Metrika
https://doi.org/10.1007/s00184-023-00939-9

An association measure for spatio-temporal time series

Divya Kappara[1,2] · Arup Bose[3] · Madhuchhanda Bhattacharjee[1,4]

Received: 19 March 2023 / Accepted: 26 November 2023

Abstract

Spatial association measures for univariate static spatial data are widely used. Suppose the data is in the form of a collection of spatial vectors, say X_{rt} where $r = 1, \dots, R$ are the regions and $t = 1, \dots, T$ are the time points, in the same temporal domain of interest. Using Bergsma's correlation coefficient ρ, we construct a measure of similarity between the regions' series. Due to the special properties of ρ, unlike other spatial association measures which test for *spatial randomness*, our statistic can account for *spatial pairwise independence*. We have derived the asymptotic distribution of our statistic under null (independence of the regions) and alternate cases (the regions are dependent) when, across t the vector time series are assumed to be independent and identically distributed. The alternate scenario of spatial dependence is explored using simulations from the spatial autoregressive and moving average models. Finally, we provide application to modelling and testing for the presence of spatial association in COVID-19 incidence data, by using our statistic on the residuals obtained after model fitting.

Keywords Bergsma's correlation · Spatial association measure · U-statistic · Spatial autoregressive model · Spatial moving average model

Research supported by J.C. Bose National Fellowship, Dept. of Science and Technology, Govt. of India.

✉ Divya Kappara
kapparadivya@gmail.com

Arup Bose
bosearu@gmail.com

Madhuchhanda Bhattacharjee
chhanda.bhatta@gmail.com

[1] University of Hyderabad, Hyderabad, India

[2] IIT Bombay, Mumbai, India

[3] Stat-Math Unit, Indian Statistical Institute, Kolkata, India

[4] University of Manchester, Manchester, UK

Published online: 23 December 2023

Mathematics Subject Classification Primary 62H20; Secondary 62F12 · 92D30 · 62H11 · 62P10 · 62M30.

1 Introduction

The primary reason behind the development of spatial association measures is the realization that it is unrealistic to assume stationarity over space. Many prominent researchers have developed such measures. One of the most widely used measures is Moran's I (Moran 1948), which is based on a global covariance representation. Cliff and Ord (1972) generalized Moran's work by demonstrating how one can test residuals of regression analysis for *spatial randomness*. They have worked out moments of Moran's I and its distributional properties under varying sampling assumptions. Geary's c (Geary 1954) is another measure that has a global differences representation.

Anselin (1995) introduced local measures of spatial association (LISA) to detect variations across space in the presence of spatial heterogeneity. Getis (2007, 2008) provide excellent reviews on the study of spatial autocorrelation. Getis and Ord (2010) introduced a family of statistics G to detect the local clusters of dependence. Using estimates of spatial autocorrelation coefficients in regression models is a widely used technique in spatial econometrics, see (Anselin 1988).

The conventional measures of spatial association are designed for static spatial data where there is only one observation per region. Suppose we have spatio-temporal data on a single variable at different time points across R spatial units, yielding concurrent spatial time series $\{X_{rt}\}$, $t = 1, \ldots, T, \ r = 1, \ldots, R$. In this case, only a few measures of spatio-temporal dependence are available in the literature. Martin and Oeppen (1975), Shao-Kuan et al. (2013) and Gao et al. (2019) etc., have extended univariate Moran's I to the Spatio-temporal Moran's I. Dubé and Legros (2013) have formulated the idea of constructing a spatio-temporal weights matrix by joining two independently constructed spatial and temporal proximity matrices. In Sect. 2.1 we give a brief discussion on general spatial association measures and their components.

Any spatial measure of association requires a *measure of similarity* between the recorded observations of a random variable, along with a choice of a *spatial proximity matrix* which accounts for the amount of geographical closeness between the regions. Some of the commonly used similarity measures are the Euclidean distance, Frechet distance, and Pearson's correlation.

We make a novel choice for the similarity measure. Bergsma (2006) introduced a correlation coefficient ρ between X and Y (both univariate), as a measure of independence such that, $\rho = 0$ if and only if X and Y are independent. Incidentally it is also a special case of “distance correlation" (Székely et al. 2007) which is defined for multivariate X and Y. From the comparative analysis of Bergsma's statistics $\hat{\rho}$ (estimate of ρ based on a sample), against other prominent statistics for testing of independence (see the arxiv article Kappara et al. 2022), it is known that $\hat{\rho}$ performs as well or better than its competitors, both in terms of power and computing efficiency. Due to these attractive features, we use it as a similarity measure. In Sect. 2.2 we give in brief the required background on this measure.

In Sect. 3, we use the pairwise ρ of the regions, and different choices for the spatial proximity matrices, to define a class of spatial association measures S_B, and an appropriate estimate $\tilde{S}_B$ of S_B. The finite sample distributional properties of global indices of spatial association are challenging to obtain in general. Often, the evaluation of the significance of these spatial association measures relies on either assuming normality or adopting a randomization approach. In the context of spatio-temporal analysis, the above mentioned studies employ empirical methods to assess the significance of the proposed measures.

In Sects. 3.1 and 3.2, we study the asymptotic distribution of $\tilde{S}_B$ as the length of time increases, assuming that the observations over time are i.i.d., while at any fixed time, they are either dependent across the regions (the non-null case) or are pairwise independent (the null case). In the first case, $\tilde{S}_B$ is asymptotically normal with appropriate centering, and a scaling by $T^{1/2}$. In the second case, $T\tilde{S}_B$ (note the different scaling, and since all pairwise ρ are 0, there is no centering) has an asymptotic distribution that involves an infinite sum of weighted i.i.d. centered chi-square variables. In both cases, the distribution depends on the unknown underlying distribution of the observations, especially through the eigenvalues of appropriate kernel functions. The case of large number of regions ($R \to \infty$) is beyond the scope of our work.

At present no distributional properties are known for $\tilde{S}_B$ when the observations are temporally dependent. This issue is important, especially from the point of view of applications. The asymptotic distribution of $\tilde{S}_B$ may be obtainable, at least for some easy spatial dependent models, with some amount of additional work. However, it is also clear that for more complex spatial models, this would require a substantial amount of work. We have considered two well-known spatially dependent models, only through simulations.

Section 4 reports all our simulation results. These show that for reasonably large number of observations, the actual finite sample distributions are well approximated by the asymptotic distributions. It is also seen that these measures and their distributions are sensitive to the choice of a spatial proximity matrix, but are not sensitive to the nature of the underlying distributions. As mentioned above, we have explored the empirical distribution of $\tilde{S}_B$ for the spatial autoregressive and the spatial moving average models.

Since the primary objective of this manuscript is to provide a global measure of independence, any local version of our statistic has not been considered. A local version can be easily developed and its asymptotic properties can be spelt out along the lines of the global statistics that we have considered.

In spite of the fact that the (limit) distribution of our statistic has been derived only when the observations are iid across the time index t, the statistic can be fruitfully used for model fitting purposes even when temporal dependence is present. This is done by first removing the temporal dependence through appropriate modelling and then using the measure on the residual series. In Sect. 5, we have presented such an application in a spatio-temporal modelling of the monthly time series COVID-19 data for the 14 districts of the state of Kerala in India.

While it may be pertinent to consider larger number of spatial locations in some situations, one should be cautious of the potential adverse effect of a large value of

R on our ability to identify dependence. It is possible that spatial dependence exists only locally and an overall measure may fail to identify it. Care also needs to be taken while choosing the proximity measure when a large number of regions are involved. See (Kappara et al. 2023) for an instance where we found a diminished sensitivity of the measure when R is large and the inverse distance proximity matrix is used.

Often there are situations where multivariate observations are available on each spatial unit. It would be interesting to ask if our approach could be extended to these situations. Noting that the distance covariance measure is available when X and Y are multivariate and that it reduces to κ when X and Y are univariate, this seems to be quite feasible. However, whether the distributional properties would still be possible to obtain remains to be seen.

2 Materials and methods

As mentioned in the introduction, in Sect. 2.1 we give a brief discussion on general spatial association measures and their components. In Sect. 2.2 we give a brief background on Bergsma's measure of covariance κ and correlation ρ. In Sect. 2.3, we describe the estimate of κ and ρ borrowed from Kappara et al. (2022). In the next section we shall use this estimate to develop the global measure of association.

2.1 Spatial association statistics

Suppose the study area is divided into R (geographical) units over which a variable is observed. Spatial association refers to the relationship between the values of the variable with respect to the proximity of the regions. The two building blocks of a spatial association statistic are two matrices, W and S.

The **spatial proximity matrix** $W = ((w_{ij}))_{1 \leq i,j \leq R}$ is based on the proximity (typically geographical proximity) of the units, and provides the spatial component. Larger weights are assigned to the pairs of regions that are spatially “more related". Getis and Ord (2010) suggests three different types of W matrices: (i) a matrix based on some theoretical notion of spatial association, such as a distance decline function, (ii) a matrix based on a geometric indicator of spatial nearness, such as the representation of contiguous spatial units, and (iii) a matrix which uses a descriptive expression of the spatial association that already exists within the data set. It is always assumed that $w_{ii} = 0$ for all i. The most commonly used W is the *adjacency matrix* where,

$$w_{ij} = \begin{cases} 1 & \text{if regions } i \text{ and } j \text{ are adjacent localities,} \\ 0 & \text{otherwise.} \end{cases}$$

Other popular choices are to take d_{ij} as the Euclidean distance, or some other distance, between the centroids of regions i and j, and then $w_{ij} = d_{ij}^{-1}$, $w_{ij} = d_{ij}^{-2}$, or the Gaussian weights $w_{ij} = \exp(-d_{ij}^2)$. One may refer to Bhattacharjee et al. (2021) for a discussion on choice of appropriate distance measure for COVID-19 data from the Indian subcontinent.

The matrix W is often *row-standardized*, so that each row sum in the matrix is equal to one. We would be working with W which are row-standardized. This will turn out to be a crucial point in our data analysis later.

The second component is a **similarity matrix** S. Suppose that we have a series of observations ${X_i}^T = \{x_{im}, m = 1, \ldots, T\}$ at each region $i = 1, 2, \ldots, R$ at T time points. The similarity matrix S is defined as $S := ((sim_{ij}))$, where sim_{ij} is some measure of similarity between X_i^T and X_j^T. For example, it could be the sample covariance between X_i^T and X_j^T. Note that these values are dependent only on any possible non-spatial relationship across the space.

Any choice of W and S yields a global **spatial autocorrelation index**

$$\Big(\sum_{i,j=1}^{R} w_{ij}\Big)^{-1} \sum_{i,j=1}^{R} w_{ij} sim_{ij}. \tag{2.1}$$

When the matrices S and W have similar structures, that is, they have high or low values together for the same pair of units i and j, we can say that there is a high degree of spatial association.

When the observations X_i^T are univariate with single observation per spatial unit, then the well-known Moran's I (Moran 1948) uses the covariance as a similarity measure. On the other hand, Gearcy's c (Geary 1954 uses squared differences as the similarity.

2.2 A measure of independence

We now discuss the similarity measure that we shall use later to develop our spatial measure of association. This measure is going to be Bergsma's correlation coefficient ρ, with a corresponding covariance κ that has the special property of being equal to zero if and only if the variables are independent. For details on its background and properties, see (Bose et al. 2023). Here we briefly give the definition of κ and ρ and their estimators, using the notations of the above article.

Let Z_1, Z_2 be i.i.d. with distribution F which has finite mean. Define

$$g_F(z) := \mathbb{E}_F[\, |z - Z| \,], \tag{2.2}$$

$$g(F) := \mathbb{E}_F[\, |Z_1 - Z_2| \,] = \mathbb{E}_F[\, g_F(Z) \,], \tag{2.3}$$

$$h_F(z_1, z_2) \;=\; -\frac{1}{2}\Big[\, |z_1 - z_2| - g_F(z_1) - g_F(z_2) + g(F) \,\Big]. \tag{2.4}$$

Note that

$$\mathbb{E}h_F(Z_1, Z_2) = 0 \;\text{ whenever }\; Z_1, Z_2 \;\text{ are i.i.d. }\; F. \tag{2.5}$$

Now, let F_1 and F_2 be the marginal distributions of a bivariate random variable (X, Y). Let (X_1, Y_1) and (X_2, Y_2) be i.i.d copies of (X, Y). Then Bergsma's covariance and correlation coefficient between X and Y are defined respectively by

$$\kappa = \kappa(X, Y) := \mathbb{E}[\; h_{F_1}(X_1, X_2) h_{F_2}(Y_1, Y_2)\;], \;\text{ and}$$
$$\rho = \rho(X, Y) := \frac{\kappa(X, Y)}{\sqrt{\kappa(X, X)\kappa(Y, Y)}}. \tag{2.6}$$

2.3 Estimates of κ

Suppose we have n observations (x_i, y_i), $1 \leq i \leq n$ from a bivariate distribution F_{12}. Bergsma has given two estimators of κ namely $\tilde{\kappa}$, and $\hat{\kappa}$ respectively based on some U and V-statistics with estimated kernels, and studied their distributional properties under independence. In Kappara et al. (2022), a third estimate κ^* was introduced, and the properties of all three covariance estimates $\tilde{\kappa}$, $\hat{\kappa}$ κ^* was discussed under dependence as well as independence. In particular, their asymptotic normality was established under dependence.

Simulation results in Kappara et al. (2022) show that the V-statistic based estimate $\hat{\kappa}$ has an upward bias. Performance-wise, $\tilde{\kappa}$ had an edge in computation time over the other two estimators. Therefore, we shall use the U-statistic based estimate $\tilde{\kappa}$ and the corresponding correlation $\tilde{\rho}$ in our subsequent developments.

Note that the kernel function h_F defined in Equation (2.4), depends on the unknown distribution function F. Its sample analogue is given by $\tilde{h}_{\hat{F}}$,

$$\tilde{h}_{\hat{F}}(z_i, z_j) = -\frac{1}{2}\Bigg[|z_i - z_j| - \frac{n}{n-1}\Big(\frac{1}{n}\sum_{k=1}^{n}|z_i - z_k| + \frac{1}{n}\sum_{k=1}^{n}|z_k - z_j|$$
$$-\frac{1}{n^2}\sum_{k,l=1}^{n}|z_k - z_l|\Big)\Bigg], \tag{2.7}$$

where $\hat{F}$ is the sample distribution function. The U-statistic type estimator of κ is defined as,

$$\tilde{\kappa} = \tilde{\kappa}(x, y) := \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \tilde{h}_{\hat{F}_1}(X_i, X_j)\tilde{h}_{\hat{F}_2}(Y_i, Y_j), \;\text{ and}$$
$$\tilde{\rho} = \tilde{\rho}(x, y) := \frac{\tilde{\kappa}(x, y)}{\sqrt{\tilde{\kappa}(x, x)\tilde{\kappa}(y, y)}}, \tag{2.8}$$

where $\hat{F}_1$ and $\hat{F}_2$ are the sample distribution functions of the $\{X_i\}$ and $\{Y_i\}$ respectively.

Kappara et al. (2022) proved that if the pairs $\{(X_i, Y_i)\}$ are i.i.d. with $\mathbb{E}_{F_{12}}[X_1^2 Y_1^2] < \infty$, then as $n \to \infty$, $n^{1/2}\big(\tilde{\kappa} - \kappa\big)$ is asymptotically normal with mean 0 and some variance δ_1. The crucial step in the proof of the above result which we shall need later, is that the leading term of $\tilde{\kappa}$ is a U-statistics, whose first projection is say H_1, and

$$n^{1/2}\big(\tilde{\kappa} - \kappa\big) = \frac{1}{2} n^{-1/2} \sum_{i=1}^{n} H_1(X_i, Y_i) + R_n, \;\text{ where }\; R_n \xrightarrow{P} 0. \tag{2.9}$$

To describe H_1 and δ_1, let

$$\begin{aligned} g_{F_{12}}(x, y) &:= \mathbb{E}_{F_{12}}\big[\, |x - X_1|\, |y - Y_1| \,\big], \\ g(F_{12}) &:= \mathbb{E}_{F_{12}}\big[\, g_{F_{12}}(X_2, Y_2) \,\big] = \mathbb{E}_{F_{12}}\big[\, |X_2 - X_1|\, |Y_2 - Y_1| \,\big]. \\ \mu_1 &:= g(F_1),\ \mu_2 := g(F_2),\ \mu_{12} := g(F_{12}),\ \text{and} \\ \mu_3 &:= \mathbb{E}_{F_{12}}\big[\, g_{F_1}(X) g_{F_2}(Y) \,\big]. \end{aligned}$$

Let $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3)$ be i.i.d. Then $H_1(\cdot, \cdot)$ and δ_1 are given by

$$\begin{aligned} H_1(x, y) &= \Big[\, g_{F_{12}}(x, y) - \mu_{12} + \mu_1(g_{F_2}(y) - \mu_2) + \mu_2(g_{F_1}(x) - \mu_1) - g_{F_1}(x) g_{F_2}(y) \\ &\quad - \mathbb{E}_{F_{12}}[\, |X_2 - x|\, |Y_2 - Y_3| \,] - \mathbb{E}_{F_{12}}[\, |X_2 - X_3|\, |Y_2 - y| \,] + 3\mu_3 \,\Big], \\ \delta_1 &= \frac{1}{4} \mathbb{V}(H_1(X_1, Y_1)). \end{aligned}$$

When $\kappa = 0$, that is X and Y are independent, the first projection H_1 in (2.9) is zero. In this case, Bergsma (2006) obtained the asymptotic distributions of $\tilde{\kappa}$ and $\hat{\kappa}$. Kappara et al. (2022), gave a detailed proof of the above result, along with the asymptotic distribution of κ^*. Again, the crucial step in that proof is to identify the leading term of the estimators, in terms of the second projections of the relevant U-statistics. Indeed,

$$\tilde{\kappa} = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} h_{F_1}(X_i, X_j) h_{F_2}(Y_i, Y_j) + R_n, \ \text{where}\ nR_n \xrightarrow{P} 0. \quad (2.10)$$

Later we shall use this weak expansion to find the asymptotic distribution of our spatial association measure when the regions are assumed to be pairwise independent.

3 The S_B measure of association

We now use κ as a similarity measure for the $\binom{R}{2}$ pairs of regions to develop a global measure of spatial association. The estimate for this global measure is directly obtained by using the corresponding estimates $\tilde{\kappa}$. We provide the asymptotic distribution of this statistic under dependence and pairwise independence of the regions, in Sects. 3.1 and 3.2 respectively.

So suppose there are R regions and let X_i denote a variable corresponding to the region $i = 1, \ldots, R$. Then we can use the similarity measure ρ, to define a global spatial association measure (**Spatial Bergsma**) as

$$S_B := S_0^{-1} \sum_{i,j=1}^{R} w_{ij} \rho(X_i, X_j), \ \text{where}\ S_0 = \sum_{i,j=1}^{R} w_{ij}. \quad (3.1)$$

Note that

$$S_0 = R, \text{ whenever } W \text{ is row-standardized.} \tag{3.2}$$

Further, $\rho(X_i, X_j) = \rho(X_j, X_i)$ and $w_{ii} = 0$ for all $i, j = 1, \ldots, R$. Therefore,

$$S_B = S_0^{-1} \sum_{1 \le i < j \le R} (w_{ij} + w_{ji}) \rho(X_i, X_j). \tag{3.3}$$

Now, suppose that we have observations ${X_i}^T = \{x_{im}, m = 1, \ldots, T\}$, such that the vectors $(x_{im}, 1 \le i \le R)$, are i.i.d. for $m = 1, \cdots, T$, with marginal distributions F_i, $1 \le i \le R$. Let $\tilde{\kappa}^{(ij)} = \tilde{\kappa}(X_i^T, X_j^T)$, and $\tilde{\rho}^{(ij)}$ denote the $\tilde{\kappa}$ and $\tilde{\rho}$, calculated for the pair of variables (X_i^T, X_j^T) using Equations (2.7) and (2.8).

We define a U-statistic based estimate of S_B as

$$\tilde{S}_B := S_0^{-1} \sum_{1 \le i < j \le R} (w_{ij} + w_{ji}) \tilde{\rho}^{(ij)}, \tag{3.4}$$

where,

$$\tilde{\rho}^{(ij)} = \tilde{\rho}(X_i^T, X_j^T) = \frac{\tilde{\kappa}(X_i^T, X_j^T)}{\sqrt{\tilde{\kappa}(X_i^T, X_i^T)\tilde{\kappa}(X_j^T, X_j^T)}}. \tag{3.5}$$

Further, with a given spatial proximity $((w_{ij}))$, we can compute this *Spatial Bergsma statistic* $\tilde{S}_B$ using Equation (3.4). The `R` code for the computation of $\tilde{S}_B$ is given in the Appendix.

3.1 Asymptotic normality of $\tilde{S}_B$ under possible dependence

The finite sample distributional properties of global indices are challenging to obtain in general. Usually, additional assumption of normality is made for this purpose, or a randomization approach is used. When T is large, asymptotic distributions are also used. Tiefelsdorf and Boots (1995) showed that the accuracy of the asymptotic distribution of global indices depend heavily on the spatial proximity matrix W and the number of regions R.

Similar to the other global indices, it is impractical to obtain the exact distribution of S_B for fixed T and hence we focus on asymptotic results. The asymptotic normality of $\tilde{S}_B$, as $T \to \infty$ can be established under appropriate assumptions. This is worked out below. This normal distribution turns out to be degenerate when the variables across regions are pairwise independent. This case is discussed in the next section.

Writing Equation (3.4) explicitly we get

$$\tilde{S}_B = \frac{1}{S_0} \sum_{1 \le i < j \le R} (w_{ij} + w_{ji}) \frac{\tilde{\kappa}(X_i^T, X_j^T)}{\left[\, \tilde{\kappa}(X_i^T, X_i^T)\tilde{\kappa}(X_j^T, X_j^T) \,\right]^{1/2}}. \tag{3.6}$$

Define an $\binom{R}{2} \times 1$ column vector $\mathbf{C}$, and a $1 \times \binom{R}{2}$ row vector $\mathbf{d}$ as

$$\mathbf{C} := \Big(\tilde{\kappa}^{(ij)}, 1 \leq i < j \leq R\Big)^{\top},$$
$$\mathbf{d} := \Big(\frac{(w_{ij} + w_{ji})}{S_0\big[\ \tilde{\kappa}(X_i^T, X_i^T)\tilde{\kappa}(X_j^T, X_j^T)\ \big]^{1/2}}, 1 \leq i < j \leq R\Big).$$

Then we can rewrite Equation (3.6) as

$$\tilde{S}_B = \mathbf{d}\mathbf{C}.$$

Define the centered and scaled column vector,

$$\tilde{\mathbf{C}} := \Big(\sqrt{T}\big(\tilde{\kappa}^{(ij)} - \kappa^{(ij)}\big), 1 \leq i < j \leq R\Big)^{\top}.$$

Define $\tilde{S}_{B,st}$ after required centering and scaling of $\tilde{S}_B$ as,

$$\tilde{S}_{B,st} := \mathbf{d}\tilde{\mathbf{C}}. \tag{3.7}$$

We shall refer to $\tilde{S}_{B,st}$ as standarized $\tilde{S}_B$. Note that for every i, ${X_i}^T$, $T \geq 1$ are i.i.d. By an application of the SLLN it follows that,

$$\begin{aligned} \tilde{\kappa}(X_i^T, X_i^T) &\xrightarrow{a.s} \mathbb{E}h_{F_i}(X_i, X_i)^2 \\ &= \sum_{k=0}^{\infty} \big(\lambda_k^{(i)}\big)^2. \end{aligned} \tag{3.8}$$

Hence

$$\mathbf{d} \xrightarrow{a.s.} \Bigg(\frac{(w_{ij} + w_{ji})}{S_0\big[\sum_{k=0}^{\infty}\big(\lambda_k^{(i)}\big)^2\big]^{1/2}\big[\sum_{k=0}^{\infty}\big(\lambda_k^{(j)}\big)^2\big]^{1/2}}, 1 \leq i < j \leq R\Bigg) = \mathbf{a} \ \text{(say)}. \tag{3.9}$$

Recall that $S_0 = R$ if W is row-standardized.

We are now ready to state the asymptotic normality result for the S_B statistic. We use the notation $H_1^{(ij)}$ to denote the kernel function H_1 corresponding to the regions (i, j).

Theorem 1 *Suppose that the vectors $\{(x_{im}, 1 \leq i \leq R)\}$, $m = 1, \cdots, T$ are i.i.d. such that $\mathbb{V}({H_1}^{(ij)}(x_{im}, x_{jm})) < \infty$, for all pairs (i, j). Then as $T \to \infty$,*

$$\tilde{S}_{B,st} = \mathbf{d}\tilde{\mathbf{C}} \xrightarrow{D} N(0, \mathbf{a}^{\top}\Sigma\mathbf{a}).$$

where $\mathbf{a}$ is as in (3.9) and the covariance matrix Σ is defined in the proof given below.

Proof Note that

$$\tilde{\kappa}^{(ij)} = \tilde{\kappa}(X_i^T, X_j^T) \tag{3.10}$$

$$= \binom{T}{2}^{-1} \sum_{1\le m<n\le T} \tilde{h}_{\hat{F}_i}(x_{im}, x_{in})\tilde{h}_{\hat{F}_j}(x_{jm}, x_{jn}). \tag{3.11}$$

For any pair of regions (i, j), by arguments given in Equation (2.9)

$$\sqrt{T}(\tilde{\kappa}^{(ij)} - \kappa^{(ij)}) = \frac{1}{2}T^{-1/2}\sum_{m=1}^{T} H_1{}^{(ij)}\big(x_{im}, x_{jm}\big) + R_T^{(ij)}, \text{ where } R_T^{(ij)} \xrightarrow{P} 0. \tag{3.12}$$

Hence by the multivariate central limit theorem,

$$\tilde{\mathbf{C}}_{\binom{R}{2}\times 1} = \Big(\sqrt{T}\big(\tilde{\kappa}^{(ij)} - \kappa^{(ij)}, 1 \le i < j \le R\Big) \xrightarrow{D} N\Big(\mathbf{0},\ \Sigma_{\binom{R}{2}\times\binom{R}{2}}\Big), \tag{3.13}$$

where, $\Sigma = ((\sigma_{i_1 j_1, i_2 j_2}))_{1\le i_1<j_1\le R, 1\le i_2<j_2\le R}$ is the covariance matrix with

$$\sigma_{i_1 j_1, i_2 j_2} = \frac{1}{4}\mathbb{COV}\Big(H_1{}^{(i_1 j_1)}\big(x_{i_1 m}, x_{j_1 m}\big), H_1{}^{(i_2 j_2)}\big(x_{i_2 m}, x_{j_2 m}\big)\Big). \tag{3.14}$$

For $i_1 = i_2 = i$ and $j_1 = j_2 = j$,

$$\sigma_{i_1 j_1, i_2 j_2} = \frac{1}{4}\mathbb{V}\Big(H_1{}^{(ij)}\big(x_{im}, x_{jm}\big)\Big) = \delta_1{}^{(ij)}. \tag{3.15}$$

Therefore by Equations (3.13) and (3.9), $\tilde{S}_{B,st} = \mathbf{d}\tilde{\mathbf{C}} \xrightarrow{D} N(0, \mathbf{a}^\top\Sigma\mathbf{a})$, completing the proof. □

3.2 Asymptotic distribution of $\tilde{S}_B$ under spatial pairwise independence

Now, suppose that the regions are pairwise independent, that is, $\tilde{\kappa}^{(ij)} = 0$ for all $1 \le i < j \le R$. Then it is known that $\delta_1{}^{(ij)}$ given in (3.15) is 0 for all $i \ne j$, and the limit distribution in the above result is degenerate at 0. In this case, we have the following asymptotic distribution result.

Theorem 2 *Suppose that the vectors $(x_{im}, 1 \le i \le R)$ are i.i.d. for $1 \le m \le T$. Further, for any $1 \le i \ne j \le R$, x_{im} and x_{jm} are independent. Let F_i be the marginal distribution of x_{im}. Suppose that h_{F_i} is square integrable with the eigen decomposition (in the L^2 sense),*

$$h_{F_i}(x, y) = \sum_{k=1}^{\infty} \lambda_k^{(i)} g_k^{(i)}(x) g_k^{(j)}(y). \tag{3.16}$$

Then as $T \to \infty$,

$$T\tilde{S}_B \xrightarrow{D} \frac{1}{S_0} \sum_{1 \le i < j \le R} \left[(w_{ij} + w_{ji}) \sum_{k,l=0}^{\infty} \frac{\lambda_k^{(i)}}{\left[\sum_{t=0}^{\infty} (\lambda_t^{(i)})^2\right]^{1/2}} \frac{\lambda_l^{(j)}}{\left[\sum_{t=0}^{\infty} (\lambda_t^{(j)})^2\right]^{1/2}} (Z_{ik,jl}^2 - 1) \right], \tag{3.17}$$

where $\{Z_{ik,jl}\}$, are i.i.d. standard normal variables, and $S_0 = R$ if W is row-standardized.

Proof Recall from (3.6) that,

$$\tilde{S}_B = \frac{1}{S_0} \sum_{1 \le i < j \le R} (w_{ij} + w_{ji}) \frac{\tilde{\kappa}(X_i^T, X_j^T)}{\sqrt{\tilde{\kappa}(X_i^T, X_i^T)\tilde{\kappa}(X_j^T, X_j^T)}}. \tag{3.18}$$

Further we have

$$\tilde{\kappa}^{(ij)} = \tilde{\kappa}(X_i^T, X_j^T) = \binom{T}{2}^{-1} \sum_{1 \le m < n \le T} \tilde{h}_{\hat{F}_i}(x_{im}, x_{in}) \tilde{h}_{\hat{F}_j}(x_{jm}, x_{jn}). \tag{3.19}$$

As mentioned earlier, due to independence, the first projections $H_1^{(ij)}$ in (3.12) are zero. We now use the second projections from Equation (2.10). We can write,

$$\tilde{\kappa}^{(ij)} = \binom{T}{2}^{-1} \sum_{1 \le m < n \le T} h_{F_i}(x_{im}, x_{in}) h_{F_j}(x_{jm}, x_{jn}) + R_T^{(ij)}, \text{ where } TR_T^{(ij)} \xrightarrow{P} 0.$$

Now we can follow the proof of the general theorem on the asymptotic distribution of a U-statistic with a general degenerate kernel. The difference is that now we have several degenerate U-statistics. Each of them converges by the limit theorem for degenerate U-statistics. See for example (Bose and Chatterjee 2018). We need to ensure the joint convergence of these statistics, and argue the independence of the $\{Z_{ik,jl}\}$ across i, j. We outline the arguments below.

Using (3.2), we can write the following approximate equation for $\tilde{\kappa}$ calculated for any pair of regions, i.e., $\tilde{\kappa}^{(ij)}$ as follows:

$$\begin{aligned} T\tilde{\kappa}^{(ij)} &\simeq \frac{2}{T}\Big[\sum_{1 \le m < n \le T} h_{F_i}(x_{im}, x_{in}) h_{F_j}(x_{jm}, x_{jn}) \Big] \\ &= \frac{2}{T}\Big[\sum_{1 \le m < n \le T} \Big[\sum_{k=0}^{\infty} \lambda_k^{(i)} g_k^{(i)}(x_{im}) g_k^{(i)}(x_{in}) \Big] \Big[\sum_{k=0}^{\infty} \lambda_k^{(j)} g_k^{(j)}(x_{jm}) g_k^{(j)}(x_{jn}) \Big] \Big] \\ &= \frac{1}{T} \sum_{m,n=1}^{T} \Big[\sum_{k=0}^{\infty} \lambda_k^{(i)} g_k^{(i)}(x_{im}) g_k^{(i)}(x_{in}) \Big] \Big[\sum_{k=0}^{\infty} \lambda_k^{(j)} g_k^{(j)}(x_{jm}) g_k^{(j)}(x_{jn}) \Big] \end{aligned}$$

$$-\frac{1}{T}\sum_{m=1}^{T}\Big[\sum_{k=0}^{\infty}\lambda_k^{(i)}g_k^{(i)}(x_{im})g_k^{(i)}(x_{im})\Big]\Big[\sum_{k=0}^{\infty}\lambda_k^{(j)}g_k^{(j)}(x_{jm})g_k^{(j)}(x_{jm})\Big]$$
$$= T_1 - T_2 \ \ (say).$$

Note that $\{g_k^{(i)}(\cdot)\}$ and $\{g_l^{(j)}(\cdot)\}$ are orthonormal functions. We explain in brief the convergence of the two terms. The second term equals

$$\begin{aligned} T_2 &= \frac{1}{T}\sum_{m=1}^{T}\Big[\sum_{k=0}^{\infty}\lambda_k^{(i)}g_k^{(i)}(x_{im})g_k^{(i)}(x_{im})\Big]\Big[\sum_{l=0}^{\infty}\lambda_l^{(j)}g_l^{(j)}(x_{jm})g_l^{(j)}(x_{jm})\Big] \\ &\xrightarrow{a.s.} \Big(\sum_{k=0}^{\infty}\lambda_k^{(i)}\Big)\Big(\sum_{l=0}^{\infty}\lambda_l^{(i)}\Big). \end{aligned}$$

The first term equals,

$$\begin{aligned} T_1 &= \frac{1}{T}\sum_{m,n=1}^{T}\Big[\sum_{k=0}^{\infty}\lambda_k^{(i)}g_k^{(i)}(x_{im})g_k^{(i)}(x_{in})\Big]\Big[\sum_{l=0}^{\infty}\lambda_l^{(j)}g_l^{(j)}(x_{jm})g_l^{(j)}(x_{jn})\Big] \\ &= \sum_{k,l=1}^{\infty}\lambda_k^{(i)}\lambda_l^{(j)}\Big[\frac{1}{T}\sum_{m,n=1}^{T} g_k^{(i)}(x_{im})g_k^{(i)}(x_{in})g_l^{(j)}(x_{jm})g_l^{(j)}(x_{jn})\Big] \\ &= \sum_{k,l=1}^{\infty}\lambda_k^{(i)}\lambda_l^{(j)}\Big[\Big(\frac{1}{\sqrt{T}}\sum_{m=1}^{T} g_k^{(i)}(x_{im})g_l^{(j)}(x_{jm})\Big)\Big(\frac{1}{\sqrt{T}}\sum_{n=1}^{T} g_k^{(i)}(x_{in})g_l^{(j)}(x_{jn})\Big)\Big] \\ &\xrightarrow{D} \sum_{k,l=0}^{\infty}\lambda_k^{(i)}\lambda_l^{(j)}Z_{ik,jl}^2. \end{aligned}$$

Therefore we have, for every pair $i < j$,

$$T\tilde{\kappa}^{(ij)} \xrightarrow{D} \sum_{k,l=0}^{\infty}\lambda_k^{(i)}\lambda_l^{(j)}\big(Z_{ik,jl}^2 - 1\big),$$

where $Z_{ik,jl}$, $1 \le k, l < \infty$ are independent standard normal variables.

Moreover, when we consider the joint convergence of $T\tilde{\kappa}^{(ij)}$, $1 \le i < j \le R$, due to the pairwise independence X_i^T and X_j^T, and the orthonormality of $\{g_k^{(i)}(\cdot)\}$ and $\{g_l^{(j)}(\cdot)\}$, the variables $Z_{ik,jl}$ are all independent of each other. Incidentally, only pairwise independence is being used here. That is,

$$\Big(T\kappa^{(ij)}, 1 \le i < j \le R\Big) \xrightarrow{D} \Big(\sum_{k,l=0}^{\infty}\lambda_k^{(i)}\lambda_l^{(j)}\big(Z_{ik,jl}^2 - 1\big), 1 \le i < j \le R\Big). \tag{3.20}$$

Now, in the denominator of (3.18) we have terms of the form $\tilde{\kappa}(X_i^T, X_i^T)$ that converge to constants given in Equation (3.8).

From Equations (3.20) and (3.8) we obtain that (3.17) holds, completing the proof. □

It may be noted that the limit distribution of $\tilde{S}_B$ given in Equation (3.17) depends on $\{F_i\}$ through the eigenvalues $\{\lambda_k^{(i)}\}$ given by Equation (3.16). For most distributions, these cannot be computed in a closed form. However, they can be numerically approximated. Moreover, even though the limit distribution depends on the underlying parent distributions through the eigenvalues $\{\lambda_k^{(i)}\}$, the self-normalisation of the eigenvalues in the limit distribution (see Equation (3.17)) hints at a robustness of the limiting behavior against changes in the underlying distributions. This is borne out by the simulations in the next section.

In the special case where $\{X_{i1}, i = 1, \ldots, R\}$ have a common distribution F, we have $F_i \equiv F$ for all i. Let $(\lambda_k^{(i)}, g_k^{(i)}) \equiv (\lambda_k, g_k)$ for all $1 \leq i \leq R$. Then

$$T\tilde{S}_B \xrightarrow{D} \frac{1}{S_0 \sum_{k=0}^{\infty} \lambda_k^2} \sum_{1 \leq i < j \leq R} \Big[\big(w_{ij} + w_{ji}\big) \sum_{k,l=0}^{\infty} \lambda_k \lambda_l \big(Z_{ik,jl}^2 - 1\big) \Big]. \tag{3.21}$$

4 Simulation study

We now explore various distributional aspects of the $\tilde{S}_B$ statistic through simulations. Since we are going to explore the COVID-19 data from the 14 districts of the state of Kerala in India in the next section, we choose $R = 14$. We assume that we have observations over a reasonable length of time, and choose $T = 50$. For a spatially extended simulation with large number of regions, see (Kappara et al. 2023).

We use three different spatial proximity matrices: (i) the lag-1 adjacency matrix of the districts (i.e. regions) (ii) distance matrix of these districts, using the average of latitude and longitude of the district headquarters, and (iii) linear connectivity matrix, which is a lag-1 adjacency matrix with the regions arranged in a line. The third W matrix is motivated by the almost linear geographical organization of the districts of Kerala from North-West to South-East.

The $\tilde{S}_B$ statistic is studied by simulating it under various scenarios in the absence and presence of spatial association. This requires obtaining samples so that we can apply formulae (3.1) and (3.3). Note that $S_0 = R$ since we are working with row-standardized W matrices.

In Sect. 4.1, we study the distributions under the null case of no dependence between the regions and also validate the asymptotic approximations obtained in (3.17) and (3.21). Note that to use the finite-discrete approximation of the asymptotic distribution, we would require the eigenvalues of the kernels for the corresponding distributions.

Then in Sect. 4.2 we study the $\tilde{S}_B$ statistic in presence of spatial dependence, using two well-known spatial processes, namely spatial autoregression and spatial moving average.

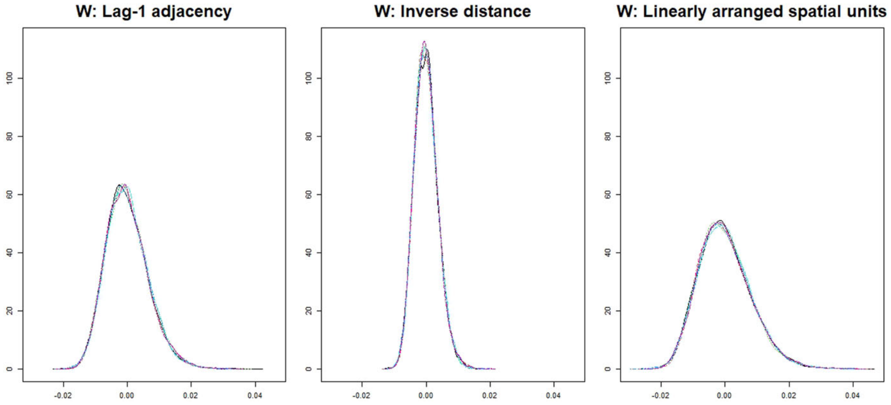

Fig. 1 Simulated null distribution of $T\tilde{S}_B$, $T = 50$, $R = 14$ for 10, 000 replicates, for six distributions. Proximity matrices W: lag-1 adjacency (left), inverse distance (middle), and linearly arranged spatial units (right)

4.1 Null case of spatial independence

In this simulation exercise, we have used a common choice of the distribution F for all R spatial locations–six different choices for F, namely normal, uniform, exponential, Laplace, logistic, and chi-square were explored. The eigenvalues $\{\lambda_k\}$ are readily available for these cases. Empirical null distributions of $T\tilde{S}_B$ were obtained using 10,000 replicates of the estimates of $\tilde{S}_B$, under spatial pairwise independence and these six choices of F.

From Fig. 1 we observe that under the null scenario of spatial independence, the finite sample distribution of $T\tilde{S}_B$ does not depend in any significant way on the distributional assumptions. However, there are visible effects of the spatial proximity matrices.

Next we validate the asymptotic behaviour of $T\tilde{S}_B$ stated in Theorem 2. As mentioned earlier this would require discrete approximations using eigenvalues. For the detailed procedure of this discrete approximation method, see (Kappara et al. 2022). We consider the finite sum based on 100 eigenvalues to approximate the asymptotic null distributions.

This theoretical asymptotic distribution is computed for the case where all regions have identical distribution (normal in this illustration). This is then compared with the simulated empirical distribution generated under the same assumption of normality for all regions.

For each choice of spatial proximity matrix, we have approximated the null distributions, as described in (3.21), based on 10, 000 samples, see Fig. 2.

Figure 2 illustrates that even with only $T = 50$, and 100 eigenvalues, the asymptotic distribution provides a good approximation to the (empirical) null distribution. Also, as before, the underlying distributional assumption appears to be less critical than the choice of the spatial proximity matrix.

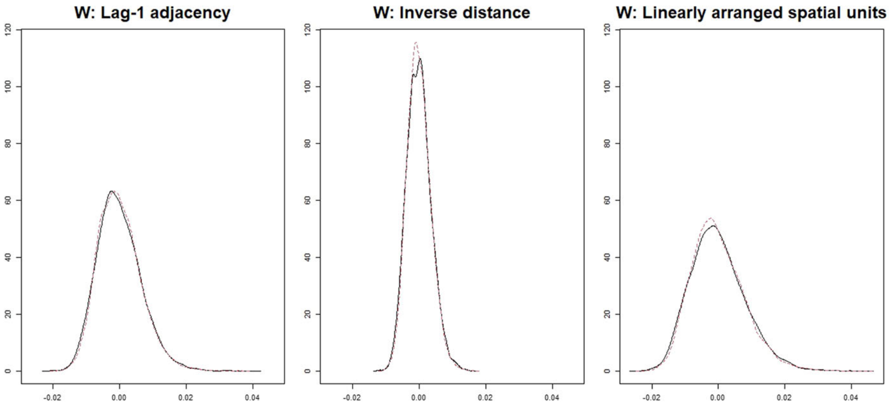

Fig. 2 Comparison of simulated null (black-solid line) and asymptotic null distributions (red-dashed-line) of $T\tilde{S}_B$ with 10, 000 replicates, based on finite discrete approximation using 100 eigenvalues for asymptotic distribution, with W matrices: lag-1 adjacency (left), inverse distance (middle), and linearly arranged spatial units (right)

4.2 Spatial dependence case

In order to illustrate the behaviour of $\tilde{S}_B$ under the presence of spatial dependence, we consider the following two general forms of spatial dependence models to simulate spatially dependent data. In simulations, we maintain the assumptions on the data given in Theorem 1, that the observations are i.i.d. across the time points.

1. Spatial moving average (SMA) dependence:

$$y = (I + \theta W)\epsilon. \tag{4.1}$$

2. Spatial autoregressive (SAR) dependence:

$$y = (I - \theta W)^{-1}\epsilon, \tag{4.2}$$

where, θ is the spatial dependence parameter, W (row standardized) corresponds to the assumed spatial proximity, and ϵ is a vector of standard normal variates throughout the simulations. Under both the models, $\theta = 0$ refers to the null case of no spatial association. Note that (Anselin and Florax (2012)) gives collection of some other spatial dependence models of higher orders.

We consider a range of values for the spatial dependence parameter θ, namely 0, 0.1, 0.25, 0.5, 0.75, and 0.9 in this illustration. We continue to use the same three choices of spatial proximity matrices described earlier, namely lag-1, inverse distance, and linear connectivity. As before we have simulated spatially autocorrelated vectors of length $T = 50$ at $R = 14$ locations.

For each choice of the W matrix and the dependence parameter θ, under the two spatial dependence models the $\tilde{S}_B$ was simulated 10,000 times. In Fig. 3 we show the distribution of $\tilde{S}_B$ under SMA and SAR models.

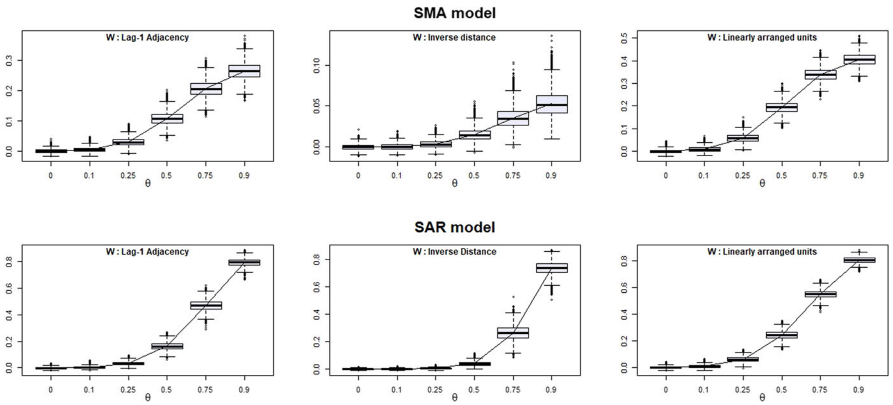

Fig. 3 Empirical distributions of $\tilde{S}_B$, R=14, T=50, 10, 000 replicates, with sample mean overlayed, dependence parameter θ on x-axis under SMA model (top), SAR model (bottom). Proximity matrices: lag-1 adjacency (left), inverse distance (middle), and linearly arranged spatial units (right)

The box plots in Fig. 3 show the distinct and monotone departure of $\tilde{S}_B$ from its null behaviour under the presence of spatial dependence. Incidentally, row-standardization of W is crucial to yield this monotonicity.

It appears that as spatial association increases, the distribution seems to become more symmetric around the median. To confirm this, we study the skewness and kurtosis of this measure in Fig. 4.

We see that for θ values close to 0, i.e., relatively close to the null, the distribution of $\tilde{S}_B$ is moderately skewed. This observation is consistent with our result in Theorem 2 that under the null, $T\tilde{S}_B$ approximately has a distribution of a centered sum of weighted chi-squares.

As the spatial parameter θ increases, the skewness decreases gradually, and kurtosis decreases and becomes close to 3 under lag-1 adjacency and linear connectivity matrices (i.e. W_1 and W_3). This is consistent with the asymptotic normality result we have obtained in Theorem 1.

Note that for lag-1 adjacency and the linearly arranged proximity matrices, the proportion of sparsity are 0.77 and 0.87. In comparison, the inverse distance W matrix has a sparsity of 0.07 only. Clearly the sparsity of the W matrix also influences the distribution of $\tilde{S}_B$.

Overall we notice, from the Figs. 3 and 4, that the presence of spatial association highly influences the first four moments of the distribution of $\tilde{S}_B$.

5 COVID-19 application

In this section, we present an application of $\tilde{S}_B$ to the COVID-19 data from the Indian state of Kerala in the southern part of India. We consider district-level monthly COVID-19 data for the 14 districts of Kerala as of October 31, 2022. The data is obtained from a crowd-sourced database https://data.incovid19.org/. We have the number of districts

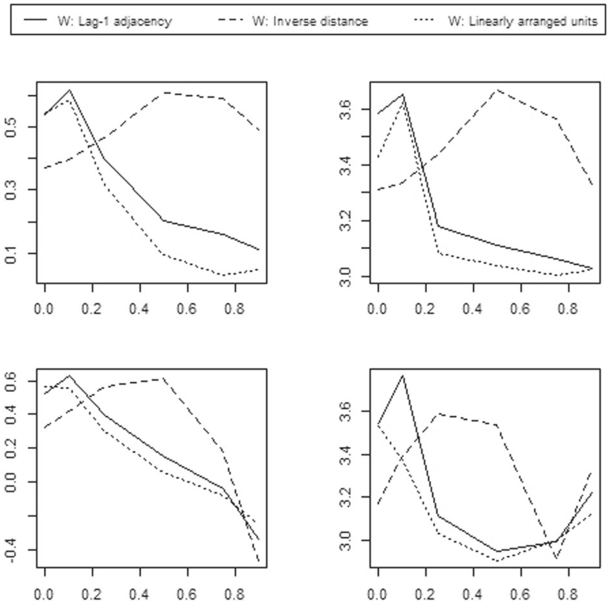

Fig. 4 Empirical skewness (left) and kurtosis (right) of $\tilde{S}_B$, plotted against spatial dependence parameter θ, under SMA model (top) and SAR model (bottom), with proximity matrix: lag-1 adjacency matrix (solid line), inverse distance matrix (dashed line), and linearly arranged spatial units (dotted line) (R=14, T=50, 10, 000 replicates)

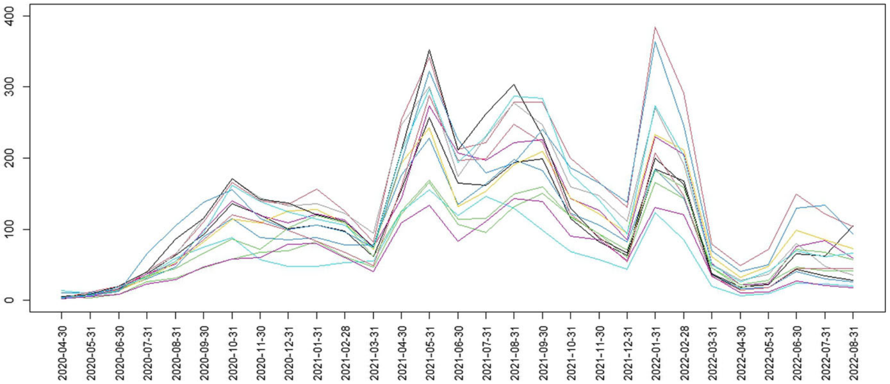

Fig. 5 District level COVID-19 time series data (square root of number of cases) for Kerala

$R = 14$ and the number of time points $T = 29$ (months). Figure 5 presents the time series of this COVID-19 incidence data.

We present the analyses here using the lag-1 adjacency matrix motivated by the works of Bhattacharjee et al. (2021). In its current form, the $\tilde{S}_B$ statistic measures spatial association in absence of temporal pattern. However, COVID-19 data naturally has temporal dependence. In order to meaningfully apply this measure, first appropriate temporal models are applied to the data, and then the residuals are considered for assessing spatial dependence.

Apart from measuring spatial dependence, we would also assess the significance of such an $\tilde{S}_B$ measure. For that, we use the asymptotic null distributions presented in Sect. 3.2 and obtain an empirical p-value of the corresponding $T\tilde{S}_B$ value. Further, using bootstrap technique we also present confidence intervals.

From Fig. 5 we observe a strong temporal pattern in the COVID-19 incidence. However, it is also apparent that there is a (spatial) consistency in this temporal behaviour.

To capture these features, we explore three models, with the first one being only temporal, and other two involving spatio-temporal modelling. By nature, the first two models are endogenous (viz. temporal AR model and spatio-temporal autoregressive model), and the third one is exogenous (viz. a spatio-temporal-gravity model).

Let x_{it}=number of new cases for the ith district during the time period $t = 1, \ldots, 29 (= T), i = 1, \ldots, 14 (= R)$. Then the models are defined as follows.

AR(3) Model. Here we fit the AR(3) models to each of the 14 district-level time series separately and extract the residuals from each series. The model is defined as,

$$x_{it} = \beta_0^{(i)} + \beta_1^{(i)} x_{i(t-1)} + \beta_2^{(i)} x_{i(t-2)} + \beta_3^{(i)} x_{i(t-3)} + \epsilon_{it}. \tag{5.1}$$

Here the superscript (i) in the parameters β represent parameters of the AR model fitted to the time series from the i-th district, $i = 1, \ldots, 14 (= R)$.

Spatio-temporal Model-1. In this model we assume that for a given region i, the count for the tth month depend on the count for the $(t-1)$th month of that region along with all the N_i neighbors of that region. That is, we regress each district's time series on its own past at lag-1 and the past of its spatial neighbors. Thus the model involves both temporal and spatial covariates.

We further assume that the count response variable follows either a Poisson distribution or a Negative Binomial distribution with mean μ_{it}. Then the linear predictor for log-mean has the following form with the spatial and temporal autoregressive terms;

$$\begin{aligned} &\log(\mu_{it}) = \beta_0^{(i)} + \beta_1^{(i)} x_{i(t-1)} + \sum_{j \in N_i} \beta_{2,j}^{(i)} w_{ij} x_{j(t-1)}, \\ &t = 1, \ldots, 29, \qquad j = 1, \ldots, 14. \end{aligned} \tag{5.2}$$

Again, the superscript (i), for the β parameters, indicates that the model is for the time series from the ith district, $i = 1, \ldots, 14 (= R)$. Here too the models are fitted to the time-series from individual regions and residuals are extracted.

Spatio-temporal Model-2. The third model is an application of a specialized spatio-temporal model proposed in Bhattacharjee et al. (2021). All covariates in this model are exogenous. They include spatial-variates, temporal-variates, and covariates such as air passenger traffic data, that includes both spatial and temporal information. This model is fitted simultaneously to all R time series due to shared parameters/covariates in the model.

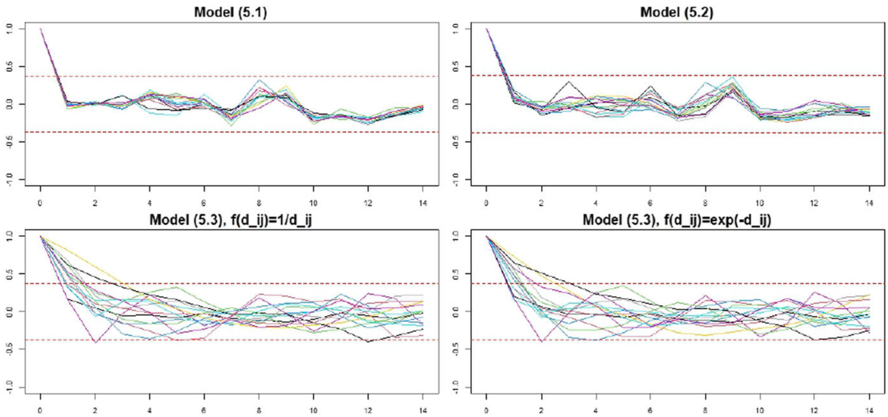

Fig. 6 Autocorrelation of residuals: Model 5.1 (top left), Model 5.2 (top right), Model 5.3 with inverse distance decay (bottom left), and exponential decay (bottom right) with 95% threshold

The state of Kerala has four operational airports at Trivandrum, Calicut, Cochin and Kannur. The data on passenger traffic at these airports have been obtained from the Airport Authority of India (AAI) source https://www.aai.aero/en/business-opportunities/aai-traffic-news. We will denote the covariates for this model as follows:

d_{ik} is the (Euclidean) distance between the i-th district center and the k-th airport.

X_{1kt} is the number of passengers arriving at the k-th airport at the t-th time point.

X_2 and X_3 are two numerical variables containing the geographical location of each district head quarter (i.e. longitude and latitude respectively).

X_4 is a categorical variable for the T time points.

Once again we assume the response variable x_{it} of COVID-19 incidence follows a negative binomial distribution, where the linear predictor for log-mean is given by,

$$\log \mu_{it} = \beta_0 + \sum_{k=1}^{4} \beta_{1,k} f(d_{ik}) X_{1kt} + \beta_2 X_2 + \beta_3 X_3 + \sum_{t=1}^{T} \beta_{4,t} I_{\{X_4=t\}}. \quad (5.3)$$

The first part of the model with parameters $\beta_{1,k}$ is adopted in the spirit of gravity models and explains the effect of the volume of air passenger traffic on the district-level COVID-19 incidence. The (inverse) distance between airports and district headquarters is captured by a distance decay function. Based on the results from Bhattacharjee et al. (2021), we use two decay functions, the inverse decay $f(d) = 1/d$, and exponential decay $f(d) = \exp(-d)$.

In Fig. 6 we present the autocorrelation functions from the 14 districts under these three models (with two distance decay functions for the third model). From the residual autocorrelations plot from an AR(3) model in Fig. 6 we observe that, in spite of fitting the models separately for the 14 districts, the residuals behave similarly. This strengthens the idea of applying spatio-temporal models to this data.

Table 1 Results on $\tilde{S}_B$ from the fitted models for COVID-19 incidence data

Model	$\tilde{S}_B$	CI	p-value
AR(3) model	0.85	(0.7278, 0.9858)	0.0127
ST model-1	0.78	(0.6163, 0.9126)	0.0182
ST model-2 ($f(d) = 1/d$)	0.09	(0, 0.0972)	0.3578
ST model-2 ($f(d) = \exp(-d)$)	0.14	(0, 0.1530)	0.3103

From Fig. 6, we also observe that the strong clustered pattern of the autocorrelation functions seems to dissipate with applications of the two spatio-temporal models. It appears to be most de-clustered for the spatio-temporal model-2.

All four sets of autocorrelation plots in Fig. 6 are within the 95% threshold for all R time-series residuals. Therefore for practical purposes, we can assume an i.i.d. structure within each of the residual series from the $R = 14$ districts. Thus the statistic $\tilde{S}_B$ can be computed based on these residuals, and we can apply our distributional results.

The above observations are further confirmed by carrying out a test of significance for the $\tilde{S}_B$ statistics (see Table 1).

For the spatio-temporal model-2 (with either distance function), we are unable to reject the null hypothesis of no spatial association. Thus we can conclude that the residuals obtained from this model are spatially pairwise independent. Therefore Model (5.3) satisfactorily explains the spatial and temporal incidence pattern of COVID-19 in Kerala.

Based on the overall conclusion on spatial association (lack thereof in the residuals after applying the spatio-temporal model with the exponential decay model 5.3), we further investigated the presence of individual pairwise dependence, if any. Accordingly in Fig. 7 we present the pairwise Bergsma correlations (see Bergsma 2006) for the 14 districts based on the residuals obtained from Model 5.3.

As was shown for the $\tilde{S}_B$ statistic, for the $\tilde{\rho}$ statistics also it is possible to assess significance (using an empirically derived cutoff of $\tilde{\rho} > 0.17$). By applying such a test, we conclude pairwise independence of most pairs of districts. There appear to be a few sporadic significant pairs, which may be due to some yet unknown factor(s).

6 Discussion

We have used a specific correlation measure and its estimate to define a global spatial measure of association. This measure is asymptotically normal when we have observations that are i.i.d. over time but are spatially dependent. In the absence of spatial dependence, more precisely when there is only pairwise spatial independence, the estimate has an asymptotic distribution involving an infinite sum of weighted i.i.d. chi-square variables. In both cases, the distribution depends on the unknown underlying distribution of the observations, especially through the eigenvalues of appropriate kernel functions.

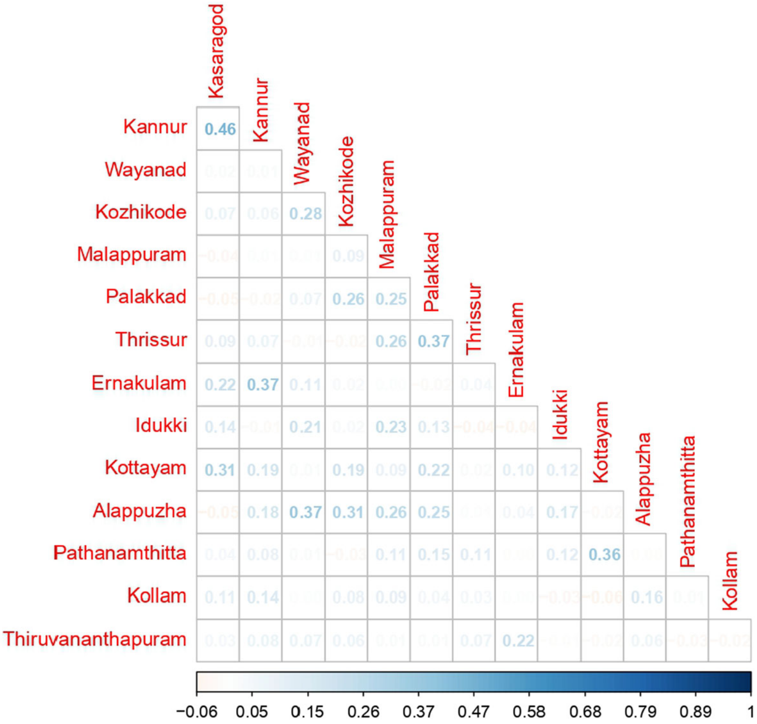

Fig. 7 Pairwise estimated Bergsma correlations for the residuals from the spatio-temporal Model (5.3) with exponential distance decay

Simulations show that for a reasonably large number of observations, the actual finite sample distributions are well approximated by the asymptotic distributions. It is also seen that these measures and their distributions are sensitive to the choice of a spatial proximity matrix, but are not sensitive to the nature of the underlying distributions.

At present, no distributional properties are known for the association estimate when the observations are also temporally dependent. In spite of this, the measure can be used for model-fitting purposes even in the presence of temporal dependence. This is done by first removing the temporal dependence through appropriate modelling and then using the measure on the residual series.

We have presented such an application for spatio-temporal modelling of COVID-19 data on the monthly time series data of the 14 districts of the Indian state of Kerala.

While considering a larger number of spatial locations may be useful in some cases, one should be cautious of their potential adverse effect on our ability to identify dependence. It is possible that spatial dependence exists only locally and an overall measure may fail to identify it. Further care needs to be taken in choosing a distance measure when working with a large number of regions. See (Kappara et al. 2023) for an instance where we find a diminished sensitivity of the measure when R is large and the inverse distance proximity matrix is used.

Another possible topic of investigation could be to consider situations where multivariate observations are available on each spatial unit. This may be feasible given that the distance covariance measure when X and Y are multivariate reduces to κ when X and Y are univariate. However, whether the distributional properties would still be possible to obtain remains to be seen.

Acknowledgements We are grateful to the Referees and the Associate Editor for their insightful comments which have led to a significant improvement in the manuscript.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix: `R` code

We give below two `R` codes. The first is for the U-statistic based estimate of Bergsma correlation. Input for this first code is bivariate data in the form of two vectors, x and y. The second is a code for computing the $\tilde{S}_B$ statistic. For this data is to be given in the format of a matrix with T rows (time points) and R columns (locations), and an $R \times R$ Spatial proximity matrix.
Bergsma's $\tilde{\rho}$

```
rho_tilde <- function(x,y)
{
n   = length(x)
X   = matrix( replicate(n, x), nrow = n); Dx = abs( X - t(X))
Y   = matrix( replicate(n, y), nrow = n); Dy = abs( Y - t(Y))
A1  = apply(Dx,1,mean)
A2  = apply(Dy,1,mean)
B1  = mean(A1)
B2  = mean(A2)
Dx1 = sweep(Dx,1,(n/(n-1))*A1)
Dx2 = sweep(Dx1,2,(n/(n-1))*A1)
Hx  = (-1/2)*(Dx2 +(n/(n-1))*B1)
Dy1 = sweep(Dy,1,(n/(n-1))*A2)
Dy2 = sweep(Dy1,2,(n/(n-1))*A2)
Hy  = (-1/2)*(Dy2 +(n/(n-1))*B2)
Hxy = Hx*Hy
I   = matrix(1,n,n)
I[lower.tri(I,diag=TRUE)] = 0
r_tilde = sum(I*Hxy)/(sqrt(sum(I*Hx*Hx)*sum(I*Hy*Hy))) #
computing rho_tilde return(r_tilde)
}
```

$\tilde{S}_B$ statistic

```
SB <- function(data, W){
r_curl = matrix(0, ncol(data), ncol(data))
for(i in 1:ncol(data)){
for(j in i:ncol(data)){
r_curl[i,j] = rho_tilde(data[,i], data[,j])
r_curl[j,i] = r_curl[i,j]
}}
Spatial_AI = sum(W*r_curl)/sum(W)
return(Spatial_AI)
}
```

References

Anselin L (1988) Spatial econometrics: methods and models, vol 4. Springer Science & Business Media, Berlin

Anselin L (1995) Local indicators of spatial association-LISA. Geogr Anal 27(2):93–115

Anselin L, Florax R (2012) New directions in spatial econometrics. Springer Science & Business Media, Berlin

Bergsma WP (2006) A new correlation coefficient, its orthogonal decomposition and associated tests of independence. Preprint arxiv.math/0604627

Bhattacharjee M, Kappara D, Bose A (2021) Modelling COVID-19 data II: spatio-temporal models with application to Kerala data. J Indian Stat Assoc, 59(2). To appear

Bose A, Chatterjee S (2018) U-statistics, M_m-estimators and resampling. Springer, Berlin

Bose A, Kappara D, Bhattacharjee M (2023) Estimation of Bergsma's covariance. J Korean Stat Soc 52(3):1–30

Cliff A, Ord K (1972) Testing for spatial autocorrelation among regression residuals. Geogr Anal 4(3):267–284

Dubé J, Legros D (2013) A spatio-temporal measure of spatial dependence: An example using real estate data. Pap Reg Sci 92(1):19–30

Gao Y, Cheng J, Meng H, Liu Y (2019) Measuring spatio-temporal autocorrelation in time series data of collective human mobility. Geo-spat Inf Sci 22(3):166–173

Geary RC (1954) The contiguity ratio and statistical mapping. Incorp Stat 5(3):115–146

Getis A (2007) Reflections on spatial autocorrelation. Reg Sci Urban Econ 37(4):491–496

Getis A (2008) A history of the concept of spatial autocorrelation: a geographer's perspective. Geogr Anal 40(3):297–309

Getis A, Ord JK (2010) The analysis of spatial association by use of distance statistics. Perspectives on spatial data analysis. Springer, Berlin, pp 127–145

Kappara D, Bose A, Bhattacharjee M (2022) Assessing bivariate independence: Revisiting Bergsma's covariance (47 pages). Preprint arXiv:2212.08921

Kappara D, Bose A, Bhattacharjee M (2023) Measuring spatial association and testing spatial independence based on short time course data (21 pages). Preprint arXiv:2303.16824 [stat.ME]

Martin RL, Oeppen J (1975) The identification of regional forecasting models using space–time correlation functions. Transactions of the Institute of British Geographers, pp 95–118

Moran PA (1948) The interpretation of statistical maps. J R Stat Soc Ser B (Methodol) 10(2):243–251

Shao-Kuan C, Wei W, Bao-Hua M, Wei G (2013) Analysis on urban traffic status based on improved spatio-temporal Moran's I. Acta Phys Sin 62(14)

Székely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. Ann Stat 35(6):2769–2794

Tiefelsdorf M, Boots B (1995) The exact distribution of Moran's I. Environ Plan A 27(6):985–999

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.