Probing the Information Theoretical Roots of Spatial Dependence Measures

Zhangyu Wang
University of California Santa Barbara, CA, USA

Krzysztof Janowicz
Faculty of Geosciences, Geography and Astronomy, University of Vienna, Austria
University of California Santa Barbara, CA, USA

Gengchen Mai
SEAI Lab, Department of Geography and the Environment, University of Texas at Austin, TX, USA
Department of Geography, University of Georgia, Atlanta, GA, USA

Ivan Majic
University of Vienna, Austria

Abstract

Intuitively, there is a relation between measures of spatial dependence and information theoretical measures of entropy. For instance, we can provide an intuition of why spatial data is special by stating that, on average, spatial data samples contain less than expected information. Similarly, spatial data, e.g., remotely sensed imagery, that is easy to compress is also likely to show significant spatial autocorrelation. Formulating our (highly specific) core concepts of spatial information theory in the widely used language of information theory opens new perspectives on their differences and similarities and also fosters cross-disciplinary collaboration, e.g., with the broader AI/ML communities. Interestingly, however, this intuitive relation is challenging to formalize and generalize, leading prior work to rely mostly on experimental results, e.g., for describing landscape patterns. In this work, we will explore the information theoretical roots of spatial autocorrelation, more specifically Moran's I, through the lens of self-information (also known as surprisal) and provide both formal proofs and experiments.

2012 ACM Subject Classification Mathematics of computing → Information theory; Information systems → Geographic information systems; Computing methodologies → Philosophical/theoretical foundations of artificial intelligence

Keywords and phrases Spatial Autocorrelation, Moran's I, Information Theory, Surprisal, Self-Information

Digital Object Identifier 10.4230/LIPIcs.COSIT.2024.9

Related Version *Full Version*: `https://arxiv.org/abs/2405.18459` [27]

Supplementary Material *Software (Source Code)*: `https://github.com/Octopolugal/Spatial-Self-Information`, archived at `swh:1:dir:808cc16c4a9af8b4d92eb57450fa13878c941075`

Funding *Zhangyu Wang*: This work was supported by the National Science Foundation under Grant No. 2033521 A1 – KnowWhereGraph: Enriching and Linking Cross-Domain Knowledge Graphs using Spatially-Explicit AI Technologies.

1 Introduction

To explain why *spatial is special* we often list the characteristics underlying spatial data and the processes that created them by pointing to classics such as the modifiable areal unit problem (MAUP) [6], spatial dependence and interaction, scale, edge effects, and so forth. However, there are many alternative formulations, some of which draw a more direct relation to our neighboring academic disciplines. For instance, when explaining Tobler's First Law of

Geography [25, 24] to our colleagues in the broader AI/ML community by introducing terms such as spatial dependence and spatial autocorrelation measures such as Moran's I, we could instead state that *spatial is special because on average a sample of spatial data contains less than expected information.*

While highlighting different aspects, e.g., omitting the explicit *nearness* from the original definition, such an information theoretic perspective is in many ways equivalent, yet opens up different avenues for understanding why spatial (data) is special from the viewpoint of neighboring but substantially larger disciplines without the need to introduce our own terminology. The information-theoretic statement above also gives rise to an entropy-based understanding of spatial dependence that translates seamlessly into common loss functions in machine learning, such as cross-entropy for classification.

Similar observations be them in space, time, or spacetime can be made about information compression. Intuitively, data, e.g., remotely sensed imagery, with a high degree of spatial autocorrelation should be easier to compress than data with almost no spatial dependence. What is true for simply run-length compression or quad-trees also holds for more abstract situations. For instance, according to lossless compression techniques such as Huffman coding [19], daily weather reports (sunny, cloudy, rainy, ...) for Santa Barbara can utilize fewer bits of information than for Vienna. Conversely, very compressible data is also more likely to show strong spatial autocorrelation.

These thoughts do not imply that measures of information compression or entropy can (or should) replace our domain-specific measures such as Moran's I, Geary's C, semivariograms, and so forth, but that it is worth exploring their commonalities (and differences). For instance, while high (Shannon) entropy implies low spatial autocorrelation, measures such as Moran's I also have an explicit notion of neighborhood encoded via their weights matrix. Hence, a binary checkboard pattern would yield $I = -1$ while a binary partition in two areas would approximate $I \approx 1$. From an entropy perspective, these two patterns are similar as the proportion of blacks and whites, e.g., cells, remains the same. Similar arguments can be made for the compression examples. However, adding the required neighborhood notion to (discrete) entropy is possible, as will be shown below for self-information (i.e., *surprisal*) for the case of classed data, be they rasters or vectors.

Interestingly, while the relationship between information theory and (spatial) autocorrelation has been noted by others before, the formalization is surprisingly challenging, leading most prior work to take a largely experimental stance. In our work here, we will provide both an experimental intuition and more formal proofs for the proposed *spatial self-information*. Summing up, exploring the information-theoretical roots of spatial dependence and, more specifically, Moran's I is worthwhile for at least the **following reasons**:

1. **Fostering cross-disciplinarity**: Our community has shown that spatially explicit machine learning models do not only increase the accuracy of (Geo)AI models when applied to geographic data but also inform and improve more general models in various domains [14, 16, 15, 17, 4, 26, 23], e.g., leading to a broad interest in location encoding methods outside of GeoAI. Conversely, researchers from the broader AI community [18] try to utilize notions such as the MAUP to study problematic coverage and representation biases in training data for image-based foundation models. Yet collaboration and reuse of prior results are sometimes hindered by our highly specific terminology and methods. Casting spatial core concepts in the shared language of information theory may mitigate these issues and also accelerate progress.
2. **Quantifying spatial patterns:** Incorporating results from information-theoretical and physical entropy (and related ideas) may open up new avenues to describe complex spatial patterns beyond what is currently available in our spatial analytics toolbox as recently demonstrated by the use of configurational entropy for complex landscapes [5].

3. **Spatial Data Science education:** Most introductory textbooks on GIS, GIScience, or Spatial Data Science barely make a connection between spatial dependence, information (e.g., image) compression, information theory, and so forth while covering all of them to at least some extent.[1] This makes it difficult for students to grasp the bigger picture, e.g., when meeting entropy again while studying spatial clustering and classification.

2 Motivation and Related Works

Spatial autocorrelation has long been a research focus for both GIScience [9, 8] and statistics [3]. Efforts have been made to develop statistics that test whether a sample of spatially distributed data is autocorrelated, i.e., against the null hypothesis that the spatial arrangement of the data is randomly generated. Commonly used statistics include Moran's I [20, 28], Geary's C [7], and so forth. Apart from autocorrelation statistics, information-theory-based measures like Batty's spatial entropy [1] and S statistics [12], which extends Moran's I by assuming the observed values are probabilities, have also been studied. However, how to relate these two types of measures lacks in-depth investigation.

As discussed in the introduction, we are motivated to connect spatial autocorrelation statistics with information-theoretic quantities like entropy and self-information, for the sake of relating theoretical concepts from different disciplines and exploring wider applications (e.g., introducing spatial autocorrelation in loss functions). More specifically, we wish to quantify the self-information, i.e., the *surprisal*, of observing a sample with a certain degree of spatial autocorrelation. The intuition is that higher spatial autocorrelation implies more regular spatial patterns, which is more surprising.

Unfortunately, research shows that there is no general relation between an autocorrelation statistic and its corresponding self-information [2]. The information-theoretic counterpart of a spatial autocorrelation statistic needs to be established case by case. In this paper, we aim at deriving that of the (global) Moran's I.

In general, we need to know the probability of observing a certain type of sample to obtain its surprisal. In our case, this means knowing the probability of observing a sample with a certain value of Moran's I. This is an under-studied topic due to the difficulty of deriving the analytical distribution of Moran's I. Instead, researchers use permutation inference to empirically compute the reference distribution. This is good enough for hypothesis tests, as we only need the p-values, but not enough for computing the self-information.

Attempts have been made to study the asymptotic behavior of Moran's I under the assumption of knowing the specific underlying stochastic process. For example, Kelejian et al [10] derived the analytical distribution of Moran's I by assuming the spatial data are generated by a linear model, and the regression disturbance is a known priori or estimated from data. This assumption is reasonable in some areas such as economics, but not necessarily appropriate in geology, urban planning, landscape, remote sensing, etc.

In many cases, the underlying stochastic process is unknown, and all we can rely on is a broad assumption of randomness. In this sense, it is a combinatorics problem with a strong relation to entropy. Some researches approach a simplified version of this problem via Shannon entropy of co-occurrence counts [13, 21]. They only consider categorical differences, i.e., whether neighboring observations are of the same class, without addressing the numerical differences that are present in classic spatial autocorrelation statistics. Cushman [5] took an important step in incorporating numerical differences by empirically revealing that

[1] O'Sullivan's and Unwin's *Geographic Information Analysis* being a rare exception.

the distribution of the total weighted distance between different-valued cells on a grid approximates the distribution of Moran's I. We are inspired by this observation and have decided to provide a more formal analytical analysis.

In short, we prove why Moran's I asymptotically follows a normal distribution, and its analytical form can be specified without sampling and estimation. This allows efficient and accurate calculation of the probability, consequently the self-information, of any spatial sample with a binary weight matrix. Without the assumptions of the underlying stochastic process, this self-information of spatial autocorrelation can be applied to a wider range of fields and is computationally friendly to be combined with learning algorithms.

3 Method

In this section we provide a comprehensive analysis of the asymptotic distribution of Moran's I with binary weights. First we formally define the problem we want to solve and necessary notations in Section 3.1, elaborate the proof from Section 3.2 to Section 3.5, and finally reach the core results of this research, i.e., the analytical approximation of Moran's I distribution specified in Theorem 8 and Theorem 9.

3.1 Problem Setup

Consider a sample of N indexed observations $\{x_1, x_2, \cdots x_N\}$. We assume the observations take limited discrete *values*, i.e., $\forall i, x_i \in C_M := \{c_1, c_2, \cdots c_M\}$, where $M \ll N$. Next, we define *value size* as $n_p := |\{x_i | x_i = c_p\}|$, i.e., the number of observations in the sample whose values are c_p. We call the set $T_M := \{(c_1, n_1), (c_2, n_2), \cdots (c_M, n_M)\}$ the *value scheme* of a sample, i.e., the discrete values and their numbers of occurrences in the sample. The value scheme measures the intrinsic variance in the sample itself regardless of the spatial arrangement. It uniquely determines the sample mean $\bar{x} = \sum_{i=1}^{N} x_i/N = \sum_{p=1}^{M} c_p n_p / \sum_{p=1}^{M} n_p$ and the sample variance $\sum_{i=1}^{N}(x_i - \bar{x})^2/N = \sum_{p=1}^{M}(c_p - \bar{x})^2 n_p / \sum_{p=1}^{M} n_p$.

The binary spatial weight is defined as $w_{i,j} = \mathbb{I}\{x_i \text{ and } x_j \text{ are neighbors}\}$. Then we can form a directed graph G with N vertices $V = \{v_1, v_2, \cdots v_N\}$ where v_i corresponds to the i-th observation x_i, and (v_i, v_j) is an edge if and only if $w_{i,j} = 1$. We further require that the degree of each vertex is a fixed number $k \ll N$ (e.g., for the conventional rook and queen weights, $k = 4$ and $k = 8$, respectively, ignoring the border and corner variables).

Recall that Moran's I is defined as:

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Moran's I values of two samples are not directly comparable when their value schemes are vastly different. For example, a grid with randomly arranged high-variance values and the same grid with checkerboard arranged low-variance values both have Moran's I ≈ 0. However, the latter case obviously demonstrates higher spatial autocorrelation and should be differentiated from the former. Therefore, throughout this paper, $\mathbb{P}(I = \alpha)$ (the probability of Moran's I being α) always refers to the conditional probability $\mathbb{P}(I = \alpha | T_M)$ (the probability of Moran's I being α given a known and fixed value scheme).

When T_M is known and fixed, $N = \sum_{p=1}^{M} n_p$, $\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j} = kN$, $\bar{x}$ and $\sum_{i=1}^{N}(x_i - \bar{x})^2$, are also known and fixed. Define $\bar{I} := \sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}(x_i - \bar{x})(x_j - \bar{x})$ as the *unscaled Moran's I*. Then $I \propto \bar{I}$. For the simplicity of discussion, in the following, we will focus on $\bar{I}$.

Our purpose is to find the distribution of $\bar{I}$ (and naturally that of I) so that for any $\alpha \in \mathbb{R}$ we know $\mathbb{P}(\bar{I} = \alpha)$. Then we define *spatial self-information* $J := -\log(\mathbb{P}(\bar{I} = \alpha))$, which quantifies the surprisal of observing a sample with a certain degree of spatial autocorrelation.

To achieve this, we will prove that the distribution of $\bar{I}$ can be asymptotically approximated by a sum of normal distributions in the rest of the section. The proof is structured as follows:

1. First, in Section 3.2, we rearrange $\bar{I}$ as a weighted sum of several random variables;
2. Then, in Section 3.3, we prove that the distributions of these random variables asymptotically follow binomial/Poisson binomial distributions under certain necessary assumptions;
3. Finally, in Section 3.4, we show that these distributions can be approximated by normal distributions, which leads to the conclusion that $\bar{I}$ also follows a normal distribution. We further derive the analytical form of this distribution in Section 3.5.

Following that, we develop a set of techniques in Section 4 that correct the error caused by violations of assumptions and conditions made during the problem set-up and the proof. Experiments demonstrate that with these corrections our approximation is robust to relaxations. It enables practical application in real-world, i.e., non-ideal situations.

3.2 Rearrangement of $\bar{I}$

▶ **Definition 1.** *A tuple (x_i, x_j) is called a pq-pair if $x_i = c_p, x_j = c_q, w_{i,j} = 1$. We say a pq-pair starts with c_p and ends with c_q.*

▶ **Definition 2.** *$S_{p,q} := \{(x_i, x_j) | x_i = c_p,\ x_j = c_q$ and $w_{i,j} = 1\}$, $c_p, c_q \in C_M$ is the set of pq-pairs. If $p \neq q$, we say $S_{p,q}$ is a different-value set; otherwise a same-value set.*

It is worth noting that the order of indices matter. In a same-value set $S_{p,p}$, (x_i, x_j) and (x_j, x_i) are counted as two different *pp*-pairs even though $x_i = x_j = c_p$.

▶ **Lemma 3** (Rearrangement of $\bar{I}$ as Cardinality of Sets)**.** *$\bar{I}$ is a weighted sum of the cardinality of all possible sets of pq-pairs. Specifically, $\bar{I} = \sum_{p,q}(c_p - \bar{x})(c_q - \bar{x})|S_{p,q}|$.*

Proof.

$$
\begin{aligned}
&\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j}(x_i - \bar{x})(x_j - \bar{x}) \\
&= \sum_{p,q}\left[\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{I}\{x_i = c_p, x_j = c_q\} w_{i,j}(x_i - \bar{x})(x_j - \bar{x})\right] \\
&= \sum_{p,q}\left[\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{I}\{x_i = c_p, x_j = c_q, w_{i,j} = 1\}(x_i - \bar{x})(x_j - \bar{x})\right] \\
&= \sum_{p,q}\left[\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{I}\{x_i = c_p, x_j = c_q, w_{i,j} = 1\}(c_p - \bar{x})(c_q - \bar{x})\right] \\
&= \sum_{p,q}(c_p - \bar{x})(c_q - \bar{x})\left[\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{I}\{x_i = c_p, x_j = c_q, w_{i,j} = 1\}\right] \\
&= \sum_{p,q}(c_p - \bar{x})(c_q - \bar{x})|S_{p,q}|
\end{aligned}
$$

◀

Now, if we know the distributions of each $|S_{p,q}|$, we can analyze the distribution of their weighted sum. We will prove that for different-value sets, $|S_{p,q}|$ asymptotically follows a binomial distribution, and for same-value sets, $|S_{p,q}|$ asymptotically follows a Poisson binomial distribution.

3.3 Asymptotic Binomial and Poisson Binomial Distributions

Let $B(p, n)$ denote the binomial distribution with probability of success p and number of trials n, $PB(p_i; i = 1, 2, \cdots, n)$ denote the Poisson binomial distribution with probabilities of success $p_1, p_2, \cdots, p_n$, and $N(\mu, \sigma^2)$ denote the normal distribution with mean μ and variance σ^2.

▶ **Lemma 4** (Probability of the Cardinality of Same-Value Sets). *If $kn_p \ll N$, then $\frac{1}{2}|S_{p,p}| \sim PB(\frac{k(t-1)}{N}; t = 1, \cdots, n_p)$, with mean $\mu_{p,p} = \frac{1}{2}(n_p - 1)\frac{kn_p}{N}$ and variance $\sigma^2_{p,p} = \frac{1}{2}(n_p - 1)\frac{kn_p}{N}\left[1 - \frac{k(2n_p-1)}{3N}\right]$.*

Proof. Let black represents c_p. Consider a one-phase graph coloring process. G is the uncolored graph. We randomly (with equal probability) color a vertex black for n_p times, resulting in a black-colored graph G_b. The probability of having l edges of same colored vertices in G_b equals the probability of $|S_{p,p}| = 2l$ because each edge of the same colored vertices will be counted twice in $|S_{p,p}|$.

To obtain the general form of the asymptotic distribution of $|S_{p,p}|$, we need an assumption that none of the colored pairs share common neighbors except themselves, because the probability of two vertices sharing common neighbors depends on the specific structure of the uncolored graph G. Say coloring a neighbor of a black vertex to be a *success*, and otherwise a $failure$. This assumption guarantees that 1) each success creates and only creates two *pp*-pairs, 2) each success adds $k - 2$ edges that have and only have one vertex colored black and each failure adds k such edges, and 3) the probability of coloring a neighbor of a black vertex is independent of the previous coloring result.

The first time of coloring will not result in any edges of the same colored vertices. For the second time, the probability of coloring a neighbor of a black vertex, i.e., success in creating two *pp*-pairs, is $\frac{k}{N-1} \approx \frac{k}{N}$. Then for the third time, the probability of coloring a neighbor of a black vertex becomes $\frac{k}{N-1}\frac{2k-2}{N-2} + \left(1 - \frac{k}{N-1}\right)\frac{2k}{N-2} = \frac{2k^2}{(N-1)(N-2)} - \frac{2k}{(N-1)(N-2)} + \frac{2k}{N-2} - \frac{2k^2}{(N-1)(N-2)} = \frac{2k(N-2)}{(N-1)(N-2)} = \frac{2k}{N-1} \approx \frac{2k}{N}$. It is easy to verify that for the t-th time, the probability of success, i.e., creating two *pp*-pairs, is approximately $\frac{(t-1)k}{N}$.

We can view the coloring process as a series of n_p independent binary trials with increasing probabilities of success. Then $\frac{1}{2}|S_{p,p}| = l$ equals the total number of success among all n_p trials, which follows a Poisson binomial distribution $PB(\frac{k(t-1)}{N}; t = 1, \cdots, n_p)$, $\mu_{p,p} = \sum_{t=1}^{n_p}\frac{(t-1)k}{N} = \frac{1}{2}(n_p - 1)\frac{kn_p}{N}, \sigma^2_{p,p} = \sum_{t=1}^{n_p}\left[\frac{(t-1)k}{N}\left(1 - \frac{(t-1)k}{N}\right)\right] = \sum_{t=1}^{n_p}\frac{(t-1)k}{N} - \sum_{t=1}^{n_p}\left(\frac{(t-1)k}{N}\right)^2 = \frac{1}{2}(n_p - 1)\frac{kn_p}{N}\left[1 - \frac{k(2n_p-1)}{3N}\right]$ asymptotically as $\frac{n_p}{N} \to 0$. ◀

▶ **Lemma 5** (Probability of the Cardinality of Different-Value Sets). *If* $n_q \ll kn_p \ll N$, *then* $|S_{p,q}| \sim B(n_q, \frac{kn_p}{N})$, *with mean* $\mu_{p,q} = n_q \frac{kn_p}{N}$ *and variance* $\sigma^2_{p,q} = n_q \frac{kn_p}{N}\left(1 - \frac{kn_p}{N}\right)$.

Proof. The proof uses similar techniques as in Lemma 4. Let black represents c_p and white represents c_q. Consider a two-phase graph coloring process. G is the uncolored graph. In the first phase, randomly (with equal probability) color n_p vertices black, resulting in a black-colored graph G_b; in the second phase, randomly (with equal probability) color n_q vertices white, resulting in a black-white-colored graph G_{bw}. The probability of having l edges of differently colored vertices in G_{bw} equals the probability of $|S_{p,q}| = l$.

Since $kn_p \ll N$, the probability that we get a G_b in which two black vertices have common neighbors is very small. Thus the assumption we mentioned in the proof of Lemma 4 can be considered valid, i.e., we can assume that in all possible outcomes of G_b, all black vertices do not share neighbors, which means in G_b there are in total kn_p edges that have and only have one vertex colored black.

Then we randomly color n_q vertices white based on G_b. Consider this process as repeatedly coloring one random vertex white for n_q times. For the first time, the probability of coloring a neighbor of a black vertex, i.e., success in creating one pq-pair, is $\frac{kn_p}{N - n_p} \approx \frac{kn_p}{N}$. For the second time, the probability becomes $\frac{kn_p}{N - n_p}\frac{kn_p - 1}{N - n_p - 1} + \left(1 - \frac{kn_p}{N - n_p}\right)\frac{kn_p}{N - n_p - 1} \approx \frac{kn_p}{N}\frac{kn_p - 1}{N} + \left(1 - \frac{kn_p}{N}\right)\frac{kn_p}{N} \approx \frac{kn_p}{N}$. That is, when $kn_p \ll N$, the second step of coloring is approximately independent of the first step with the same probability of success. It can be easily verified that this approximate independence and equal probability holds for all n_q steps as long as $n_q \ll kn_p$. The intuition is simple: if you draw a dozen out of thousands of black and white balls, the color of each draw is almost independent of other draws with equal possibilities. Subsequently, the second phase of coloring can be viewed approximately as n_q i.i.d. binary trials with probability of success $\frac{kn_p}{N}$. Each successful trial adds a black-white edge to G_{bw}. We immediately know for any G_b, $|S_{p,q}| \sim B(n_q, \frac{kn_p}{N})$, $\mu_{p,q} = n_q \frac{kn_p}{N}, \sigma^2_{p,q} = \frac{kn_p n_q}{N}(1 - \frac{kn_q}{N})$ asymptotically as $\frac{n_p}{N} \to 0$ and $\frac{n_q}{N} \to 0$ because the sum of i.i.d. binary random variables is a binomial random variable. ◀

3.4 Normal Approximation of Binomial and Poisson Binomial Distributions

Whereas we have demonstrated that $\bar{I}$ is a weighted sum of binomial and Poisson binomial random variables, the analytical form of its distribution can not be simply found. Instead, if we approximate the binomial and Poisson binomial distributions with normal distributions, by the fact that normal distributions are stable distributions, the weighted sum can also be approximated by a normal distribution.

By the De Moivre-Laplace theorem [22], a binomial distribution with a relatively large probability of success (e.g., > 0.1) can be well approximated by a normal distribution with the same mean and variance. Therefore, we have:

▶ **Lemma 6** (Normal Approximation for Different-Value Sets). *If* $n_q \ll kn_p \ll N$, *then* $|S_{p,q}| \sim N\left(n_q \frac{kn_p}{N}, n_q \frac{kn_p}{N}\left(1 - \frac{kn_p}{N}\right)\right)$ *approximately.*

In addition, in [11], the authors demonstrate that a Poisson binomial distribution consisting of n independent binary trials with mean μ and variance σ^2 can be approximated by a normal distribution $N(\mu - \frac{1}{2}, \sigma^2)$ when n is sufficiently large. According to [11], we have:

▶ **Lemma 7** (Normal Approximation for Same-Value Sets). *If $kn_p \ll N$ and n_p is sufficiently large, then $|S_{p,p}| \sim N\Big((n_p-1)\frac{kn_p}{N} - 1, 2(n_p-1)\frac{kn_p}{N}\left[1 - \frac{k(2n_p-1)}{3N}\right]\Big)$ approximately.*

According to Lemma 6 and Lemma 7, both $|S_{p,q}|$ and $|S_{p,p}|$ can be approximated as normal distributions. Additionally, we know that for a normal random variable $Y \sim N(\mu, \sigma^2)$ and any constants $a_1, a_2, a_1 \neq 0$, $a_1Y + a_2 \sim N(a_1\mu + a_2, a_1^2\sigma^2)$. Since Lemma 3 shows that $\bar{I}$ is a weighted sum of $|S_{p,q}|$ and $|S_{p,p}|$, we can derive that $\bar{I}$ can be also approximated with a normal distribution.

3.5 Analytical Approximation of the Distribution of $\bar{I}$

We know from the discussions above that $\bar{I}$ approximately follows a normal distribution. The parameters we need to specify are its mean and variance.

▶ **Theorem 8** (Approximate Mean of $\bar{I}$). *Given T_M, the approximate mean of $\bar{I}$ is*

$$\tilde{\mu}_{\bar{I}} = \sum_{p \neq q}(c_p - \bar{x})(c_q - \bar{x})\mu_{p,q} + \sum_{p}(c_p - \bar{x})^2\mu_{p,p} \tag{1}$$

where $\mu_{p,q} = \min(n_p, n_q)\frac{k\max(n_p, n_q)}{N}$, $\mu_{p,p} = (n_p - 1)\frac{kn_p}{N} - 1$

Proof. This is a simple consequence of the fact that for normal random variables, the mean of the sum equals the sum of the means. The minimum and the maximum functions are included to satisfy the condition that $n_q \ll kn_p$ in Lemma 5. ◀

It is more complicated to derive the variance. We know for independent normal random variables, the variance of the sum equals the sum of variances. However, this independence requirement does not hold for all $|S_{p,q}|$. When the spatial arrangement of $M-1$ values is known, the spatial arrangement of the remaining value is automatically known. That is, for any $1 \leq r \leq M$, $\sum_{p\neq r, q\neq r}|S_{p,q}|$ and $\sum_{p=r \text{ or } q=r}|S_{p,q}|$ are correlated. Call c_r the *background value* and the other values *foreground values*. The following theorem states that if there is a background value that a sufficiently large proportion of samples takes, we can analytically approximate the variance of $\bar{I}$.

▶ **Theorem 9** (Approximate Variance of $\bar{I}$). *Given T_M, let $c_{r_{\max}}$ be the value that has the largest value size $n_{r_{\max}}$. If $\frac{n_{r_{\max}}}{N}$ is sufficiently large, the approximate variance of $\bar{I}$ is*

$$\begin{aligned}\tilde{\sigma}^2_{\bar{I}} = &\sum_{p\neq q\neq r_{\max}}\left[(c_p - \bar{x})(c_q - \bar{x}) - 2(c_p - \bar{x})(c_{r_{\max}} - \bar{x}) + (c_{r_{\max}} - \bar{x})^2\right]^2\sigma^2_{p,q} \\ &+ \sum_{p\neq r_{\max}}\left[(c_p - c_{r_{\max}})^2\right]^2\sigma^2_{p,p}\end{aligned} \tag{2}$$

where

$$\sigma^2_{p,q} = \min(n_p, n_q)\frac{k\max(n_p, n_q)}{N}\left(1 - \frac{k\max(n_p, n_q)}{N}\right)$$

$$\sigma^2_{p,p} = 2(n_p - 1)\frac{kn_p}{N}\left[1 - \frac{k(2n_p-1)}{3N}\right]$$

Proof. Consider generating a sample by an $M-1$ phase graph coloring process, given T_M. Choose c_r as the background value. In each phase, we select an index i from 1 to M except r without replacement and fill n_i vertices with the color of c_i. Regardless of how we select the values, after $M-1$ phases, the way to color the remaining n_r vertices is now fixed – i.e., the cardinality of the sets related to the last value is fully determined by the cardinality of other sets.

Formally, given any $1 \leq r \leq M$ and $p \neq r$,

$$kn_p = \sum_q |S_{p,q}| = |S_{p,r}| + \sum_{q \neq r} |S_{p,q}|$$

because $\sum_q |S_{p,q}|$ is the total number of pairs that start with c_p, i.e., the edges in the directed graph G that starts with v_p. Then

$$|S_{p,r}| = kn_p - \sum_{q \neq r} |S_{p,q}| \tag{3}$$

By symmetry, $|S_{r,p}| = |S_{p,r}|$. Now consider $|S_{r,r}|$. Similarly,

$$\begin{aligned} kN &= \sum_{p,q} |S_{p,q}| = |S_{r,r}| + \sum_{p \neq r \text{ or } q \neq r} |S_{p,q}| \\ &= |S_{r,r}| + \sum_{q \neq r} |S_{r,q}| + \sum_{p \neq r} |S_{p,r}| + \sum_{p \neq r, q \neq r} |S_{p,q}| \end{aligned}$$

Then

$$|S_{r,r}| = kN - \left(\sum_{q \neq r} |S_{r,q}| + \sum_{p \neq r} |S_{p,r}| + \sum_{p \neq r, q \neq r} |S_{p,q}| \right) \tag{4}$$

When $\frac{n_r}{N}$ is sufficiently large, the $M-1$ phases of graph coloring can be considered independent (similar intuition as in the proof of Lemma 5). Thus $|S_{p,q}|, p \neq r, q \neq r$, i.e. the cardinality of sets restricted to the foreground values, are approximately independent normal random variables. If we can represent $\bar{I}$ as a weighted sum of the cardinality of these sets, the variance of $\bar{I}$ can then be represented as a weighted sum of their variances.

▶ **Lemma 10** (Rearrangement of $\bar{I}$ as Cardinality of Sets restricted to Foreground Values)**.** *$\bar{I}$ is a weighted sum of the cardinality of all possible sets of pq-pairs restricted to the foreground values. Specifically, let c_r be the background value, then*

$$\bar{I} = Q + \sum_{p \neq r, q \neq r} \left[(c_p - \bar{x})(c_q - \bar{x}) - 2(c_p - \bar{x})(c_r - \bar{x}) + (c_r - \bar{x})^2 \right] |S_{p,q}|$$

where $Q := (c_r - \bar{x})^2 kN + 2 \sum_{p \neq r} \left[(c_p - \bar{x})(c_r - \bar{x}) - (c_r - \bar{x})^2 \right] kn_p$ is a constant.

Proof.

$$\begin{aligned} \bar{I} &= \sum_{p,q} (c_p - \bar{x})(c_q - \bar{x}) |S_{p,q}| \\ &= (c_r - \bar{x})^2 |S_{r,r}| + \sum_{q \neq r} (c_r - \bar{x})(c_q - \bar{x}) |S_{r,q}| + \sum_{p \neq r} (c_p - \bar{x})(c_r - \bar{x}) |S_{p,r}| \\ &\quad + \sum_{p \neq r, q \neq r} (c_p - \bar{x})(c_q - \bar{x}) |S_{p,q}| \end{aligned}$$

Use Equation 4,

$$\begin{aligned}\bar{I} = (c_r - \bar{x})^2 \left[kN - \left(\sum_{q \neq r} |S_{r,q}| + \sum_{p \neq r} |S_{p,r}| + \sum_{p \neq r, q \neq r} |S_{p,q}| \right) \right] \\ + \sum_{q \neq r} (c_r - \bar{x})(c_q - \bar{x}) |S_{r,q}| + \sum_{p \neq r} (c_p - \bar{x})(c_r - \bar{x}) |S_{p,r}| \\ + \sum_{p \neq r, q \neq r} (c_p - \bar{x})(c_q - \bar{x}) |S_{p,q}|\end{aligned}$$

Use symmetry $|S_{r,p}| = |S_{p,r}|$,

$$\begin{aligned}\bar{I} &= (c_r - \bar{x})^2 kN - (c_r - \bar{x})^2 \left(2 \sum_{p \neq r} |S_{p,r}| + \sum_{p \neq r, q \neq r} |S_{p,q}| \right) \\ &\quad + 2 \sum_{p \neq r} (c_p - \bar{x})(c_r - \bar{x}) |S_{p,r}| + \sum_{p \neq r, q \neq r} (c_p - \bar{x})(c_q - \bar{x}) |S_{p,q}| \\ &= (c_r - \bar{x})^2 kN + 2 \sum_{p \neq r} \left[(c_p - \bar{x})(c_r - \bar{x}) - (c_r - \bar{x})^2 \right] |S_{p,r}| \\ &\quad + \sum_{p \neq r, q \neq r} \left[(c_p - \bar{x})(c_q - \bar{x}) - (c_r - \bar{x})^2 \right] |S_{p,q}|\end{aligned}$$

Use Equation 3,

$$\begin{aligned}\bar{I} &= (c_r - \bar{x})^2 kN + 2 \sum_{p \neq r} \left[(c_p - \bar{x})(c_r - \bar{x}) - (c_r - \bar{x})^2 \right] \left(kn_p - \sum_{q \neq r} |S_{p,q}| \right) \\ &\quad + \sum_{p \neq r, q \neq r} \left[(c_p - \bar{x})(c_q - \bar{x}) - (c_r - \bar{x})^2 \right] |S_{p,q}|\end{aligned}$$

Finally, merge like terms with respect to $|S_{p,q}|$,

$$\begin{aligned}\bar{I} &= (c_r - \bar{x})^2 kN + 2 \sum_{p \neq r} \left[(c_p - \bar{x})(c_r - \bar{x}) - (c_r - \bar{x})^2 \right] kn_p \\ &\quad - 2 \sum_{p \neq r} \left[(c_p - \bar{x})(c_r - \bar{x}) - (c_r - \bar{x})^2 \right] \sum_{q \neq r} |S_{p,q}| \\ &\quad + \sum_{p \neq r, q \neq r} \left[(c_p - \bar{x})(c_q - \bar{x}) - (c_r - \bar{x})^2 \right] |S_{p,q}| \\ &= (c_r - \bar{x})^2 kN + 2 \sum_{p \neq r} \left[(c_p - \bar{x})(c_r - \bar{x}) - (c_r - \bar{x})^2 \right] kn_p \\ &\quad - 2 \sum_{p \neq r, q \neq r} \left[(c_p - \bar{x})(c_r - \bar{x}) - (c_r - \bar{x})^2 \right] |S_{p,q}| \\ &\quad + \sum_{p \neq r, q \neq r} \left[(c_p - \bar{x})(c_q - \bar{x}) - (c_r - \bar{x})^2 \right] |S_{p,q}| \\ &= (c_r - \bar{x})^2 kN + 2 \sum_{p \neq r} \left[(c_p - \bar{x})(c_r - \bar{x}) - (c_r - \bar{x})^2 \right] kn_p \\ &\quad + \sum_{p \neq r, q \neq r} \left[(c_p - \bar{x})(c_q - \bar{x}) - 2(c_p - \bar{x})(c_r - \bar{x}) + (c_r - \bar{x})^2 \right] |S_{p,q}| \\ &= Q + \sum_{p \neq r, q \neq r} \left[(c_p - \bar{x})(c_q - \bar{x}) - 2(c_p - \bar{x})(c_r - \bar{x}) + (c_r - \bar{x})^2 \right] |S_{p,q}|\end{aligned}$$

◁

As Q remains fixed given T_M and $|S_{p,q}|, p \neq r, q \neq r$ are approximately independent, the following relation holds by linearity:

$$\begin{aligned}
\mathrm{Var}(\bar{I}) &= \mathrm{Var}(Q + \sum_{p\neq r, q\neq r} \left[(c_p-\bar{x})(c_q-\bar{x}) - 2(c_p-\bar{x})(c_r-\bar{x}) + (c_r-\bar{x})^2\right] |S_{p,q}|) \\
&= \mathrm{Var}(\sum_{p\neq r, q\neq r} \left[(c_p-\bar{x})(c_q-\bar{x}) - 2(c_p-\bar{x})(c_r-\bar{x}) + (c_r-\bar{x})^2\right] |S_{p,q}|) \\
&\approx \sum_{p\neq r, q\neq r} \mathrm{Var}(\left[(c_p-\bar{x})(c_q-\bar{x}) - 2(c_p-\bar{x})(c_r-\bar{x}) + (c_r-\bar{x})^2\right] |S_{p,q}|) \\
&= \sum_{p\neq r, q\neq r} \left[(c_p-\bar{x})(c_q-\bar{x}) - 2(c_p-\bar{x})(c_r-\bar{x}) + (c_r-\bar{x})^2\right]^2 \mathrm{Var}(|S_{p,q}|) \\
&= \sum_{p\neq q\neq r} \left[(c_p-\bar{x})(c_q-\bar{x}) - 2(c_p-\bar{x})(c_r-\bar{x}) + (c_r-\bar{x})^2\right]^2 \mathrm{Var}(|S_{p,q}|) \\
&\quad + \sum_{p\neq r} \left[(c_p-\bar{x})^2 - 2(c_p-\bar{x})(c_r-\bar{x}) + (c_r-\bar{x})^2\right]^2 \mathrm{Var}(|S_{p,p}|) \\
&= \sum_{p\neq q\neq r} \left[(c_p-\bar{x})(c_q-\bar{x}) - 2(c_p-\bar{x})(c_r-\bar{x}) + (c_r-\bar{x})^2\right]^2 \mathrm{Var}(|S_{p,q}|) \\
&\quad + \sum_{p\neq r} \left[(c_p-c_r)^2\right]^2 \mathrm{Var}(|S_{p,p}|)
\end{aligned}$$

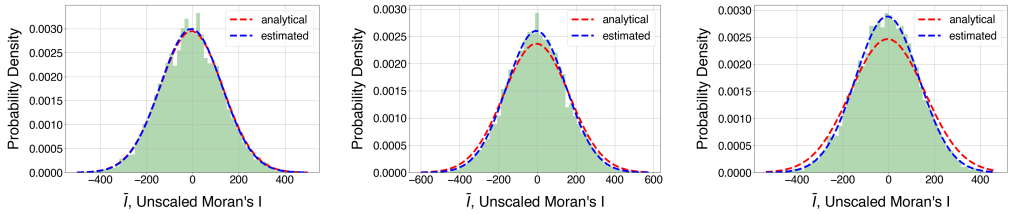
Insert the formulae of $\mathrm{Var}(|S_{p,q}|)$ and $\mathrm{Var}(|S_{p,p}|)$ from Lemma 6 and Lemma 7 and Theorem 9 is proved. To best satisfy the condition of approximate independence, the background value should have the largest value size. ◀

4 Analysis of Approximation Accuracy and Robustness on Synthetic Data

For the sake of mathematical preciseness, we have introduced many assumptions and conditions during the problem setup and the proof. It is necessary to systematically investigate how violations of these assumptions and conditions may affect the accuracy of approximation. In the following, by using a series of experiments on synthetic data, we demonstrate that most of them can be relaxed while our approximation remains sufficiently accurate. This implies that the spatial self-information J is numerically robust for practical use.

As a quick recap, the assumptions and conditions used in our theoretical derivations are listed below in the order they are introduced:

- **Assumption 1**: the spatial weights are binary (0-1). See Section 3.1.
- **Assumption 2**: the observations take a small number of discrete values. See Section 3.1.
- **Assumption 3**: the observations have equal numbers of neighbors. See Section 3.1.
- **Assumption 4**: no pairs of same-colored vertices share common neighbors other than themselves. See Section 3.3.
- **Condition 1**: for deriving the asymptotic distributions of $|S_{p,q}|$ and $|S_{p,p}|$, it is required that $n_q \ll kn_p \ll N$. See Section 3.3.
- **Condition 2**: n_p is required to be sufficiently large to have good normal approximations of the asymptotic binomial and Poisson binomial distributions. See Section 3.4.
- **Condition 3**: $n_{r_{\max}}/N$ is required to be sufficiently large to ensure approximate independence of normal random variables. See Section 3.5.



(a) high independence, $b = 65\%$.
$\tilde{\mu}_{\bar{I}} = -5.21$, $\tilde{\sigma}_{\bar{I}} = 135.11$
$\hat{\mu}_{\bar{I}} = -4.51$, $\hat{\sigma}_{\bar{I}} = 132.84$

(b) mid independence, $b = 45\%$.
$\tilde{\mu}_{\bar{I}} = -4.45$, $\tilde{\sigma}_{\bar{I}} = 168.46$
$\hat{\mu}_{\bar{I}} = -5.56$, $\hat{\sigma}_{\bar{I}} = 153.99$

(c) low independence, $b = 25\%$.
$\tilde{\mu}_{\bar{I}} = -4.39$, $\tilde{\sigma}_{\bar{I}} = 161.66$
$\hat{\mu}_{\bar{I}} = -3.94$, $\hat{\sigma}_{\bar{I}} = 139.85$

Figure 1 Histograms of $\bar{I}$ of 10,000 randomly generated 40×40 grids using rook's distance. From (a) to (c) the proportion of background b decreases, i.e., the level of independence decreases. The blue lines represent the estimated normal distributions from the histograms of $\bar{I}$. The red lines represent the analytical approximations based on Theorem 8 and 9.

Among all these terms, only Assumption 1 is mandatory. We will analyze how and to what extent the other assumptions and conditions can be relaxed in the following subsections, starting from the most critical ones.

The statistics used to measure the accuracy of approximation are: the analytical mean $\tilde{\mu}_{\bar{I}}$, the analytical standard deviation $\tilde{\sigma}_{\bar{I}}$, the sample mean $\hat{\mu}_{\bar{I}}$, the sample standard deviation $\hat{\sigma}_{\bar{I}}$, the standardized difference of mean $|\tilde{\mu}_{\bar{I}} - \hat{\mu}_{\bar{I}}|/\tilde{\sigma}_{\bar{I}}$, the standardized difference of standard deviation $|\tilde{\sigma}_{\bar{I}} - \hat{\sigma}_{\bar{I}}|/\hat{\sigma}_{\bar{I}}$, and the KL-divergence from the analytical normal distribution to the empirical distribution. For all experiments, we randomly sample 10,000 40×40 grids for 10 times and report both the means and the standard deviations of the statistics.

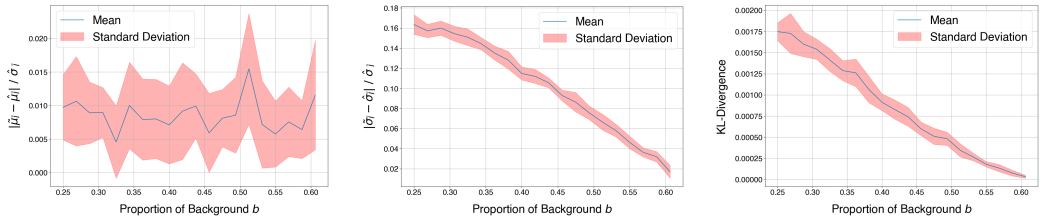
4.1 Relaxation of Condition 3: Violation of Approximate Independence

The approximation accuracy is mostly dependent on the satisfaction of the independence condition in Theorem 9. The level of independence is measured by $b = n_{r_{\max}}/N$, the proportion of background. The higher b, the higher independence.

Figure 1 demonstrates that when the independence condition is satisfied, the analytical normal approximation perfectly fits the empirical distribution. More specifically, in the high independence case, the analytical approximation passes the Kolmogorov-Smironov Test with a p-value of 0.20, which is significantly larger than 0.05. That means our approximation is statistically indistinguishable from the actual distribution. As the proportion of background decreases, the analytical standard deviation becomes increasingly overestimated, due to the violation of the independence condition. However, Figure 1 and Figure 2 also show, though the approximate variance becomes inaccurate, the approximate mean remains extremely accurate: the absolute difference between the analytical mean and the empirical mean is consistently below 2% of the empirical standard deviation regardless of the level of independence. This is very useful for designing loss functions: given the mean of a normal distribution, even though we do not have its exact variance, the **relative** difference between any two values is known, i.e., we can still accurately compute the direction of the gradients.

4.2 Relaxation of Assumption 3: Different Numbers of Neighbors

It is common that not all observations have the same number of neighbors. For example, for grid data with rook's weights, the border observations only have 3 neighbors and corner observations only 2, instead of 4. There are two cases of violation: (1) the total number of neighbors remains kN, but not equally distributed, or (2) the total number of neighbors is larger or smaller than kN.

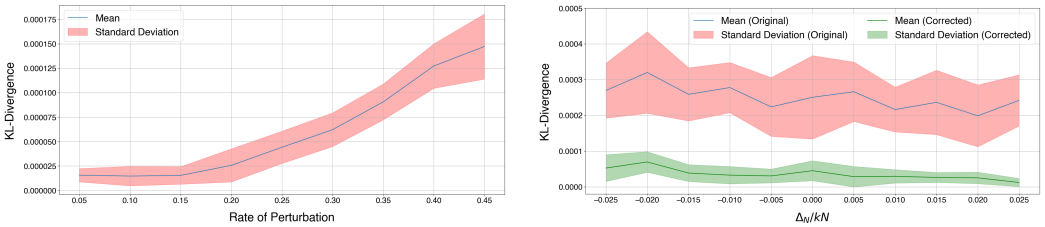


(a) standardized difference of the mean of $\bar{I}$.

(b) standardized difference of the standard deviation of $\bar{I}$.

(c) KL-divergence to the empirical distribution.

Figure 2 The relation between the approximation accuracy and the level of independence (measured by $b = n_{r_{\max}}/N$, the proportion of background values). At each level of independence, we repeatedly sample 10,000 40×40 grids randomly for 10 times. (a) and (b) plot the standardized difference between the analytical mean/standard deviation and the empirical mean/standard deviation. (c) plots the KL divergence from the analytical approximation to the empirical distribution.



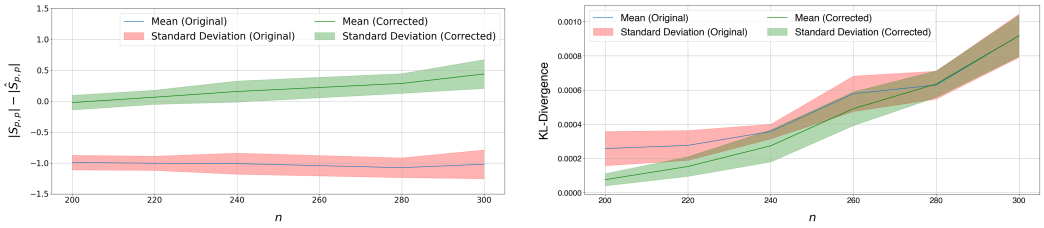
(a) perturbation of weight matrix.

(b) change of total number of neighbors.

Figure 3 (a) The relation between the approximation accuracy and the level of perturbation in the number of neighbors, measured by the rate of perturbation. (b) The relation between the approximation accuracy and the level of change in total number of neighbors, measured by the change rate Δ_N/kN.

The first case can be seen as a perturbation of the weight matrix. To investigate this, given a rate of perturbation ρ, we randomly select ρkN 1s and ρkN 0s in the weight matrix and flip their values (0 to 1 and vice versa). This is equivalent to randomly deleting and adding equal amount of edges in the graph. After such perturbation, while the total number of neighbors remains the same, the equal-number-of-neighbor assumption is violated. We generate a random sample of grids, perturb the weight matrix at increasing rates of perturbation, derive the analytical approximation according to the perturbed weight matrices, and compute the KL-divergence respectively. In Figure 3a we plot how the approximation accuracy changes as the rate of perturbation increases. When the rate of perturbation is under 0.15, the KL-divergence remains under 2.5×10^{-5}, which is comparable to the high-independence KL-divergence values in Figure 2c. It indicates that a mild violation of the equal-number-of-neighbor assumption shall not introduce significant approximation error.

In the second case, the violation leads to systematic overestimation or underestimation of $|S_{p,q}|$. Let the difference between the total number of neighbors and kN be Δ_N, each $|S_{p,q}|$ can be corrected by multiplying the scaling factor $1 - \Delta_N/kN$. To verify this, we randomly select Δ_N 1s/0s from the weight matrix and flip their values, which increases/decreases the total number of neighbors by Δ_N. In Figure 3b we plot how the approximation accuracy changes with Δ_N when applying and not applying the scaling factor. We can see the corrected approximation has consistently low KL-divergence ($< 1 \times 10^{-4}$). It means that we can safely use the approximation with spatial weights that have edge cases (like the borders and corners in the rook/queen weights).



(a) difference from the empirical $|\hat{S}_{p,p}|$.

(b) KL-divergence to the empirical distribution.

Figure 4 (a) The uncorrected analytical $|S_{p,p}|$ is constantly underestimated, while the corrected analytical $|S_{p,p}|$ approximates the empirical $|\hat{S}_{p,p}|$ better. (b) The relation between the KL-divergence and n_p. The larger the n_p, the more common neighbors, and the worse approximation accuracy.

4.3 Relaxation of Assumption 4: Common Neighbors

The assumption in Lemma 7 that none of the colored vertices share common neighbors is used to derive the probability of success. When this assumption holds, each success creates exactly 2 same-value pairs and $k-2$ candidate vertices which we can color to obtain another 2 same-value pairs. Intuitively, it assumes that colored vertices do not form large clusters. When n_p is small, the assumption is valid because the colored vertices are scattered. However, when n_p is large, $|S_{p,p}|$ will be underestimated because coloring a common neighbor creates more than 2 same-value pairs.

We find that multiplying each $|S_{p,p}|$ with a scaling factor $(|S_{p,p}|-1)/|S_{p,p}|$ corrects the underestimation. The intuition is that, as $|S_{p,p}|$ gets large, the violation of having common neighbors has a decreasing effect on the approximate probability of success. To verify this, we generate 40×40 grids with rook's weights that have one background value and three foreground values. Set the value sizes of all the foreground values to be the same number n from 200 to 300 by a step of 20, and investigate how the analytical $|S_{p,p}|$ differs from the empirical $|\hat{S}_{p,p}|$ as n increases. Figure 4a demonstrates that the corrected analytical $|S_{p,p}|$ is an extremely accurate approximation of the empirical $|\hat{S}_{p,p}|$, and Figure 4b demonstrates that such improvement results in better overall KL-divergence from the approximation to the empirical distribution. The up-going trend in Figure 4b is irrelevant to the assumption of no common neighbors. Instead, it is a result of the violation of the independence condition as increasing n means a smaller proportion of background, which makes the approximation of standard deviations less accurate.

4.4 Trivial Relaxations

Some of the assumptions and conditions are trivial to satisfy. We summarize them as follows:

Relaxation of Assumption 2. Data with continuous values and excessively many discrete values can be approximated by discretization (i.e., bucketization or binning). The approximation accuracy is dependent on the granularity of the buckets.

Relaxation of Condition 1. This condition is automatically satisfied when the independence condition is satisfied and k is not extremely large. In Lemma 5 we use the minimum/maximum functions. That guarantees $n_q < n_p$. Then, as k is usually a sufficiently large integer (e.g., the number of nearest neighbors), $n_q \ll kn_p$. Besides, when the independence condition is satisfied, $n_{r_{\max}}/N$ is sufficiently large, i.e. $N - n_{r_{\max}} \ll N$, thus $n_p < N - n_{r_{\max}} \ll N$. Then, as long as k is not excessively large, $kn_p \ll N$.

Relaxation of Condition 2. Violation of this condition does not significantly affect the approximation accuracy because when n_p is small, its total contribution to the analytical mean and variance is also small.

5 Applications on Real-World Data

To further demonstrate the potential for practical application of our proposed method, we demonstrate how it can be applied to measure the surprisal of slope data via spatial self-information J.

The slope dataset[2] used here is obtained from the European Union statistics organization and covers the area of EU. The values in the data show slope values (i.e., 0 deg - 90 deg) that are normalized into the range of 0 - 250. We split the original data into tiles of size 1000×1000, which represents a relatively homogeneous region, and further split the tiles into 50×50 patches. Due to the relatively small size of the patches, we bucketize the slope values with bin size 20 to avoid values that only appear very few times, i.e, merging values from 0 to 19 as 1, values from 20 to 39 as 2, and so forth until values from 240 to 250 as 13. On the bucketized patches we compute the Moran's I and the spatial self-information J.

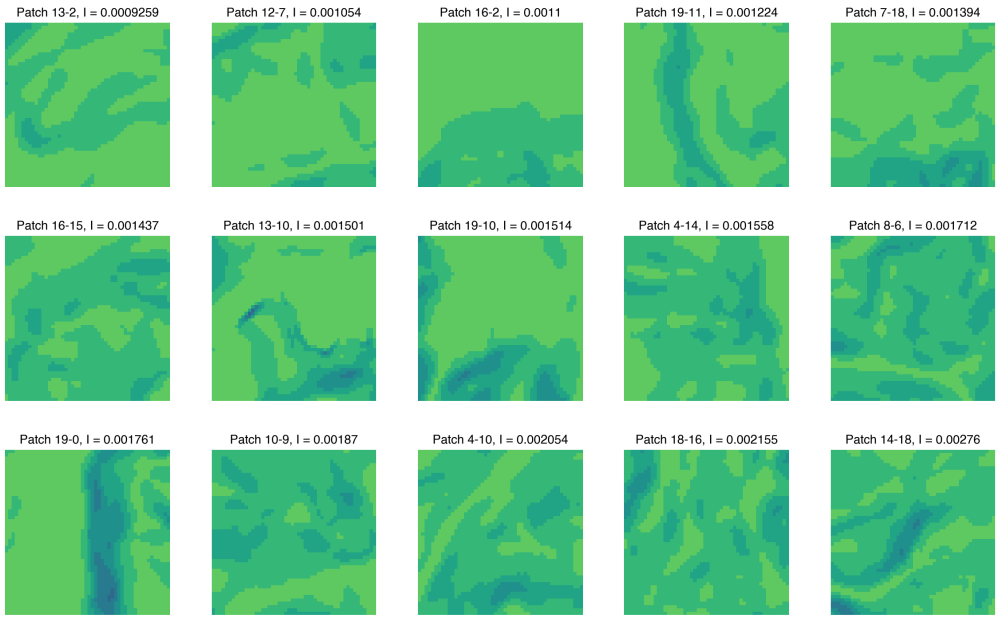


Figure 5 Slope patches in ascending order of Moran's I from left to right and top to bottom.

As we have pointed out, Moran's I statistics with different value schemes are not directly comparable. We only know that statistically, a Moran's I value that is significantly above $1/(N-1)$ or below $-1/(N-1)$ indicates a presence of spatial autocorrelation, which in our case equals $1/2500 = 0.0004$. Instead, the spatial self-information provides a unified measure of spatial autocorrelation from the perspective of information content, regardless of the value schemes as long as the independence assumption holds. To demonstrate this, we randomly select 15 patches whose proportion of background ranges from 0.55 to 0.65, which we know

[2] https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/elevation/eu-dem/slope

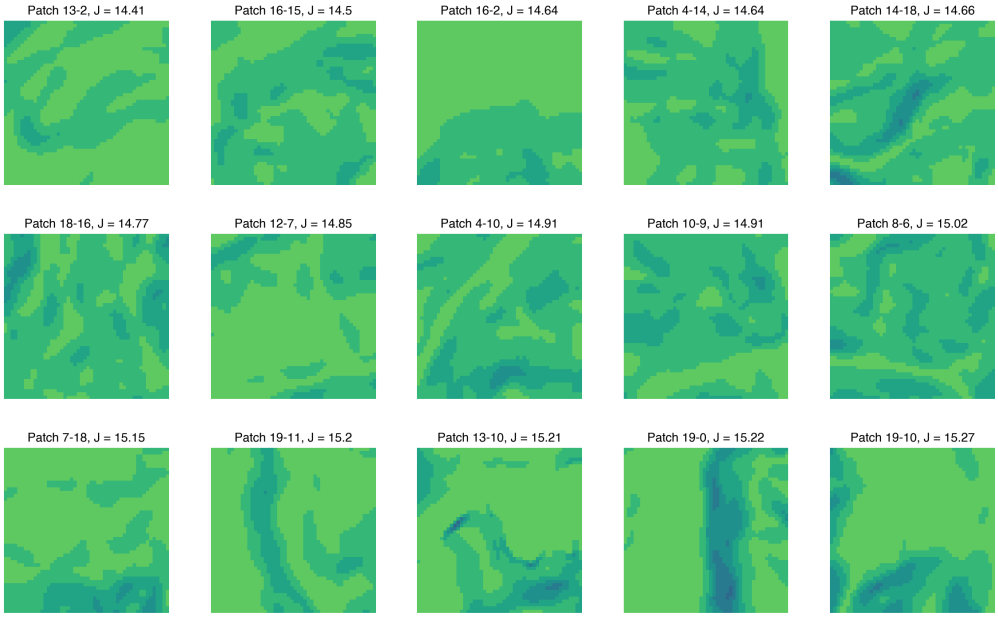


Figure 6 Slope patches in ascending order of surprisal from left to right and top to bottom.

from Figure 1a guarantees high independence. We do not require their value schemes to be identical (which is impossible). In Figure 5, the patches are plotted in ascending order of Moran's I. And in Figure 6, the same patches are plotted, but in ascending order of spatial self-information J. We can see the two orders are not the same. For example, Patch 19-0 shows a sharp spatial pattern where a strip of mountains stands against a wide plane. In Figure 5, however, it has lower positive Moran's I than Patch 10-9, which is more scattered and should have weaker spatial autocorrelation. This phenomenon occurs because Patch 10-9 has larger areas of deeper green (higher values), which inflates the variance of Moran's I with its value scheme. Our spatial self-information J, on the contrary, is able to correctly capture the strength of spatial autocorrelation. In Figure 6, Patch 19-0 has the second largest spatial self-information (15.22) compared to Patch 10-9 (14.91), meaning it is $e^{15.22-14.9} = 1.36$ times easier to observe Patch 10-9 than Patch 19-0, i.e., Patch 19-0 is more surprising.

6 Conclusions and Future Work

In this paper, we theoretically derive the asymptotic analytical distribution of global Moran's I in the case of binary weights, under a series of broad randomness assumptions. We further develop a comprehensive set of techniques that efficiently computes the approximate probability and self-information of a spatial sample and corrects the error caused by violations of assumptions. Both synthetic and real-world experiments show that our approximation remains accurate and robust even if the assumptions and conditions are not ideally satisfied. Our research provides practical means to measure the information loss in spatially distributed data due to the presence of spatial autocorrelation with applications in spatial data analysis, GeoAI models, and general machine learning/deep learning. For future work, it is worth exploring to **(1)** relax the independence assumption to enable more accurate approximation on highly scattered spatial data such as maps of POIs, **(2)** derive a non-binary weight version, and **(3)** study different settings such as continuous value surfaces and continuous entropy.

Finally, while our work was centered around Moran's I, similar ideas likely generalize to related concepts such as the semivariogram which could be expressed as increased entropy by distance.

References

1 Michael Batty. Spatial Entropy. *Geographical Analysis*, 6(1):1–31, 1974. doi:10.1111/j.1538-4632.1974.tb01014.x.

2 François Chapeau-Blondeau. Autocorrelation versus entropy-based autoinformation for measuring dependence in random signal. *Physica A: Statistical Mechanics and its Applications*, 380:1–18, 2007.

3 AD Cliff and JK Ord. Model building and the analysis of spatial pattern in human geography. *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(3):297–328, 1975.

4 Elijah Cole, Grant Van Horn, Christian Lange, Alexander Shepard, Patrick Leary, Pietro Perona, Scott Loarie, and Oisin Mac Aodha. Spatial implicit neural representations for global-scale species mapping. In *International conference on machine learning*, pages 6320–6342. PMLR, 2023.

5 Samuel Cushman. Calculation of Configurational Entropy in Complex Landscapes. *Entropy*, 20(4):298, April 2018. doi:10.3390/e20040298.

6 A Stewart Fotheringham and David WS Wong. The modifiable areal unit problem in multivariate statistical analysis. *Environment and planning A*, 23(7):1025–1044, 1991.

7 Robert C Geary. The contiguity ratio and statistical mapping. *The incorporated statistician*, 5(3):115–146, 1954.

8 Arthur Getis. *Spatial Autocorrelation*, pages 255–278. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. doi:10.1007/978-3-642-03647-7_14.

9 Michael F Goodchild. *Spatial Autocorrelation*, volume 47 of *Concepts and Techniques in Modern Geography*. Geo Books, 1986.

10 Harry H. Kelejian and Ingmar R. Prucha. On the asymptotic distribution of the moran i test statistic with applications. *Journal of Econometrics*, 104(2):219–257, 2001. doi:10.1016/S0304-4076(01)00064-1.

11 Yili Hong. On computing the distribution function for the poisson binomial distribution. *Computational Statistics & Data Analysis*, 59:41–51, March 2013. doi:10.1016/j.csda.2012.10.006.

12 Anders Karlström and Vania Ceccato. A new information theoretical measure of global and local spatial association. MPRA Paper 6848, University Library of Munich, Germany, August 2000. URL: https://ideas.repec.org/p/pra/mprapa/6848.html.

13 Didier G Leibovici. Defining spatial entropy from multivariate distributions of co-occurrences. In *Spatial Information Theory: 9th International Conference, COSIT 2009 Aber Wrac'h, France, September 21-25, 2009 Proceedings 9*, pages 392–404. Springer, 2009.

14 Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9596–9606, 2019.

15 Gengchen Mai, Krzysztof Janowicz, Yingjie Hu, Song Gao, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. A review of location encoding for geoai: methods and applications. *International Journal of Geographical Information Science*, 36(4):639–673, 2022.

16 Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. Multi-scale representation learning for spatial feature distributions using grid cells. In *International Conference on Learning Representations*, 2020.

17 Gengchen Mai, Yao Xuan, Wenyun Zuo, Yutong He, Jiaming Song, Stefano Ermon, Krzysztof Janowicz, and Ni Lao. Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:439–462, 2023.

18 Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

19 Alistair Moffat. Huffman coding. *ACM Computing Surveys (CSUR)*, 52(4):1–35, 2019.

20 P. A. P. Moran. Notes on Continuous Stochastic Phenomena. *Biometrika*, 37(1/2):17–23, 1950. doi:10.2307/2332142.

21 Jakub Nowosad and Tomasz F Stepinski. Information theory as a consistent framework for quantification and classification of landscape patterns. *Landscape Ecology*, 34:2091–2101, 2019.

22 S Papoulis. *Probability, Random Variables and Stochastic Processes by Athanasios*. Boston: McGraw-Hill, 2002.

23 Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. Geographic location encoding with spherical harmonics and sinusoidal representation networks. In *The Twelfth International Conference on Learning Representations*, 2023.

24 Waldo Tobler. On the first law of geography: A reply. *Annals of the Association of American Geographers*, 94(2):304–310, 2004.

25 Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

26 Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36, 2024.

27 Zhangyu Wang, Krzysztof Janowicz, Gengchen Mai, and Ivan Majic. Probing the information theoretical roots of spatial dependence measures, 2024. arXiv:2405.18459.

28 Neil Wrigley. Spatial processes: models and applications. *The Geographical Journal*, 148(3):383–385, 1982. URL: http://www.jstor.org/stable/633177.