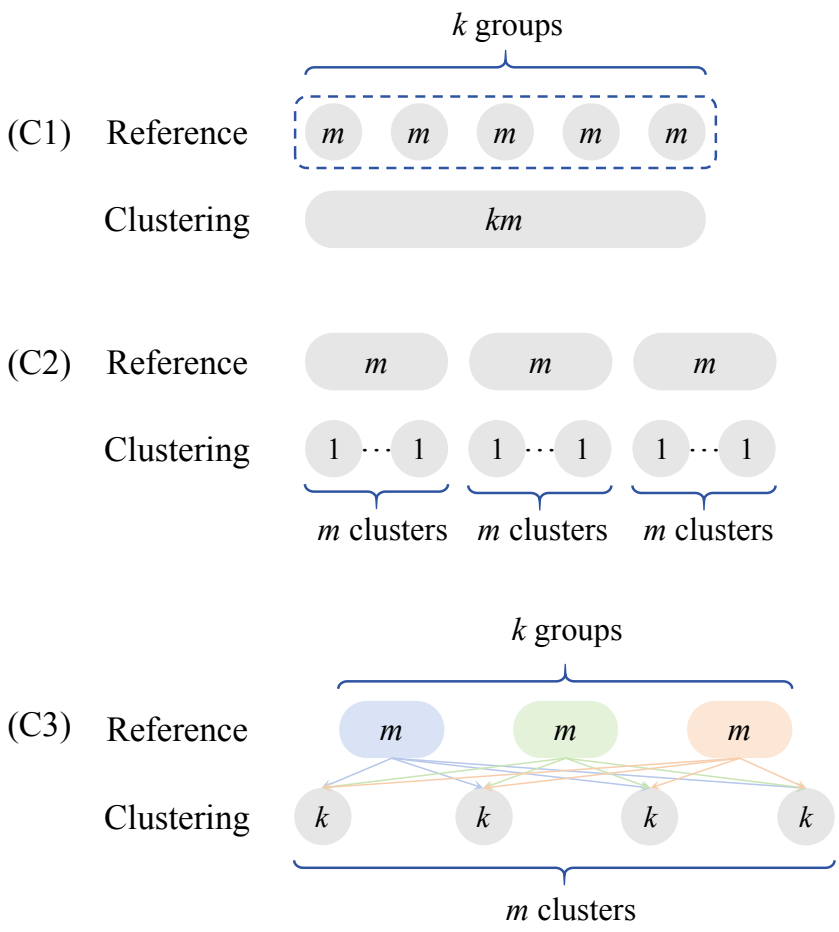


Figure S5: Illustrations of the three operations (C1)–(C3) applied to the reference partition, each resulting in a distinct clustering partition.

Table S1: Forms of spRI and its limits for operation cases (A1)-(A4), (B1)-(B3), and (C1)-(C3).

Cases	spRI	$m \to +\infty$	$k \to +\infty$
(A1)	$\frac{k^2-2-\frac{k}{m}+\frac{2}{m^2}\sum_{i\in U_r, j\in U_{r'}} f_{ij}}{k^2-\frac{k}{m}}$	$\frac{k^2-2+2\mu_{f,U}(r,r')}{k^2}$	1
(A2)	$\frac{(2k^2-1)-\frac{2k}{m}+\frac{4}{m^2}\sum_{i\in V_c, j\in V_{c'}} h_{ij}}{2k^2-\frac{2k}{m}}$	$\frac{2k^2-1+\mu_{h,V}(c,c')}{2k^2}$	1
(A3)	$\frac{k^2-1-\frac{k-1}{m}+\frac{2}{m^2}\sum_{i\neq j\in U_r} h_{ij}}{k^2-\frac{k}{m}}$	$\frac{k^2-1+\mu_{h,U}(r,r)}{k^2}$	1
(A4)	$\frac{m^2-\frac{3m}{k}-1+\frac{3}{k}+\frac{2}{k^2}\sum_{i\neq j\in V_{k+1}} f_{ij}+\frac{2}{k^2}\sum_{i\in V_{k+1}, j\in U(i)\setminus\{i\}} h_{ij}}{m^2-\frac{m}{k}}$	1	$\frac{m^2-1+\phi_V}{m^2}$
(B1)	$\frac{k^2m-k-4+\frac{2}{m}\sum_{j\in V_{2c'}\setminus\{a\}} f_{aj}+\frac{2}{m}\sum_{j\in V_{1c}\setminus\{a\}} h_{aj}}{k^2m-k}$	1	1
(B2)	$\frac{2k^2-\frac{2k}{m}-2+\frac{4}{m^2}\sum_{i\in V_{1c}, j\in V_{1c'}} f_{ij}+\frac{4}{m^2}\sum_{i\in V_{2c}, j\in V_{2c'}} h_{ij}}{2k^2-\frac{2k}{m}}$	$\frac{2k^2-2+\mu_{f,V_1}(c,c')+\mu_{h,V_2}(c,c')}{2k^2}$	1
(B3)	$\frac{k^2-\frac{k}{m}-4+\frac{2}{m^2}\sum_{i\in U_t, j\in U_{t'}} f_{ij}+\frac{2}{m^2}\sum_{i\in U_r, j\in U_{r'}} h_{ij}}{k^2-\frac{k}{m}}$	$\frac{k^2-4+2\mu_{f,U}(t,t')+2\mu_{h,U}(r,r')}{k^2}$	1
(C1)	$\frac{1-\frac{1}{m}+\frac{2}{k}\sum_{r<r'}\frac{1}{m^2}\sum_{i\in U_r, j\in U_{r'}} f_{ij}}{k-\frac{1}{m}}$	$\frac{1+\frac{2}{k}\sum_{r<r'}\mu_{f,U}(r,r')}{k}$	$\bar{\mu}_U$
(C2)	$\frac{k^2-k+\sum_r\frac{2}{m^2}\sum_{i\neq j\in U_r} h_{ij}}{k^2-\frac{k}{m}}$	$\frac{k^2-k+\sum_r\mu_{h,U}(r,r)}{k^2}$	1
(C3)	$\frac{(k-1)(m-1)+\frac{2}{km}[\sum_r\sum_{i\neq j\in U_r} h_{ij}+\sum_c\sum_{i\neq j\in V_c} f_{ij}]}{km-1}$	$\frac{k-1+\frac{1}{k}\sum_r\mu_{h,U}(r,r)}{k}$	$\frac{m-1+\frac{1}{m}\sum_c\mu_{f,V}(c,c)}{m}$

(B2) Split different groups in $\mathbf{U}$ into two clusters of equal size (assuming m is even), respectively. Denote the two clusters in $\mathbf{V}_1$ and $\mathbf{V}_2$ by $(V_{1c}, V_{1c'})$ and $(V_{2c}, V_{2c'})$, respectively.

(B3) Merge two distinct groups in $\mathbf{U}$. Denote the two groups that are merged into a single cluster in $\mathbf{V}_1$ by U_r and $U_{r'}$, and those that are merged in $\mathbf{V}_2$ by U_t and $U_{t'}$.

In the third circumstance, we consider the following three operations on the reference partition that result in a clustering partition $\mathbf{V}$, which are illustrated in Figure S5. For each operation, the values of spRI between $\mathbf{U}$ and $\mathbf{V}$ as well as its limits when either m or k tends to infinity are shown in the rows (C1)-(C3) in Table S1.

(C1) Merge all groups in the reference partition into a single cluster. We let

$$\bar{\mu}_U = \lim_k \frac{2}{k^2m^2}\sum_{r<r'}\sum_{i\in U_r, j\in U_{r'}} f_{ij},$$

which is assumed to exist.

(C2) Split all groups in the reference partition into km clusters, each containing one object.

(C3) Reorganize all objects into m "opposite" clusters, each containing k objects that come from the k different groups in the reference partition.

Web Appendix F: Calculations in the sensitivity of spRI

The spRI using the subsample $\tau\Omega$, denoted by spRI_τ, is

$$\text{spRI}_\tau = \frac{\frac{1}{2}n(n-\frac{1}{\tau}) + \sum_{r=1}^{R}\sum_{c=1}^{C} n_{rc}^2 - \frac{1}{2}\left[\sum_{r=1}^{R} n_{r\cdot}^2 + \sum_{c=1}^{C} n_{\cdot c}^2\right] + \sum_{\substack{i,j\in\tau\Omega \\ (i,j)\text{:type "s×g"}}} \frac{f_{ij}}{\tau^2} + \sum_{\substack{i,j\in\tau\Omega \\ (i,j)\text{:type "g×s"}}} \frac{h_{ij}}{\tau^2}}{\frac{1}{2}n(n-\frac{1}{\tau})}.$$

For the whole object set Ω, the spRI is

$$\text{spRI} = \frac{\frac{1}{2}n(n-1) + \sum_{r=1}^{R}\sum_{c=1}^{C} n_{rc}^2 - \frac{1}{2}\left[\sum_{r=1}^{R} n_{r\cdot}^2 + \sum_{c=1}^{C} n_{\cdot c}^2\right] + \sum_{\substack{i,j\in\Omega \\ (i,j)\text{:type "s×g"}}} f_{ij} + \sum_{\substack{i,j\in\Omega \\ (i,j)\text{:type "g×s"}}} h_{ij}}{\frac{1}{2}n(n-1)}.$$

Subsequently, the difference in spRI between the subsample $\tau\Omega$ and the whole sample Ω is given by

$$\text{spRI}_\tau - \text{spRI} = \frac{(\frac{1}{\tau}-1)\sum_{r=1}^{R}\sum_{c=1}^{C} n_{rc}^2 - \frac{1}{2}(\frac{1}{\tau}-1)\left[\sum_{r=1}^{R} n_{r\cdot}^2 + \sum_{c=1}^{C} n_{\cdot c}^2\right] + \Delta F_\tau + \Delta H_\tau}{\frac{1}{2}n(n-1)(n-\frac{1}{\tau})}, \tag{6}$$

where

$$\Delta F_\tau = \frac{n-1}{\tau^2} \sum_{\substack{i,j\in\tau\Omega \\ (i,j)\text{:type "s×g"}}} f_{ij} - (n-\frac{1}{\tau}) \sum_{\substack{i,j\in\Omega \\ (i,j)\text{:type "s×g"}}} f_{ij},$$

$$\Delta H_\tau = \frac{n-1}{\tau^2} \sum_{\substack{i,j\in\tau\Omega \\ (i,j)\text{:type "g×s"}}} h_{ij} - (n-\frac{1}{\tau}) \sum_{\substack{i,j\in\Omega \\ (i,j)\text{:type "g×s"}}} h_{ij}.$$

The spRI difference in Equation (6), $\text{spRI}_\tau - \text{spRI}$, measures the sensitivity of spRI to the subsampling proportion τ. As expected, it approaches zero as the proportion τ increases to one. Let $\Delta\bar{f}$ denote the difference between the mean value of f_{ij} over all type "s×g" pairs of objects in $\tau\Omega$ and that in Ω, and $\Delta\bar{h}$ represents the difference between the mean value of h_{ij} over the type "g×s" pairs of objects in $\tau\Omega$ and that in Ω. Recall that the type "s×g" object pair number and the type "g×s" object pair number in Ω are N_{21} and N_{12}, respectively. According to Equations (1) and (2) in the main text, the corresponding numbers of such pairs in $\tau\Omega$ are $\tau^2 N_{21}$ and $\tau^2 N_{12}$, respectively. The following proposition provides an upper bound on the spRI difference.

Proposition S2. *Assume that* $\frac{|N_{21}\Delta\bar{f}+N_{12}\Delta\bar{h}|}{n}$ *converges to zero as* n *increases to infinity. Then the subsampling approximation error is given by* $|spRI_\tau - spRI| \leq 3\frac{1/\tau-1}{n-1/\tau}$ *when* n *is large enough.*

Proof. The Equation (6) can be rewritten as

$$\text{spRI}_\tau - \text{spRI} = \frac{\frac{1}{\tau}-1}{n-\frac{1}{\tau}} \cdot \frac{\sum_{r=1}^{R}\sum_{c=1}^{C} n_{rc}^2 - \frac{1}{2}\left[\sum_{r=1}^{R} n_{r\cdot}^2 + \sum_{c=1}^{C} n_{\cdot c}^2\right]}{\frac{1}{2}n(n-1)} + \frac{\Delta F_\tau + \Delta H_\tau}{\frac{1}{2}n(n-1)(n-\frac{1}{\tau})},$$

We denote the first term on the right hand side of this equation as T_1 and the second term as T_2. By the definition of RI, we have $|T_1| \leq \frac{\frac{1}{\tau}-1}{n-\frac{1}{\tau}}(1-\text{RI})$. Next, we aim to derive an upper bound for T_2. Define that

$$\bar{f}_\Omega = \frac{1}{N_{21}} \sum_{\substack{i,j\in\Omega \\ (i,j)\text{:type “s×g”}}} f_{ij}, \quad \bar{f}_{\tau\Omega} = \frac{1}{\tau^2 N_{21}} \sum_{\substack{i,j\in\tau\Omega \\ (i,j)\text{:type “s×g”}}} f_{ij};$$
$$\bar{h}_\Omega = \frac{1}{N_{12}} \sum_{\substack{i,j\in\Omega \\ (i,j)\text{:type “g×s”}}} h_{ij}, \quad \bar{h}_{\tau\Omega} = \frac{1}{\tau^2 N_{12}} \sum_{\substack{i,j\in\tau\Omega \\ (i,j)\text{:type “g×s”}}} h_{ij}.$$

Since functions f and h take values in the interval $[0,\alpha]$, it is easy to verify that $\bar{f}_\Omega$, $\bar{f}_{\tau\Omega}$, $\bar{h}_\Omega$ and $\bar{h}_{\tau\Omega}$ all lie in the range of $[0,1]$. Then ΔF_τ and ΔH_τ can be reformulated as

$$\Delta F_\tau = N_{21}\left[(n-1)\bar{f}_{\tau\Omega} - (n-\frac{1}{\tau})\bar{f}_\Omega\right],$$
$$\Delta H_\tau = N_{12}\left[(n-1)\bar{h}_{\tau\Omega} - (n-\frac{1}{\tau})\bar{h}_\Omega\right].$$

Recall that $\Delta\bar{f} = \bar{f}_{\tau\Omega} - \bar{f}_\Omega$ and $\Delta\bar{h} = \bar{h}_{\tau\Omega} - \bar{h}_\Omega$. This gives us $\Delta F_\tau = N_{21}\left[(\frac{1}{\tau}-1)\bar{f}_\Omega + (n-1)\Delta\bar{f}\right]$ and $\Delta H_\tau = N_{12}\left[(\frac{1}{\tau}-1)\bar{h}_\Omega + (n-1)\Delta\bar{h}\right]$. Substituting these values into T_2, we obtain that

$$\begin{aligned} |T_2| &= \frac{|(\frac{1}{\tau}-1)(N_{21}\bar{f}_\Omega + N_{12}\bar{h}_\Omega) + (n-1)(N_{21}\Delta\bar{f} + N_{12}\Delta\bar{h})|}{\frac{1}{2}n(n-1)(n-\frac{1}{\tau})} \\ &\leq \frac{|(\frac{1}{\tau}-1)(N_{21}\bar{f}_\Omega + N_{12}\bar{h}_\Omega)|}{\frac{1}{2}n(n-1)(n-\frac{1}{\tau})} + \frac{|(n-1)(N_{21}\Delta\bar{f} + N_{12}\Delta\bar{h})|}{\frac{1}{2}n(n-1)(n-\frac{1}{\tau})}, \end{aligned} \tag{7}$$

where the inequality follows from the triangle inequality. Denote the first and second term on the right hand side of Inequality (7) as $|T_{21}|$ and $|T_{22}|$, respectively. For T_{21}, since $\bar{f}_\Omega, \bar{f}_{\tau\Omega}, \bar{h}_\Omega, \bar{h}_{\tau\Omega}$ are bounded in the interval $[0,1]$, we have

$$|T_{21}| = \frac{(\frac{1}{\tau}-1)}{(n-\frac{1}{\tau})} \cdot \frac{|N_{21}\bar{f}_\Omega + N_{12}\bar{h}_\Omega|}{\frac{1}{2}n(n-1)} \leq \frac{(\frac{1}{\tau}-1)}{(n-\frac{1}{\tau})}(1-\text{RI}),$$

where the last inequality follows from the definition of RI. For T_{22}, since we assume that $\lim_{n\to\infty}\frac{|N_{21}\Delta\bar{f}+N_{12}\Delta\bar{h}|}{n} = 0$, there exists a positive integer n_0 such that when $n > n_0$, $\frac{|N_{21}\Delta\bar{f}+N_{12}\Delta\bar{h}|}{n} \leq \frac{1}{2}\left(\frac{1}{\tau}-1\right)$. This implies that

$$|T_{22}| = \frac{|(n-1)(N_{21}\Delta\bar{f} + N_{12}\Delta\bar{h})|}{\frac{1}{2}n(n-1)(n-\frac{1}{\tau})} \leq \frac{(\frac{1}{\tau}-1)}{(n-\frac{1}{\tau})}$$

Substituting the results into Inequality (7), we have

$$|T_2| \leq |T_{21}| + |T_{22}| \leq \frac{(\frac{1}{\tau}-1)}{(n-\frac{1}{\tau})}(1-\text{RI}) + \frac{(\frac{1}{\tau}-1)}{(n-\frac{1}{\tau})} \leq 2\frac{(\frac{1}{\tau}-1)}{(n-\frac{1}{\tau})}$$

Finally, combining the upper bounds on $|T_1|$ and $|T_2|$, we can obtain

$$|\text{spRI}_\tau - \text{spRI}| \leq |T_1| + |T_2| \leq \frac{\frac{1}{\tau}-1}{n-\frac{1}{\tau}}(1-\text{RI}) + 2\frac{(\frac{1}{\tau}-1)}{(n-\frac{1}{\tau})} \leq 3\frac{(\frac{1}{\tau}-1)}{(n-\frac{1}{\tau})}$$

□

Web Appendix G: Detailed steps in the permutation test

Given a significance level δ, the permutation testing procedure consists of the following four steps.

Step1 Randomly permute the reference labels and the clustering labels of all the objects, respectively. This ensures that the number of groups/clusters and the number of objects within each group/cluster remain the same as those in the original reference and clustering partitions. In other words, R, C, $n_{r\cdot}$, and $n_{\cdot c}$ do not change with the permutations.

Step2 Compute the simulated spARI value based on the permuted reference and clustering, denoted by $\text{spARI}_{\text{sim}}$.

Step3 Repeat steps 1 and 2 T times to obtain T simulated spARI values.

Step4 Compute the proportion of simulated values satisfying $|\text{spARI}_{\text{sim}} - \psi| \geq |\text{spARI}_{\text{obs}} - \psi|$. This proportion is taken as the p-value of this testing. If the p-value is smaller than the significance level δ, we have enough evidence to reject the null hypothesis, indicating that the observed spARI is significantly different from ψ.

The detailed steps in the algorithm are presented in Algorithm 1. Specifically, the notation $\text{spARI}(\vec{u}, \vec{v}, \mathbf{Z})$ denotes the computation of spARI given the reference labels $\vec{u} = (u_i)_{i=1}^n$, the clustering labels $\vec{v} = (v_i)_{i=1}^n$, and the spatial coordinate matrix $\mathbf{Z}$. The indicator function $\mathbb{I}\left\{\text{spARI}_{\text{sim}}^{(t)} > \text{spARI}_{\text{obs}}\right\}$ equals one when $\text{spARI}_{\text{sim}}^{(t)} > \text{spARI}_{\text{obs}}$ and equals zero otherwise. We also added an R function "perm_test" to automatically execute this permutation testing procedure.

Moreover, we incorporated an additional simulation case to provide some evidences that the proposed permutation test described above can control type I error well. In this scenario, the synthetic dataset consists of 200 objects, with the reference partition assigning them into two groups of sizes 1 and 199, respectively, and the clustering partition separating them into two clusters of sizes 90 and 110, respectively. The observed spARI value is close to zero (-0.0002), and the corresponding p-value derived from the R function "perm_test" is 0.59, indicating that we do not have strong evidences to reject the null hypothesis spARI = 0 in the significance level $\delta = 0.1$. Furthermore, we repeated the permutation test 100 times under the same setting, and the resulting p-values range from 0.43 to 0.64, none of which leads to an incorrect rejection of the null hypothesis. This demonstrates that the proposed hypothesis testing in Algorithm 1 effectively controls type I error.

Algorithm 1: Permutation test for spARI

Input: Reference labels $\vec{u} = (u_i)_{i=1}^n$, clustering labels $\vec{v} = (v_i)_{i=1}^n$, spatial coordinate matrix $\mathbf{Z}$, number of permutations T, significance level δ

Output: Decision on whether to reject H_0: spARI $= 0$

Compute observed spARI: $\text{spARI}_{\text{obs}} \leftarrow \text{spARI}(\vec{u}, \vec{v}, \mathbf{Z})$;

for $t = 1$ **to** T **do**

  Generate random permutations $\rho_1^{(t)}$ of reference labels and $\rho_2^{(t)}$ of clustering labels ;

  $\vec{u}^{(t)} \leftarrow (\rho_1^{(t)}(u_i))_{i=1}^n, \;\; \vec{v}^{(t)} \leftarrow (\rho_2^{(t)}(v_i))_{i=1}^n$;

  Compute simulated spARI: $\text{spARI}_{\text{sim}}^{(t)} \leftarrow \text{spARI}(\vec{u}^{(t)}, \vec{v}^{(t)}, \mathbf{Z})$;

Compute p-value: $p^* \leftarrow \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\left\{\text{spARI}_{\text{sim}}^{(t)} > \text{spARI}_{\text{obs}}\right\}$;

Result: Reject H_0 if $p^* < \delta$

Web Appendix H: Computational efficiency of spARI

We investigated the computational efficiency of spARI by varying the number of objects. Specifically, we generated ten simulated datasets with the object number n from 1,000 to 10,000 in increments of 1,000. In each dataset, all objects were partitioned into four reference groups, and half of them were set to be misclustered, which forms the clustering partition. Figures S6(a) and (b) demonstrate the boxplots of the computational time of spARI in seconds across ten replicates for each value of n. The numerical experiments were run on a MacBook Air powered by Apple M4 CPU with 16GB of RAM. We observe that in the setting where the spatial coordinates are provided, the runtime remains below one second for $n \leq 7,000$. If we use a precomputed distance matrix as input, the runtime stays under one second even for $n \leq 8,000$. Here, a distance matrix records the pairwise distances among all objects. Overall, employing a distance matrix can lead to a faster calculation compared to using spatial coordinates, which is expected as the computation of pairwise distances from coordinates becomes increasingly time-consuming when n is large. Accordingly, the R package of spARI supports the direct input of a distance matrix by users.

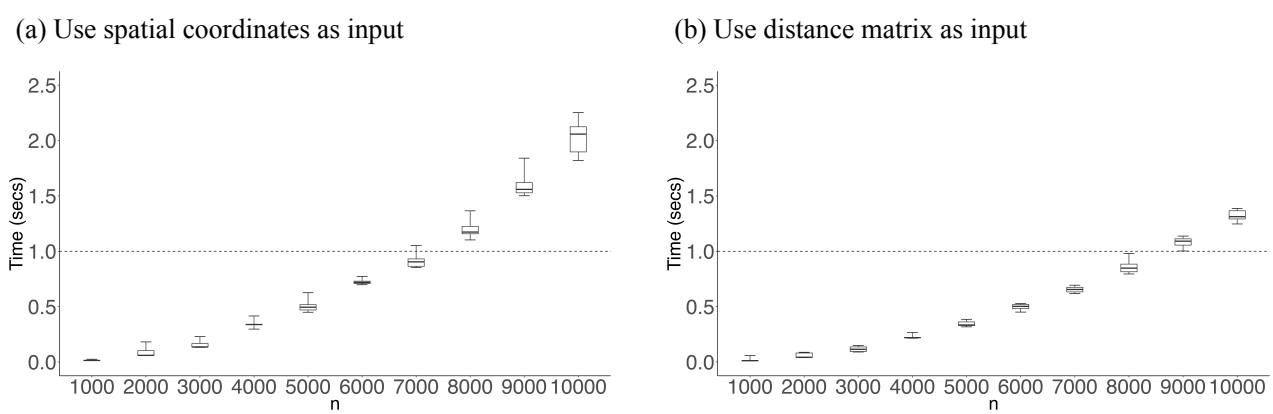


Figure S6: The boxplots of computational time of spARI in seconds for different object numbers when users provide (a) the spatial coordinates and (b) the distance matrix.

When $n = 100,000$, the full dense distance matrix requires over 16GB of memory, making spARI incomputable in common laptops. To address this issue, we may store the distance matrix in a sparse format. We simulate a large-scale scenario by generating a synthetic ST dataset with $n = 100,000$ cells. The spatial coordinates of these cells were uniformly sampled from the $[0, 1] \times [0, 1]$ region, and each cell was assigned to one of 15 distinct cell types. Due to the large size of the dataset, we focused only on the local spatial neighborhood of these cells. For example, we only computed the Euclidean distances for each cell's five nearest neighbors. Since the distance matrix need to be symmetric, a distance between cells i and j was recorded in the sparse distance matrix if either i or j was among the other's five nearest neighbors. The resulting sparse matrix contained approximately 600,000 nonzero entries. We then randomly selected 10,000 cells and manually assigned their cell type labels to construct one clustering partition. The reference partition, the clustering partition, and the sparse distance matrix was passed to the R function to compute the spARI value (with all other arguments set to default values). We repeated the computational process ten times, and the median runtime is 4.8 seconds. This shows that spARI is scalable to very large-scale ST data if users are able to provide the distance matrix in a sparse format.

Web Appendix I: Biological interpretations for clustering results of BayesSpace and BANKSY

BayesSpace incorporates a Markov random field prior to induce spatial dependence among neighboring spots, which encourages them to be assigned to the same cluster. This may lead to integral spatial domains that are more consistent with the layered organization of the cortex. Maynard et al. (2021) reported that the marker gene *CCK* is highly expressed across the annotated Layers 2, 3 and 6, suggesting that these layers may share transcriptomic features or functional similarities. This observation may help explain why BayesSpace identified the cluster C4 that combines spots from these three layers. In contrast, BANKSY tends to produce smaller, spatially fragmented regions, as marked by the dashed circles in Figure 2(a) of the main text. From a biological perspective, such fragmentation may disrupt the regular layered architecture of the cortex in the ground truth. Therefore, the biological interpretations support our results that BayesSpace achieves a higher spARI value due to its better ability to capture spatial compactness than BANKSY.

In addition, we conducted permutation tests for the spARI values obtained from BayesSpace and BANKSY. As previously reported, their spARI values are 0.3594 and 0.3421, respectively. Both clustering results yielded p-values zero, providing strong evidences of significant associations with the reference partition. These results confirm that each method captures some meaningful spatial structures relative to the ground truth.

We also applied the hypothesis testing to the clustering results in the third simulation study (Section 4.3 in the main text). Recall that the spARI values of the two clustering results are 0.672 and 0.632. Implementing the permutation test based on Algorithm 1, both the clustering results obtained p-values of zero, strongly indicating their significant associations with the reference partition.

Web Appendix J: Application of spARI to a geographical dataset

We use the social media geographical dataset in Kang et al. (2022) to demonstrate the application of spRI and spARI. This dataset contains 7,604 records, each with longitude and latitude geographical coordinates. The ground truth for position categories is shown in Figure S7(a) and serves as the reference partition, where the blue dots represent workplaces and the yellow dots denote home positions. Figure S7(b) demonstrates the clustering result estimated by Kang et al. (2022), along with its corresponding values of RI, spRI, ARI, and spARI.

Figure S7: (a) The reference partition of the 7,604 social media records with geographical coordinates. (b) The estimated clustering partition. The values of RI, spRI, ARI, and spARI for the clustering result is shown at the right of panels.

Moreover, spARI is also compatible with spatial data characterized by an undirected graph structure, where each node corresponds to an object and edges denote connections. In this setting, the spatial proximity between any two objects is represented by an adjacency matrix, where the (i, j)-th entry equals one if objects i and j are connected and equals zero otherwise. This 0-1 symmetric adjacency matrix can be viewed as a special case of a distance matrix.

References

Kang, Y., K. Wu, S. Gao, I. Ng, J. Rao, S. Ye, F. Zhang, and T. Fei (2022). STICC: a multivariate spatial clustering method for repeated geographic pattern discovery with consideration of spatial contiguity. *International Journal of Geographical Information Science 36*(8), 1518–1549.

Maynard, K. R., L. Collado-Torres, L. M. Weber, C. Uytingco, B. K. Barry, S. R. Williams, J. L. Catallini, M. N. Tran, Z. Besich, M. Tippani, et al. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience 24*(3), 425–436.