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, which is referred to as region segmentation for spot level ST data and spatial cell typing for single-cell ST data. An accurate ST clustering result can greatly benefit downstream biological analyses. As various ST clustering approaches are proposed in recent years, quantifying and comparing their clustering accuracy become important in the benchmarking study. However, the commonly used metric adjusted Rand index (ARI) totally ignores the invaluable spatial locations of objects (spots or cells) in ST data. To address this issue, we propose a spatially aware Rand index (spRI) as well as its adjusted version (spARI) that incorporate the spatial distance information into the clustering evaluation. Specifically, when comparing two partitions, spRI provides a pair of objects that are in the same cluster of one partition but are in different clusters of the other partition (called disagreement pair) with a weight relying on the distance of the two objects, while original Rand index offers this disagreement pair with an identical zero weight. This spatially aware feature of spRI successfully differentiates disagreement object pairs based on their distinct distances, which cannot be achieved by Rand index. The spRI is then adjusted to correct for random chances such that its expectation takes on the zero value under an appropriate random null model, resulting in spARI. Statistical properties of spRI and spARI are thoroughly discussed. The applications of spARI to simulation study and two ST datasets demonstrate its better utilities than 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, spatial transcriptomics, Rand index, hypergeometric distribution

*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 technologies based on in situ hybridization, in situ sequencing, and spatial barcoding (Williams et al., 2022; Moses and Pachter, 2022; Du et al., 2023). 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 during data mining to understand and investigate the solid tissue spatial heterogeneity from the perspective of transcriptomics especially for cancer (De Zuani et al., 2024) and brain (Piwecka et al., 2023). 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 which 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 benefit greatly 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 distances of spots/cells) to improve accuracy, thus offering more insights into the complex cellular microenvironment and disease molecular mechanisms (Park et al., 2023). Considering a big 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; Hu 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 try 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, 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 two partitions are identical. The commonly used adjusted Rand index (ARI) (Hubert and Arabie, 1985) is a variant of Rand index that is corrected for random chance. By counting common object numbers of two partitions in a contingency table and assuming a generalized hypergeometric distribution for the table entry number with fixed marginal numbers, ARI has an expectation zero and is bounded above by one.

In addition, Cover (1999) proposes the normalized mutual information (NMI) to measure the dependence between two partitions in the information theory framework. This metric normalizes the mutual information score to account for cluster size effects, ensuring that the value ranges between zero and one, where one indicates a perfect clustering alignment and zero denotes no agreement. The V-measure (Rosenberg and Hirschberg, 2007) is a conditional entropy-based metric by balancing homogeneity and completeness, where V stands for "validity" describing the goodness of a clustering. The homogeneity measures the degree that samples within each cluster tend to share the same class label of another partition, while the completeness quantifies the degree that samples with the same class label tend to be grouped into a single cluster. The V-measure is defined as a weighted harmonic mean of homogeneity and completeness. Despite the widespread use of ARI, NMI, or V-measure in evaluating the clustering accuracy, all these metrics ignore an invaluable characteristic of the ST partition—the spatial locations of objects. Therefore, there is an urgent need to propose a subtle *spatially aware* clustering metric that can sufficiently 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 clustering metric that directly makes 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 Rand index and NMI by accounting for the cell type hierarchy. Using Rand index as an example, for a disagreement cell pair, its contribution to the weighted Rand index (wRI) 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 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 the spatial grouping discrepancy (SGD) metric specifically designed for ST data. SGD evaluates both label accuracy and spatial structure consistency of the detected anomalous tissue domains by bootstrap resampling and the Wasserstein distance, enabling the measurement of spatial distribution discrepancies between predicted and reference labels. Moreover, Xu et al. (2024) introduce multi-SGD as an extension of SGD to evaluate discrepancies across multiple domain types. Nevertheless, SGD is primarily for anomalous tissue region detection and is not entirely compatible with the general ST clustering accuracy quantification in our task.

In this article, 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. Motivated by 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 this disagreement pair. This spatially aware feature of spRI successfully differentiates disagreement object pairs based on their distinct distances, which cannot be achieved by Rand index. We adjust spRI to correct for random chances such that its expectation takes on zero value under an appropriate random null model, which results in spARI. We further investigate the statistical property of spRI and spARI, which includes the analytical forms between two similar clusterings, the sensitivity to subsampling, and the statistical hypothesis testings. The applications of spARI to simulation study and two ST datasets demonstrate its better utilities than ARI in evaluating ST clustering methods, which prefers a more spatially coherent clustering.

2 Method

2.1 Review of Rand index and its adjusted version

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; Hu 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., both $i, j \in U_r$ 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., both $i, j \notin U_r$ 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 listed in Table 1.

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

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 $\mathbf{V}$.

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

$$\mathrm{RI} = \frac{N_{11} + N_{22}}{N},$$

which is exactly the proportion of agreement pairs among all the N object pairs. Moreover, through some simple algebra, RI can be rewritten as $N_{11}/N_{\cdot 1} \cdot N_{\cdot 1}/N + N_{22}/N_{\cdot 2} \cdot N_{\cdot 2}/N$, where

$N_{11}/N_{\cdot 1}$ is defined as the positive predictive value and $N_{22}/N_{\cdot 2}$ the negative predictive value in Wu and Wu (2020). Thus, RI can be also seen as a weighted average of the positive predictive value and the negative predictive value, where the weights are $N_{\cdot 1}/N$ and $N_{\cdot 2}/N$, respectively.

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		$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot C}$	$n_{\cdot\cdot} = n$

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

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

$$
\begin{aligned}
N_{12} &= \sum_{r=1}^{R} \sum_{1 \le c_1 < c_2 \le C} \binom{n_{rc_1}}{1} \binom{n_{rc_2}}{1} = \sum_{r=1}^{R} \sum_{1 \le c_1 < c_2 \le C} n_{rc_1} n_{rc_2} \qquad (1) \\
&= \frac{1}{2} \left[\sum_{r=1}^{R} \left(\sum_{c=1}^{C} n_{rc} \right)^2 - \sum_{r=1}^{R} \sum_{c=1}^{C} n_{rc}^2 \right] = \frac{1}{2} \left[\sum_{r=1}^{R} n_{r\cdot}^2 - \sum_{r=1}^{R} \sum_{c=1}^{C} n_{rc}^2 \right], \\
N_{21} &= \sum_{c=1}^{C} \sum_{1 \le r_1 < r_2 \le R} \binom{n_{r_1 c}}{1} \binom{n_{r_2 c}}{1} = \sum_{c=1}^{C} \sum_{1 \le r_1 < r_2 \le R} n_{r_1 c} n_{r_2 c} \qquad (2) \\
&= \frac{1}{2} \left[\sum_{c=1}^{C} \left(\sum_{r=1}^{R} n_{rc} \right)^2 - \sum_{r=1}^{R} \sum_{c=1}^{C} n_{rc}^2 \right] = \frac{1}{2} \left[\sum_{c=1}^{C} n_{\cdot c}^2 - \sum_{r=1}^{R} \sum_{c=1}^{C} n_{rc}^2 \right].
\end{aligned}
$$

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,

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

To account for the randomness in the contingency table 2, Hubert and Arabie (1985) sets 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 or row sums ($n_{\cdot c}$ or $n_{r\cdot}$) fixed. Under this assumption, the marginal distribution of n_{rc}

is actually a hypergeometric distribution, and it turns out that (Hubert, 1977; Fowlkes and Mallows, 1983; Klastorin, 1985; 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 $\mathbf{V}$. In this case, $R = C = 1$ and $n_{1\cdot} = n_{\cdot 1} = n$, and we do not consider this naive case $\mathbb{E}(\mathrm{RI}) = 1$ for convenience in the following definition. The adjusted Rand index (ARI) (Hubert and Arabie, 1985) then has the following form,

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

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 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

$$\mathrm{RI} = \frac{\sum_{1\le i<j\le n} W(i,j)}{N}. \tag{4}$$

We observe that for any disagreement pair, its weight in RI is identical to zero, so this weight function cannot make a 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. To remove the effect of spatial location unit, we suggest normalizing 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 g. 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 Equation (4), we define the spatially aware Rand index (spRI) as

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

Following the same random model used in ARI where column/row sums $(n_{\cdot c}/n_{r\cdot})$, 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 can let $p_{ij} = p$ and $q_{ij} = q$. Subsequently, owing to the independence between reference $\mathbf{U}$ and clustering $\mathbf{V}$, for any object pair, the probability of 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 can be derived as follows,

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

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 $p = q = 1$, which corresponds to the naive case $(R = C = 1, n_{1\cdot} = n_{\cdot 1} = n)$. We do not consider this naive case for convenience in the following definition and Proposition 1. Based on Hubert and Arabie (1985), we define the adjusted version of spRI, named spARI, as

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

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).

In practice, when we evaluate the clustering performances of two methods against one reference, we aim to investigate whether one method with a higher spRI than the other method can still give a higher spARI. The following proposition answers the question.

Proposition 1. *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

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

As mentioned earlier, unless in the naive case $R = C = 1$, 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 (5) by $(1 - \mathbb{E}(\text{spRI}_1))(1 - \mathbb{E}(\text{spRI}_2))$, resulting in

$$[\text{spRI}_2 - \mathbb{E}(\text{spRI}_2) + (\xi - \eta)]\,[1 - \mathbb{E}(\text{spRI}_2)] > [\text{spRI}_2 - \mathbb{E}(\text{spRI}_2)]\,[1 - \mathbb{E}(\text{spRI}_2) - \eta]\,. \tag{6}$$

After simple algebra and using the positivity of $1 - \mathbb{E}(\text{spRI}_2)$, we can obtain that the inequality (6) holds if and only if

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

□

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 1, 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 $\xi > \frac{1-\text{spRI}_2}{1-\mathbb{E}(\text{spRI}_2)}\eta$ is satisfied. This provides a practical guideline about when spRI and spARI can reach the same evaluation conclusion for two ST clustering methods.

2.3 More results for spRI and spARI

2.3.1 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), suppose there are $R = k$ groups in the initial clustering (e.g., reference) and each group has m objects, so there are a total of $n = km$ objects. Table 1 considers the spRI values between the initial clustering and a new clustering formed by simple operations. Table 2 provides the spRI values between two clustering, each of which is created by applying the same operation to different parts of the same initial clustering. Table 3 shows the spRI in some standard settings including all-objects-in-one-cluster, one-object-in-one-cluster, as well as the opposite clustering.

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. 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 Table 3.

(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.

Cases	spRI	$m \to +\infty$	$k \to +\infty$
A1	$\frac{k^2-2-k/m+2/m^2 \sum_{i \in U_r, j \in U_{r'}} f_{ij}}{k^2-k/m}$	$\frac{k^2-2+2\mu_{f,U}(r,r')}{k^2}$	1
A2	$\frac{(2k^2-1)-2k/m+4/m^2 \sum_{i \in V_c, j \in V_{c'}} h_{ij}}{2k^2-2k/m}$	$\frac{2k^2-1+\mu_{h,V}(c,c')}{2k^2}$	1
A3	$\frac{k^2-1-(k-1)/m+2/m^2 \sum_{i \neq j \in U_r} h_{ij}}{k^2-k/m}$	$\frac{k^2-1+\mu_{h,U}(r,r)}{k^2}$	1
A4	$\frac{m^2-3m/k-1+3/k+2/k^2 \sum_{i \neq j \in V_{k+1}} f_{ij}+2/k^2 \sum_{i \in V_{k+1}, j \in U(i) \setminus \{i\}} h_{ij}}{m^2-m/k}$	1	$\frac{m^2-1+\phi_V}{m^2}$

Table 3: Forms of spRI and its limits for operation cases A1-A4.

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 $\mathbf{U}$. We then compute the spRI values between $\mathbf{V}_1$ and $\mathbf{V}_2$ as well as its limits in Table 4.

(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.

(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 $U_{t'}$.

Cases	spRI	$m \to +\infty$	$k \to +\infty$
B1	$\frac{k^2m-k-4+2/m \sum_{j \in V_{2c'} \setminus \{a\}} f_{aj}+2/m \sum_{j \in V_{1c} \setminus \{a\}} h_{aj}}{k^2m-k}$	1	1
B2	$\frac{2k^2-2k/m-2+4/m^2[\sum_{i \in V_{1c}, j \in V_{1c'}} f_{ij}+\sum_{i \in V_{2c}, j \in V_{2c'}} h_{ij}]}{2k^2-2k/m}$	$\frac{2k^2-2+\mu_{f,V_1}(c,c')+\mu_{h,V_2}(c,c')}{2k^2}$	1
B3	$\frac{k^2-k/m-4+2/m^2 \sum_{i \in U_t, j \in U_{t'}} f_{ij}+2/m^2 \sum_{i \in U_r, j \in U_{r'}} h_{ij}}{k^2-k/m}$	$\frac{k^2-4+2\mu_{f,U}(t,t')+2\mu_{h,U}(r,r')}{k^2}$	1

Table 4: Forms of spRI and its limits for operation cases B1-B3.

In the third circumstance, we consider the following three operations on the reference partition that result in a clustering partition $\mathbf{V}$. 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 Table 5.

(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.

Cases	spRI	$m \to +\infty$	$k \to +\infty$
C1	$\frac{1-1/m+2/k\sum_{r<r'} 1/m^2 \sum_{i\in U_r, j\in U_{r'}} f_{ij}}{k-1/m}$	$\frac{1+2/k\sum_{r<r'} \mu_{f,U}(r,r')}{k}$	$\bar{\mu}_U$
C2	$\frac{k^2-k+\sum_r 2/m^2 \sum_{i\neq j\in U_r} h_{ij}}{k^2-k/m}$	$\frac{k^2-k+\sum_r \mu_{h,U}(r,r)}{k^2}$	1
C3	$\frac{(k-1)(m-1)+\frac{2}{km}[\sum_r \sum_{i\neq j\in U_r} h_{ij}+\sum_c \sum_{i\neq j\in V_c} f_{ij}]}{km-1}$	$\frac{k-1+\frac{1}{k}\sum_r \mu_{h,U}(r,r)}{k}$	$\frac{m-1+\frac{1}{m}\sum_c \mu_{f,V}(c,c)}{m}$

Table 5: Forms of spRI and its limits for operation cases C1-C3.

2.3.2 The sensitivity of spRI for subsampling

The value of spRI may change when we focus on a subset of all objects, and this subsampling technique is often adopted when we cluster millions of spots/cells. Hence, it is critical 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. Subsequently, the spRI using the subsample $\tau\Omega$, denoted by spRI_τ, becomes

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

For the whole object set Ω, the spRI is

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

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

$$\text{spRI}_\tau - \text{spRI} = \frac{(\frac{1}{\tau}-1)\sum_{r=1}^{R}\sum_{c=1}^{C} n_{rc}^2 - \frac{1}{2}(\frac{1}{\tau}-1)\left[\sum_{r=1}^{R} n_{r\cdot}^2 + \sum_{c=1}^{C} n_{\cdot c}^2\right] + \Delta F_\tau + \Delta H_\tau}{\frac{1}{2}n(n-1)(n-\frac{1}{\tau})}, \quad (7)$$

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 (7), $\text{spRI}_\tau - \text{spRI}$, measures the sensitivity of spRI to the subsampling proportion τ. As expected, it approaches zero as the proportion τ increases to one.

2.3.3 Hypothesis testings for spARI

In most ST clustering literature, only one single clustering index is calculated and reported for method evaluations. As the clustering result often suffers from some randomness in the method such as random initializations or stochastic algorithms, it is helpful to carry out statistical hypothesis testings for these indices, especially for the proposed index spRI or spARI. We denote the spARI value using the observed partition by $\text{spARI}_{\text{obs}}$, and we aim to test the null hypothesis $H_0 : \text{spARI} = \psi$ versus the alternative hypothesis $H_a : \text{spARI} \neq \psi$. Without any assumption about the probability distribution of spARI, we adopt the permutation test (Pitman, 1937) to implement the nonparametric test. Given a significance level δ, the 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.

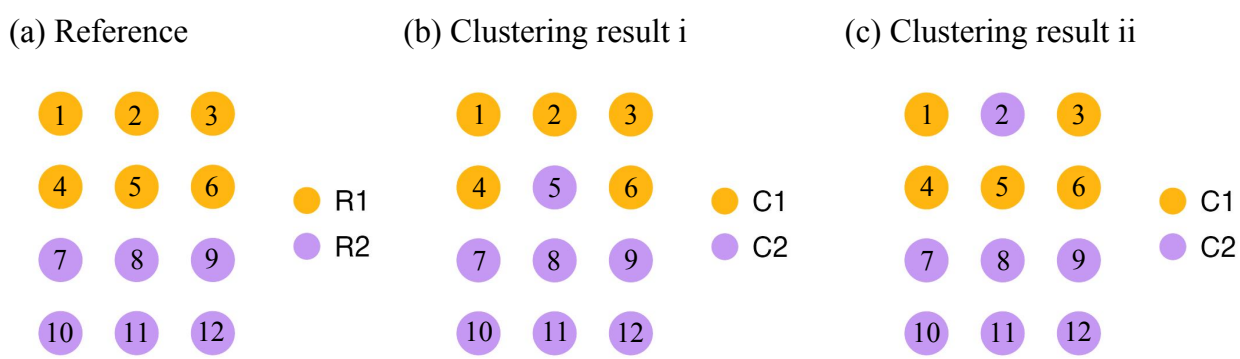
Furthermore, if we have two clustering results by two competing methods, respectively, we aim to investigate whether their corresponding spARI values are significantly different. In this case, we have the null hypothesis $H_0 : \text{spARI}_1 - \text{spARI}_2 = 0$ versus the alternative hypothesis $H_a : \text{spARI}_1 - \text{spARI}_2 \neq 0$. We denote the observed spARI of the two clustering partitions by $\text{spARI}_{1,\text{obs}}$ and $\text{spARI}_{2,\text{obs}}$, respectively. Next, we randomly permute the labels of objects in the following three partitions—the reference and the two clustering, giving simulated values $\text{spARI}_{1,\text{sim}}$ and $\text{spARI}_{2,\text{sim}}$. Repeat this simulation process T times. Finally, we compute the proportion of times that satisfy $|\text{spARI}_{1,\text{sim}} - \text{spARI}_{2,\text{sim}}| \geq |\text{spARI}_{1,\text{obs}} - \text{spARI}_{2,\text{obs}}|$ as the p-value. If it is smaller than δ (e.g., $\delta = 0.05$), we have strong statistical evidences to reject the null hypothesis $H_0 : \text{spARI}_1 - \text{spARI}_2 = 0$.

2.3.4 The R package for computing spRI and 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 for users. The main function of the R package is "spARI", which takes the following inputs—the reference partition ("r_labels"), the clustering partition ("c_labels"), and the spatial coordinates of corresponding objects ("coords"). The main function includes a built-in step for normalizing spatial coordinates as described in Section 2.2. Users only need to input the two partition results along with the spatial coordinates, and then the main function efficiently computes the spRI and spARI values. When applied to a dataset with 100 objects, the total computation time for spRI and spARI is approximately 0.001 seconds on a MacBook Pro with an Intel Core i5 CPU at 2GHz and 16GB of RAM.

3 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 $H_{g\times s}$.



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

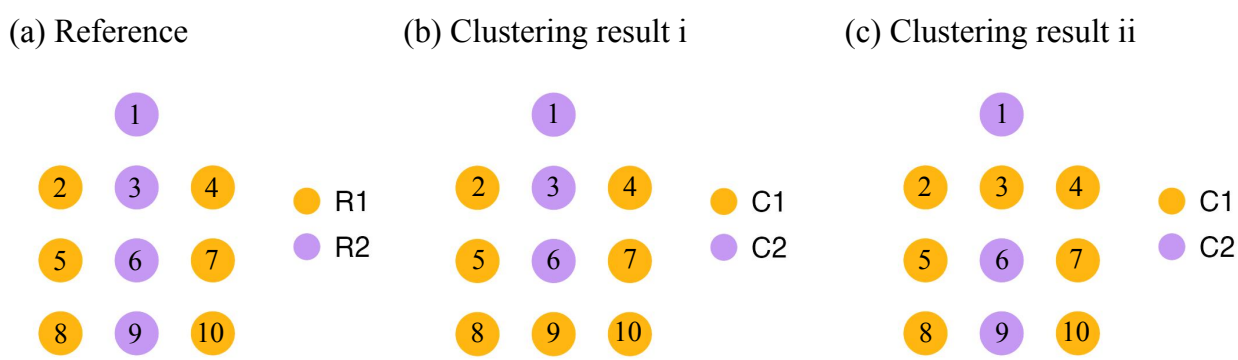
Figure 1: (a) The reference partition of the 12 objects. (b) The clustering partition by method i. (c) The clustering partition by method ii. The specific values of RI, ARI, spRI, and spARI for the two clustering results are demonstrated at the bottom of panels.

3.1 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 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 Figures 1(b) and (c), 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, 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. 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 $F_{sg} = 2.065$ than that of clustering i because object number 2 is more far away from other objects in cluster C2 compared to object number 5. On the other hand, the 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 Proposition 1), clustering result i also achieves a higher spARI value.



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

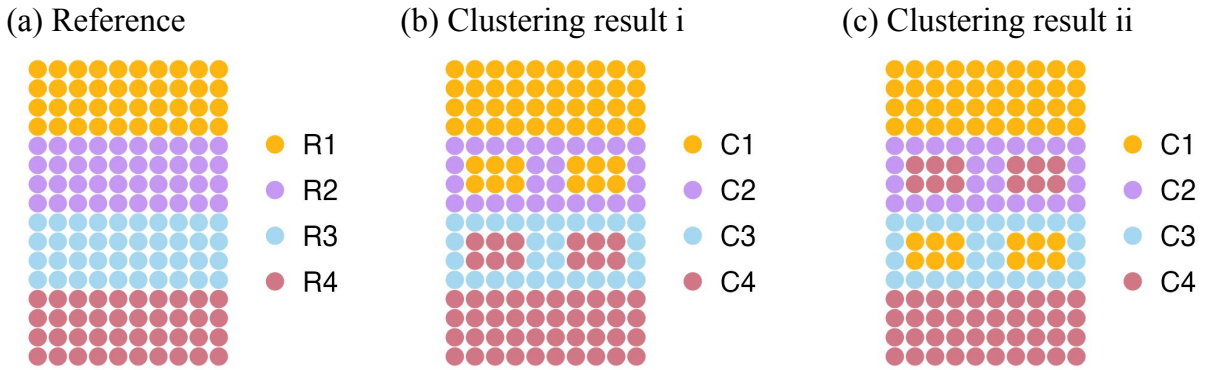
Figure 2: (a) The reference partition of the 10 objects. (b) The clustering partition by method i. (c) The clustering partition by method ii. The specific values of RI, ARI, spRI, and spARI for the two clustering results are demonstrated at the bottom of panels.

3.2 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 2(a), 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 Figure 2(b) and (c), 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 Proposition 1.

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.



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 3: (a) The reference partition of the 160 objects. (b) The clustering partition by method i. (c) The clustering partition by method ii. The specific values of RI, ARI, spRI, and spARI for the two clustering results are demonstrated at the bottom of panels.

3.3 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;

see Figure 3(a). Each layer contains 40 objects. Two clustering results represent the partitions obtained by two ST clustering methods, as shown in Figures 3(b) and (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 these misclustered objects and the 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 ST clustering method i shows more efforts to keep regions integral than ST clustering method ii.

4 Real applications

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.

4.1 Region segmentation evaluation in human dorsolateral prefrontal cortex data

The human dorsolateral prefrontal cortex (DLPFC) is a functional area of the human brain and involves 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. Figure 4 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 Figures 5(b) and (c) 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 Table 6.

Based on the reference illustrated in Figure 5(a), we observe that both BayesSpace and BANKSY identify the major part of Layer 1, but exhibit unsatisfactory performances in detecting Layers 2 and 3. For Layers 4, 5 and 6, neither method captures the complete

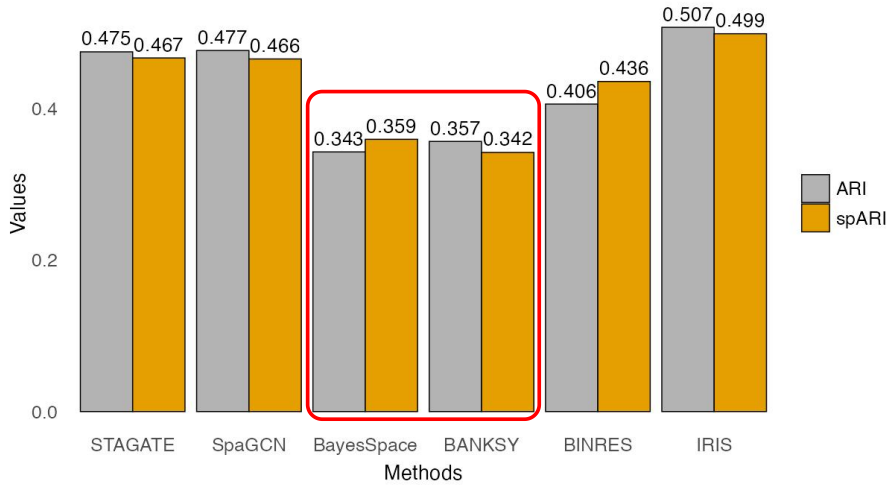


Figure 4: Barplots of ARI and spARI values for the six region segmentation methods—STAGATE, SpaGCN, BayesSpace, BANKSY, BINRES, and IRIS. The ARI/spARI values are displayed at the top of bars.

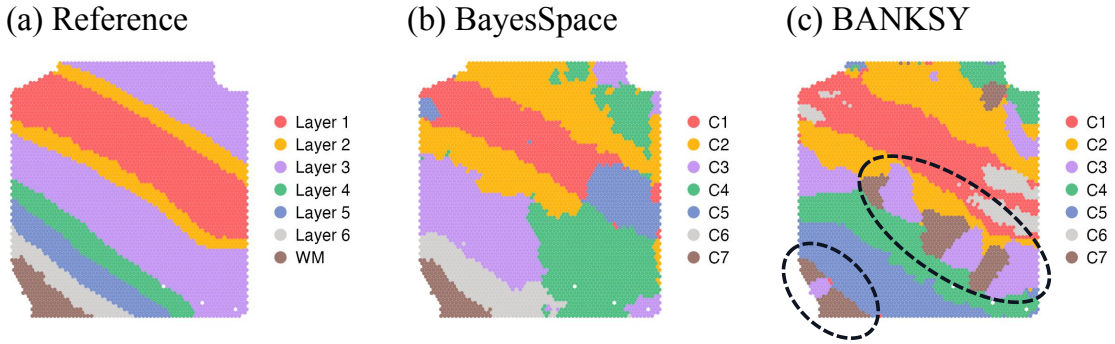


Figure 5: (a) The reference annotations of the DLPFC 151509 section, including six DLPFC layers and the white matter (WM). (b) The region segmentation by BayesSpace. (c) The region segmentation by BANKSY.

three-layer structure well. Quantitatively, Table 6 shows that BANKSY has a larger number of agreement pairs than BayesSpace. Nevertheless, the dashed domains in Figure 5(c) 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 in the evaluation. This demonstrates that spARI prefers the ST clustering method that can yield more compact spatial structures.

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
BayesSpace	0.7760	**0.0301**	**0.0550**	**0.8611**	0.7833	**0.3593**
BANKSY	**0.7874**	0.0294	0.0420	0.8588	**0.7854**	0.3420

Table 6: Related quantities used to calculate the spARI for BayesSpace and BANKSY.

4.2 Cell typing evaluation in mouse visual cortex STARmap* data

STARmap* is an imaging-based ST technique that captures gene expression patterns of single cells. This dataset consists of 1207 cells and 1020 genes. The ARI/spARI barplots of cell typing methods are demonstrated in Figure 6. 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.

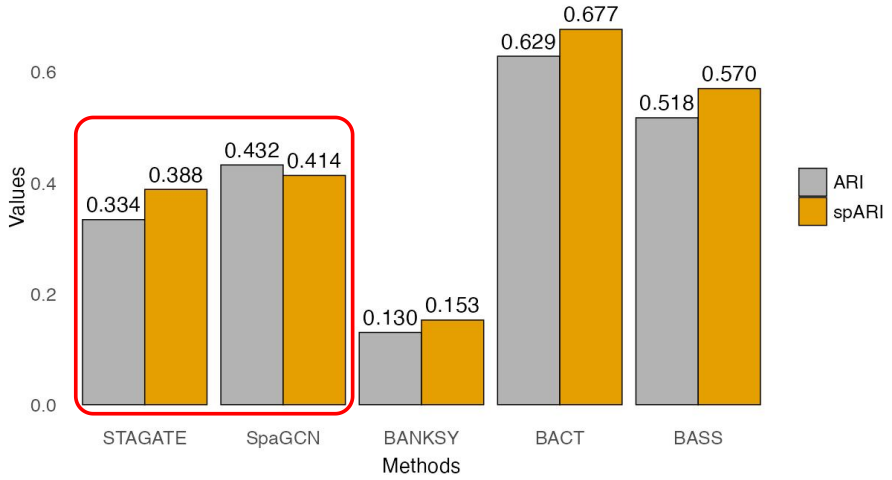


Figure 6: Barplots of ARI and spARI values for the five cell typing methods—STAGATE, SpaGCN, BANKSY, BACT, and BASS. The ARI/spARI values are displayed at the top of bars.

To look into what makes this difference, we provide the reference and the clustering results by STAGATE and SpaGCN in Figure 7 as well as related quantities during the calculations for their spARI. We can observe that in Figure 7 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, resulting in larger values of $H_{g\times s}$ and $F_{s\times g}$ for STAGATE as shown in Table 7. Nevertheless, as spRI is a trade off between RI and spatial proximity, the increases in $H_{g\times s}$ and $F_{s\times g}$ for STAGATE cannot fill the relatively large gap in RI, the spRI of STAGATE is still lower than that of SpaGCN. To put it another way, in Figure 7, 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.

Methods	$\#\{s\times s, g\times g\}/\binom{n}{2}$	$H_{gs}/\binom{n}{2}$	$F_{sg}/\binom{n}{2}$	spRI	$\mathbb{E}$(spRI)	spARI
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**

Table 7: Related quantities used to calculate the spARI for STAGATE and SpaGCN.

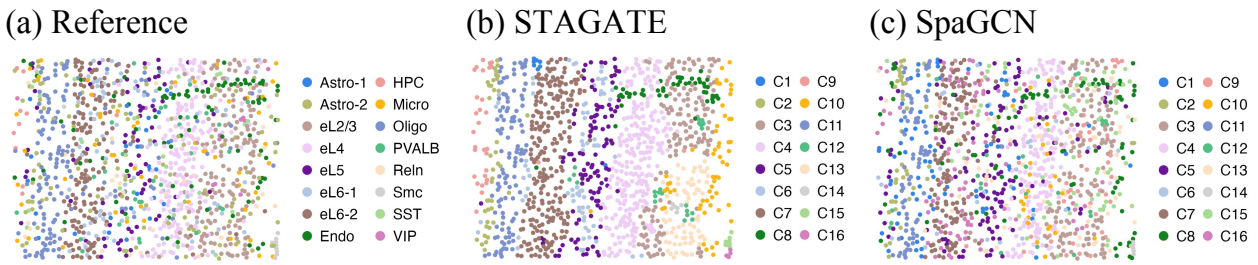


Figure 7: (a) The reference annotations of the STARmap* data, including sixteen cell types. (b) The cell typing by STAGATE. (c) The cell typing by SpaGCN.

Furthermore, we compute the differences in spRI and $\mathbb{E}(\text{spRI})$ between the region segmentation results of STAGATE and SpaGCN, and obtain the values of $\xi = -0.0086$ and $\eta = -0.0098$, which satisfied $\xi = -0.0086 < -0.0058 = \frac{1-\text{spRI}_{\text{SpaGCN}}}{1-\mathbb{E}(\text{spRI}_{\text{SpaGCN}})}\eta$. According to Proposition 1, 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.

5 Conclusion

In this paper, we propose the spatially aware adjusted Rand index, named spARI, to evaluate the ST clustering methods. Built upon the widely adopted ARI that only uses the partitions to provide a clustering accuracy index, spARI further incorporates the spatial information to assess the ST clustering accuracy in a more reasonable way. 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, such as the forms between two similar clusterings, the sensitivity to subsampling, and the statistical hypothesis testings. 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. Therefore, we envision that spARI will be an effective and statistically sound index for any problems associated with spatial clustering evaluations.

Acknowledgments

This research was supported by National Natural Science Foundation of China with grant number 12471275 and 72271060.

References

Chidester, B., T. Zhou, S. Alam, and J. Ma (2023). SpiceMix enables integrative single-cell spatial modeling of cell identity. *Nature Genetics*, 1–11.

Cover, T. M. (1999). *Elements of information theory.* John Wiley & Sons.

De Zuani, M., H. Xue, J. S. Park, S. C. Dentro, Z. Seferbekova, J. Tessier, S. Curras-Alonso, A. Hadjipanayis, E. I. Athanasiadis, M. Gerstung, et al. (2024). Single-cell and spatial transcriptomics analysis of non-small cell lung cancer. *Nature Communications 15*(1), 4388.

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.

Du, J., Y.-C. Yang, Z.-J. An, M.-H. Zhang, X.-H. Fu, Z.-F. Huang, Y. Yuan, and J. Hou (2023). Advances in spatial transcriptomics and related data analysis strategies. *Journal of Translational Medicine 21*(1), 330.

Fowlkes, E. B. and C. L. Mallows (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association 78*(383), 553–569.

Guo, Y., B. Zhu, C. Tang, R. Rong, Y. Ma, G. Xiao, L. Xu, and Q. Li (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, E. B. Lee, R. T. Shinohara, and M. Li (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.

Hu, Y., M. Xie, Y. Li, M. Rao, W. Shen, C. Luo, H. Qin, J. Baek, and X. M. Zhou (2024). Benchmarking clustering, alignment, and integration methods for spatial transcriptomics. *Genome Biology 25*(1), 212.

Hubert, L. (1977). Nominal scale response agreement as a generalized correlation. *British Journal of Mathematical and Statistical Psychology 30*(1), 98–103.

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

Klastorin, T. D. (1985). The p-median problem for cluster analysis: A comparative test using the mixture model approach. *Management Science 31*(1), 84–95.

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*, 1–14.

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.

Moses, L. and L. Pachter (2022). Museum of spatial transcriptomics. *Nature Methods 19*(5), 534–546.

Park, H.-E., S. H. Jo, R. H. Lee, C. P. Macks, T. Ku, J. Park, C. W. Lee, J. K. Hur, and C. H. Sohn (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.

Piwecka, M., N. Rajewsky, and A. Rybak-Wolf (2023). Single-cell and spatial transcriptomics: deciphering brain complexity in health and disease. *Nature Reviews Neurology 19*(6), 346–362.

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

Rosenberg, A. and J. Hirschberg (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 410–420.

Shi, X., Y. Yang, X. Ma, Y. Zhou, Z. Guo, C. Wang, and J. Liu (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, L. Lin, Y.-C. Chang, E. M. L. Teo, J. Aow, 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, X. Wan, Z. Lin, A. R. Wu, and C. Yang (2023). Integrating spatial and single-cell transcriptomics data using deep generative models with SpatialScope. *Nature Communications 14*(1), 7848.

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.

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, H. Feng, H. Wu, and X. Sun (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, X.-Y. Zhang, and Y. Zhao (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, S. R. Williams, C. R. Uytingco, S. E. Taylor, P. Nghiem, J. H. Bielas, and R. Gottardo (2021). Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*, 1–10.