Author for correspondence:
Maurizio Porfiri
e-mail: mporfiri@nyu.edu

An information-theoretic approach to study spatial dependencies in small datasets

Maurizio Porfiri[1,2,3] and Manuel Ruiz Marín[3]

[1] Department of Mechanical and Aerospace Engineering, New York University Tandon School of Engineering, Brooklyn NY, USA, [2] Department of Biomedical Engineering, New York University Tandon School of Engineering, Brooklyn NY, USA, [3] Department of Quantitative Methods, Law and Modern Languages, Technical University of Cartagena, Cartagena, Murcia, Spain

From epidemiology to economics, there is a fundamental need of statistically-principled approaches to unveil spatial patterns and identify their underpinning mechanisms. Grounded in network and information theory, we establish a nonparametric scheme to study spatial associations from limited measurements of a spatial process. Through the lens of network theory, we relate spatial patterning in the dataset to the topology of a network on which the process unfolds. From the available observations of the spatial process and a candidate network topology, we compute a mutual information statistic that measures the extent to which the measurement at a node is explained by observations at neighbouring nodes. For a class of networks and linear autoregressive processes, we establish closed-form expressions for the mutual information statistic in terms of network topological features. We demonstrate the feasibility of the approach on synthetic datasets comprising 25 to 100 measurements, generated by linear or nonlinear autoregressive processes. Upon validation on synthetic processes, we examine datasets of human migration under climate change in Bangladesh and motor vehicle deaths in the United States of America. For both these real datasets, our approach is successful in identifying meaningful spatial patterns, begetting statistically-principled insight into the mechanisms of important socioeconomic problems.

THE ROYAL SOCIETY
PUBLISHING

1. Introduction

The identification and quantification of spatial dependencies in ensembles of random variables is a ubiquitous problem in science and engineering [1–3]. From epidemiology [4] to ecology [5], meteorology [6], agriculture [7], and economics [8], there is a pressing need to elucidate spatial relationships in real datasets and offer insight into the underlying mechanisms of spatial association. For example, models of vegetation distributions can be improved by incorporating spatial dependencies of environmental variables [5]. Likewise, accounting for spatial dependencies in the rain gauge network enhances the accuracy of precipitation models [6].

As noted by Getis [2], who borrowed language from the seminal work of Cliff and Ord [9, 10], two key recurrent questions in spatial analysis are: "Is the spatial pattern displayed by the phenomenon significant in some sense and therefore worth interpreting?" and "Can we obtain any information on the processes which have produced the observed pattern from an analysis of the mapped distribution of the phenomenon?"

Toward addressing these questions, different approaches have been proposed to measure statistical dependencies in a spatial process. One of the first and most celebrated tool to study spatial relationships is Moran's spatial autocorrelation index [11], which measures the spatial correlation among different locations in space. The index is designed to detect positive or negative spatial autocorrelation according to the interaction scheme among the variables. Based on Moran's index, different parametric tests for spatial dependence have been formulated over the years, including the maximum-likelihood methodology [12], instrumental variables tests [13], and Lagrange multiplier tests [14,15].

While these tests largely rely on linear dependence among the variables, the growing interest toward nonlinear spatial modeling has fueled the introduction of a number of nonparametric tests. For example, the test formulated by De Graaff *et al.* [16] offers a spatial variant of the Broock-Dechert-Scheinkman (BDS) test for time-series analysis [17], in the form of a misspecification test for spatial regression. Likewise, López *et al.* [18] extended prior work on time-series analysis by Matilla-García and Ruiz Marín [19] to construct a simple test of spatial independence that uses concepts in symbolic dynamics and information theory. The τ test conceived by Brett and Pinkse [20] examines spatial independence through the use of characteristic functions, under the premise that a set of random variables are independent if the joint characteristic function is equal to the product of the marginals.

Although many of these approaches successfully unveiled hidden patterns of spatial association in real datasets, they often rely on strong assumptions that may limit their broad application. These limitations could include the restrictive premise of linear interactions [11–15], prior knowledge about the underlying spatial process whose features must be estimated [12–15], and large datasets to afford reliable conclusions [16,18].

It is particularly important to draw inferences regarding spatial association with small datasets, comprising only 25 to 100 measurements. For example, the study of human migration [21] and policy making [22] must often rely on measurements at the coarse level of districts, regions, states, or countries. Existing nonparametric methodologies are inherently data hungry; for example, the study of the population distribution by López *et al.* [18] relied on population data for all the 257 regions of the European Union at Nuts II level.

Toward the formulation of an association scheme that is tailored to the study of small datasets, we frame our approach in the context of networks [23,24]. In this vein, we view the spatial process as a finite ensemble of random variables, each related to a node in the network. Links in the network encapsulate spatial dependencies in the process, whereby the value of one random variable depends on the value of its neighbours. Spatial independence corresponds to a disconnected network, in which nodes are isolated from each other. On the contrary, spatial association translates into a connected network, whose topology is indicative of the inner workings of spatial patterning. Cogently incorporating the adjacency matrix of the network in the statistical test is expected to contribute two main advantages: i) clarifying the mechanisms

underneath spatial association, whereby spatial association will be tied to a specific topology, and ii) mitigating the possibility of spurious rejections of the null hypothesis of independence.

Under some assumptions of spatial stationarity, we propose an information-theoretic test of spatial association that only requires a single realization of the random variables and a candidate network topology. From these realizations and the candidate topology, for every node of the network we calculate the local mean of its neighbours. By calculating mutual information [25] between the nodal variable and the mean of its neighbours, we ultimately test the independence of the nodes from their neighbours. Should the candidate topology have no overlap with the true interaction pattern, mutual information will be equal to zero; on the other hand, nonzero values of mutual information will identify candidate networks that capture some of the true interaction between the nodal variables. The estimation of the mutual information statistic can be carried out by binning the variables under consideration, thereby facilitating the study of small datasets [26,27]. In principle, the test is model-free, whereby it does not require knowledge about the nature of the spatial process, which could be linear or nonlinear.

We provide rigorous analytical results for the validity of the proposed information-theoretic association scheme for the case of k-circulant networks [28] and linear spatial autoregressive (SAR) processes [1,3], for which we exactly fulfill the spatial stationarity condition and can compute mutual information in closed-form. Using insight from the functional dependence of mutual information on the network topology, we introduce a perturbation analysis that extends our claims to k-regular networks, where all the nodes share the same out-degree but may have different in-degree [23]. For both cases, we prove that the mutual information statistic is controlled by topological features of the network (average number of directed and undirected links connected to any network node) and by the parameter that measures the intensity of the spatial dependence in the SAR process.

Working with synthetic data, we demonstrate the feasibility of the proposed association scheme on linear and nonlinear SAR processes, from 25 to 100 measurements. Our results point at a predictable effect of the network topology and size on the success of the approach, whereby decreasing the connectivity of the network or increasing its size are found to have a beneficial effect on the rejection rate of the null hypothesis of spatial independence. Upon gaining confidence in the approach, we examine two real datasets: human migration under climate change in Bangladesh [21] and motor vehicle deaths in the United States of America (USA) [22]. For both datasets, the proposed association scheme is successful in identifying meaningful spatial patterns, which offer statistically-principled insight into the mechanisms of important socioeconomic problems.

The remainder of the paper is organized as follows. In Section 2, we provide the theoretical backdrop of the paper, thereby introducing the proposed information-theoretic association scheme. In Section 3, we present analytical results for linear SAR processes over circulant networks, which we extend to k-regular networks in Section 4. In Section 5, we illustrate our approach on synthetic and real datasets. Section 6 concludes the paper, summarizing our main findings and identifying avenues of future research.

2.   Information-theoretic association scheme

We seek to elucidate the structure underlying an ensemble of N scalar measurements $\{\tilde{y}_i\}_{i=1}^N$ of a given spatial process. Each measurement is acquired at a different physical location and is affected by inherent uncertainty associated with sampling. Overall, these values could be viewed as single realizations of N random variables $y_1, \ldots, y_N$, which are, in general, not independent.

Toward the formulation of a parsimonious information-theoretic approach that could draw inferences from the available measurements, we hypothesize strong stationarity with respect to the variable index [29]. This assumption is equivalent to the classical assumption of strong (or strict) stationarity of a traditional spatial process, which requires that the process is invariant under translation of its coordinates, or, more precisely, that the joint distribution of any ensemble of its variates depends only on the relative positions of their locations [29]. The relationship

between this instance of stationarity and other, weaker, forms of stationarity are discussed in [29], including second-order (or weak or wide-sense) stationarity and intrinsic stationarity.

Hence, for any set of indices $i_1, \ldots, i_q$, with $q \leq N$, and any shift $h < N$, we satisfy the following equality for the joint probability distribution:

$$\text{Prob}\left(y_{i_1} \leq \tilde{y}_{i_1}, \ldots, y_{i_q} \leq \tilde{y}_{i_q}\right) = \text{Prob}\left(y_{i_1+h} \leq \tilde{y}_{i_1}, \ldots, y_{i_q+h} \leq \tilde{y}_{i_q}\right) \tag{2.1}$$

where the index summation is intended to be congruent modulo N. Within the analogy with classical stationary processes, adding h represents a mere translation, which, due to the finiteness of the node set, is carried out modulo N (for example, a shift of one will cause N to be shifted to 1, as if the N nodes were ordered along a ring). By advocating (2.1), we can define a common scalar random variable Y, whose realizations are the available measurements and whose distribution is the same of any marginal distribution

$$\text{Prob}\left(Y \leq \tilde{Y}\right) = \text{Prob}\left(y_i \leq \tilde{y}_i\right) \tag{2.2}$$

for any $i = 1, \ldots, N$.

A possible way to generate a dataset with the assumed spatial structure is through a nonlinear SAR process of the following form:

$$y_i = f_i\left(y_{i+1}, \ldots, y_N, y_1, \ldots, y_{i-1}\right) + e_i \tag{2.3}$$

for $i = 1, \ldots, N$, where the smooth nonlinear functions $f_1, \ldots, f_N$ satisfy

$$f_i\left(y_{i+1}, \ldots, y_N, y_1, \ldots, y_{i-1}\right) = f_j\left(y_{j+1}, \ldots, y_N, y_1, \ldots, y_{j-1}\right) \tag{2.4}$$

for any index pair $i, j = 1, \ldots, N$, and $e_1, \ldots, e_N$ are independent, identically distributed (i.i.d.) random variables with common random variable $E \sim \mathcal{N}(0, 1)$ (normal distribution with zero mean and unit variance)[1]. Equation (2.3) describes the coupling between the random variables, by prescribing how the whole system nonlinearly determines the value of any given random variable.

Condition (2.4) and the fact that added noises are i.i.d. ensure the strong stationarity in (2.1), by practically imposing the invariance of the spatial process with respect to any shift. For example, setting $i = 1$ and $j = 3$, we have $f_1(y_2, ..., y_N) = f_3(y_4, ..., y_N, y_1, y_2)$, so that node 1 is affected by nodes $2, 3, \ldots$ in the same way that node 3 is affected by nodes $4, 5, \ldots$.

Each of the nonlinear combinations on the right hand side of (2.3) can be viewed as a random variable ϕ_i, with $i = 1, \ldots, N$. By construction, these variables are identically distributed and their common random variable is denoted as Φ, such that

$$Y = \Phi + E \tag{2.5}$$

The pattern of the interaction in (2.3) could be represented by a directed, unweighted network [23], whose node set is $\{1, \ldots, N\}$ and whose edge set is encapsulated by an adjacency matrix $A \in \{0, 1\}^{N \times N}$, such that

$$\frac{\partial f_i}{\partial y_j} = 0, \quad \text{if } A_{ij} = 0 \tag{2.6}$$

for any $i, j = 1, \ldots, N$. Hence, a link from i to j (that is, $A_{ij} = 1$) corresponds to y_j entering f_i in (2.4), thereby leaving a footprint on y_i. In this vein, the adjacency matrix only encodes the presence/absence of an interaction, whose intensity is given by the nonlinear function. Through the lens of a graph, space should be intended as an abscissa on the graph. Hence, a shift in the index of the nodes corresponds to a translation of the spatial coordinates, so that the strong stationarity assumption introduced earlier implies the invariance of any joint distribution under any possible shift.

Although the generic random variable y_i in (2.3) could be influenced by any other random variable of the ensemble, in many applications it is tenable that such an influence is limited to a small number of neighbours $k \ll N$, such that only k elements will be different from zero in each

---

[1]The arguments of the function f_i are always $N - 1$; for $i = 1$, the list ends at y_N, for $i = 2$, it ends at y_1, and so on.

row of A. For example, the number of neighbours k could be related to the size of the physical neighbourhood where the interaction is expected to occur in geography or the number of peer institutions in spatial econometrics [3].

Given condition (2.4), the adjacency matrix has a circulant structure [28], which is completely encoded in a vector $a = [a_1, \cdots, a_N]^{\mathrm{T}} \in \{0, 1\}^N$. The first element of this vector is zero as self-loops are not included, and the number of nonzero entries is equal to the number of neighbours k. We construct the adjacency matrix of the network, A, by setting its first row to be equal to a^{T}, and then shifting cyclically one position of a^{T} to the right to obtain the successive rows. Hence,

$$A = \mathrm{circ}(a) = \begin{bmatrix} a_1 & a_2 & \cdots & \cdots & a_N \\ a_N & a_1 & \cdots & \cdots & a_{N-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_2 & a_3 & \cdots & \cdots & a_1 \end{bmatrix} \tag{2.7}$$

Our objective is to establish an information-theoretic association scheme that takes as inputs the measurements $\{\tilde{y}_i\}_{i=1}^N$ and a candidate interaction $f_i^\star$ for $i = 1, \ldots, N$ (satisfying (2.4) with $f_i^\star$ replacing f_i), and outputs an indication of the extent to which the candidate interaction is similar to the true interaction. The candidate interaction is, in turn, associated with another circulant adjacency matrix $A^\star \in \{0, 1\}^{N \times N}$, whose nonzero elements determine the corresponding interaction pattern. For $i = 1, \ldots, N$, we write $\phi_i^\star = f_i^\star(y_{i+1}, \ldots, y_N, y_1, \ldots, y_{i-1})$ for the random variables constructed from the candidate interactions. Due to the strong stationarity of (2.1), all these variables are identically distributed, and the common random variable is written as $\Phi^\star$.

From the measurements $\{\tilde{y}_i\}_{i=1}^N$, we estimate the joint probability density function of Y and $\Phi^\star$. Upon knowledge of such a probability density function, we can test the independence of these random variables through their mutual information

$$I(Y; \Phi^\star) = h(Y) + h(\Phi^\star) - h(Y, \Phi^\star) = h(Y) - h(Y \mid \Phi^\star) \tag{2.8}$$

where $h(\cdot)$ is the (differential) entropy [25]. Mutual information is a nonnegative quantity that is equal to zero if and only if the two variables are independent [25]. Hence, if the candidate interaction pattern $A^\star$ has no overlap with the true interaction pattern A, mutual information will be zero. Likewise, a nonzero value of mutual information will signal that $A^\star$ is capturing some of the true interaction among the random variables.

Importantly, we can show that the largest value of the mutual information (2.8) corresponds to the candidate interaction matching the true interaction. To prove this claim, we follow these steps. First, we evaluate $I(Y; \Phi)$ by using (2.5) as

$$I(Y; \Phi) = h(Y) - h(Y \mid \Phi) = h(Y) - h(E) \tag{2.9}$$

where $h(E) = \frac{1}{2} \log(2\pi e)$. Next, we study $I(Y; \Phi^\star)$. By adding and subtracting $\phi_i^\star$ on the right hand side of (2.3) for $i = 1, \ldots, N$ and introducing the mismatch $\xi_i = \phi_i - \phi_i^\star$, whose common random variable is Ξ, we determine

$$I(Y; \Phi^\star) = h(Y) - h(Y \mid \Phi^\star) = h(Y) - h(E + \Xi \mid \Phi^\star) \tag{2.10}$$

Given that the entropy of the sum of two random variables is bounded from below by the entropy of any of the summands [30], we have

$$I(Y; \Phi^\star) \leq h(Y) - h(E) = I(Y; \Phi) \tag{2.11}$$

In the absence of any prior knowledge or insight into plausible interactions, $f_i^\star$ can be simply taken as the average of all the measurements of the nodes that could be connected to node i; in this case, we would replace $\phi_i^\star$ with $m_i^\star$ and $\Phi^\star$ with $M^\star$ to specify that we are taking a local

"mean" of candidate neighbours, that is,

$$f_i^\star (y_{i+1}, \ldots, y_N, y_1, \ldots, y_{i-1}) = \frac{1}{k} \sum_{j=1}^{N} A_{ij}^\star y_j \equiv m_i^\star \tag{2.12}$$

where k is the number of neighbours involved in the averaging process.

In addition, we often deal with small datasets (N from 25 to 100 measurements), for which it could be more convenient to work with discrete representations of the probability density functions through binning than implementing nonconventional estimators [26,27]. For these small datasets, the number of bins n used to estimate the mutual information statistic should not exceed three. In fact, quantifying mutual information requires the estimation of a bidimensional distribution with n^2 possible outcomes: for these outcomes to be statistically distinguishable and to obtain a reliable estimation of the distribution, the sample size should be between five or 10 times the number of outcomes [31]. In our analysis, we use equal-bin-count histograms, instead of equal-bin-width histograms when binning continuous measurements [32]. This selection is largely based on the mitigate estimation errors due to low counts in some of the bins or disproportionate counts between multiple bins. As an alternative to this traditional plug-in estimation, one may pursue alternative, heuristic approaches that are based on some prior knowledge of the distribution [33].

Under the premise of (2.12) and working with discrete representations, we would focus our analysis on the computation of the mutual information between Y and M, that is,

$$I(Y; M^\star) = H(Y) + H(M^\star) - H(Y, M^\star) \tag{2.13}$$

where $H(\cdot)$ is the discrete entropy [25].

Hence, when dealing with real datasets, we proceed as follows. First, we assume to have knowledge regarding a candidate interaction pattern $A^\star$ with $k^\star$ neighbours, which may come, for instance, from similarity between additional traits in the dataset. For example, should we be interested in the vegetation distribution or precipitation predictions, we could propose neighbours to be locations which are geographically close (that is, the additional trait would be the geographic location). Second, for the choice of $k^\star$ and a discrete representation of the dataset with cardinality n, we estimate the mutual information statistic in (2.13). The main steps of the computations are shown in Fig. 1.

While the chosen adjacency matrix should in principle be circulant to ensure strong stationarity, we propose that the approach could be applied to k-regular directed networks that generally have a wider range of application in practical settings [24]. This proposition rests upon theoretical results based on perturbation theory and numerical findings, which are presented in Section 4.

Once the value of $I(Y; M^\star)$ is obtained, we statistically test the following null and alternative hypotheses to determine whether $A^\star$ is a valid association scheme for the dataset:

$$I(Y; M^\star) = 0 \text{ against } I(Y; M^\star) > 0 \tag{2.14}$$

This is tested through a random permutation test in which we compute B boostrap realizations by shuffling each time the measurements and computing the corresponding mutual information. From these bootstrap realizations, we create a surrogate distribution and we reject the null hypothesis if $I(Y; M^\star)$ is in the right tail of the distribution with a p-value lower than a chosen significance value (typically 5%).

3. Analytical results for linear averaging over circulant networks

We illustrate the approach for the study of a linear SAR process, for which (2.3) becomes

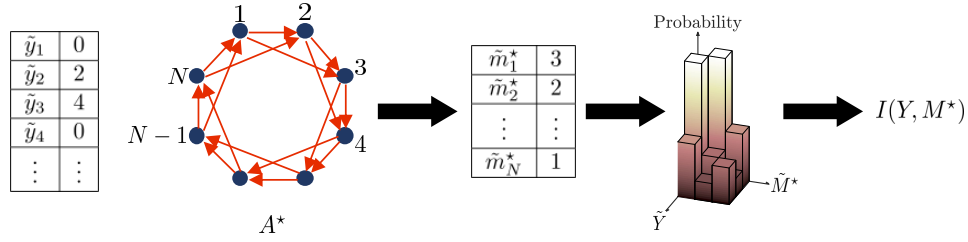$$y_i = \rho \sum_{j=1}^{N} W_{ij} y_j + e_i \tag{3.1}$$

Figure 1: Schematic of the computation of the mutual information statistic in (2.13), which takes as input N available measurements $\{\tilde{y}_i\}_{i=1}^N$ and a candidate interaction pattern $A^\star$. The computation is based on the following steps: i) for each node, we calculate the mean value of the measurements in neighboring nodes, according to $A^\star$, to obtain $\{\tilde{m}_i^\star\}_{i=1}^N$; ii) both $\{\tilde{y}_i\}_{i=1}^N$ and $\{\tilde{m}_i^\star\}_{i=1}^N$ are grouped into a small number of bins n to estimate the joint distribution $\text{Prob}\left(Y \leq \tilde{Y}, M^\star \leq \tilde{M}^\star\right)$; and iii) mutual information in (2.13) is evaluated. In this example, $A^\star$ is generated from $a = [0, 1, 1, 0, \cdots, 0]^T$; note that a link from node i to node j (for example, links from node 1 to nodes 2 and 3) indicates that node j is supposed to leave a footprint on node i (for example, nodes 2 and 3 on node 1), which we capture by computing the mean (for example, nodes 2 and 3 are used to compute the mean for node 1).

where ρ is a common weighting parameter measuring the intensity of the dependency among the measurements and the weight matrix W, typically used in the technical literature [34], is simply obtained by scaling the adjacency matrix of the underlying circulant network A by k (that is, $W = A/k$). Due to the linearity of the model, the fact that the added noise is normal implies that all the random variables $y_1, \ldots, y_N$ are also normal [1,3]. Their expected value is always zero and their covariance determines the mutual information $I(Y; M)$, whose computation is the objective of this Section.

Being a circulant matrix, W is diagonalizable with eigenvalues given by [28]

$$\lambda_i = \frac{1}{k} \sum_{r=1}^N a_r x^{(r-1)(i-1)}, \quad i = 1, \ldots, N \tag{3.2}$$

where $x = \exp(2\pi\mathrm{I}/N)$ and $\mathrm{I} = \sqrt{-1}$ is the imaginary unit. The complex eigenvector associated with the i-th eigenvalue is

$$V_i = \frac{1}{\sqrt{N}} \left[1, x^{(i-1)}, \cdots, x^{(N-1)(i-1)}\right]^T \tag{3.3}$$

which is independent of the specific entries of the vector a. By juxtaposing the eigenvectors column-wise, we obtain the similarity matrix $V \in \mathbb{C}^{N \times N}$. The inverse of this matrix $Q = V^{-1}$ can be easily computed by juxtaposing row-wise the vectors

$$Q_i = \frac{1}{\sqrt{N}} \left[1, s^{(i-1)}, \cdots, s^{(N-1)(i-1)}\right]^T \tag{3.4}$$

where $s = x^{-1}$.

By using the similarity matrix V, we invert the original linear SAR process in (3.1) to explicitly compute $y = [y_1, \cdots, y_N]^T$ in terms of $e = [e_1, \cdots, e_N]^T$, that is, we can write a closed-form expression for the matrix F mapping e to y,

$$F = (I - \rho W)^{-1} \tag{3.5}$$

Specifically, we establish

$$F = V(I - \rho \Lambda)^{-1} V^{-1} \tag{3.6}$$

where Λ is the diagonal matrix that collates all the eigenvalues of W. By replacing the expressions for the eigenvectors and eigenvalues, the ij-th entry of $F = V(I - \rho\Lambda)^{-1}V^{-1}$ can be compactly written as

$$F_{ij} = \gamma(i - j) \tag{3.7}$$

where we introduced the N-periodic function $\gamma : \mathbb{Z} \to \mathbb{R}$,

$$\gamma(\Delta) = \frac{1}{N} \sum_{r=1}^{N} \frac{\exp\left(\frac{2\pi I(r-1)\Delta}{N}\right)}{1 - \frac{\rho}{k} \sum_{s=1}^{N} a_s \exp\left(\frac{2\pi I(r-1)(s-1)}{N}\right)} \tag{3.8}$$

Given that noise has unit variance, the covariance matrix of the multivariate normal process y is

$$\Sigma_y = FF^{\mathrm{T}} \tag{3.9}$$

which, component-wise, reads

$$(\Sigma_y)_{ij} = \sum_{r=1}^{N} \gamma(i - r)\gamma(j - r) \tag{3.10}$$

for $i, j = 1, \ldots, N$. From the covariance matrix, one can compute the covariance of all the marginal distributions for $y_1, \ldots, y_N$ and the local means $m_1, \ldots, m_N$, where $m_i = \sum_{j=1}^{N} W_{ij}y_j$ for $i = 1, \ldots, N$. The covariance of the common random variable Y is simply given by the diagonal elements of Σ_y, which are equal to

$$\Sigma_Y = \sum_{r=0}^{N-1} \gamma^2(r) \tag{3.11}$$

The covariance of the common random variable M is computed by looking at the off-diagonal elements that are involved in the mean, thereby leading to

$$\Sigma_M = \frac{1}{k}\Sigma_Y + \frac{2}{k^2} \sum_{r=1}^{N} \sum_{s=r+1}^{N} a_r a_s \sum_{t=1}^{N} \gamma(r - t)\gamma(s - t) \tag{3.12}$$

Note that the first term on the left hand side is the average of all the variances of the corresponding nodes and the second term accounts for covariance between the nodes.

Hence, the (differential) entropy of Y is given by [25]

$$h(Y) = \frac{1}{2} \log(2\pi e \Sigma_Y) \tag{3.13}$$

with e being Napier's constant. Rather than computing the joint entropy of Y and M and the entropy of M to calculate $I(Y; M)$, we compute the differential entropy of Y given M. The latter is simply the differential entropy of the noise, that is,

$$h(Y \mid M) = \frac{1}{2} \log(2\pi e) \tag{3.14}$$

Mutual information of Y and M is thus equal to

$$I(Y; M) = h(Y) - h(Y \mid M) = \frac{1}{2} \log(\Sigma_Y) \tag{3.15}$$

where the variance of Y is given in (3.11), with the function γ in (3.8).

Interestingly, for small values of the parameter ρ, it is possible to calculate an approximate expression of mutual information by taking a MacLaurin expansion in ρ up to the second order. More specifically, for $\rho \ll 1$, expanding (3.8), we find

$$\gamma(\Delta) \simeq \delta_{\Delta,0} + \frac{\rho}{k} a_{N+1-\Delta} + \frac{\rho^2}{k^2} \sum_{r=1}^{N} a_r a_{2-r-\Delta} \tag{3.16}$$

for $\Delta = 0, \ldots, N - 1$, where $\delta_{(\cdot,\cdot)}$ is the Kronecker delta and the last summation should be intended to be congruent modulo N. Here, we have used the identity $\frac{1}{N} \sum_{r=1}^{N} \exp\left(\frac{2\pi I(r-1)\Delta}{N}\right) =$

$\delta_{\Delta,0}$. Using this McLaurin expansion, we can approximate the variance in (3.11) as follows:

$$\Sigma_Y \simeq 1 + \frac{\rho^2}{k} + 2\frac{\rho^2}{k^2}\sum_{r=2}^{N} a_r a_{2-r+N} \tag{3.17}$$

where we have accounted for the lack of self-loops, that is, $a_1 = 0$.

The last term on the right hand side of (3.17), $\sum_{r=2}^{N} a_r a_{2-r+N}$, corresponds to the number of bidirectional links connected to any node in the network, u, whereby any of the listed products between the elements of the vector a identifies the pair of potential directed links between two nodes. For example, $a_2 a_N$ identifies the occurrence of a link from node 1 to node 2 and of a link between node 2 and node 1. The variance of M could be equivalently computed.
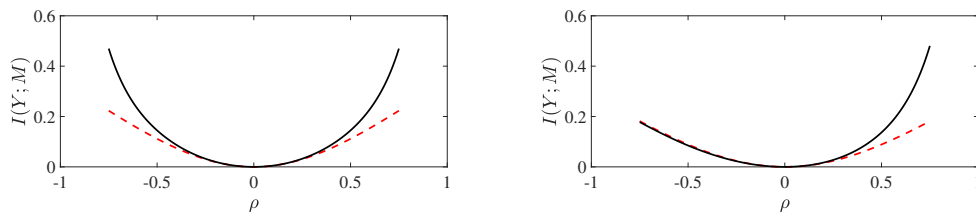

Figure 2: Mutual information (3.15), computed using the exact variance Σ_Y in (3.11) (solid black) or its second-order approximation in (3.17) (dashed red), over 10-node circulant networks with $a = [0, 1, 0, 0, 0, 0, 0, 0, 0, 0]^{\mathrm{T}}$ (left panel) and $a = [0, 1, 1, 0, 0, 0, 0, 0, 1, 0]^{\mathrm{T}}$ (right panel), as a function of the intensity of the spatial dependence ρ.

To illustrate the value of the approximate solution, in Fig. 2, we compare the values of mutual information obtained by using the exact variance Σ_Y in (3.11) or its second-order approximation in (3.17), over a wide range of ρ. We consider two representative circular networks: i) a cycle in which each node is only connected to the next one, and ii) a denser network obtained as the union of the cycle and an undirected network where each node is bidirectionally connected to the next to next one. For the cycle $k = 1$ and $u = 0$, while for the denser network $k = 3$ and $u = 2$. In both cases, the approximate solution is in close agreement with the exact one for a wide range of ρ. Interestingly, we register an effect of the sign of ρ for the denser network that should be attributed to higher order topological features, beyond the number of bidirectional links that is captured by the approximate solution.

4. Perturbation analysis for linear averaging over k-regular networks

The analysis presented thus far is strictly applicable only to circulant networks. As we examine arbitrary k-regular networks, exact guarantees of stationarity and closed-form results for variances and mutual information become not feasible. In this case, each node may have a different in-degree, thereby straining the assumption of stationarity of any related spatial process. A potential line of approach to extend the analysis of the SAR process in (3.1) to k-regular networks is through a perturbation analysis, performed upfront with respect to the weighting parameter ρ.

For any choice of the matrix W in (3.1), through a second order perturbation in ρ, we can write F as follows [3]:

$$F \simeq I + \rho W + \rho^2 W^2 \tag{4.1}$$

The corresponding covariance matrix in (3.9) is

$$\Sigma_y \simeq I + \rho \left(W + W^{\mathrm{T}} \right) + \rho^2 \left[W W^{\mathrm{T}} + W^2 + \left(W^{\mathrm{T}} \right)^2 \right] =$$
$$I + \frac{\rho}{k} \left(A + A^{\mathrm{T}} \right) + \frac{\rho^2}{k^2} \left[A A^{\mathrm{T}} + A^2 + \left(A^{\mathrm{T}} \right)^2 \right] \quad (4.2)$$

Considering a k-regular network, each row of the adjacency matrix A contains exactly k ones (identifying links from the node associated with a particular row to any other node in the network) and all the diagonal elements of the adjacency matrix are zero. Under these premises, we can offer a simple interpretation of each of the summands for the generic ij-th entry of the covariance matrix, up to the second order in ρ, based on classical results in network theory [35].

Specifically, the term in ρ/k is equal to one if there is a link from i to j or from j to i, is two if i and j are connected by an undirected link, and is zero otherwise. In the term in ρ^2/k^2, we identify three contributions: i) $\left(A A^{\mathrm{T}} \right)_{ij}$ is equal to the total number of nodes in the network, excluding i and j, to which either i or j are connected; ii) $\left(A^2 \right)_{ij}$ is the total number of paths of length two from i to j; and iii) $\left(\left(A^{\mathrm{T}} \right)^2 \right)_{ij}$ is the total number of path of length two from j to i.

To ascertain the stationarity of the process, we should compare the diagonal terms, which are equal to

$$\Sigma_{y_i} \simeq 1 + \frac{\rho^2}{k^2} (k + 2u_i) \quad (4.3)$$

where u_i is the number of undirected links connected to node i, that is, $u_i = \sum_{j=1}^{N} A_{ij} A_{ji}$. We note that the leading order expansion is in ρ^2, which is the lowest order one should retain to obtain a nontrivial expansion. In general, u_i will depend on i, thereby challenging the stationarity of the process. While acknowledging this as a limitation in the use of a stationary model, we present numerical results that support the premise of a stationary process, within a first degree of a approximation.

Not only do we require the diagonal terms to be close to each other, but also we should enforce that the local means are drawn from the same distribution. The variance of the mean at the i-th node is

$$\Sigma_{m_i} = \frac{1}{k^2} \sum_{j=1}^{N} A_{ij} \left(\Sigma_y \right)_{jj} + \frac{1}{k^2} \sum_{j,r=1}^{N} A_{ij} A_{ir} \left(\Sigma_y \right)_{jr} \quad (4.4)$$

Although a second order expansion in ρ is possible by simply replacing from (4.2), we present the leading order expansion in ρ

$$\Sigma_{m_i} \simeq \frac{1}{k} + \frac{\rho}{k^3} c_i \quad (4.5)$$

where $c_i = 2 \sum_{j,r=1}^{N} A_{ij} A_{ir} A_{jr}$ is the number of links between any two neighbours of node i. Again, in general this quantity might vary throughout the network, questioning the stationarity of the process, which is, however, supported by numerical results summarized below.

To advocate the assumption of stationarity, we assume that u_i are approximately all equal to their mean value, $\overline{u}$. Hence, all the available realizations of $\tilde{y}_1, \ldots, \tilde{y}_N$ and their corresponding mean values $\tilde{m}_1, \ldots, \tilde{m}_N$ can be viewed as realizations from two random variables Y and M, for which we can calculate a leading order expansion of mutual information as a function of ρ. By directly using (3.15) with (4.3), we establish

$$I(Y; M) \simeq \frac{1}{2} \log \left(1 + \frac{\rho^2}{k^2} (k + 2\overline{u}) \right) \quad (4.6)$$

We illustrate the feasibility of applying the stationarity assumption for k-regular network through numerical simulations. For any value of k, we generate 1,000 networks by randomly (uniformly) placing $N = 50$ points in $[0, 1]^2$ and connecting each point with the k closest

neighbours, according to a Euclidean distance. Other choices could also be contemplated to generate k-regular networks [36]; the selection considered herein is motivated by standard practice in spatial analysis where neighbourhoods are often defined on the basis of geographic proximity. We examine different values of ρ in the linear SAR process in (3.1). For a given k and ρ, we calculate the covariance matrix in (3.9) for all the generated random networks.
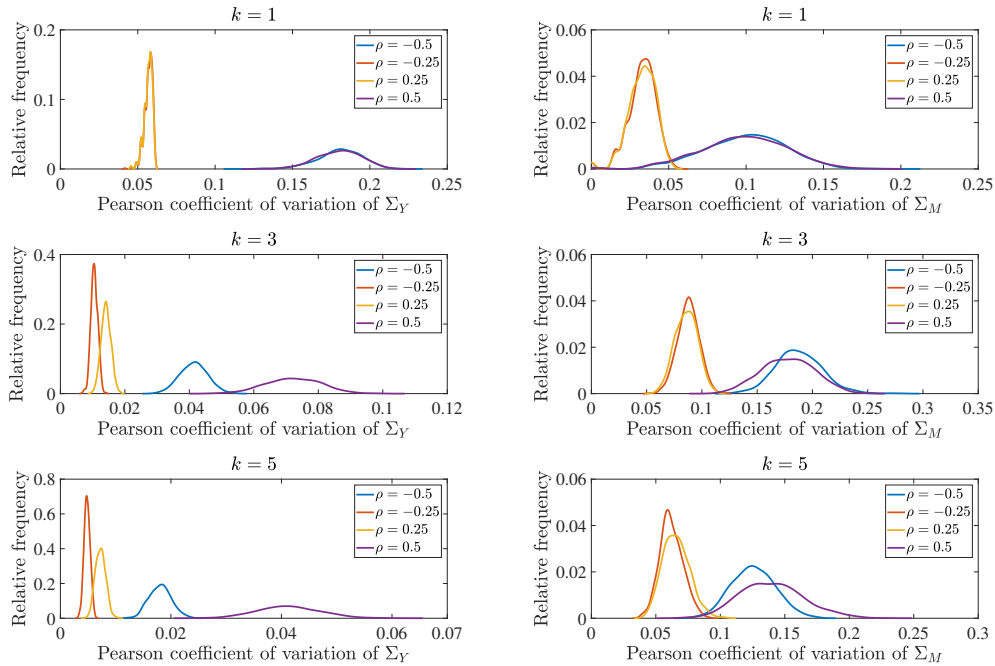
Figure 3: Estimated distributions of the Pearson coefficient of variation of the variance of Y (left column) and M (right column), obtained from 1,000 realizations of k-regular networks. Computations are performed for $k = 1, 3,$ and 5 and different intensities of spatial association $\rho = -0.5, -0.25, 0, 0.25,$ and 0.5.

By taking the mean value of the elements on the diagonal and their standard deviation, we calculate a Pearson coefficient of variation for the variance of Y for each network realization. Aggregating the 1,000 realizations, we compute a distribution for such a Pearson coefficient, which is shown in Fig. 3. From the same covariance matrix, we use (4.4) to compute N values for the covariance of the mean, from which we calculate the distribution of the Pearson coefficient of variation of the variance of M, which is also shown in Fig. 3. In agreement with (4.3) and (4.5), the Pearson coefficients of variance of both Y and M are smaller for small values of ρ, pointing at more stable variances that would better support the validity of the stationarity assumption. The parameter ρ has a stronger effect on the Pearson coefficient of Y than M, which is also in agreement with (4.3) and (4.5) that beget a leading order dependence of order two and one, respectively. With respect to the dependency on k, we register a predictable improvement in the stability of the variance for a larger number of neighbours, which would, in fact, contribute to a stronger averaging effect of the SAR process.

Albeit (4.6) is based on the estimation of the complete probability density function, an equivalent trend is obtained when working with discrete distributions that would be more appropriate to use when working with a small dataset as in Fig. 3. To demonstrate this claim, in Fig. 4, we numerically compute $I(Y; M)$ for the same dataset considered in Fig. 3, using $n = 2$ and
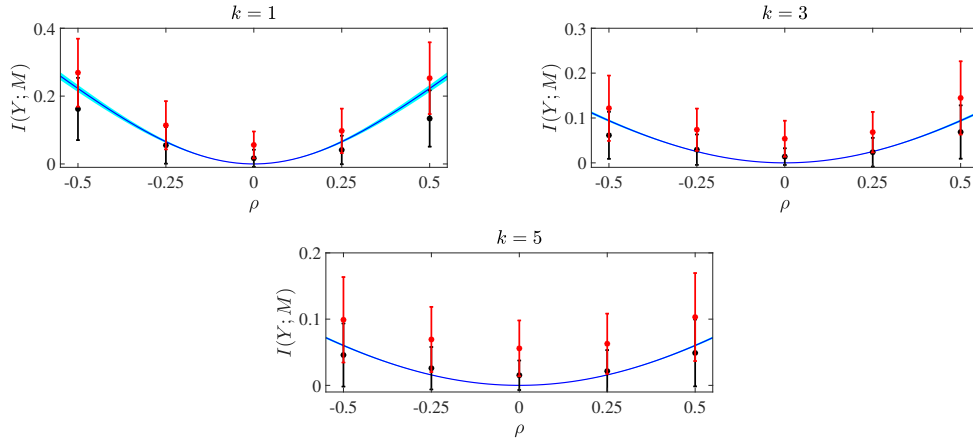
Figure 4: Mutual information, numerically estimated from (2.13) and theoretically evaluated using perturbation theory (4.6). Numerical computations are based on 1,000 realizations of k-regular networks and i.i.d. additive errors $\sim \mathcal{N}(0,1)$. Computations are performed for $k = 1, 3$, and 5, different intensities of spatial association $\rho = -0.5, -0.25, 0, 0.25$, and 0.5, and number of bins $n = 2$ (red) and $n = 3$ (black). Numerical values are presented as a mean plus/minus a standard deviation, while theoretical predictions as a mean along with a shaded region associated with the standard deviation in the average number of undirected links in the network across the 1,000 realizations.

3 bins. More specifically, for each condition and network realization, we independently draw the noise from a normal variable with zero mean and unit variance and compute realizations $\{\tilde{y}_i\}_{i=1}^N$ in (3.1). We separately bin these measurements and the corresponding mean values $\{\tilde{m}_i\}_{i=1}^N$ using one or two quantiles that divide the distributions into two or three equal parts, respectively. Fig. 4 shows the estimated mean value of the mutual information $I(Y; M)$ with an error bar of plus/minus one standard deviation, superimposed to (4.6), where we have adjusted the value of $\overline{u}$ to encompass variations across the 1,000 realizations.

5. Illustration of the approach on synthetic and real datasets

We demonstrate the approach on a range of cases, spanning synthetic data of linear and nonlinear SAR processes and real data on human migration [21] and motor vehicle deaths in the USA [22].

(a) Synthetic datasets from linear and nonlinear spatial autoregressive processes

As a first validation of the proposed approach, we consider the linear SAR process in (3.1) for k-regular networks, generated by randomly (uniformly) placing N points in $[0,1]^2$ and connecting each point with the k closest neighbours. The noise variables have i.i.d. normal distributions with zero mean and unit variance.

Within a detailed parametric study, we consider the following values of N, ρ, and k: $\{25, 50, 100\}$, $\{-0.9, -0.6, -0.3, 0, 0.3, 0.6, 0.9\}$, and $\{1, 3, 5\}$, respectively. For each parameter configuration, we run 1,000 Monte Carlo simulations, in which we independently draw the network topology and the additive noise. For each simulation, we implement our statistical test in (2.14) using $B = 10,000$ bootstrap realizations and we record the number of times we reject the null hypothesis with a confidence level of 5%. We refer to the ratio between the number of

instances in which we reject the hypothesis of independence and the number of Monte Carlo simulations as the power of the test.

ρ	-0.9	-0.6	-0.3	0	0.3	0.6	0.9
$N=25$							
$k=1, n=2$	0.998	0.746	0.220	0.052	0.193	0.659	0.971
$k=3, n=2$	0.466	0.244	0.075	0.036	0.106	0.407	0.844
$k=5, n=2$	0.270	0.163	0.078	0.053	0.063	0.223	0.699
$k=1, n=3$	1.000	0.770	0.208	0.047	0.194	0.682	0.997
$k=3, n=3$	0.505	0.250	0.096	0.050	0.089	0.297	0.864
$k=5, n=3$	0.239	0.156	0.078	0.054	0.060	0.131	0.522
$N=50$							
$k=1, n=2$	1.000	0.931	0.315	0.023	0.253	0.895	1.000
$k=3, n=2$	0.811	0.390	0.120	0.020	0.127	0.544	0.991
$k=5, n=2$	0.433	0.228	0.087	0.022	0.052	0.356	0.923
$k=1, n=3$	1.000	0.975	0.424	0.045	0.326	0.964	1.000
$k=3, n=3$	0.911	0.575	0.170	0.035	0.181	0.740	0.999
$k=5, n=3$	0.595	0.327	0.156	0.055	0.096	0.477	0.986
$N=100$							
$k=1, n=2$	1.000	0.999	0.713	0.033	0.672	1.000	1.000
$k=3, n=2$	0.989	0.782	0.282	0.033	0.299	0.921	1.000
$k=5, n=2$	0.882	0.525	0.185	0.031	0.165	0.781	1.000
$k=1, n=3$	1.000	1.000	0.762	0.054	0.732	1.000	1.000
$k=3, n=3$	0.999	0.929	0.362	0.056	0.367	0.973	1.000
$k=5, n=3$	0.938	0.634	0.217	0.058	0.223	0.859	1.000

Table 1: Rejection rates of the null hypothesis of independence of the mutual information statistical test in (2.14) over 1,000 Monte Carlo simulations of a linear SAR process over a k-regular network with i.i.d. additive errors $\sim \mathcal{N}(0,1)$, using $B=10,000$ bootstrap realizations at a 5% confidence level. Computations are performed for $k=1,3,$ and 5, different intensities of spatial association $\rho=-0.9,-0.6,-0.3,0,0.3,0.6,$ and 0.9, and number of bins $n=2$ and 3.

Table 1 shows the rejection rate of the mutual information statistic of the SAR process (3.1) for the above mentioned configurations. As expected, the power of the test increases with the intensity of the spatial dependence (absolute value of ρ), which facilitates the detection of an association between the observations. Notably, the power of the test is not insensitive to the sign of ρ, especially for $|\rho|\geq 0.3$. This evidence echoes the findings in Fig. 2 about circular networks, hinting at a salient role of higher order topological features on the mutual information statistic.

Increasing the size of the dataset has also a beneficial effect on the power of the test, improving the robustness of the estimation of mutual information. Albeit secondary, the effect of the number of bins depend on the sample size. For small dataset, increasing the number of bins hinders the estimation of mutual information, which may confound the statistical test. For larger dataset, we register the opposite behaviour, whereby more bins allow for a finer resolution of the statistic.

Interestingly, the power of the test decreases when the connectivity of the adjacency matrix increases, that is, when k increases. This effect should be ascribed to the diffusive nature of the SAR process, which will reduce information-rich variations within the observations. We conclude by commenting on the conservativeness of the test, which is evidenced by the low rejection rates when $\rho=0$. In this case, the observations are generated under the null of independence and the test is successful in minimizing the number of false positives, which are close to the nominal level of 5%.

To explore the accuracy of the approach to study nonlinear spatial processes, we examine a second synthetic dataset generated using a nonlinear SAR process of the form

$$y_i = \frac{1}{2} \sum_{j=1}^{N} W_{ij} y_j + \eta \sum_{j=1}^{N} W_{ij} y_j^3 + e_i \tag{5.1}$$

for any $i = 1, \ldots, N$. Similar to the linear dataset considered above, we focus on k-regular networks and normal noise with zero mean and unit variance. We perform the analysis for $B = 10,000$ bootstrap realizations at a 5% confidence level, for the following values of N, η, and k: $\{25, 50, 100\}$, $\{0.025, 0.05, 0.075\}$, and $\{1, 3, 5\}$, respectively. The proposed nonlinear process can be viewed as an extension of the linear SAR process in (3.1) with a linear part that is characterized by moderate intensity, $\rho = 0.5$, and a nonlinear cubic part that is modulated by the parameter η.

η	0.025	0.050	0.075
$N = 25$			
$k = 1, n = 2$	0.468	0.504	0.528
$k = 3, n = 2$	0.362	0.429	0.434
$k = 5, n = 2$	0.231	0.302	0.306
$k = 1, n = 3$	0.631	0.622	0.574
$k = 3, n = 3$	0.351	0.360	0.386
$k = 5, n = 3$	0.153	0.210	0.212
$N = 50$			
$k = 1, n = 2$	0.836	0.875	0.876
$k = 3, n = 2$	0.533	0.607	0.678
$k = 5, n = 2$	0.307	0.444	0.493
$k = 1, n = 3$	0.951	0.969	0.966
$k = 3, n = 3$	0.671	0.786	0.798
$k = 5, n = 3$	0.449	0.590	0.655
$N = 100$			
$k = 1, n = 2$	0.999	0.999	0.998
$k = 3, n = 2$	0.909	0.956	0.971
$k = 5, n = 2$	0.737	0.861	0.907
$k = 1, n = 3$	1.000	1.000	1.000
$k = 3, n = 3$	0.958	0.992	0.996
$k = 5, n = 3$	0.821	0.923	0.967

Table 2: Rejection rates of the null hypothesis of independence of the mutual information statistical test in (2.14) over 1,000 Monte Carlo simulation of the nonlinear SAR process (5.1) over a k-regular network with i.i.d. additive errors $\sim \mathcal{N}(0, 1)$, using $B = 10,000$ bootstrap realizations at a 5% confidence level. Computations are performed for $k = 1, 3$ and 5, different intensities of spatial association $\eta = 0.025, 0.050$, and 0.075, and number of bins $n = 2$ and 3.

Table 2 reports the rejection rates of the nonlinear SAR process in (5.1) for the combination of parameters described above. Similar to the linear dataset examined above, the power increases with the size of the dataset and it reduces with the number of neighbours. The number of bins has a secondary role, which is, again, moderated by the size of the dataset. In favor of the use of the proposed approach in the presence of richer nonlinear interactions, we register a positive dependence on the strength of the nonlinearity, whereby higher rejection rates are attained for larger values of η.

Together, Tables 1 and 2 indicate that the proposed mutual information statistic is successful in detecting spatial dependencies in small datasets without the need of an underlying mathematical model. For datasets comprising at least 50 observations and moderate intensities of spatial association (linear or nonlinear), the approach can lead to rejection rates above 95%.

(b) Human migrations under climate change in Bangladesh

As a first real dataset on which to test the viability of the proposed information-theoretic approach, we examine internal migrations in Bangladesh under climate change. Bangladesh has the third largest population living in low elevation coastal zones, whose density grows at the fastest rate. Coastal flooding is expected to radically impact the economy of the country and the lives of millions of people. The recent study by Davis *et al.* [21] has put forward a data-driven model of human migration that adapts the diffusion model by Simini *et al.* [37] to predict internal migration in Bangladesh driven by sea level rise (SLR).

The approach by Davis *et al.* [21] takes as input SLR projections for different representative concentration pathway (RCP) scenarios from the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [38], current population in each of the 64 zilas (districts) from the Gridded Population of the World (GPWv4) dataset [39], and population projections by the United Nations [40] to estimate the number of people in each zila that will be displaced due to climate change. Then, it applies a diffusion model [37] to estimate internal migration in Bangladesh, in the form of local fluxes between zilas that are controlled by mutual distances between zilas, variations in populations across zilas, and the number of people which must leave the zila due to flooding.
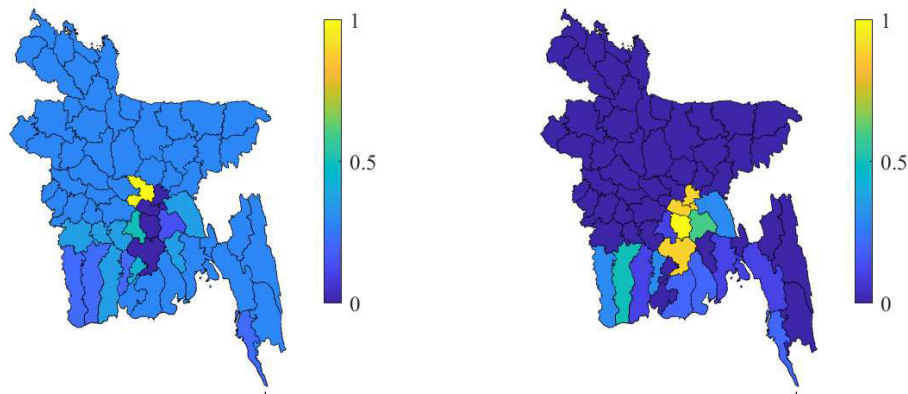

Figure 5: Analysis of 2050 internal migration in Bangladesh due to inundations for a representative concentration pathway of 8.5 and sea level rise of 0.3 m. Net migration (left column) - difference between arriving and departing migrants - and total number of people forced to leave their zila due to flooding (right column). Each variable is normalized between zero and one for all the zilas. The minimum and maximum values for net migration are $(-95,004; 207,374)$ and for the total number of migrants are $(0; 114,642)$ – note that the largest net migration is in the capital zila of Dacca. Colored maps were created with the MATLAB Mapping toolbox, using data from www.arcgis.com.

Here, we seek to unfold associations between zila-level predictions of human migration and either mutual distances between zilas or similarities in the severity of expected inundations between zilas. More specifically, we work with the estimated net migration (Net: difference between arriving and departing migrants) for year 2050 under an RCP of 8.5 and an SLR of 0.3 m, included as Supplementary Information in [21] and depicted in Fig. 5. We examine two alternative explanations for the observed migration pattern. First, we consider the association between net migration $\{\text{Net}_i\}_{i=1}^{64}$ and mutual distance between zilas $\{d_{ij}\}_{i,j=1}^{64}$, computed by taking the distance between their centroids (shared by the authors of [21] through private communication). Second, we study the association between the observed net migration and the difference in the total number of people that are expected displaced from their district $\{|T_i - T_j|\}_{i,j=1}^{64}$; in Fig. 5, we display the total number of people who are expected to leave their zila as reported in the Supplementary Information of [21].

To perform the mutual information statistical test in (2.14), we define the $k^\star$ neighbours of a zila from either $\{d_{ij}\}_{i,j=1}^{64}$ or $\{|T_i - T_j|\}_{i,j=1}^{64}$, according to the specific explanation which we plan to test. For each potential explanation, we test two different binning numbers $n = 2$ and 3 that divide the distributions into two or three equal parts, respectively, and three possible levels of network connectivity $k^* = 1, 3$, and 5. The permutation test is performed on $B = 10,000$ bootstrap realizations for both the potential explanations. However, since there are several districts which are expected to remain unaffected by inundations, there are several zeros in $\{|T_i - T_j|\}_{i,j=1}^{64}$ that could lead to arbitrariness in the assignment of neighbours. To address this issue, we amend the permutation test to include $1,000$ iterations per bootstrap realization in which we perturb the values of $\{T_i\}_{i=1}^{64}$ away from zero. From the $1,000$ iterations, we select the one that leads to the smallest value of mutual information, thereby reducing the likelihood of false positives.

With respect to the association between net migration and geographic distance, we reject the null of independence at a 5% confidence level only with $n = 3$ bins and $k^* > 1$ ($k^* = 3$, $p = 0.0057$; $k^* = 5$, $p = 0.0020$). With respect to the association between net migration and similarity in the severity of the inundations, we fail to reject the null of independence at a 5% confidence level for any combination of nearest neighbours k^* and number of bins n. Overall, these results point at a strong effect of geography in the prediction of migration patterns under future climate change, which is, in fact, a feature of the diffusion model adopted by Davis *et al.* [21]. Based on the original diffusion model in [37], displaced individual will stop in the closest location which could offer benefits that are higher than what could be available in his/her own zila, thereby leading to a strong association between geography and net migration. The lack of an association between net migration and the total number of people displaced is likely due to the fact that only 21 out of 64 zilas are expected to be impacted by inundation, thereby challenging the inference of a structure on such a small dataset.

(c) Motor vehicle deaths in the United States of America

To further demonstrate the viability of the proposed information-theoretic approach to unfold spatial associations, we consider a second real dataset, consisting of motor vehicle deaths in the USA [22]. More specifically, we seek to examine whether state-by-state motor vehicle deaths from 1980 to 2009 are associated with either the spatial proximity of the states or similarities in their political ideology. (Political ideology quantifies the average location of the elected officials in each state on a liberal-to-conservative continuum [41].) An equivalent research question has been addressed by Abaid *et al.* [22] through a linear correlation analysis, without the support of a statistically-principled methodology as the one proposed herein.

To facilitate comparisons of state-level data, we work with motor vehicle deaths per 100 million vehicle miles of travel MVD at a yearly resolution for the 50 states, as shown in Fig. 6 for three representative years in the dataset (made available by the authors of [22]). By considering this variable rather than the absolute count of fatalities, we control for over-emphasis of states which have higher populations or more roads. We contemplate two alternative explanations for the observed yearly patterns of motor vehicle deaths. Similar to the human migration dataset, we first
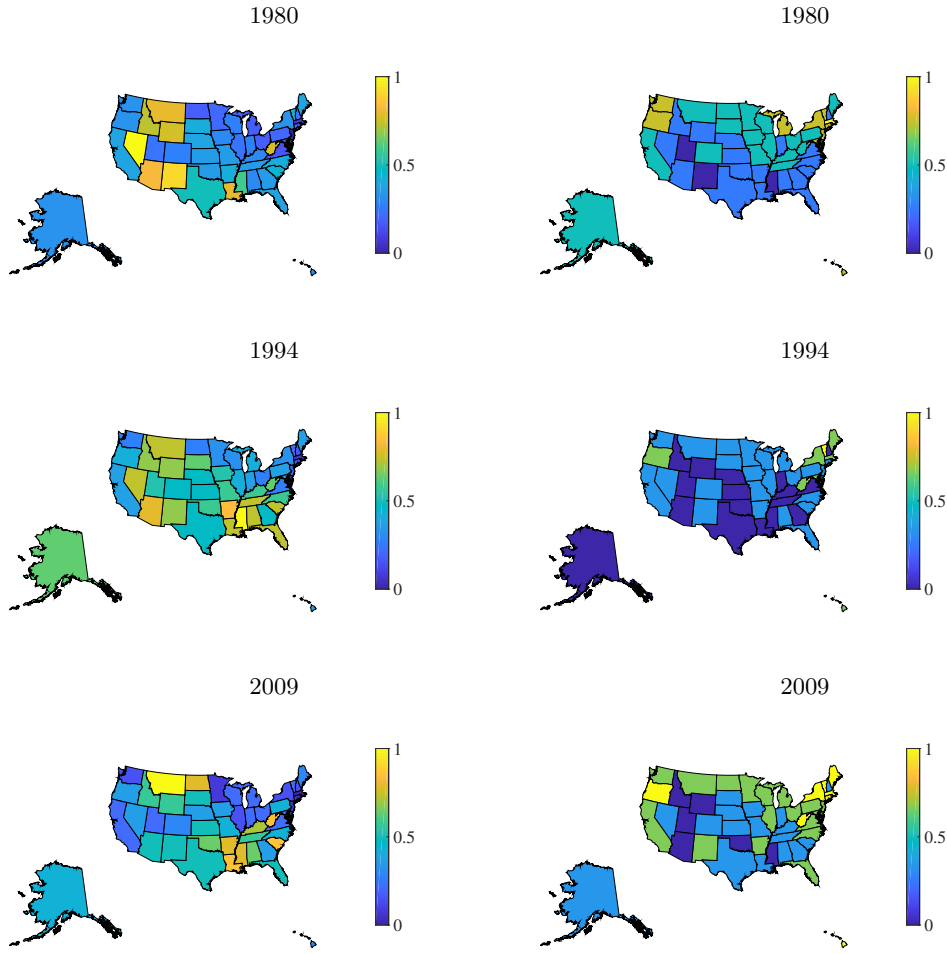
Figure 6: Snapshots of motor vehicle death per 100 million vehicle miles of travel MVD (left column) and ideology index I (right column) for three representative years: 1980, 1994, and 2009. Each variable is normalized between zero and one for all states independently for each year, where the ranges of MVD and I are $(2.23; 5.67)$ and $(0.58; 0.90)$ for 1980, respectively; $(0.89; 2.77)$ and $(0.61; 0.94)$ for 1994, respectively; and $(0.61; 2.00)$ and $(0.63; 0.98)$ for 2009, respectively. All colored maps were created with the MATLAB Mapping toolbox.

propose an association between $\{\mathrm{MVD}_i\}_{i=1}^{50}$ and the distance between the centroids of the states $\{d_{ij}\}_{i,j=1}^{50}$ (made available by the authors of [22]). Second, we test the possibility that $\{\mathrm{MVD}_i\}_{i=1}^{50}$ could be associated with similarities in the political ideology of the states $\{|I_i - I_j|\}_{i,j=1}^{50}$, where ideology of each state varies from zero (least liberal) to one (most liberal), as shown in Fig. 6. Data on political ideology are included as Supplementary Information in [22]; notably, ideology varies over time, thereby requiring the identification of neighbours for each year.

The mutual information statistical test in (2.14) is computed for the two k^*-regular network configurations associated with geographic and ideological distances at three different connectivity levels of $k^* = 1, 3,$ and 5. Like the human migration dataset, we explore two different binning numbers $n = 2$ and 3 that divide the distributions into two or three equal parts,

Year	$n=2$			$n=3$		
	$k^*=1$	$k^*=3$	$k^*=5$	$k^*=1$	$k^*=3$	$k^*=5$
1980	0.2145	0.2677	0.5917	0.0669	0.2207	0.1027
1981	0.0587	**0.0034**	**0.0024**	**0.0055**	**0.0258**	**0.0000**
1982	**0.0086**	0.2652	**0.0180**	**0.0080**	**0.0008**	**0.0332**
1983	0.0616	0.1527	**0.0235**	**0.0053**	**0.0081**	**0.0079**
1984	0.0930	**0.0259**	**0.0263**	**0.0008**	**0.0067**	**0.0001**
1985	**0.0105**	0.1188	**0.0185**	0.0615	**0.0012**	**0.0033**
1986	**0.0171**	**0.0383**	0.0918	**0.0033**	**0.0350**	**0.0009**
1987	0.0767	**0.0033**	0.2385	**0.0021**	0.2785	**0.0034**
1988	0.4773	0.2958	0.5130	**0.0248**	0.3159	0.0672
1989	**0.0144**	0.2663	0.1995	0.0752	**0.0243**	0.0502
1990	0.4430	0.4347	0.0979	0.0841	0.1152	0.0597
1991	0.4338	0.0862	0.2506	0.1545	**0.0389**	**0.0083**
1992	0.7232	0.3275	0.2631	**0.0445**	**0.0404**	**0.0356**
1993	0.0568	0.0886	0.0877	**0.0334**	**0.0314**	**0.0038**
1994	**0.0131**	**0.0272**	**0.0365**	0.1870	**0.0103**	**0.0023**
1995	0.5004	0.0771	0.2860	0.1032	0.1469	**0.0256**
1996	0.0812	**0.0430**	**0.0369**	0.0584	**0.0022**	**0.0029**
1997	0.0748	0.1693	**0.0073**	**0.0075**	**0.0087**	**0.0422**
1998	**0.0139**	0.1124	**0.0319**	0.1519	**0.0092**	**0.0488**
1999	0.2374	**0.0036**	**0.0347**	**0.0026**	**0.0015**	**0.0025**
2000	0.0721	0.1384	**0.0328**	**0.0108**	**0.0100**	**0.0032**
2001	0.2406	0.1168	0.1287	**0.0347**	0.0503	0.0501
2002	0.5178	0.3339	**0.0415**	**0.0204**	**0.0115**	**0.0015**
2003	0.2151	0.1093	0.2852	0.1270	**0.0416**	0.1214
2004	0.2640	0.2668	**0.0368**	0.1905	**0.0086**	**0.0219**
2005	0.4650	0.1219	0.3068	**0.0354**	**0.0006**	0.0637
2006	0.4844	0.2201	0.2988	**0.0097**	**0.0120**	0.0599
2007	0.2344	0.2532	**0.0050**	**0.0368**	**0.0351**	**0.0032**
2008	0.7673	**0.0266**	0.1140	**0.0048**	**0.0465**	**0.0031**
2009	0.4956	0.2421	**0.0056**	**0.0242**	**0.0067**	**0.0012**

Table 3: p-values of a random permutation test for the null of independence with 10.000 bootstrap realizations using the mutual information statistic for the variable MVD (motor vehicle deaths per 100 million vehicle miles of travel) from 1980 to 2009, for different number of neighbours ($k^\star = 1, 3$, and 5) and number of bins ($n = 2$ and 3). The network configuration is prescribed by the geographic distance between the states, such that each state is connected to the closest k^* states.

respectively. The permutation test is again performed on $B = 10,000$ bootstrap realizations for both configurations.

Tables 3 and 4 summarize the results of the statistical analysis, in terms of the p-values of motor vehicle deaths per 100 million vehicle miles of travel from 1980 to 2009, for the network configurations construed from geographic and ideological distances, respectively. In bold, we underscore the cases in which the test rejects the null hypothesis of independence at a 5% confidence level.

Overall, statistical results in Tables 3 and 4 support the qualitative observations in [22], demonstrating a strong association between state spatial proximity and motor vehicle deaths. The proposed information-theoretic test rejects the null of independence at a 5% confidence level

Year	$n=2$			$n=3$		
	$k^*=1$	$k^*=3$	$k^*=5$	$k^*=1$	$k^*=3$	$k^*=5$
1980	0.5017	0.4565	0.1595	**0.0002**	**0.0040**	0.0727
1981	**0.0013**	**0.0000**	**0.0005**	**0.0100**	**0.0052**	**0.0093**
1982	0.3224	0.5466	0.0540	0.2277	**0.0181**	**0.0188**
1983	1.0000	0.3295	0.3666	**0.0308**	0.0676	0.1413
1984	0.1420	0.3982	0.4955	0.8829	0.1678	0.5909
1985	0.0665	0.5188	0.1541	0.2396	0.1878	0.2542
1986	0.6992	**0.0420**	0.3618	0.0619	0.1815	0.1031
1987	0.6047	0.3366	1.0000	0.5900	0.4880	0.8829
1988	0.7004	0.4407	0.4966	0.6990	1.0000	0.2859
1989	0.1374	0.0590	0.1557	0.4382	0.3502	0.4305
1990	0.5039	0.9134	0.4836	0.4202	0.3790	0.3569
1991	**0.0014**	**0.0039**	**0.0044**	**0.0071**	**0.0013**	**0.0029**
1992	0.4990	0.0588	0.4801	0.5201	0.1794	0.0672
1993	0.2007	0.3799	0.3490	0.3577	**0.0206**	0.1323
1994	**0.0014**	**0.0298**	**0.0053**	0.0611	**0.0019**	**0.0137**
1995	1.0000	0.1662	0.1592	0.2470	0.0745	0.3993
1996	0.1535	0.3180	**0.0053**	0.3996	**0.0026**	0.2308
1997	**0.0012**	0.0512	**0.0001**	**0.0024**	**0.0003**	**0.0042**
1998	0.0624	0.5699	**0.0188**	**0.0132**	**0.0265**	**0.0022**
1999	0.2014	0.3027	0.3607	0.2358	0.1865	0.3327
2000	0.0281	0.4338	**0.0059**	0.1073	**0.0009**	0.0503
2001	0.1610	0.5172	0.1788	0.0681	0.1864	0.2331
2002	0.1652	0.7066	**0.0223**	0.2482	**0.0069**	0.2329
2003	0.5045	0.9109	0.0689	0.4075	0.0751	0.2779
2004	0.1477	0.1366	0.4983	0.4718	0.1927	0.5193
2005	0.1789	0.2302	0.1700	**0.0281**	0.0804	0.1111
2006	0.3428	0.3769	**0.0213**	0.0944	**0.0021**	0.1511
2007	**0.0284**	0.1707	0.1705	**0.0341**	**0.0245**	0.1282
2008	0.6951	0.5387	0.3539	0.3143	0.3856	0.0726
2009	1.0000	0.1134	0.3545	1.0000	1.0000	0.1626

Table 4: p-values of a random permutation test for the null of independence with 10.000 bootstrap realizations using the mutual information statistic for the variable MVD (motor vehicle deaths per 100 million vehicle miles of travel) from 1980 to 2009, for different number of neighbours ($k^* = 1, 3$, and 5) and number of bins ($n = 2$ and 3). The network configuration is prescribed by the ideological distance between the states, such that each state is connected to the k^* states whose political ideology is the most similar.

for 28 out of the 30 years for some combination of number of neighbours k^* and number of bins n. Only in 1980 and 1990, the null of independence is not rejected for any combination of k^* and n. As proposed by Abaid *et al.* [22] this association may be explained by the composition of the transportation infrastructure and the legal environment in the USA. With respect to the highway infrastructure, states that are geographically close are connected by interstate highways between urban centers and low traffic volume roads outside built-up areas. Hence, it is tenable that motor vehicles deaths in states that are geographically close will tend to be similar. With respect to the legal environment, proximal states are likely to learn from each other best practices in traffic safety and cooperate on laws regarding traffic fatalities.

Our statistical test also identifies a predictable association between motor vehicle deaths and political ideology, especially when three or five neighbours are considered in the network assembly. However, compared to the dependence on geographic proximity, the association with political ideology is weaker, whereby only for 15 of the 30 years we reject the null hypothesis of independence for some combination of $k^\star$ and n. As proposed in [22], a more liberal ideology contributes to greater number of laws targeting behaviours in different state contexts, thereby linking patterns of motor vehicle deaths to similarities in the ideology index. States with widely different political ideologies are likely to differ in motor vehicles deaths, whereby they would embrace different approaches to deal with risks and protective factors for traffic fatalities.

6. Conclusions

Discovering the emergence of spatial patterns and pinpointing their causes are critical areas of research in science and engineering. This paper puts forward a model-free, statistically-principled approach to study spatial dependencies in small datasets. Grounded in information and network theory, the proposed approach takes as inputs available observations and a candidate interaction network, and outputs a mutual information statistic that measures the extent to which the candidate interaction is similar to the true interaction.

For linear spatial autoregressive properties and a class of networks (circulant and k-regular), we have established closed-form results for the mutual information statistic, in terms of salient topological features and intensity of the spatial dependence in the underlying process. To some extent, these analytical results complement existing findings on time-series analysis, such as prior work by Hahs and Pethel [42] and Smirnov [43], who presented exact results for information-theoretic measures in autoregressive temporal processes.

Unique to spatial analysis is the complexity of the interaction among the random variables that comprise the process [34]. In the words of Cliff and Ord [44], "the variate of a time series is influenced only by past values, while for a spatial process, dependence extends in all directions." Hence, for a temporal process, we can utilize the time-axis to order the interactions among the observations, which are, in general, confined to local time-histories. In the case of a spatial process, the underlying network could favor nonlocal interactions among the variables, which might even challenge the stationarity of the process.

Across a range of synthetic datasets, we have demonstrated the viability of the proposed association scheme in the study of small datasets. For linear and nonlinear spatial autoregressive processes, we have shown that 50 observations could be sufficient for the proposed approach to support the accurate inference of spatial patterns. Not only is the association scheme successful in identifying spatial patterns when truly present in the process, but also it begets a minimal number of false positives when implemented on observations generated by independent variables. Increasing the intensity of the spatial dependence in either linear or nonlinear datasets favors the accuracy of the association scheme, offering evidence in favor of the model-free use of the approach on arbitrary datasets.

With respect to real datasets, we have demonstrated the association scheme on human migration data under climate change in Bangladesh [21] and motor vehicle deaths in the USA [22]. For both datasets, the association scheme is successful in isolating valid explanations for spatial patterning. In the study of human migration, we successfully uncovered a tendency of people to migrate toward neighbouring districts, which should have been expected based on the diffusion model used by the authors to predict human displacement [37]. In the analysis of motor vehicle deaths, we offered strong evidence in favor of the correlation analysis in [22] that pointed at a strong association between geographic proximity of states and motor vehicle deaths.

There are two main avenues of future research that we plan to pursue. First, the proposed approach could be extended to support multivariate spatial analysis, potentially extending the line of work in [45]. Therein, the authors introduced an information-theoretic approach to isolate asymmetric relationships between two spatial processes, under a known network of interaction assembled on the basis of geographic proximity. It is tenable that the proposed association scheme

could offer an appropriate tool to select optimal interaction networks on which to examine the proposed asymmetric relationships.

Second, we foresee the possibility of extending the approach to study spatiotemporal patterns in small datasets. Along this research direction, the main technical challenge is likely in how to parsimoniously discretize the process to afford accurate inferences for small datasets. This may be achieved through symbolic representations of spatiotemporal features that could increase the power of the association scheme without exacerbating the problem of estimating salient information-theoretic quantities. If successful, the extension of the approach to spatiotemporal processes could support new endeavors in causality analysis [46,47] at a yet-to-be-explored mesoscale, where the researcher will neither seek to detail granular interactions between nodes (which typically require prohibitively large datasets), nor to describe the macroscopic response of the system (which may be offering limited insight into the phenomenon under investigation).

References

1. J. P. Elhorst, *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. Springer, 2014, vol. 479.
2. A. Getis, "Reflections on spatial autocorrelation," *Regional Science and Urban Economics*, vol. 37, no. 4, pp. 491–496, 2007.
3. J. LeSage and R. K. Pace, *Introduction to Spatial Econometrics*. Chapman and Hall/CRC, 2009.
4. J. Lessler, H. Salje, M. K. Grabowski, and D. A. Cummings, "Measuring spatial dependence for infectious disease epidemiology," *PLoS ONE*, vol. 11, no. 5, 2016.
5. J. Miller, J. Franklin, and R. Aspinall, "Incorporating spatial dependence in predictive vegetation models," *Ecological Modelling*, vol. 202, no. 3-4, pp. 225–242, 2007.
6. J. P. Hughes and P. Guttorp, "Incorporating spatial dependence and atmospheric data in a model of precipitation," *Journal of Applied Meteorology*, vol. 33, no. 12, pp. 1503–1515, 1994.
7. P. McCullagh and D. Clifford, "Evidence for conformal invariance of crop yields," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 462, no. 2071, pp. 2119–2143, 2006.
8. R. Basile, "Regional economic growth in europe: A semiparametric spatial dependence approach," *Papers in Regional Science*, vol. 87, no. 4, pp. 527–544, 2008.
9. A. D. Cliff, "Spatial autocorrelation," Tech. Rep., 1973.
10. A. D. Cliff and J. K. Ord, *Spatial Processes: Models & Applications*. Taylor & Francis, 1981.
11. P. A. Moran, "Notes on continuous stochastic phenomena," *Biometrika*, vol. 37, no. 1/2, pp. 17–23, 1950.
12. L. Anselin, *Spatial Econometrics: Methods and Models*. Springer Science & Business Media, 2013, vol. 4.
13. H. H. Kelejian and D. P. Robinson, "A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model," *Papers in Regional Science*, vol. 72, no. 3, pp. 297–312, 1993.
14. ——, "Spatial correlation: a suggested alternative to the autoregressive model," in *New Directions in Spatial Econometrics*. Springer, 1995, pp. 75–95.
15. L. Anselin and R. Moreno, "Properties of tests for spatial error components," *Regional Science and Urban Economics*, vol. 33, no. 5, pp. 595–618, 2003.
16. T. De Graaff, R. J. Florax, P. Nijkamp, and A. Reggiani, "A general misspecification test for spatial regression models: dependence, heterogeneity, and nonlinearity," *Journal of Regional*

Science, vol. 41, no. 2, pp. 255–276, 2001.

17. W. A. Broock, J. A. Scheinkman, W. D. Dechert, and B. LeBaron, "A test for independence based on the correlation dimension," *Econometric Reviews*, vol. 15, no. 3, pp. 197–235, 1996.

18. F. López, M. Matilla-García, J. Mur, and M. R. Marín, "A non-parametric spatial independence test using symbolic entropy," *Regional Science and Urban Economics*, vol. 40, no. 2-3, pp. 106–115, 2010.

19. M. Matilla-García and M. R. Marín, "A non-parametric independence test using permutation entropy," *Journal of Econometrics*, vol. 144, no. 1, pp. 139–155, 2008.

20. C. Brett and J. Pinkse, "Those taxes are all over the map! A test for spatial independence of municipal tax rates in British Columbia," *International Regional Science Review*, vol. 20, no. 1-2, pp. 131–151, 1997.

21. K. F. Davis, A. Bhattachan, P. DâĂŹOdorico, and S. Suweis, "A universal model for predicting human migration under climate change: examining future sea level rise in Bangladesh," *Environmental Research Letters*, vol. 13, no. 6, p. 064030, 2018.

22. N. Abaid, J. Macinko, D. Silver, and M. Porfiri, "The effect of geography and citizen behavior on motor vehicle deaths in the United States," *PLoS ONE*, vol. 10, no. 4, 2015.

23. E. Estrada and P. A. Knight, *A First Course in Network Theory*. Oxford University Press, USA, 2015.

24. M. Barthélemy, "Spatial networks," *Physics Reports*, vol. 499, no. 1-3, pp. 1–101, 2011.

25. T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

26. P. Hall and S. C. Morton, "On the estimation of entropy," *Annals of the Institute of Statistical Mathematics*, vol. 45, no. 1, pp. 69–88, 1993.

27. L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

28. P. J. Davis, *Circulant Matrices*. American Mathematical Society, 2013.

29. D. E. Myers, "To be or not to be... stationary? That is the question," *Mathematical Geology*, vol. 21, no. 3, pp. 347–362, 1989.

30. I. Kontoyiannis and M. Madiman, "Sumset and inverse sumset inequalities for differential entropy and mutual information," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4503–4514, 2014.

31. V. Rohatgi, "An introduction to probability theory and mathematical," *Statistics (New York: Wiley, 1976)*, p. 27, 1976.

32. P. Sulewski, "Equal-bin-width histogram versus equal-bin-count histogram," *Journal of Applied Statistics*, pp. 1–20, 2020.

33. J. Montalvão, R. Attux, and D. Silva, "A pragmatic entropy and differential entropy estimator for small datasets," *Journal of Communication and Information Systems*, vol. 29, no. 1, 2014.

34. B. Fingleton, "Spatial autoregression," *Geographical Analysis*, vol. 41, no. 4, pp. 385–391, 2009.

35. E. Estrada and D. J. Higham, "Network properties revealed through matrix functions," *SIAM Review*, vol. 52, no. 4, pp. 696–714, 2010.

36. G. Brinkmann, "Generating regular directed graphs," *Discrete Mathematics*, vol. 313, no. 1, pp. 1–7, 2013.

37. F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, no. 7392, pp. 96–100, 2012.

38. T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P. M. Midgley *et al.*, "Climate change 2013: The physical science basis," *Contribution of working group I to the fifth assessment report of the Intergovernmental Panel on Climate Change*, vol. 1535, 2013.

39. I. Ciesin, "Gridded population of the world, version 4 (gpwv4): Population count. palisades, NY: NASA socioeconomic data and applications center (SEDAC)," *Center for International Earth Science Information Network (CIESIN) Columbia University*, 2016.

40. U. Nations, "World population prospects: The 2015 revision, key findings and advance tables. New York: United Nations, Department of Economic and Social Affairs PD," *Population Division*, 2015.

41. F. S. Berry and W. D. Berry, "Innovation and diusion models in policy research," in *Theories of the Policy Process*. Routledge, 2018, pp. 263–308.

42. D. W. Hahs and S. D. Pethel, "Transfer entropy for coupled autoregressive processes," *Entropy*, vol. 15, no. 3, pp. 767–788, 2013.

43. D. A. Smirnov, "Spurious causalities with transfer entropy," *Physical Review E*, vol. 87, no. 4, p. 042917, 2013.

44. A. Cliff and J. K. Ord, "The problem of spatial autocorrelation," *London Papers in Regional Science*, 1969.

45. M. Herrera, J. Mur, and M. Ruiz, "Detecting causal relationships between spatial processes," *Papers in Regional Science*, vol. 95, no. 3, pp. 577–594, 2016.

46. T. Bossomaier, L. Barnett, M. Harré, and J. T. Lizier, "An introduction to transfer entropy," *Cham: Springer International Publishing*, pp. 65–95, 2016.

47. M. Weilenmann and R. Colbeck, "Analysing causal structures with entropy," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 473, no. 2207, p. 20170483, 2017.