Scaling geospatial data from the perspective of complexity: Exploring the scaling behaviour of the entropogram

Wen-Bin Zhang[a,b,c], Yong Ge[a], Shengjie Lai[b], Peter. M. Atkinson[c]

[a]*State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China*

[b]*WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton SO17 1BJ, UK*

[c]*Lancaster Environment Center, Faculty of Science and Technology, Lancaster University, Lancaster LA1 4YW, UK*

A fundamental challenge in geospatial data science is to determine how a property, or its characterisation, changes with a change in the scale of measurement. However, except for geostatistical regularization of the variogram, which is theoretically well established, the scaling behaviour of a wide range of alternative measures of spatial association remains unclear. This limits the ability to make inferences at scales beyond the scale of measurement. The scaling behaviour of the recently introduced entropogram function also remains unclear. However, since the entropogram is essentially the generalisation of the variogram to categorical spatial variables, the possibility to derive a scaling model for the entropogram exists. Here, the scaling behaviour of the entropogram based on the scale effect of Shannon entropy is derived, providing a theoretical basis for the regularization of the entropogram. To validate the developed regularization model for the entropogram, a series of multiscale data was generated. Both theoretical derivation and experimental results showed that the entropogram is scale-invariant, under certain conditions for the generation of the categorical data. This research, thus, generalises the entropogram to changes in measurement scale and, thereby, increases our ability to characterise spatial data and make inferences about the underlying dynamic process. It also provides a reference for the interactions between patterns and processes at different scales.

Keywords: Entropogram; geospatial data; scale effect; spatial association

1. Introduction

Geospatial data are commonly measurements or representations of spatially varying properties, and can be conceptualised as realisations of geographical processes with intrinsic spatial association properties. Measurements are inevitably associated with a measurement scale, referred to as the spatial support, which filters the possible information about the real world. Here, the support refers to the space on which an observation is defined and measured, with a given size, geometry and orientation (Matheron, 1963). For example, gridded population data represent population counts on grid lattices, where each grid cell represents the support of each population count. Given that measurements are associated with a support of give size (e.g., volume, area, length), the spatial variation in geographical data is generally scale-dependent (Goodchild, 2011). For geographical studies, the effect of the measurement scales of both input data and model (i.e., an approximate characterisation of the geographical process of interest operating at specific scales) will consequently affect the relevance and credibility of results derived from the model output. Therefore, once measurement has occurred, to characterise the information of interest it may be necessary to generalise the information (i.e., spatial association) from the fixed observation scale to other scales.

The most common solution to generalise information to other scales is to transfer the measurements or observations to other measurement supports. With the popularity of machine learning, artificial neural networks have been used widely to change the spatial support of Earth science data in recent years (Jia et al., 2019; Zhang et al., 2021; Sdraka et al., 2022). However, whilst interesting and with high accuracy in general, these approaches are limited because they require direct measurement at the scale of interest, either for the covariates or the target variable. In turn, it is not surprising that attention is currently being directed at ways to improve the quality of training data (Atkinson, 2013), which express our ability to represent the spatial character of the property of interest on the target support. Under this circumstance, the information at the scale of interest, generalised from the measurement scale, is *de facto* information incorporated from the additional data source. Despite incorporating more information providing an effective way to change the support of geospatial data in practice (Pu and Bonafoni, 2023), our understanding of the scale effect, and change of scale, is not necessarily increased. Therefore, there exists a more general desire to understand the effects of the support on data acquired through measurement (Ge et al., 2019). As the support is a filter on reality for all knowledge acquired through measurement, it acts as a fundamental limit on what can be known about the real world, and as such it deserves significant attention. To capture the spatial association information at measurement scales that have not yet been observed, the scaling behaviours of a particular characterization of spatial association should be modelled, which requires an understanding of how spatial association varies across scales.

Currently, there are a basket of measures or descriptors of spatial association, including Moran's *I*, Geary's *c* and the variogram. The values of these measures will vary with a change of scale (or support) (Wiens, 1989; Wu, 2004). However, the mechanisms of the relevant scale transformations are still unclear, except for the variogram. The variogram in geostatistics may be the most well-known tool to describe spatial variation as a function of lag (a vector in distance and direction). It is defined as the (semi-)variance of the difference between variables at two locations separated by different lags. In contrast to measures such as Moran's *I*, the scaling behaviour of the modelled variogram with a change of support has a solid theoretical foundation, referred to in geostatistics as regularization (Clark, 1977). Taking the average semi-variance of point pairs covered by the two supports as the semi-variance between measurements over two supports defines *de facto* the scaling behaviour of the variogram (Jupp et al., 1988).

As a generalisation of the variogram to cover categorical variables, the entropogram was recently developed by Zhang et al. (2023a) to describe the spatial association of geographical variables from the perspective of complexity. However, the scaling behaviour of the entropogram has not yet been defined. In this article, we investigate, for the first time, the effect of the support on the entropogram theoretically (i.e., regularization of the entropogram). Then, both numerical simulations and real-world experiments are conducted to validate the derived scaling behaviours of the entropogram. Finally, some remarks and conclusions are provided.

2. Scaling behaviours of spatial association

2.1 Variogram and entropogram

Geographical elements, for example, properties of the soil and land surface, such as soil moisture and land cover types, are either quantitatively (i.e., referring to continua) or qualitatively (i.e., referring to categories) measured to create geographical variables, representing different locations (Ge et al., 2019). Geospatial data, then, are generally a collection of measurements (states) for a quantitative (qualitative) geographical variable distributed across space. The observed measurements over space can be considered as a realisation of a second-order stationary random function (RF), where each location is characterised by a random variable (RV) that describes the behaviour of the studied geographical variable at that location. The collective behaviours of the RVs over space approximately substitute for an unknown underlying spatial, dynamic process that generates the property of interest.

As a defining property of the RF, spatial association plays a crucial role in increasing our understanding of the corresponding geographical dynamic process and, consequently, in making predictions about the geographical variable of interest. Specifically, the second-order stationary RF Z consists of RVs $Z(\mathbf{s})$ at each location $\mathbf{s}$, where each RV $Z(\mathbf{s})$ takes a value $z(\mathbf{s})$ as its realisation. The variogram is the most well-known approach to describe the spatial association of the RF,

$$\gamma(\mathbf{h}) = \gamma(Z(\mathbf{s}), Z(\mathbf{s}-\mathbf{h})) = \frac{1}{2}\mathrm{E}\left[\left(Z(\mathbf{s}) - Z(\mathbf{s}-\mathbf{h})\right)^2\right], \quad (1)$$

However, the variogram cannot be applied to qualitative (i.e., specifically categorical) geospatial data directly. The variogram is built on the difference between measurements, which is undefined for the multiple states of categorical geographical variables. For example, the feature space distance between land cover types (e.g., grass and water) is not well defined. Under this circumstance, the entropogram is proposed to describe the spatial association of qualitative, categorical geographical variables based on the frequency domain of the RVs of the RF. Specifically, in analogy to the variogram, the entropogram is built on the mutual information between two locations at lag $\mathbf{h}$ apart,

$$\begin{aligned}\tau(\mathbf{h}) &= H(Z(\mathbf{s})) + H(Z(\mathbf{s}-\mathbf{h})) - H(Z(\mathbf{s}), Z(\mathbf{s}-\mathbf{h})) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} p\big(Z(\mathbf{s}) = x_i, Z(\mathbf{s}-\mathbf{h}) = x_j\big) \ln\left(\frac{p\big(Z(\mathbf{s})=x_i, Z(\mathbf{s}-\mathbf{h})=x_j\big)}{p(Z(\mathbf{s})=x_i)\,p\big(Z(\mathbf{s}-\mathbf{h})=x_j\big)}\right), \quad (2)\end{aligned}$$

where $p\big(Z(\mathbf{s}) = x_i, Z(\mathbf{s}-\mathbf{h}) = x_j\big)$ is the probability mass of the *pair* of locations $\mathbf{s}$ and $\mathbf{s}-\mathbf{h}$ being state x_i and x_j, and $p(Z(\mathbf{s}) = x_i)$ and $p\big(Z(\mathbf{s}-\mathbf{h}) = x_j\big)$ are the probability mass at locations $\mathbf{s}$ and $\mathbf{s}-\mathbf{h}$, respectively, being states x_i and x_j. Note that x_i, $i = 1, \ldots, n$, represents the states shared by the categorical RVs. For simplicity, $p(Z(\mathbf{s}) = x_i)$ is noted as $p(x_i)$ in the remaining part. Despite the entropogram being conceived as the counterpart of the variogram for qualitative spatial variables, the entropogram can also be applied to quantitative spatial variables by discretizing the continuous values into qualitatively different states.

The entropogram captures different information compared to the variogram. Figure 1 gives details of the entropogram as well as its difference to the indicator variogram (the extension of the variogram to the binary components of categorical spatial data). We first generated a continuous Gaussian RF with spatial covariance function $C(h) = \exp(-h/18)$. Then, the continuous map was transferred into a 2-state categorical RF, where state 1 represents values smaller than the mean value of the continuous map and state 2 the remainder. The spatial association of the 2-state categorical RF was modelled by the entropogram and the indicator variogram, respectively (see Figure 1(c)). The indicator variograms for both states are the same after their transformation into binary data (0 represents absence and 1 represents presence of the state). The entropogram depicts the complexity of the co-occurrence for pairs of locations regarding the whole state space over all spatial lags, whilst the variogram describes the spatial dissimilarity for each state over the spatial lags. Figure 1(d-e) shows that the conditional probabilities for both states are around 0.5 through lags 5 to 10, representing a more complex joint distribution of the states for pairs of locations compared to other lags. The entropogram in Figure 1(c) reflects this situation appropriately by producing smaller values at lags from 5 to 10. Then, for the larger spatial lags >10, the conditional probabilities of one state base on another state are increased. Despite the dissimilarity increasing under this circumstance, the correlation between states is the same as for smaller spatial lags <5. It is apparent from Figure 1(c) that the variogram fails to depict the correlation between states; while the entropogram fills the gap and successfully characterizes the correlation intensity across spatial lags.

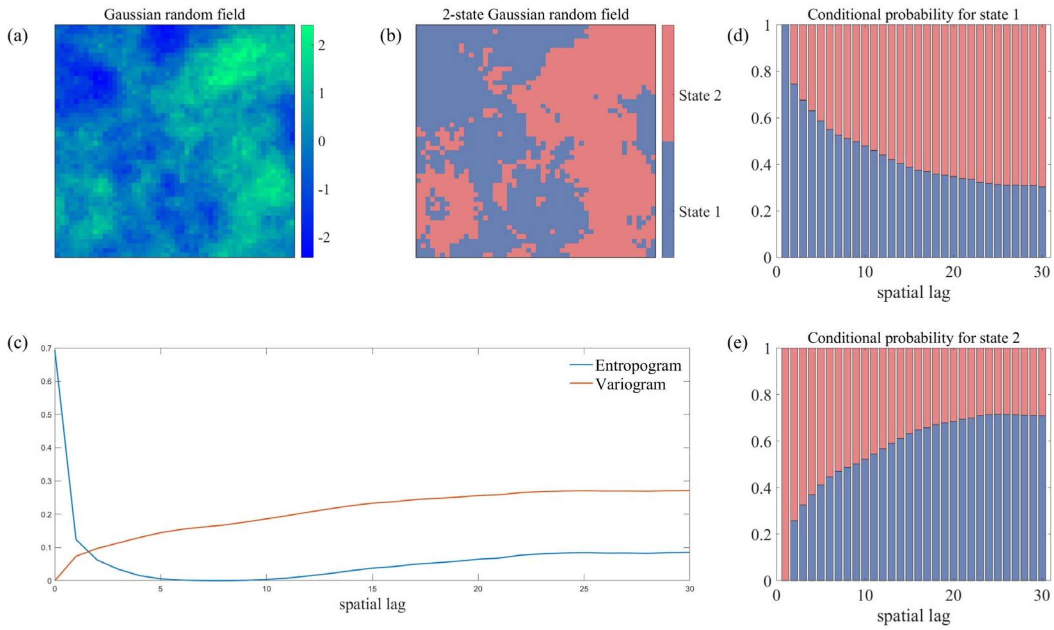

Figure 1. (a) Continuous Gaussian Random Field. (b) Categorical Random Field with two states obtained from (a). State 1 represents values smaller than the mean, and state 2 represents values greater than or equal to the mean. (c) Comparison between the entropogram and the indicator variogram of (b). (d) The conditional probability for state 1 across spatial lags. (e) The conditional probability for state 2 across spatial lags.

2.2 Regularization

2.2.1 Variogram

Spatial structure is essential to interpreting the observed geographical data, and it is commonly represented by a spatial covariance function or variogram. However, when the data are measured on a different support, the observed spatial structure may be changed due to an altered interaction between the underlying spatial generating process and the measurement support. Measurement imposes a specific scale of measurement on the data (see Figure 2(a)), considering $\boldsymbol{z} = f(\mathbf{s}, \mathbf{R})$, where $\boldsymbol{z}$ is the data, $\mathbf{s}$ is the sampling framework and $\mathbf{R}$ is the underlying property defined on a given space (e.g., 2D, Euclidean). Taking spatial covariance functions as an example, the (auto)covariance between two points $\mathbf{s}_1$ and $\mathbf{s}_2$ is

$$\mathrm{Cov}(\mathbf{s}_1, \mathbf{s}_2) = E\big[\big(f(\mathbf{s}_1) - m(\mathbf{s}_1)\big)\big(f(\mathbf{s}_2) - m(\mathbf{s}_2)\big)\big], \qquad (3)$$

where $m(\mathbf{s}_1)$ and $m(\mathbf{s}_2)$ are the expectation of $f(\mathbf{s}_1)$ and $f(\mathbf{s}_2)$, respectively. When the support of the variable changes (e.g., from point to area, see Figure 2(b)), the covariance for the values associated on the areas $\mathbf{D}_1$ and $\mathbf{D}_2$ is

$$\mathrm{Cov}(\mathbf{D}_1, \mathbf{D}_2) = \frac{1}{L(\mathbf{D}_1)L(\mathbf{D}_2)} \int_{\mathbf{D}_1} \int_{\mathbf{D}_2} \mathrm{Cov}(\mathbf{s}_1, \mathbf{s}_2) d\mathbf{s}_1 d\mathbf{s}_2, \qquad (4)$$

where $L(*)$ is the Lebesque measure. In other words, the covariance between two areas $\mathbf{D}_1$ and $\mathbf{D}_2$ is considered as the average covariance between all point pairs within $\mathbf{D}_1$ and $\mathbf{D}_2$ (Jupp et al., 1988). Under the assumption of second-order stationarity, the spatial (auto)covariance between any two points $\mathbf{s}_1$ and $\mathbf{s}_2$ depends only on their coordinates lag separation vector $\mathbf{h} = \mathbf{s}_1 - \mathbf{s}_2$. It follows that

$$\mathrm{Cov}(\mathbf{D}_1, \mathbf{D}_2) = \frac{1}{L(\mathbf{D}_1)L(\mathbf{D}_2)} \int_{\boldsymbol{R}} I_{\mathbf{D}_1,\mathbf{D}_2}(\mathbf{h}) \mathrm{Cov}(\mathbf{h}) d\mathbf{h}, \qquad (5)$$

where $I_{\mathbf{D}_1,\mathbf{D}_2}(\mathbf{h}) = L(\mathbf{D}_1 \cap (\mathbf{D}_2 - \mathbf{h}))$ indicates the portion of point pairs having a lag of $\mathbf{h}$, i.e., the area of overlap between $\mathbf{D}_1$ and $\mathbf{D}_2 - \mathbf{h}$, where $\mathbf{D}_2 - \mathbf{h}$ (i.e., $\mathbf{D}_1$) represents an area $\mathbf{D}_2$ shifted by the vector $\mathbf{h}$. In this way, Eq. (5) calculates the weighted average of the point-support-based covariances of various lags $\mathbf{h}$ by the corresponding proportion of the point pairs. However, the associated probability mass distribution of the entropogram makes it infeasible to implement Eq. (5) for non-point support-based observations of categorical spatial data.

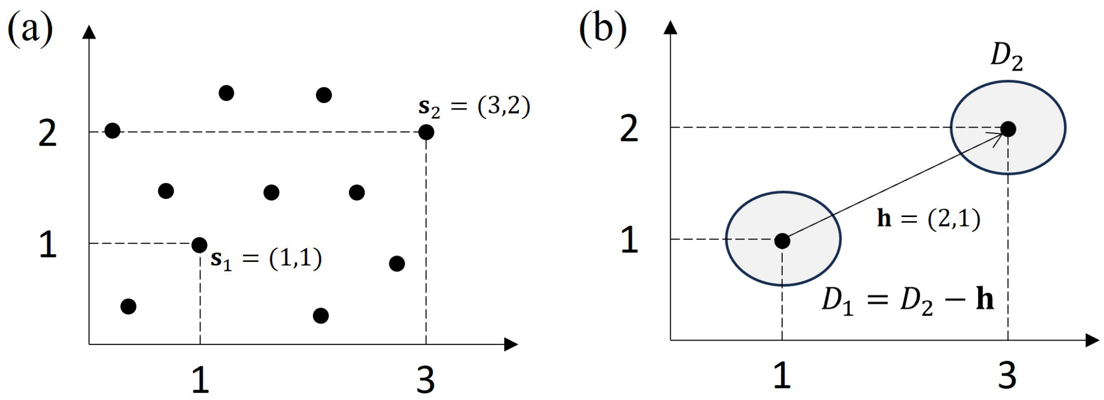

Figure 2. Illustration of the regularization of variogram. (a) Point-support-based data sampling. (b) Area-support-based data.

2.2.2 Entropogram

Theoretically, the measurement associated with an area is an average of all possible measurements associated to the points included within the area, leading the RF defined on an areal support to be more regular than a hypothetical point-based RF equivalent. Such an average eliminates the variation among the points within the area, and the covariance function calculated from the areal variable, of necessity, has a smaller variance than that calculated from the equivalent point variable. This phenomenon of change in characteristic of covariance as the support of the data changes is known as regularization in geostatistics. Under this situation, the complexity of the integrated RF is also smaller than that of the RF representing points or, more generally, a finer measurement support. However, the complexity characterized by Shannon entropy should be measured by discretization in practice as the true probability distribution is unknown. In the remaining part of this section, the variation of the entropogram as the support changes is derived.

Given that the entropogram is built on the Shannon entropy, we first discuss how the Shannon entropy varies across scales. For individual RVs, the Shannon entropy increases with the size of support. That is,

$$\begin{aligned} H_\Delta(Z) &= -\sum_{x_i} p(x_i) \ln\left(\frac{p(x_i)}{\Delta}\right) \\ &= -\sum_{x_i} p(x_i) \ln p(x_i) + \ln \Delta, \\ &= H(Z) + \ln \Delta \end{aligned} \quad (6)$$

as the corresponding state space may be smoothed with more randomness due to the mixture of states (Batty, 1974). Here, $H(Z)$ is the entropy over a point support, also referred to as differential entropy, $H_\Delta(Z)$ is the discrete entropy with support Δ which is the area of the support of Z, or say discretization scale. Now, we have Eq. (6) which transfers the point-based differential entropy to any scale. For example, consider that the differential entropy for a uniform distribution U(0,1) is 0. If we discretize U(0,1) with discretization scale 0.5 into p[(0,0.5)]=1/2 and p[(0.5,1)]=1/2, the Shannon's entropy is $-\log_2 \frac{1}{2} = 1$. Based on Eq. (6), we can reconstruct the differential entropy by $-\log_2(\frac{1}{2}/(\Delta = 0.5)\,) = 0$, which is the differential entropy for a continuous RV with uniform distribution U(0,1). In fact, the scale-derived difference between the differential entropy and the discrete entropy in Equation (6) (i.e., $\ln(\Delta)$) reflects the relationship between the probability spaces at the different scales.

Next, we can take Eq. (6) into Eq. (2), that is, the formula of the entropogram with point support, to represent the entropogram with the discretization scale as shown in Eq. (6). Specifically, consider that two RVs $Z(\mathbf{s}_0)$ and $Z(\mathbf{s}_0 - \mathbf{h})$ are sampled from locations $\mathbf{s}_0$ and $\mathbf{s}_0 - \mathbf{h}$ with support $\mathbf{D}_1$ and $\mathbf{D}_2$, respectively. Their entropogram should be defined as

$$\begin{aligned}\tau_{\Delta_1\Delta_2}(\mathbf{h}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} p(x_i, x_j)\ln\left(\frac{p(x_i,x_j)/(\Delta_1\Delta_2)}{p(x_i)p(x_j)/\Delta_1\Delta_2}\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} p(x_i, x_j)\ln\left(\frac{p(x_i,x_j)}{p(x_i)p(x_j)}\right) \\ &= \tau(\mathbf{h})\end{aligned}, \quad (7)$$

where $\tau_{\Delta_1\Delta_2}(\mathbf{h})$ represents the spatial association between RVs at locations a lag $\mathbf{h}$ apart in terms of support $\mathbf{D}_1$ and $\mathbf{D}_2$, respectively. $\ln(\frac{p(x_i,x_j)}{\Delta_1\Delta_2})$, $\ln(\frac{p(x_i)}{\Delta_1})$ and $\ln(\frac{p(x_j)}{\Delta_2})$ are the scale-dependent information content for RVs $\mathbf{Z} = (Z(\mathbf{s}_0),\ Z(\mathbf{s}_0 - \mathbf{h}))$, $Z(\mathbf{s}_0)$ and $Z(\mathbf{s}_0 - \mathbf{h})$, respectively. We can clearly see that the entropogram with support $\mathbf{D}_1$ and $\mathbf{D}_2$ is equal to the point-support based entropogram after simplification, where the scale-derived difference $\ln(\Delta)$ has been eliminated.

Different to the variogram, we proved that the entropogram is generally scale-invariant (i.e., the spatial association captured in the frequency domain is independent of scale). In fact, Eq. (6) can be regarded as a scale interpretation model, in which the support is introduced as a correction for information at a specific scale. However, after normalizing the information on co-occurrence (x_i, x_j) as in Eq. (7), the effect of the support on the information is reverted.

3 Experimental results

3.1 Shannon entropy

3.1.1 Univariate case

Three generally used probability distributions (Gaussian, Uniform and Weibull distribution) were used as the basis for this analysis. In Figure 3(a), for individual RVs, the distribution of their realizations, or say measurements, can be described exactly by their intrinsic probability distributions (the blue area). However, when the measurement support is defined as an interval or area instead, measurements of a geographical element are treated as the average of the realizations of the RVs within the measurement support. Under this situation, the geospatial data can be considered as a general description of the real world, but filtered by a sampling strategy with a defined measurement scale. Therefore, for each distribution, we first generate a fixed number of independent samples. Then, we take the mean value of the independent samples as one realization of the merged RVs associated with the support covering the fixed number of points. With the expansion of measurement scale, the probability distribution of the RV representing the property of interest tends to behave increasingly as a Gaussian distribution regardless of the intrinsic distribution, due to central limit theorem. In probability theory, the central limit theorem establishes that, in many situations, for identically distributed independent observations, the distribution of the sample mean tends towards the standard normal distribution even if the original variables themselves are not normally distributed.

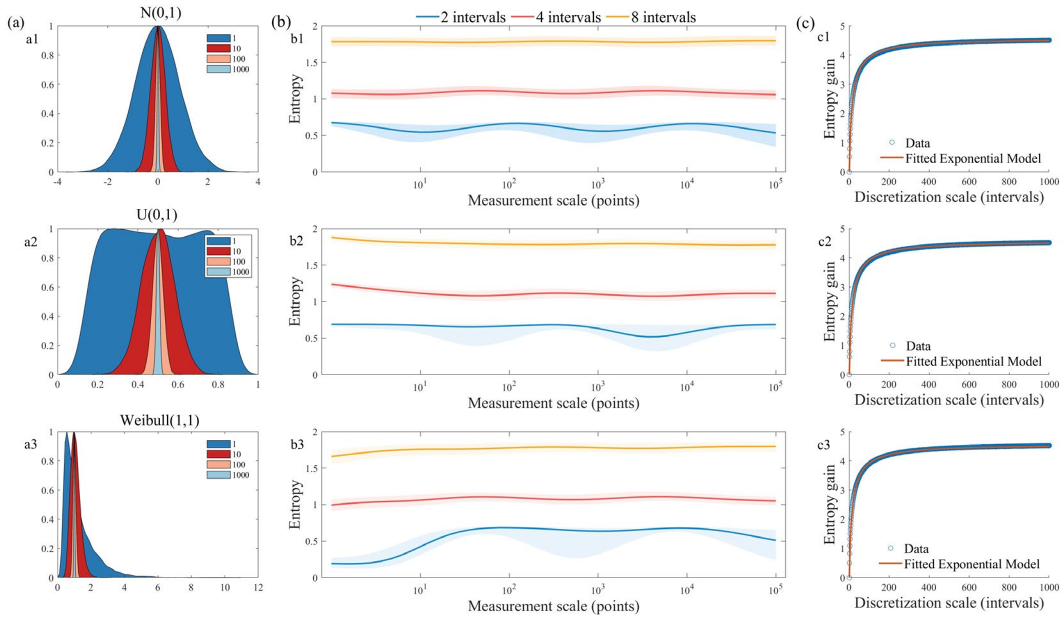

Figure 3. The scale effect of Shannon entropy for univariate data. (a) The probability distribution of N(0,1), U(0,1) and Weibull(1,1) for different measurement scales, respectively. (b) The effect of measurement scale on the Shannon entropy with fixed discretization scale. The *x*-axis refers to measurement scale representing how many points were covered by the corresponding measurement support. The blue line refers to using 2 equal-length intervals to discretize the continuous values; the red line refers to using 4 equal-length intervals to discretize the continuous values; and the yellow line refers to using 8 equal-length intervals to discretize the continuous values. (c) The entropy increases exponentially to the limit with an increase in discretization scale.

Although the probability distribution of a geographical variable at a location varies with a change of support, we find that the discrete Shannon entropy is relatively stable across measurement scales with a fixed discretization scale. The discretization scale, here, is defined as the number of intervals used to discretize the continuous probability distribution for the estimation of Shannon entropy. Here, we group the identically distributed independent samples into different kinds of categories for each measurement scale. The Shannon entropy of the geographical variable on different measurement scales is consequently estimated by the frequency distribution of the assumed categories. It is apparent from Figure 3(b) that the Shannon entropy does not change significantly as the support changes. This may be caused by the resultant asymptotically Gaussian distribution for the expanding measurement support.

In contrast to the above, the discretization scale does play a role in impacting the estimation of Shannon entropy. Specifically, the discrete Shannon entropy increases with the discretization scale (i.e., the number of intervals). When the number of intervals increases, given that the support of the probability distribution is fixed, the variation within each interval is reduced while the inter-interval variation is increased. The complexity of the probability distribution of the intervals, thus, increases in response to the reduction of the interval size, leading to an increase in the information contained in the corresponding probability distribution, represented as a larger value of Shannon entropy. The effect of increasing the number of intervals on the discrete approximation to continuous entropy is plotted comprehensively in Figure 3(c). The relationship between discrete Shannon entropy y and the number of intervals x is $y = 5(1 - \exp(-0.09\ln^2(x)))$. Note that in practice the discretization scale is naturally predefined and unchangeable for discrete probability distributions.

3.1.2 Bivariate case

The scale effect of the Shannon entropy for the joint probability distribution of two or more variables is the same as the situation for a single variable. Figure 4(a) demonstrates the sample distribution of the standard bivariate Gaussian distribution. When integrating different numbers of identically distributed independent bivariate observations into different measurement supports, the distributions of the synthetic non-point-support samples are relatively stable across the gradually narrowed definition support of the probability distribution. The relationship between the discrete Shannon entropy y and the number of intervals for each variable x is now $y = 7(1 -$

$\exp(-0.15\ln^2(x)))$ (see Figure 4(f)). Note that there are x^2 areas used to estimate the discrete Shannon entropy. Specifically, we divided the space into four, 16 and 64 areas, respectively, to produce the frequency distributions and calculate the discrete Shannon entropy in Figure 4(e) for different supports.

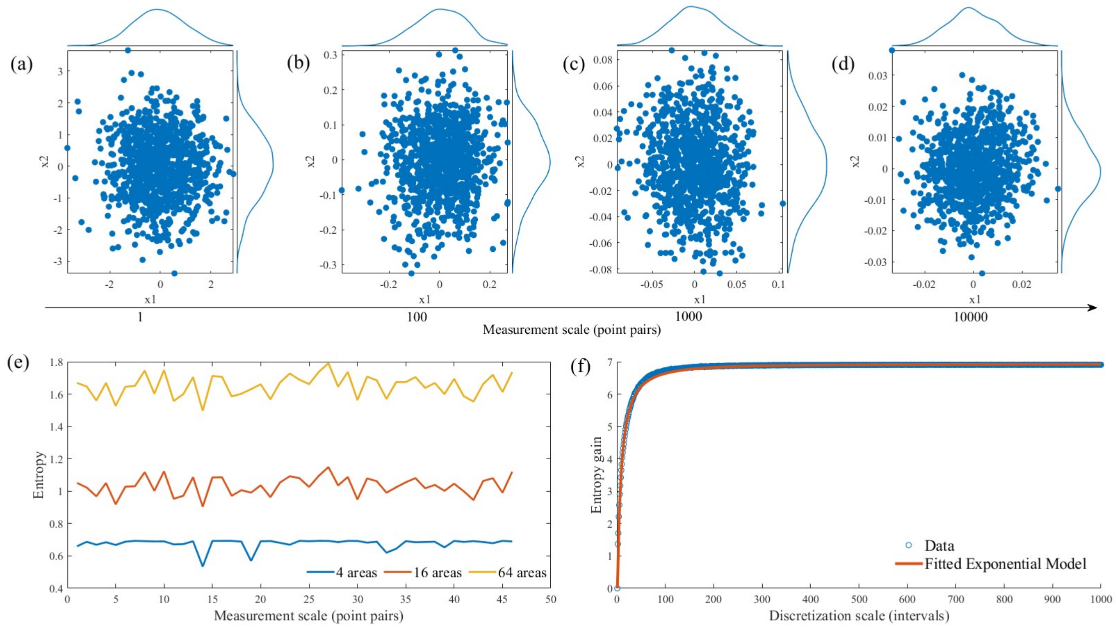

Figure 4. The scale effect of Shannon entropy for independent bivariate data. (a-d) Realizations of the standard bivariate Gaussian distribution for measurement scales of 1, 100, 1000 and 10000 pairs of points. (e) The effect of measurement scale on the Shannon entropy with fixed discretization scale. (f) The entropy increases exponentially to the limit with an increase in discretization scale.

Figure 5 demonstrates the scale effect of Shannon entropy in a more geographical way. Consider an increase in the support as increasing convolution of the elements across the support. The measurement over the support can then be captured by the continuous average of any two adjacent measurements across the spatial continua until there is only one measurement over the support of interest. In this way, the scaling behaviour of integrating two adjacent measurements can be generalized to any support. For each location, each measurement can be modelled by a Gaussian RV where the mean is the measurement, and the variance is the measurement uncertainty due to the measurement instruments. Figure 5(d) shows that the Shannon entropy increases with higher discretization scale to a limit, whilst the change of support does not affect the Shannon entropy of the resultant measurements.

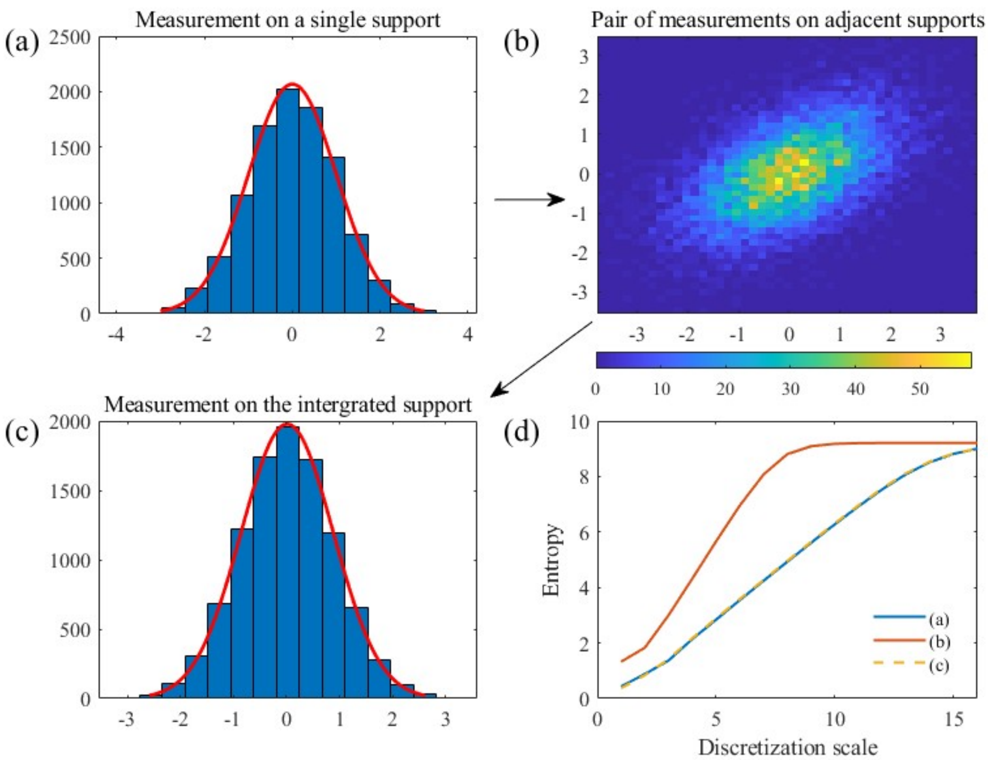

Figure 5. The scale effect of Shannon entropy for correlated bivariate data. Note that the blue line in (d) is covered by the yellow line. The discretization scale is defined by the number of intervals used to discretize the continuous probability distribution for the estimation of Shannon entropy.

3.2 Entropogram

Given that the entropogram is built upon the classic Shannon entropy, under a fixed discretization scale, the entropogram is naturally scale-invariant, that is, invariant regarding the measurement scale (or size of support). Recall that the discretization scale is usually predefined and constant in a given study. To investigate the impact of measurement scale on the entropogram, three second-order stationary Gaussian processes were conceived with different degrees of heterogeneity, as shown in Figure 6(a-c). The corresponding covariance functions are plotted in Figure 6(d-f).

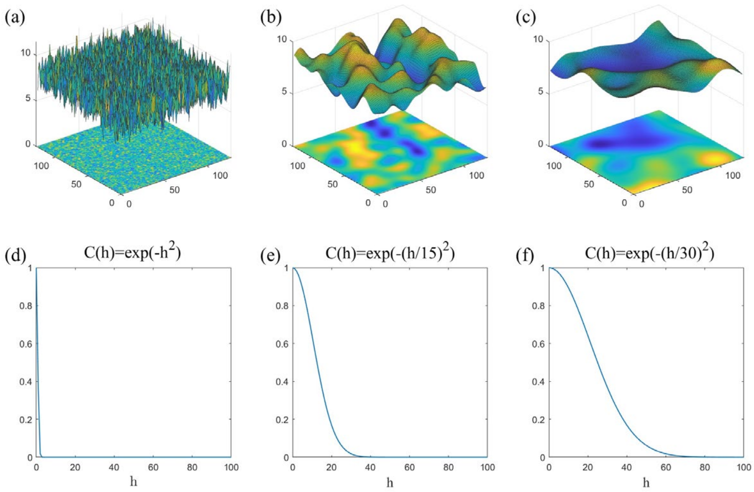

Figure 6. Simulated geospatial data. (a) Highly heterogeneous scene. (b) Moderately heterogeneous scene. (c) Homogeneous scene. (d-f) The covariance functions used to generate the geospatial data of (a), (b), and (c), respectively.

The geospatial data in Figure 6(a-c) were iteratively convoluted into hierarchical scales by a 2x2 mean window, to examine the effect of measurement scale. Note that to create the required categorical data at different scales, we first simulated continuous data at a range of scales and then discretized them, instead of aggregating an initial fine-resolution categorical dataset. The discretization scales of two and three intervals were investigated for the entropogram also. Under the discretization scale of two intervals, the geospatial data are *de facto* modelled as two qualitatively different states, wherein one state represents values smaller than the mean of the whole dataset and the other state represents values greater or equal to the mean. The geospatial data of three intervals are similarly produced by regrouping the values into three states based on three equal length intervals dividing the original values. The artificially generated multiscale geospatial datasets are plotted in Figure 7.

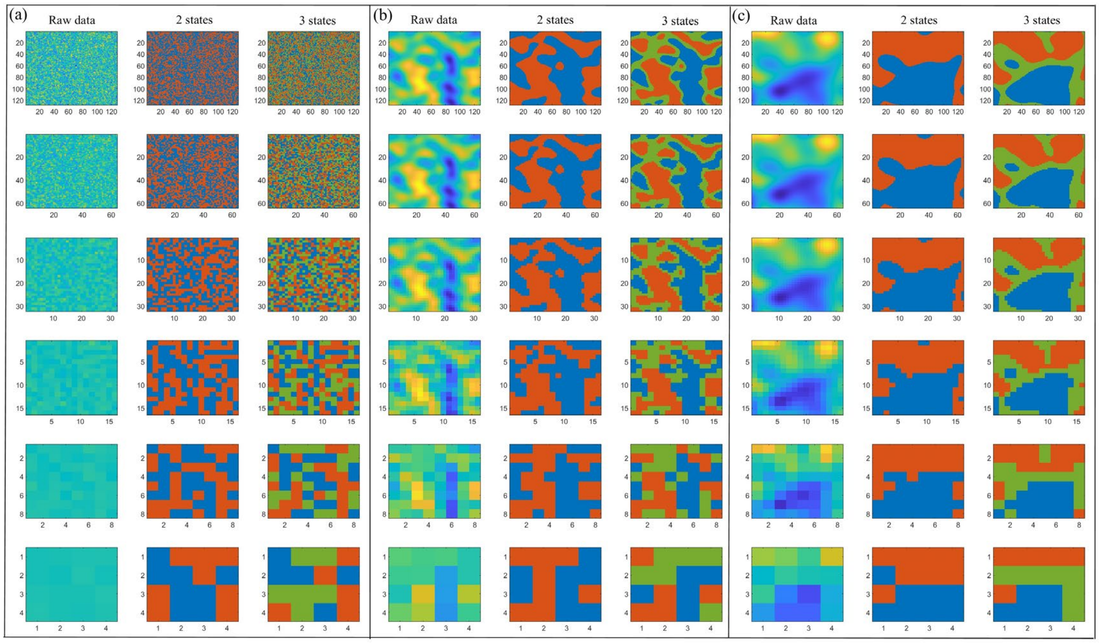

Figure 7. Artificially generated multiscale geospatial datasets. (a) Highly heterogeneous scene. (b) Moderately heterogeneous scene. (c) Homogeneous scene.

Inherited from the scaling stability of the Shannon entropy in terms of measurement scale, the entropy-based entropogram is intrinsically scale-invariant as the support changes. On the other hand, the entropy gain with an increase in the discretization scale makes the values of the entropogram positively related to the discretization scale of the geospatial data. The scaling behaviour of the entropogram is illustrated in Figure 8. Regardless of the characteristics of the dynamic process (i.e., Figure 8(a), (b) and (c)), the entropogram is stable across measurement scales, for both discretization scales of two intervals (i.e., Figure 8(a1), (b1), and (c1)) and three intervals ((i.e., Figure 8(a2), (b2), and (c2))), respectively. Note that with an increase in measurement scale, the number of simulated measurements will decrease, as the spatial extent is fixed. Thus, for a measurement scale of 4x4 lattices, the entropogram may be unreliable due to limited data, resulting in a difference to the entropogram defined at other scales. This echoes the important issue in remote sensing of how to choose a suitable measurement scale for the spatial property of interest given a fixed study spatial extent (cf. Atkinson and Lloyd 2021), which is beyond the scope of this study.

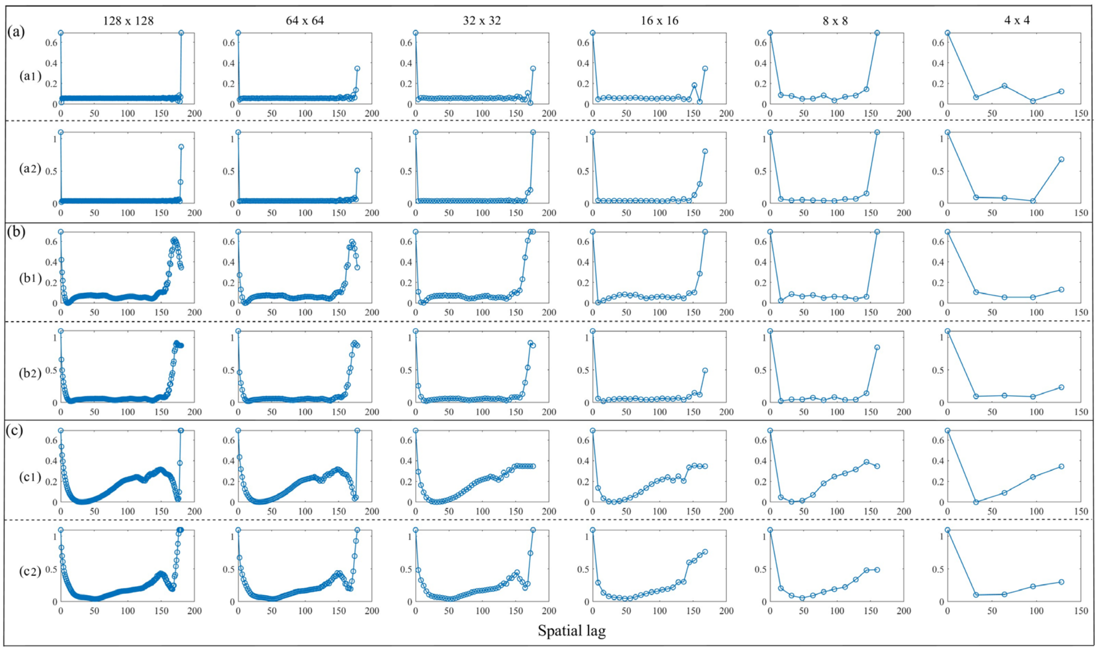

Figure 8. The scaling behaviour of the entropogram. (a) Highly heterogeneous scene. (b) Moderately heterogeneous scene. (c) Homogeneous scene. a1, b1 and c1 are of discretization scale of two. a2, b2 and c2 are of discretization scale of three.

The entropogram is determined by the complexity of the conditional probabilities between the RVs on pairs of points. To demonstrate the scale-invariant property of the entropogram in detail regarding the measurement scale, the conditional probabilities of discretization scale two are plotted in Figure 9. Under this condition, the entropogram refers to the correlation between the two states corresponding to the discretization scale. Given that the realization of the conditioned RV is known, it can be found that the conditional probabilities across spatial lags are consistent as the support changes. When the discretization scale is three, the conditional probabilities remain stable for different measurement scales. The conditional probabilities in Figure 9(g) for measurement scale of 4x4 lattices and spatial lag of 256 look different as the number of data may be too small. This again stresses the importance of the study scale, that is, spatial extent. The measurement-scale independence of the conditional probabilities shows that the entropogram is measurement scale-invariant and particularly the correlation between states can be extended to other measurement scales under the same discretization scale. In addition, such measurement scale-independence of the entropogram as well as its embedded conditional probabilities are independent to the spatial patterns of the geospatial data, given that the three columns in Figure 9 from left to right correspond to the highly heterogeneous (Figure 6(a)), moderately heterogeneous (Figure 6(b)), and homogeneous scene (Figure 6(c)), respectively.

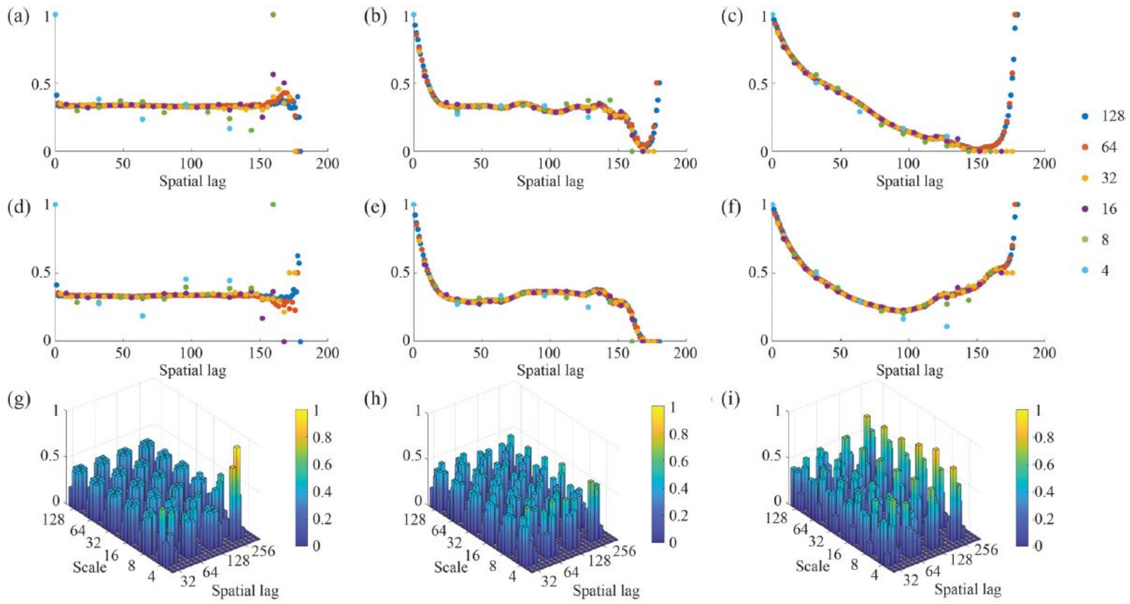

Figure 9. Conditional probabilities over different measurement scales and discretization scales. (a-c) The conditional probability of the realization of one RV is state one given that the realization of another RV is state one. (d-f) The conditional probability of the realization of one RV is state two given that the realization of another RV is state two. (h-g) The conditional probability patterns of the three states for discretization scale of three. Three columns from left to right correspond to the highly heterogeneous, the moderately heterogeneous, and the homogeneous scene, respectively.

Finally, we examined the scale independence of the entropogram using real-world digital elevation data around the City of London. The elevation data were produced by WorldPop (Lloyd et al., 2019) at a spatial resolution of 3 arc-seconds (approximately 100 m at the Equator), and the value of each grid cell represents its elevation above sea level (in meters). Figure 10(a) illustrates the original raw data and the corresponding discretized category map. We then aggregated the 3 arc-second continuous data to 6 arc-second continuous data and discretized it into three class with equal-length intervals again, see Figure 10(b). The entropogram of the 3 arc-second and 6 arc-second discretized elevation data in Figure 10(c) show that spatial association characterized by the entropogram was relatively stable across the two different scales (i.e., 3 arc-second and 6 arc-second), for the different spatial lags.

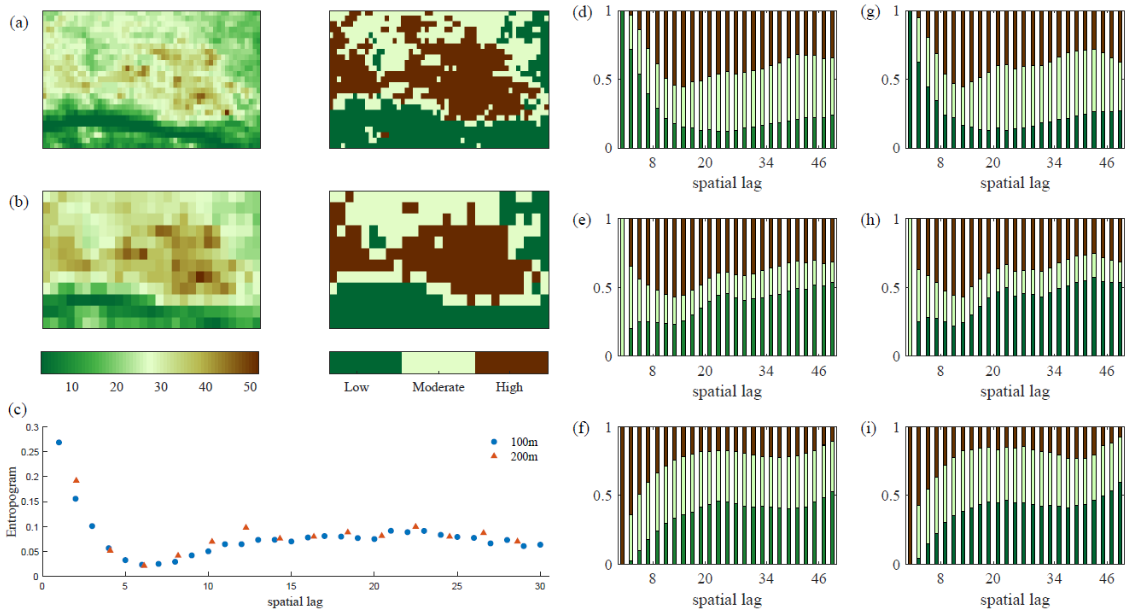

Figure 10. Digital elevation map of the City of London, UK with (a) 100 m resolution and (b) ~200 m resolution. The continuous elevation was discretized into three classes (High, Moderate and Low elevation) by equal-length intervals. (c) Entropogram of the discretized digital elevation maps (a) and (b). (d-f) The conditional probability of different states, i.e., Low, Moderate and High, conditioned on states Low, Moderate and High, respectively, for 100 m resolution. (g-i) The conditional probability of different states, i.e., Low, Moderate and High, conditioned on states Low, Moderate and High, respectively, for ~200 m resolution.

In addition to the entropogram *per se*, we also demonstrate the conditional probabilities for the discretized Low, Moderate and High elevation states across different spatial lags for the 3 arc-second and 6 arc-second discretized elevation data, respectively. For example, Figure 10 (d) refers to the probability of different states conditioned on state Low for the 3 arc-second discretized data, and the first bar on the left shows that given a pixel is state Low, another pixel a distance 0 apart will be Low always with 100% probability. This is natural since they are the same pixel when the lag is 0. With an increase of lag, the incidence probability of another pixel being Low decreases, with a corresponding increasing probability of state Moderate and High. Figure 10 (e)-(f) are the conditional probabilities conditioned on states Moderate and High, respectively, for the 3 arc-second discretized data; whilst Figure 10 (g)-(i) illustrate the conditional probabilities conditioned on states Low, Moderate and High, respectively, for the corresponding 6 arc-second discretized data.

For different scales, we find that the pattern of conditional probabilities is relatively stable at the two different scales. Figure 11 further shows the correlation between the variation of conditional probabilities across the two scales for each category. To avoid coincidence, we used a bootstrap to estimate the distribution of the correlation coefficients. Besides, Table 1 also provides the corresponding Pearson's correlation coefficients with confidence intervals estimated by t-statistic. All the distributions produce a large correlation between patterns across scales, while the probabilities conditioned on the same state show the largest correlation coefficients close to 1. This may be because of the first law of geography, where the same state tends to be present around the same state leading to more robust patterns of the corresponding conditional probabilities across spatial lags. The difference may be caused by the aggregation algorithm (cubic interpolation) used to produce the 6 arc-second data, and the rounded distance between pairs of pixels.

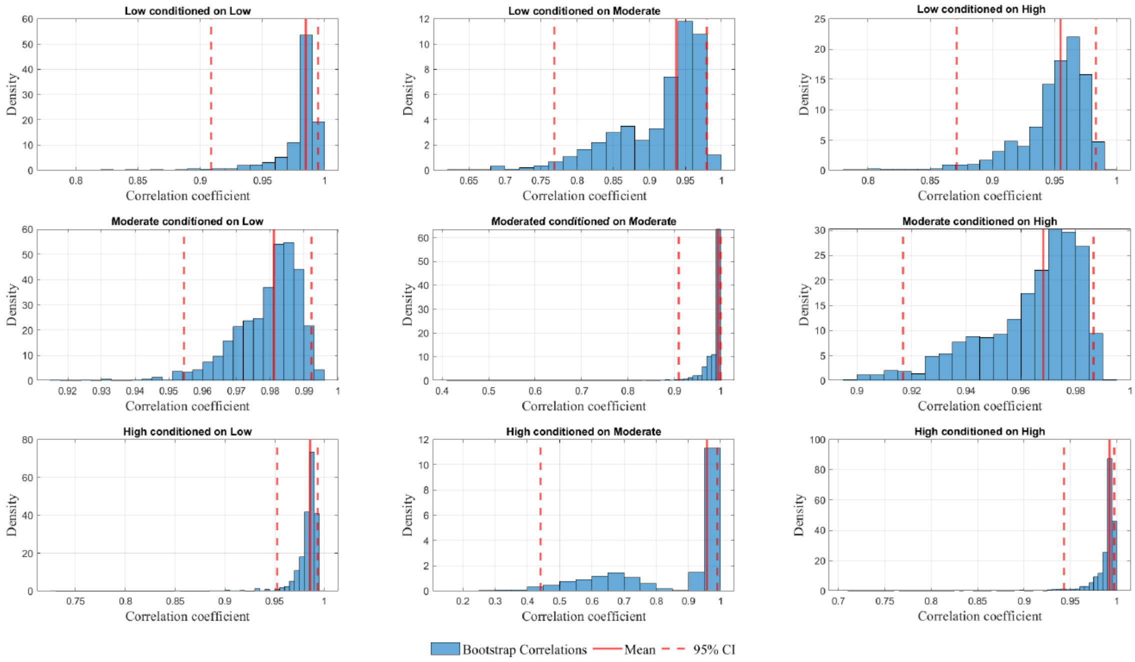

Figure 11. Correlation coefficients between the conditional probabilities for the discretized Low, Moderate and High elevation state across different spatial lags for the 3 arc-second and 6 arc-second discretized elevation data, respectively.

Table 1. Correlation coefficients between the conditional probabilities for the discretized Low, Moderate and High elevation state across different spatial lags for the 3 arc-second and 6 arc-second discretized elevation data, respectively. The confidence intervals were estimated by t-statistic. Mean [Lower 95%CI, Upper 95%CI].

	Conditions		
States	Low	Moderate	High
Low	0.9848 [0.9652, 0.9934]	0.9374 [0.8613, 0.9724]	0.9549 [0.8990, 0.9802]
Moderate	0.9568 [0.9810, 0.9917]	0.9945 [0.8613, 0.9976]	0.9682 [0.9281, 0.9861]
High	0.9858 [0.9676, 0.9938]	0.9567 [0.9029, 0.9810]	0.9922 [0.9821, 0.9966]

4. Discussion

Earth's surface space can be regarded as a giant, complex system and its dynamics are extremely closely related to the activities of human beings (Steffen, 2020). The spatial information we require about the Earth’s surface has a great potential span of temporal and spatial resolution (De Boer, 1992). The properties observed at one scale, and the principles or laws that are established, may still be valid at another scale, may be similar, or may need to be corrected. Therefore, when observing and interpreting patterns, attention should be paid to understanding the measurement process as well as the underlying geographical process of interest (Dodge, 2021). While semantic interpretation of observed patterns may be informative of the underlying process, the scale effect should be considered when making inferences about process, or studying the relationship between pattern and process (Wu & Levin, 1994; Woodcock & Strahler, 1987).

The scaling behaviour of geospatial data refers to the patterns and processes in geoscience data showing different characteristics or properties at different spatial scales (Atkinson and Tate, 2000; Goodchild, 2011). This is because the same phenomenon can be represented in different ways at different levels of detail and scales. For example, in the study of land use and land cover, a forested area may appear to be a single, homogeneous unit from a broad regional perspective, but upon closer examination, from a more detailed, localized perspective, the same area may contain different tree species and land uses, such as agricultural land or urban construction land. The scaling behaviour of geospatial data may, thus, have an important impact on scientific inferences and policy decisions. For example, the choice of spatial scale may affect the interpretation of analysis results and patterns, such as the identification of hot spots, clusters or outliers (Xu, Croot, and Zhang, 2021; Walter, Tillyer, and Acolin, 2023). Besides, the effects of environmental change may also vary depending on the scale of measurement, as different processes and feedback mechanisms operate at different scales, such as land-use change or climate change (Yang, Huang, and Liu, 2020).

The value of quantifying spatial information is difficult to convey. Arguably, spatial information has been relatively neglected compared to a predominant focus in the literature on accuracy, especially when introducing new methods of prediction. However, spatial information is equally important (Zhang et al., 2023b). For example, when predicting at different spatial resolutions the spatial information in the predicted images changes, with commonly finer spatial resolutions containing greater information (i.e., spatial detail), but with a lower accuracy per pixel. Thus, when the spatial resolution varies it is important to quantify both the spatial information and the accuracy.

The entropogram provides an information-based alternative solution to the variogram with which to characterize the spatial association between pairs of locations. It quantifies the shared information between locations, based on the well-established Shannon entropy. We studied how the entropogram behaves when the underlying spatial support changes, with some surprising consistency found across scales revealed for the first time. This analysis represents a crucial first step to establishing how it might be possible to regularize the entropogram, that is, to map the entropogram model from one support to another without new data. Regularization of the variogram is well-established, but this is not so for the entropogram. A further potential application of our findings is that it may be possible to specify a specialised spatial interpolation model through the interpretation of information, with extension to multi-scale analysis and data aggregation (Turner et al., 1989; Wang et al., 2020; Fotheringham & Sachdeva, 2022).

In spatial analysis, continuous geospatial data may have greater explanatory ability after discretization (Cao et al., 2013). However, the discretization method, or choice of discretization level, may affect the scaling behaviours of the entropogram. This means that the choice of discretization method has a strong effect on the result, where the scale-invariant property of the entropogram may only be valid under specific conditions. Specifically, using different discretization methods can affect the resultant entropogram of the discretized geospatial data. For example, the same data point may be allocated a different discretization class under different discretization criteria. Nonetheless, the scale independence of the entropogram is not affected by using different discretization methods, as long as the same discretization standards are maintained across scales (i.e., keeping the proportional structure of the discretized classes the same across scales).

In this article, we first generated continuous data at a range of scales and then discretised these data into categories, and kept the proportional structure of the discretized states the same across scales. This was done, first, to match the most common situation in the real world, which is that categorical data are often produced by measuring, and then discretizing, a geographical process. This is true, for example, for land cover and land use where a remotely sensed image is first obtained of reflectance, which is then discretised to land use-land cover categories. Over a limited range the number of categories would reasonably be expected to remain constant. The alternative would be to aggregate a fixed, fine-resolution categorical representation, but this is less common in practice. The second reason for choosing this approach is that it allowed us to demonstrate some underlying properties of the nature of scaling information (i.e., using the entropogram) as opposed to variation (i.e., using the variogram), as well as to demonstrate the utility of the entropogram as a measure of spatial association. Future research should be undertaken to explore the scaling behaviour of the entropogram under alternative choices of simulation, scaling and discretization methods.

5. Conclusion

Spatial association, as a general descriptive feature of spatial patterns, plays a crucial role in scaling geospatial data. However, except for the variogram, the scaling behaviours of commonly used spatial association measures is usually nonlinear and lacks a strict theoretical basis. This means that geostatistical change-of-support theory is the *de facto* method for scaling Earth science data, based on regularization of the variogram. The recently introduced entropogram represents an alternative to the variogram suitable for characterizing the information in categorical variables. However, until now the relationship between the entropogram and the spatial support, and, thus, how to model the effect of a change in support on the entropogram, was not clear. Regularization of the entropogram, from both a theoretical and experimental point of view, is derived in this article. We found that the entropogram is surprisingly scale-invariant under certain conditions about the way in which categories are derived. That is, the values as well as the conditional probabilities of the entropogram remain relatively unchanged as the measurement support changes. The generalisation of the entropogram to changes in the scale of measurement extends the utility of this relatively new measure of spatial association.

6. Data and code availability

The data and code that support the findings of this study are openly available in figshare at https://figshare.com/s/39d4a02c33d8d92a7af7.

Reference

Atkinson, P.M. and Tate, N.J., 2000. Spatial scale problems and geostatistical solutions: a review. The Professional Geographer, 52(4), pp.607-623.

Atkinson, P.M., 2013. Downscaling in remote sensing. International Journal of Applied Earth Observation and Geoinformation, 22, pp.106-114.

Atkinson, P.M. and Lloyd, C.D., 2021. Geostatistical models and spatial interpolation. In Handbook of regional science (pp. 1813-1827). Berlin, Heidelberg: Springer Berlin Heidelberg.

Batty, M., 1974. Spatial entropy. Geographical analysis, 6(1), pp.1-31.

Clark, I., 1977. Regularization of a semivariogram. Computers & Geosciences, 3(2), pp.341-346.

Cao F, Ge Y, Wang J F., 2013. Optimal discretization for geographical detectors-based risk assessment. GIScience & Remote Sensing, 50(1), pp.78-92.

De Boer, D.H., 1992. Hierarchies and spatial scale in process geomorphology: a review. Geomorphology, 4(5), pp.303-318.

Dodge, S., 2021. A data science framework for movement. Geographical Analysis, 53(1), pp.92-112.

Fotheringham, A.S. and Sachdeva, M., 2022. Scale and local modeling: new perspectives on the modifiable areal unit problem and Simpson's paradox. Journal of Geographical Systems, pp.1-25.

Goodchild, M.F., 2011. Scale in GIS: An overview. Geomorphology, 130(1-2), pp.5-9.

Ge, Y., Jin, Y., Stein, A., Chen, Y., Wang, J., Wang, J., Cheng, Q., Bai, H., Liu, M. and Atkinson, P.M., 2019. Principles and methods of scaling geospatial Earth science data. Earth-science reviews, 197, p.102897.

Jia, Y., Ge, Y., Chen, Y., Li, S., Heuvelink, G.B. and Ling, F., 2019. Super-resolution land cover mapping based on the convolutional neural network. Remote Sensing, 11(15), p.1815.

Jupp, D.L., Strahler, A.H. and Woodcock, C.E., 1988. Autocorrelation and regularization in digital images. I. Basic theory. IEEE Transactions on Geoscience and Remote Sensing, 26(4), pp.463-473.

Lloyd, C. T., Chamberlain, H., Kerr, D., Yetman, G., Pistolesi, L., Stevens, F. R., … Tatem, A. J., 2019. Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. Big Earth Data, 3(2), 108–139. https://doi.org/10.1080/20964471.2019.1625151

Matheron, G., 1963. Principles of geostatistics. Economic geology, 58(8), pp.1246-1266.

Pu, R. and Bonafoni, S., 2023. Thermal infrared remote sensing data downscaling investigations: An overview on current status and perspectives. Remote Sensing Applications: Society and Environment, 29, p.100921.

Steffen, W., Richardson, K., Rockström, J. et al., 2020. The emergence and evolution of Earth System Science. Nature Reviews Earth & Environment, 1, 54–63. https://doi.org/10.1038/s43017-019-0005-6

Sdraka, M., Papoutsis, I., Psomas, B., Vlachos, K., Ioannidis, K., Karantzalos, K., Gialampoukidis, I. and Vrochidis, S., 2022. Deep learning for downscaling remote sensing images: Fusion and super-resolution. IEEE Geoscience and Remote Sensing Magazine, 10(3), pp.202-255.

Turner, M.G., O'Neill, R.V., Gardner, R.H. and Milne, B.T., 1989. Effects of changing spatial scale on the analysis of landscape pattern. Landscape ecology, 3(3), pp.153-162.

Woodcock, C.E. and Strahler, A.H., 1987. The factor of scale in remote sensing. Remote sensing of Environment, 21(3), pp.311-332.

Wiens, J.A., 1989. Spatial scaling in ecology. Functional ecology, 3(4), pp.385-397.

Wang, P., Wang, L., Leung, H. and Zhang, G., 2020. Super-resolution mapping based on spatial–spectral correlation for spectral imagery. IEEE Transactions on Geoscience and Remote Sensing, 59(3), pp.2256-2268.

Wu, J., Levin, S. A., 1994. A spatial patch dynamic modeling approach to pattern and process in annual grassland. Ecological Monographs, 64(4), pp.447-464.

Wu, J., 2004. Effects of changing scale on landscape pattern analysis: scaling relations. Landscape ecology, 19(2), pp.125-138.

Walter, R.J., Tillyer, M.S. and Acolin, A., 2023. Spatiotemporal crime patterns across six US cities: Analyzing stability and change in clusters and outliers. Journal of Quantitative Criminology, 39(4), pp.951-974.

Xu, H., Croot, P. and Zhang, C., 2021. Discovering hidden spatial patterns and their associations with controlling factors for potentially toxic elements in topsoil using hot spot analysis and K-means clustering analysis. Environment International, 151, p.106456.

Yang, H., Huang, J. and Liu, D., 2020. Linking climate change and socioeconomic development to urban land use simulation: Analysis of their concurrent effects on carbon storage. Applied Geography, 115, p.102135.

Zhang, X., Ge, Y., Ling, F., Chen, J., Chen, Y. and Jia, Y., 2021. Graph convolutional networks-based super-resolution land cover mapping. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, pp.7667-7681.

Zhang, W.B., Ge, Y., Bai, H., Jin, Y., Stein, A. & Atkinson, P.M., 2023a. Spatial Association from the Perspective of Mutual Information, Annals of the American Association of Geographers, DOI: 10.1080/24694452.2023.2209629

Zhang, Z., Song, Y., Luo, P. and Wu, P., 2023b. Geocomplexity explains spatial errors. International Journal of Geographical Information Science, 37(7), pp.1449-1469.