International Journal of Geo-Information

MDPI

Article

Adaptive Geometric Interval Classifier

Shuang Li and Jie Shan *

School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA; li4132@purdue.edu
* Correspondence: jshan@purdue.edu

Abstract: Quantile, equal interval, and natural breaks methods are widely used data classification methods in geospatial analysis and cartography. However, when applied to data with skewed distributions, they can only reveal the variations of either high frequent values or extremes, which often leads to undesired and biased classification results. To handle this problem, Esri provided a compromise method, named geometric interval classification (GIC). Although GIC performs well for various classification tasks, its mathematics and solution process remain unclear. Moreover, GIC is theoretically only applicable to single-peak (single-modal), one-dimensional data. This paper first mathematically formulates GIC as a general optimization problem subject to equality constraint. We then further adapt such formulated GIC to handle multi-peak and multi-dimensional data. Both thematic data and remote sensing images are used in this study. The comparison with other classification methods demonstrates the advantage of GIC being able to highlight both middle and extreme values. As such, it can be regarded as a general data classification approach for thematic mapping and other geospatial applications.

Keywords: data classification; thematic mapping; optimization; cartography; geospatial analysis

Citation: Li, S.; Shan, J. Adaptive Geometric Interval Classifier. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 430. https://doi.org/10.3390/ijgi11080430

Academic Editors: Florian Hruby and Wolfgang Kainz

Received: 31 May 2022
Accepted: 25 July 2022
Published: 31 July 2022

1. Introduction

The objective of classification is to group the data into several classes for an aggregated presentation and characterization [1]. The data to be classified can be associated with an areal unit (e.g., administrative boundaries) or a particular geographic location (e.g., remote sensing images) for thematic mapping. Each class should have small within-class scatter, while different classes usually have larger between-class scatter. The widely used data classification methods in cartography science are quantile, equal interval, and natural breaks [2–4]. Two parameters are often used to describe the classification results, i.e., the class intervals (or breaks) and number of classes. The number of classes for these classification methods is usually predefined. The quantile method aims to make each class contain the same number of data samples (elements). The equal interval method divides the range of data values into equal parts. The natural breaks method determines breaks where data values have large differences. Although the idea of determining class intervals based on data distribution has existed for a long time [5], the above methods are most suitable for data with specific distribution patterns [6], such as linear or uniform. However, these methods emphasize the variations either in the middle of the data or at the extremes (ends) of the data, and often cause misleading classification results and presentation, especially for continuous data with skewed distributions [7]. The skewed continuous distribution is common in many natural and social phenomena. For example, in urban environments, there are much more small-sized blocks than large-sized ones, and much more cities with small populations than mega-cities with very large populations [8]. More effective classification methods that can accommodate skewed data should be developed.

Considering these drawbacks, Esri had a compromise, proprietary solution that is more applicable to continuous data, named smart quantile or geometric interval classification (GIC) [9]. It partitions sequential data values to produce different classes with geometric intervals. The ratio between adjacent intervals is called the geometric coefficient, which is a

constant or the inverse of this constant [10]. Under this requirement, the proprietary GIC approach intends to minimize the squared sum of the number of data samples per class so that the number of samples in each class is approximately equal. GIC works well for the classification and visualization of heavily skewed continuous data [11]. Comparing to the GIC, the quantile method only considers the number of samples per class. As such, it could result in samples with similar values being classified into adjacent classes, whereas samples with much different values being classified into the same class [12]. A comparison of ArcGIS-based breaks obtained by smart quantile (i.e., the GIC) and quantile method is shown in Figure 1, where the histogram is from the blue band of a Landsat-8 image. The information entropy is maximal when the number of samples in each class is the same [13,14]. In the quantile method, the number of data samples in each class is equal, whereas GIC produces more classes at the ends of the data values while keeping a reasonable division in the middle of the distribution. Furthermore, the introduction of the inverse of the geometric coefficient can highlight changes in both the middle part and the extreme parts of the data [15,16]. The in-class differences of low-frequent values are also reduced by smaller intervals for the corresponding classes.

Figure 1. Comparison of the class breaks (tics at the bottom) for quantile (**top**) and GIC (**bottom**) methods for the blue band of a Landsat-8 image. The bar charts are the histogram of the image.

With the above-mentioned advantages, GIC has been applied to various classification tasks, such as producing flood potential maps [17–19], suitability maps for planning [20], creating crime density maps [21] and soil erosion hazard zones [22], and finding urban heat islands [23]. Huan et al. (2012) built an index system and applied the geometric interval, equal interval, natural breaks, and quantile methods to map the groundwater vulnerability to nitrate based on these indices. The results showed that GIC obtained more accurate and rational vulnerability maps than other classification methods. Costache et al. classified the flash-flood potential index into five classes using the natural breaks, quantile, equal interval, and geometric interval classification methods [7]. Better classification results are achieved by GIC and natural breaks, where the areas classified with a high potential for floods contained more torrential pixels.

The Esri's procedure of this method remain unclear to public. To the best knowledge of the authors, there was only one published work that tried to implement GIC by Python-based coding [24]. The upper limits of all classes were defined by a geometric series. The geometric coefficient was calculated based on the ratio of the lowest and highest sample values and remained constant for each class. The classification results were compared with

those of the equal interval, quantile, standard deviation, natural breaks, and logarithmic scale methods. It was found that GIC achieved a more realistic results than other methods for right-skewed data [12,24]. However, the reported algorithm is only applicable to data where the number of samples decreases as the data value increases and is right-skewed, such as archaeological data.

In summary, the existing knowledge gaps on GIC include: (1) no or minimal documented theoretic formulation; (2) no corresponding mathematical solution to the problem; (3) unable to handle data with multiple peaks or multi-modals; and (4) generality to multi-dimensional data, such as images. Our contributions can be outlined as follows. Firstly, the GIC method is mathematically formulated as an optimization problem and solved iteratively to obtain the geometric intervals and class breaks. We investigate the applicability and properties of this formulation by using different kinds of data, such as population data and remote sensing images. Secondly, the setting of the geometric coefficient has been extended to make GIC to be adaptive, or AGIC, to handle data with multi-peak distributions and multi-dimensions. To demonstrate such properties of the adaptive GIC, we compare the results of the GIC with other classification methods commonly used in thematic mapping, including the equal interval, quantile, and natural breaks.

2. The Geometric Interval Classifier

2.1. Formulation of GIC

Given ascending sorted N data samples,

$$s_i \in S = \{s_1, s_2, \cdots, s_N\} \tag{1}$$

Our task is to classify them into K classes by determining the breaks, b_{c-1} and b_c ($c = 1, 2, \ldots, K$), which are the lower and upper bounds (breaks) of the data range for class c, respectively. Among the break points or simply breaks, $b_0 = s_1$ and $b_K = s_N$ are, respectively, the minimum and maximum data values. The breaks and corresponding class intervals are illustrated in Figure 2. In the GIC method, the class intervals $\delta_1, \delta_2, \ldots, \delta_K$ form one or more geometric series, whose geometric coefficient is either q or its inverse q^{-1}.

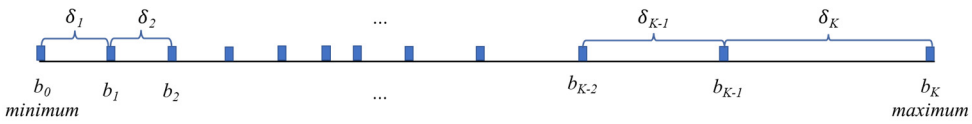
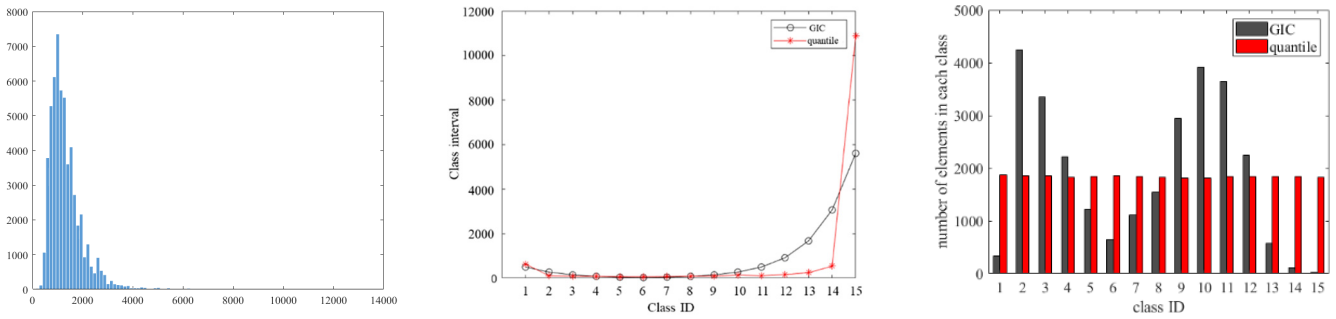


Figure 2. Break points and the corresponding intervals for geometric interval classifier.

The objective of GIC is to balance the number of data samples in each class under the constraint that the class intervals form geometric series. The numbers of samples with extreme values are usually smaller than other samples. To make the sample size of each class similar, the intervals of the classes with more frequent values should be small, while the intervals of the classes with extreme values should be large. The idea of formulating GIC is to adjust the class interval so that the number of samples in each class is similar. At the same time, the distribution of the data shall be considered. For data with a single-peak distribution, extreme data values at the ends (tails) of the histogram should be divided into more classes, which reduces information loss and provides a more detailed interpretation of the extreme values. Figure 3 shows an example of Indiana population data where there is a peak on the left of the histogram. The corresponding class intervals for GIC and quantile methods implemented by ArcGIS become smaller first and then larger to balance the number of samples in each class. The variation of class intervals obtained by GIC is smoother than that of the quantile method. Comparing the number of samples per class in Figure 3c, the quantile method can achieve a more balanced class size. The GIC can be a compromise between the equal interval and the quantile methods [16].



(**a**) (**b**) (**c**)

Figure 3. Histogram of Indiana population (**a**), class intervals (**b**), and number of samples (**c**) for each class determined by the quantile and GIC methods.

2.2. Solution of GIC

We introduce an optimization framework to determine the parameters of GIC. The optimization problem can be defined as:

Objective function:

$$\min \sum_{c=0}^{K-1} count(b_c \leq \forall s_i < b_{c+1})^2 \tag{2}$$

Subject to:

$$b_0 + mq^0 + mq^{-1} + mq^{-2} + \cdots + mq^{-I} + mq^{-I+1} + \cdots + mq^{K-2I+1} = b_K \tag{3}$$

$I \in [2, 3, \ldots, K-1]$ indicates where to inverse the geometric coefficient, whereas $I = 1$ or K means no inverse occurs. This is a nonlinear optimization problem subject to an equality constraint. The constraint is multiplied by a Lagrange multiplier and added to the objective function, which leads to a differentiable Lagrangian function [25]. This optimization problem can be solved iteratively. The basic idea is to convert a constrained problem into a form such that the derivative test of an unconstrained problem can still be applied.

The current GIC implementation needs a pre-defined number of classes K. However, this information may not be known in many applications. Hence, we further modify the GIC to automatically determine K. Based on the idea that the variations within a class should be small, we set a threshold to the coefficient of variation cv_c [26] within each class c as

$$cv_c = \frac{\sigma_c}{\mu_c} \quad (\mu_c \neq 0) \tag{4}$$

where μ_c and σ_c are, respectively, the mean and standard deviation of the data values of class c. We introduce this as a constraint to determine the number of classes when it is not given.

Such formulated GIC will determine its parameters through optimization, $m, q,\ I$, as well as K, when it is not predefined. The calculation will start with an initial number of classes. After solving the optimization problem, the cv_c of each class is calculated. When cv_c is larger than a predefined threshold θ_{cv}, the corresponding class is divided into two classes with equal intervals and the number of classes should be increased by 1. A group of new break points is obtained, which is then used as the initial values for the next iteration. This process stops until the cv_c of all classes are less than the threshold θ_{cv}. It should be noted that, when the threshold is large, the categories are not divided, while when the threshold is small, the number of classes are large.

3. Adaptive GIC

The above section formulated GIC method and provided a solution accordingly. This section will extend the above formulation to handle multiple-peak, multiple dimensional data.

3.1. Adapting GIC for Multi-Peak Data

In the above GIC formulation, the geometric coefficient q can only be inversed at most once, which limits the applicability of GIC to data with more complex distributions. The geometric interval usually decreases first and then increases. This property means that GIC is suitable for data with a distribution of one peak in the middle of the histogram. For data with multiple peaks, we need to modify the formulation and develop a new approach to determine where and how many times to inverse the geometric coefficient.

The formulation of Equation (3) only considers the scenario of one geometric series with geometric quotient q. However, the general situation is that this idea can be used recursively multiple times, i.e., there can be more than one geometric series with geometric coefficient of either q or its inverse q^{-1}, and each of them may start at any place in the given data samples.

We propose an algorithm that determines the inverse position based on the inhomogeneity of the data distribution. Two sets of breaks can be obtained by equal interval and quantile, denoted as b_c^e and b_c^q ($c = 1, 2, \ldots, K$), respectively. For data with multi-peak distribution, the geometric coefficient is inversed at the I_j-th interval for j-th times when the $b^q_{I_j-1} > b^e_{I_j-1}$ and $b^q_{I_j+1} < b^e_{I_j+1}$ (or $b^q_{I_j-1} < b^e_{I_j-1}$ and $b^q_{I_j+1} > b^e_{I_j+1}$), which can be described as

$$\left(b^q_{I_j-1} - b^e_{I_j-1}\right) \times \left(b^q_{I_j+1} - b^e_{I_j+1}\right) < 0,\ 2 \leq\ I_j\ \leq K-1,\ 1 \leq\ j\ \leq i \tag{5}$$

This condition can determine how many times in total the geometric coefficient should be inversed.

The interval for class c can be calculated as

$$\delta_c = mq^{\sum_{j=1}^{i}(-1)^j(I_j - I_{j-1}) + (-1)^{i+1}(c-I_i)},\ 1 \leq c \leq K \tag{6}$$

where $\delta_1 = m$ is the first interval. The geometric coefficient has been inversed $i \in [0,\ c-2]$ times when it reaches class c. $I_j \in [2,\ c-1]$ ($j \in [1,\ i]$) represents where to inverse the geometric coefficient for the j-th time, and the ratio of the two adjacent intervals δ_c/δ_{c-1} is $q^{(-1)^j}$ for $c \in [I_{j-1}+1, I_j]$. I_0 is set to 1 and $\sum_{j=1}^{i}(-1)^j\left(I_j - I_{j-1}\right)$ is set to 0 when $i = 0$, which indicates that there is no inverse of the geometric coefficient.

Figure 4 illustrates an example of eight classes and the geometric coefficient is inversed at $I_1 = 5$ and $I_2 = 6$. For the case of inversing q for two times ($i = 2$), the equality constraint of Equation (3) for the optimization problem should be modified as

$$\begin{gathered} b_0 + mq^0 + \cdots + mq^{-I_1} + mq^{-I_1+1} + \cdots + mq^{I_2-2I_1+1} + \\ mq^{I_2-2I_1} + \cdots + mq^{2I_2-2I_1-K+1} = b_K \end{gathered} \tag{7}$$

where quotient q is inversed at class I_j ($j = 1,\ 2$).

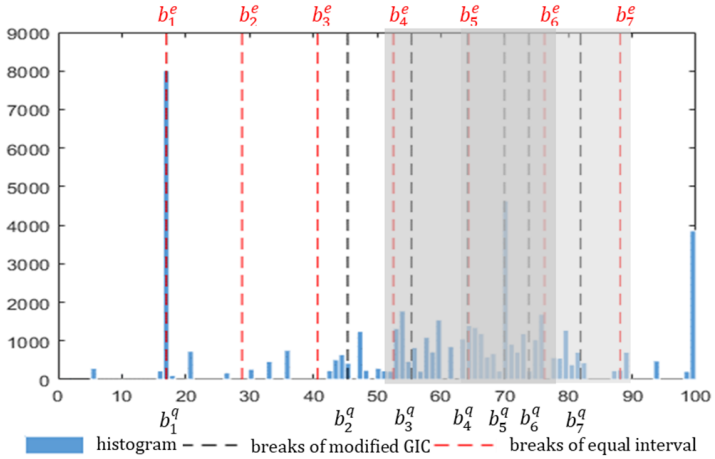


Figure 4. Illustration of determining where and how many times to inverse q in the adaptive GIC for multi-peak data.

3.2. Adapting GIC for Multi-Dimensional Data

The GIC formulation above is designed for classifying data of one dimension, which is the data values of the samples. When dealing with multi-dimensional data, i.e., the sample data has multiple attributes, we form a feature space by calculating the Euclidean norm [27] of these attributes. The definition of the Euclidean norm for 3D data $x \in N_1 \times N_2 \times N_3$ is

$$s_{ij} = \sqrt{\sum_{k=1}^{N_3} x^2(i,j,k)}\ ,\ i \in \{1,\ N_1\},\ j \in \{1, N_2\} \tag{8}$$

where $x(i,\ j,\ k)$ is a sample or pixel in the k-th band of the image with a dimension $N_1 \times N_2$, and N_3 is the number of bands. s_{ij} is the transformed feature, which in this case is simply the squared sum of pixel values in a multispectral image. The break points and geometric coefficient will then be generated for this transformed feature s_{ij}.

4. Test Data and Experiment Design

4.1. Test Data

This study used different data at various scales to evaluate the performance of the adaptive GIC method, including GIS population data and remote sensing images. Figure 5 shows the population per census tract of the State of Indiana, population per zip unit of the City of Chicago. Additionally, to better demonstrate the effectiveness of GIC for multi-peak data, simulated population data for the State of Georgia with multi-peak distribution were created by specifying the population of each county in Georgia (Figure 5c). To investigate the effectiveness of GIC on remote sensing images, we chose the blue band (band 1), cirrus band (band 9), and thermal band (band 10) of Landsat 8. Furthermore, the RGB composite of a Landsat 8 image was employed to show the classification performance of the adaptive GIC for multi-dimensional data. The Landsat 8 images of a field on the border of the Warren County in Indiana and the Vermilion County in Illinois are shown in Figure 6.

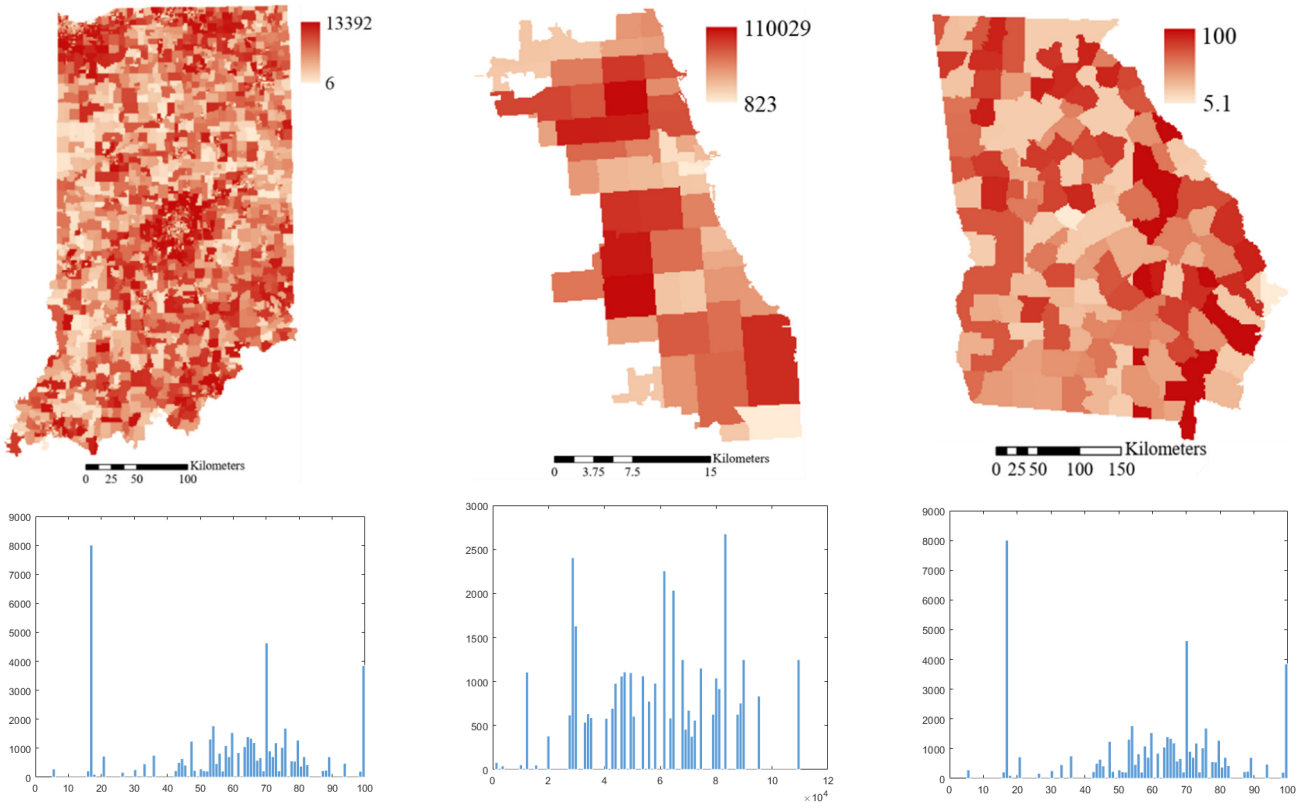


Figure 5. Population of Indiana (**a**), Chicago (**b**), and simulated population of Georgia (**c**) and their histograms.

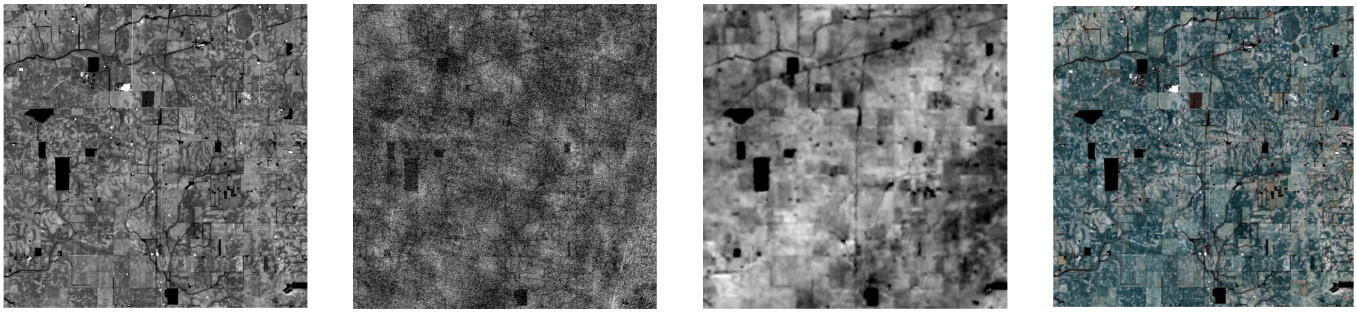


Figure 6. The blue (**a**), cirrus (**b**), thermal (**c**) bands, and RGB composite (**d**) of a Landsat 8 image in Indiana.

4.2. Experimental Design

The experimental settings in this work are summarized in Table 1. The performance of GIC was evaluated on the population of Indiana, Chicago, and the simulated population of Georgia, as well as the blue band, cirrus band, and thermal band of a Landsat 8 image. The initial number of classes was set to 5 and the threshold of coefficient of variation (θ_{cv}) was selected empirically as 0.25. For comparison, other classification methods, including the equal interval, quantile [28], and natural breaks [4,29], were also applied through ArcGIS Pro. The number of classes was set to 10 for all methods in the comparison experiments. The effectiveness of the adaptive GIC for multi-peak data was evaluated on the population data and the Landsat image. The number of times that q needs to be inversed was automatically

determined, which can be 0, 1 or more. The RGB composite of the Landsat 8 image was employed to assess the performance of adaptive GIC for multi-dimensional data.

Table 1. Experimental settings.

Method	Data	Description
GIC	Population of Indiana, Chicago, and simulated population of Georgia	(1) Apply GIC to classify population and Landsat data by solving the optimization problem. (2) Compare the results of GIC with those from equal interval, quantile, and natural breaks.
	Landsat 8 single band images	
Adaptive GIC	Population of Indiana, Chicago, and simulated population of Georgia Landsat 8 single band images	(1) Adaptively determine the number of times that q is inversed (2) Study the classification of multi-peak data.
	Landsat 8 RGB composite	(1) Perform GIC on multi-dimensional data.

5. Results and Discussion

5.1. Single-Peak Data

Figure 7 shows the classification maps for the population data using GIC. The population data are classified into eight or five classes, with larger class IDs representing more population. The GIC results clearly demonstrate the distribution of population over these areas. Additionally, Table 2 also lists the break points of our results and the ones from the GIC method in ArcGIS Pro (Esri proprietary solution); they are similar and consistent.

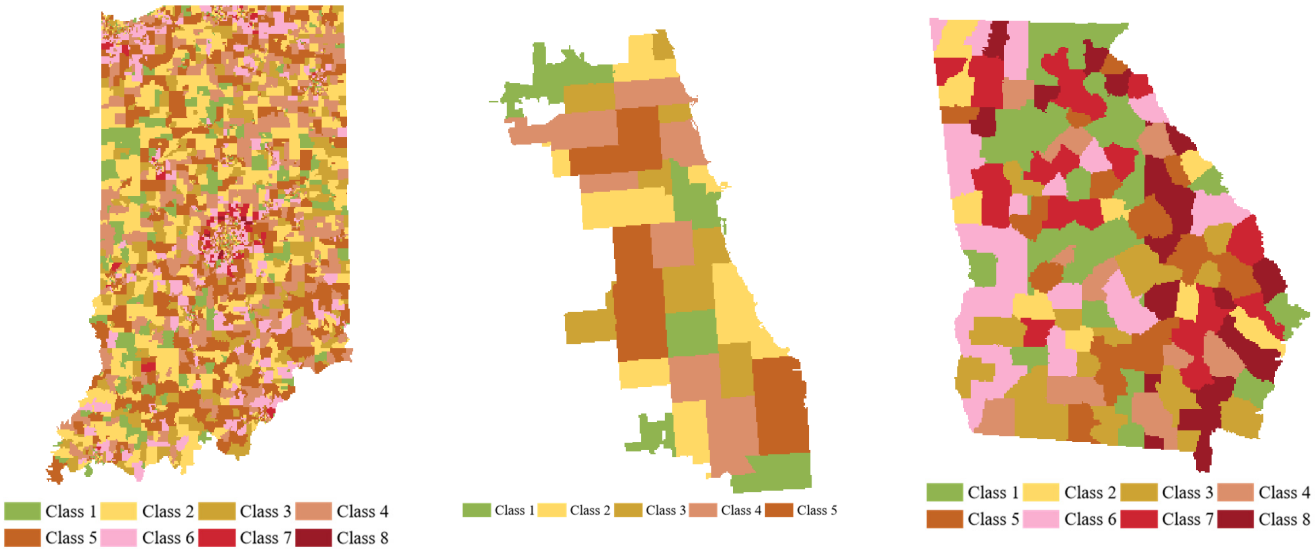


Figure 7. Classification results of population data using GIC for Indiana (**a**), Chicago (**b**), and simulated population data of Georgia (**c**). The larger the class ID, the more the population.

Table 2. Break points of GIC for populations of Indiana, Chicago, and simulated population of Georgia.

Indiana		Chicago		Georgia	
GIC-ArcGIS	**GIC-Author**	**GIC-ArcGIS**	**GIC-Author**	**GIC-ArcGIS**	**GIC-Author**
6.00	6.00	823.00	823.00	5.10	5.10
672.71	676.00	30,479.18	30,644.84	31.57	30.06
981.11	986.67	49,394.02	49,478.53	47.82	46.07
1123.76	1130.73	61,457.98	61,372.77	57.79	56.33
1432.16	1441.41	80,372.82	80,206.47	63.91	62.91
2098.87	2111.41	110,029.00	110,029.00	67.66	67.13
3540.18	3556.33			73.78	73.72
6656.04	6672.45			83.75	83.98
13,392.00	13,392.00			100.00	100.00

Figure 8 shows the classification maps of the single-band remote sensing images using GIC. The histograms of the blue, cirrus and thermal bands are very different. The histograms of the blue band and thermal are, respectively, right and left skewed, whereas the cirrus band is overall Gaussian, though with multiple peaks. Through the GIC classification, different objects can be distinguished by GIC based on the pixel values of the blue and thermal bands, such as water, tree, grass, bare soil, and building. For the cirrus band, GIC can also recognize the pixels that are affected by clouds. It should be noted that the geometric coefficient is expectedly inversed at the peaks of the histogram. As the result, the class of frequent values has a smaller interval, while the class of extreme values has a larger interval. Comparing the quantile method, the formulation and solution of the GIC indeed allows for a more balanced number of samples in each class and at the same time ensures that samples with similar values are classified into the same class. Additionally, the variation of intervals is very smooth (as geometric sequences), making it suitable to classify continuous data, e.g., remote sensing images.

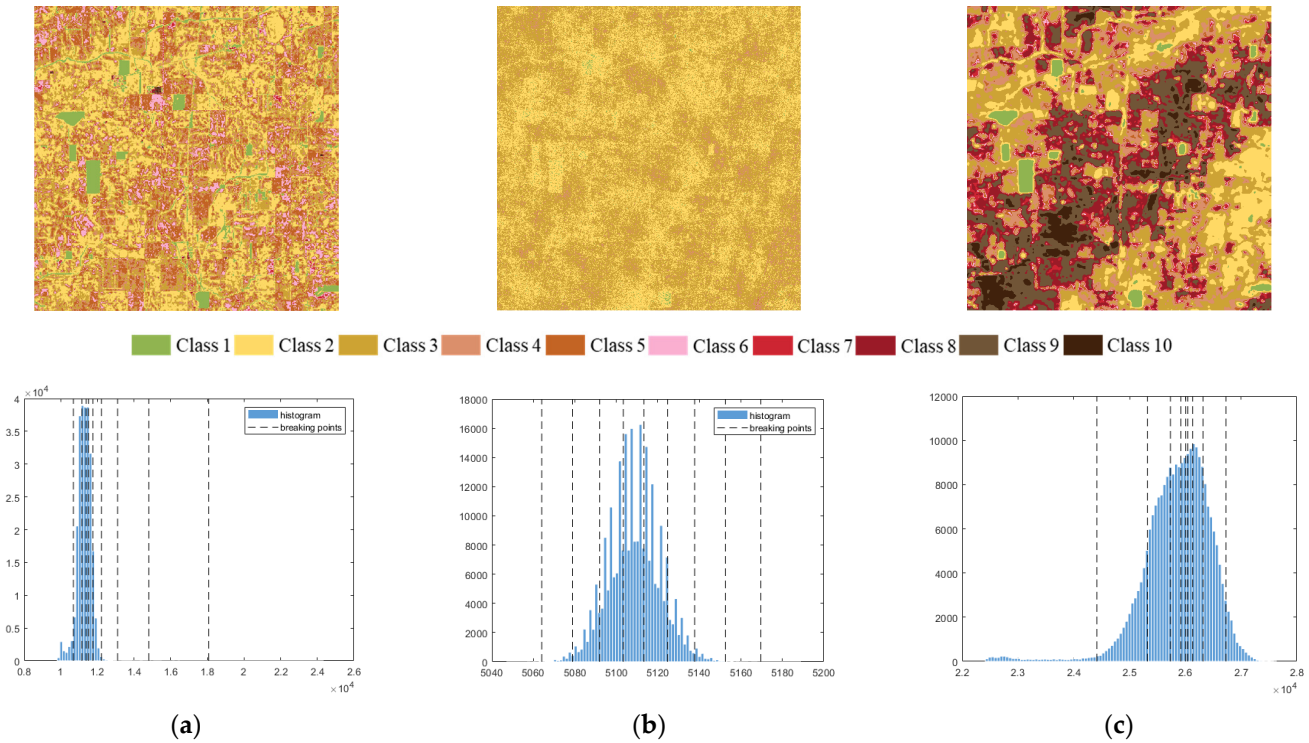


Figure 8. GIC classification results of the Landsat 8 images and corresponding break points for the blue band (**a**), cirrus band (**b**), and thermal band (**c**).

We further examined the performance of GIC in terms of the distribution characteristics of the data. As shown in Table 3, the skewness and kurtosis [30] of the population data and remote sensing images ertr calculated to describe the distribution of the data. The classification performance of GIC was quantitatively evaluated by calculating the between-class (S_B) and within-class scatter (S_W) and their ratios (S_B/S_W). When analyzing the classification results and corresponding skewness and kurtosis, GIC is more performant for data with skewness that is far from 0 (normal distribution) and with kurtosis that differs greatly with 3 (normal distribution). It is noticed that GIC demonstrates superiority and suitability when working with non-normal distributed data, i.e., single-peak and multi-peak data with a completely different skewness and kurtosis from the normal distribution. Both the variations of high frequent values and low frequent extremes can be visible in the classification maps.

Table 3. Fisher's linear discriminant, including S_W, S_B, and S_B/S_W, of the GIC classification results in terms of skewness and kurtosis of the population data and remote sensing images.

	Population			**Landsat 8**		
	Indiana	**Chicago**	**Georgia**	**Blue**	**Cirrus**	**Thermal**
skewness	3.360	−0.030	−0.390	2.820	0.070	−1.520
kurtosis	30.080	2.330	2.360	68.910	3.160	8.520
S_W	11.323	134.140	41.689	12.681	124.240	191.160
S_B	189.130	4897.800	4892.400	183.640	1791.400	3332.200
S_B/S_W	16.703	36.513	117.360	14.481	14.419	174.315

5.2. Multi-Peak Data

When exploring the scenario that the distribution of the data has multiple peaks, we applied the conditions (Equation (5)) to determine whether and where the geometric coefficient q need to be inversed. For data whose distribution has multiple clumps of values (peaks), i.e., the Indiana population data, the thermal band of the Landsat 8 image, and the simulated Georgia population data, the classification maps, and the corresponding break points are shown in Figure 9. It can be noticed that the geometric coefficient q tends to be inversed when there is a peak in the histogram, and the number of inversions is almost equal to the number of peaks of the data. This makes the number of samples for each class more balanced, compared with the results obtained by the current Esri GIC, which inverses q for only once. For example, compared to Figure 8c, the classification map of the thermal band shown in Figure 9c shows more details with more balanced number of samples for each class. The samples of the dominant classes are further classified into different classes. Given the same number of classes, the adaptive GIC for multi-peak data allows for a more detailed consideration of the data distribution while keeping the number of samples in each class balanced.

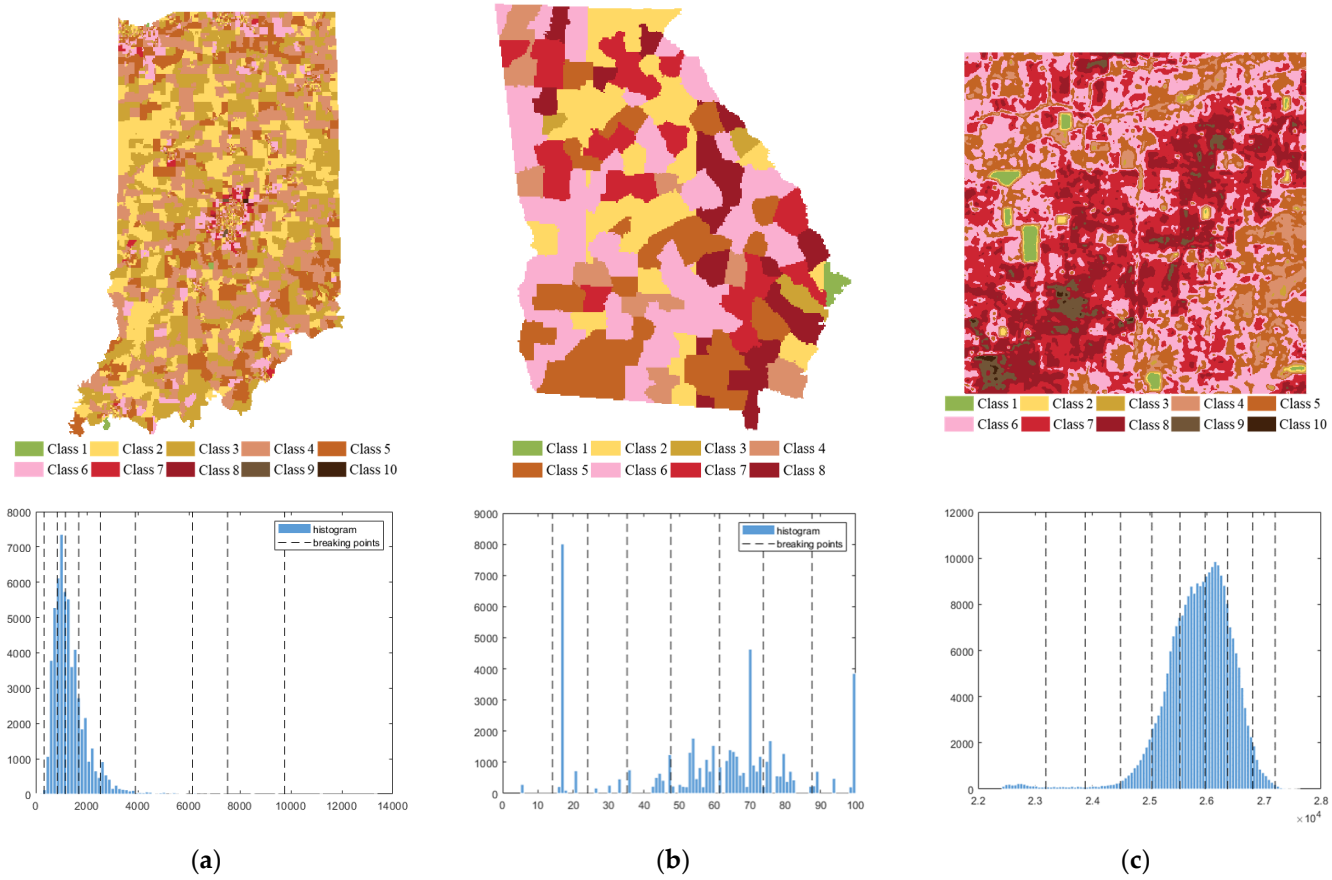


Figure 9. Classification results of the population of Indiana (**a**), the simulated population of Georgia (**b**), the thermal band (**c**), and the break points using adaptive GIC for multi-peak data.

5.3. Multi-Dimensional Data

To deal with multi-dimensional data, we transformed the RGB composite into a new feature space by calculating the Euclidean norm of each pixel. Figure 10 shows the feature map and the classification results using the adaptive GIC. This feature maintains the representative information of higher dimensions, i.e., the spectral domain, and reduce the dimensionality to make GIC applicable. Most objects in this study area can be distinguished, such as water, tree, grass, and bare soil. The Euclidean norm provides an effective way to extract the variations from data values of high dimensions. The variations of both the frequent small data values and less frequent extremes, as shown in Figure 10b, can be highlighted. This result indicates the advantage of the adaptive GIC, which leads to a balanced number of samples for each class as well as high separability of the classes.

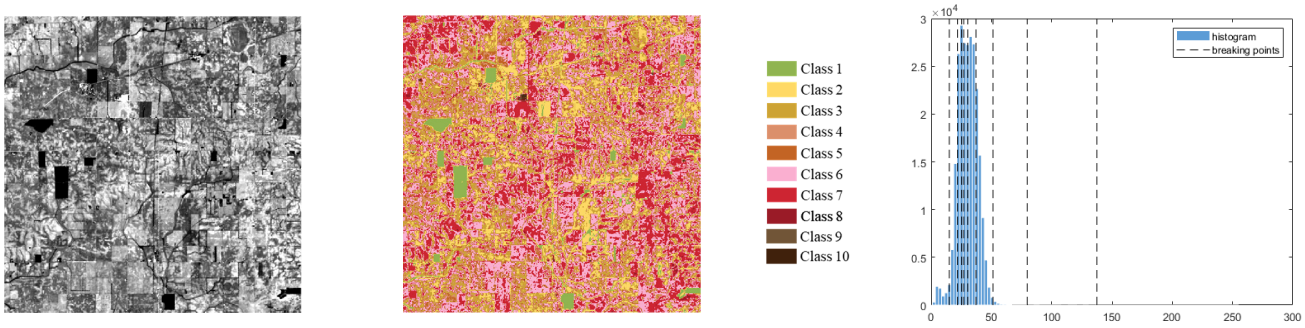


Figure 10. GIC classification results of the Landsat 8 RGB composite. Euclidean norm of the composite (**a**), classification map (**b**), and corresponding breakpoints shown in the histogram (**c**).

The spatial resolution can also affect the classification results. The difference in spatial resolution may lead to a change in data distribution characteristics, which in turn may have an effect in the classification outcome. To investigate the effect of spatial resolution on the performance of GIC, the RGB composition of a Landsat 8 image and an aerial RGB color image for the same study area are selected (Figure 11). This area is part of West Lafayette, Indiana. Both images are scaled to 0–255 gray levels and converted to the feature space in Section 3.2. Figure 11 shows the classification maps and break points on the histograms and Table 4 provides the results of Fisher's discriminant analysis. The skewness for the Landsat 8 image and the aerial image is, respectively, 2.5077 and 0.0899, while the kurtosis is, respectively, 16.1109 and 2.6783. The Landsat 8 medium resolution data are highly right-skewed and have kurtosis that is much larger than 3 (i.e., away from the normal distribution), whereas the high-resolution aerial data have a skewness close to 0 and kurtosis close to 3, which is more similar to the Gaussian distribution. Nevertheless, it can be noticed that the aerial image exhibits multiple peaks in its distribution and the Landsat 8 image has a highly right-skewed, largely single-peak distribution.

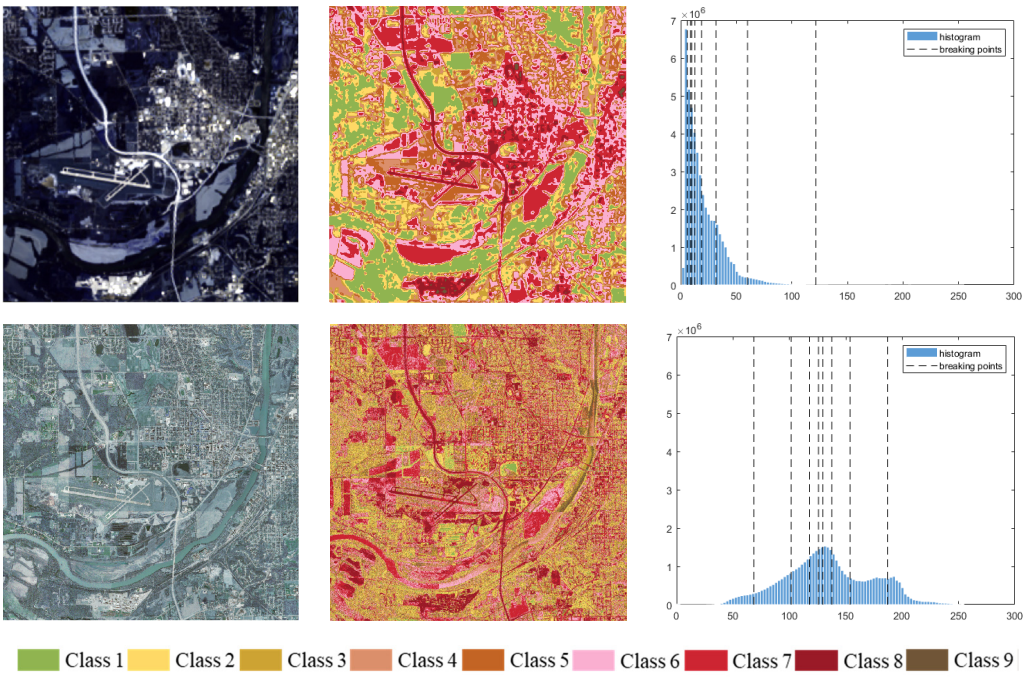


Figure 11. GIC classification results for the Landsat 8 image (**top**) and aerial RGB image (**bottom**) over the West Lafayette area. From left to right: RGB composite, classification result, and corresponding break points on the histograms.

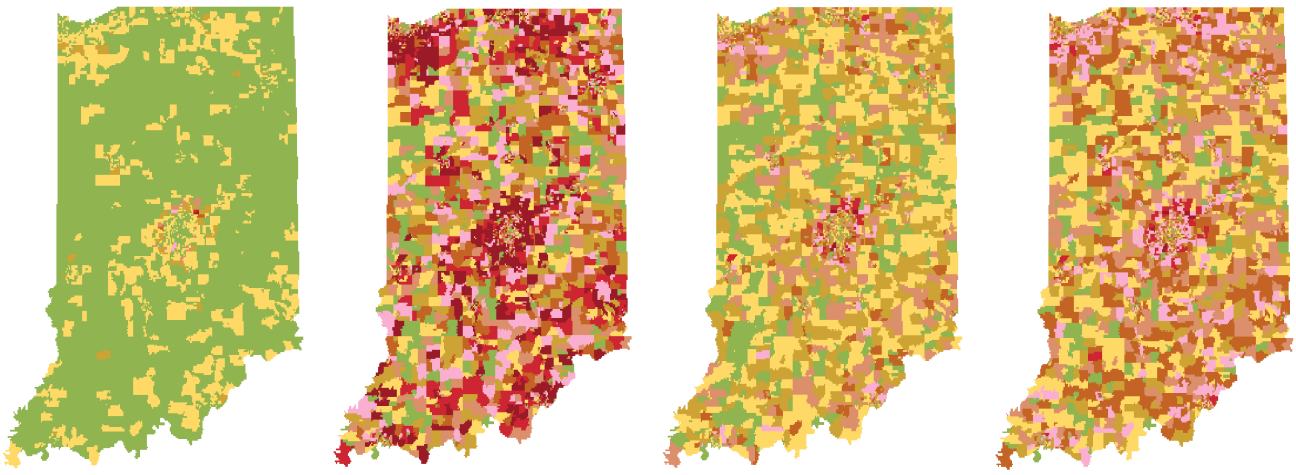
Table 4. Fisher's linear discriminant analysis, including the within-class scatter metric (S_W), between-class scatter metric (S_B), S_B/S_W, and information entropy of the classification results of GIC for the RGB composites of the Landsat 8 and aerial images.

	Landsat 8	**Aerial Image**
S_W	1.392×10^4	4.392×10^4
S_B	2.032×10^5	1.052×10^6
S_B/S_W	1.460×10^1	**23.960**
Entropy	2.839	**3.053**

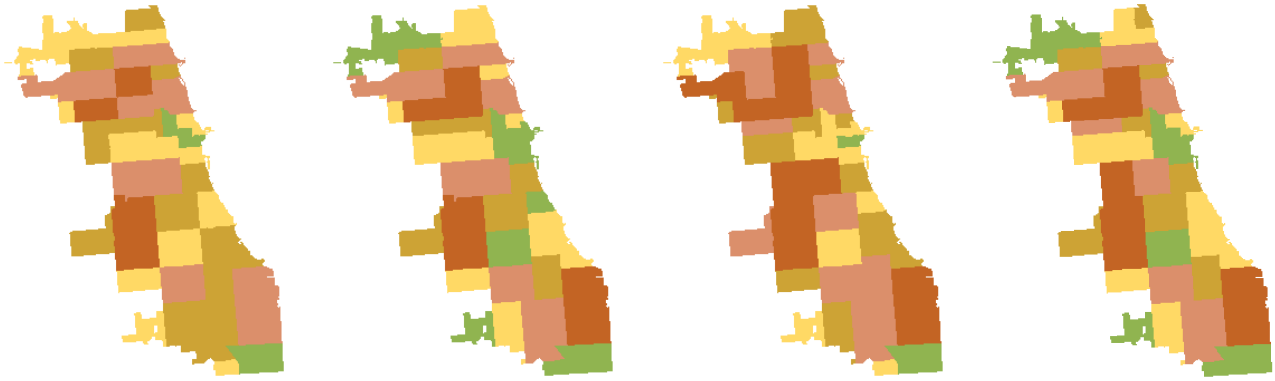
As expected, the high-resolution (aerial image) classification result reveals more details of the objects. Comparing the within- and between-class scatters, as shown in Table 4, the within- and between-class scatters (S_W and S_B) of the high-resolution classification results are higher than those of the medium resolution results (the Landsat 8 image). The ratio between S_W and S_B is also higher for the high-resolution results, indicating a higher separability between different objects. The high-resolution classification result also demonstrates higher information entropy than from the medium resolution image, a fact that means a more balanced number of samples for each class. Comparing the histograms and break points, the distribution of gray values in the Landsat 8 image is highly right skewed with high kurtosis, while that of the aerial image is more symmetric with smaller kurtosis.

5.4. Comparison and Discussion

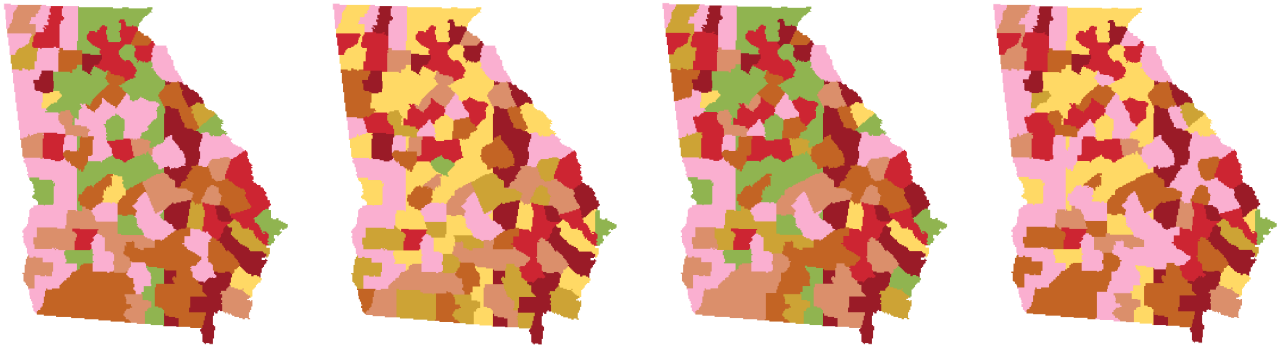
The classification results of comparison methods, including the equal interval, quantile, and natural breaks methods, are provided in Figure 12. The Fisher's linear discriminant analysis [31] was performed to measure the separability of different classes, and the within-class scatter matrix (S_W), between-class scatter matrix (S_B) and S_B/S_W are shown in Table 5. For the classification results of each study dataset, the two highest S_B/S_W ratios are bolded. GIC has a better performance for the population of Indiana, the blue and thermal bands of Landsat 8 image, while the natural breaks method performs best for the population of Chicago and Georgia, as well as the cirrus band. These results verify that the GIC is more appropriate for non-normally distributed data with high skewness and kurtosis. In fact, GIC can be regarded as a compromise of the natural breaks method and quantile. It resolves the problem of the quantile method by taking the distribution characteristics of data into consideration. Additionally, GIC avoids the potential of natural breaks for obtaining class breaks on extreme values with low frequency. The equal interval method achieves better results on the population data of Chicago and Georgia. It often highlights extreme values and de-emphasizes the samples with high frequent values [31,32]. Hence, it is more suitable for data with small skewness, such as the population data of Chicago and Georgia. The classification results of the quantile method are inferior to others for most datasets, except for the cirrus band. These misleading results are because that quantile only considers the number of samples and ignores the distribution of the data.



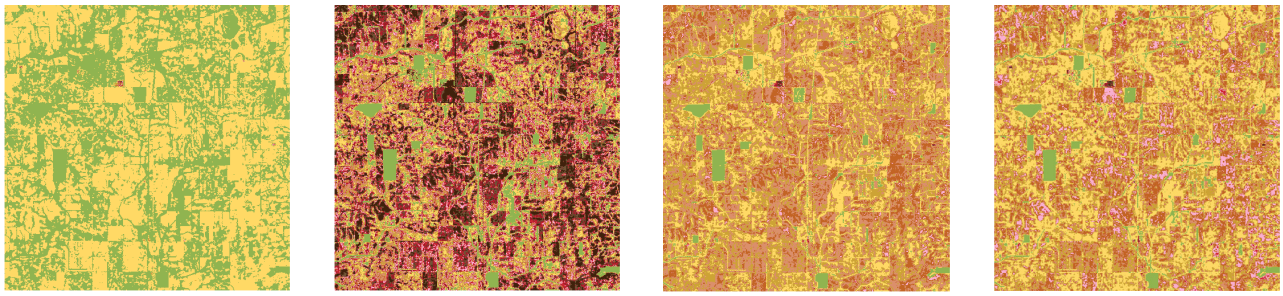
Indiana Population



Chicago Population



Georgia Population



Blue band

Figure 12. *Cont.*

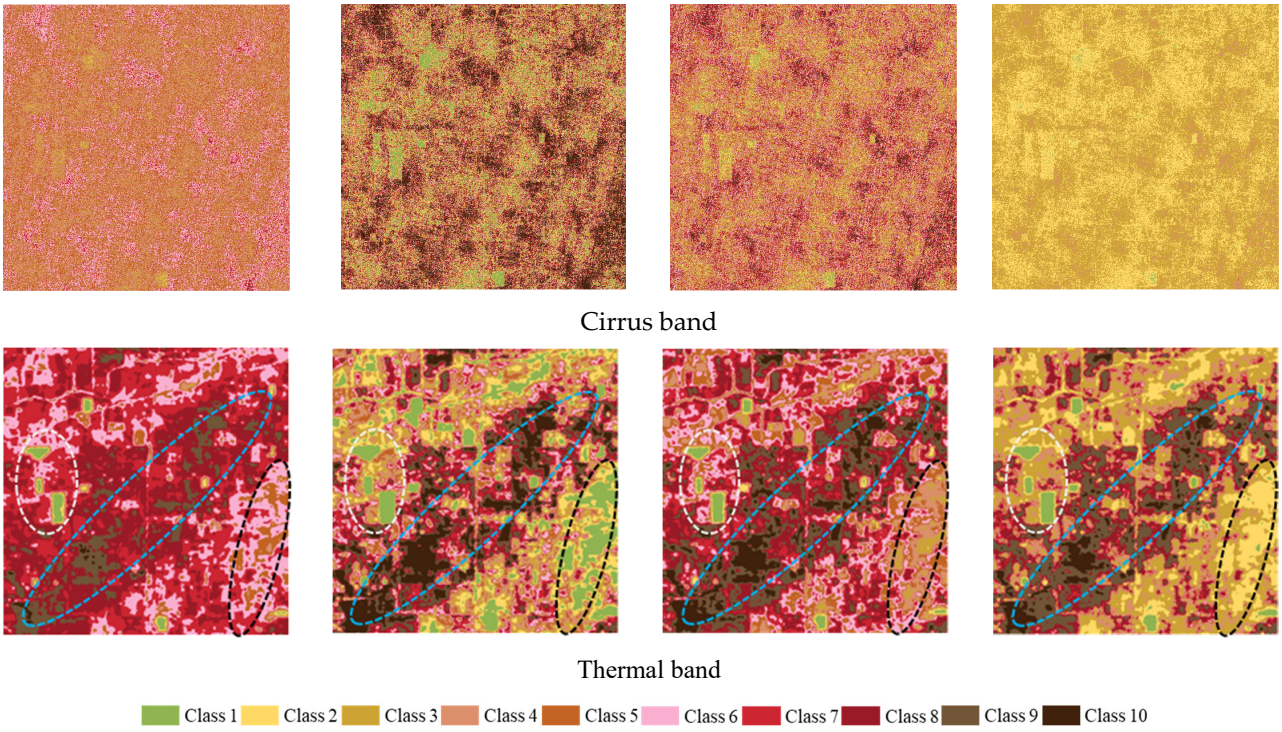


Figure 12. Comparison results using the equal interval (**a**), quantile (**b**), natural breaks (**c**), and GIC (d) methods for the population data of Indiana, Chicago, and simulated population data of Georgia, and the blue, cirrus, and thermal bands of the Landsat image.

Table 5. Fisher's linear discriminants, including the within-class metric (S_W), between-class metric (S_B) and the ratio S_B/S_W for different classification methods.

		Population Indiana	Population Chicago	Population Georgia	Blue Band	Cirrus Band	Thermal Band
equal interval	S_W	3.469×10^1	1.299×10^2	4.915×10^1	6.344×10^1	1.840×10^2	1.926×10^2
	S_B	1.658×10^2	4.902×10^3	4.885×10^3	1.329×10^2	1.732×10^3	3.331×10^3
	S_B/S_W	4.778	$\mathbf{3.773 \times 10^1}$	$\mathbf{9.939 \times 10^2}$	2.095	9.413	1.729×10^1
quantile	S_W	3.881×10^1	1.313×10^2	1.689×10^2	4.566×10^1	8.382×10^1	5.143×10^2
	S_B	1.617×10^2	4.901×10^3	4.765×10^3	1.507×10^2	1.832×10^3	3.009×10^3
	S_B/S_W	4.165	3.731×10^1	2.821×10^1	3.300	$\mathbf{2.185 \times 10^1}$	5.851
natural breaks	S_W	7.712	1.239×10^2	3.721×10^1	1.241×10^1	5.022×10^1	8.614×10^1
	S_B	1.928×10^2	4.908×10^3	4.897×10^3	1.839×10^2	1.865×10^3	3.437×10^3
	S_B/S_W	$\mathbf{2.499 \times 10^1}$	$\mathbf{3.960 \times 10^1}$	$\mathbf{1.316 \times 10^2}$	$\mathbf{1.482 \times 10^1}$	$\mathbf{3.715 \times 10^1}$	$\mathbf{3.990 \times 10^1}$
GIC	S_W	1.132×10^1	1.341×10^2	4.169×10^1	1.268×10^1	1.242×10^2	1.912×10^2
	S_B	1.891×10^2	4.898×10^3	4.892×10^3	1.836×10^2	1.791×10^3	3.332×10^3
	S_B/S_W	1.670×10^1	3.651×10^1	1.174×10^2	$\mathbf{1.448 \times 10^1}$	1.442×10^1	$\mathbf{1.743 \times 10^1}$

Furthermore, the information entropy was also calculated for the classification results obtained by GIC and three other comparison methods, as shown in Table 6. The theory of maximum entropy is that a uniform distribution has the largest entropy [33,34]. In the classification problem, the information entropy is maximal when the data are classified into several classes with equal probability and the size of each class is the same. Given a probability distribution, the information entropy is defined as $H_p = -p(x)\lg p(x)$, where $p(x)$ represents the probability of occurrence of x [35]. The classification results of GIC and quantile have the highest information entropy, which illustrates that the number of samples

in each class is balanced. In contrast to the quantile method, GIC takes into account the distribution of the data in addition to the size of classes. The inverse of the geometric coefficient highlights the variation in both the middle values and the extreme values and can reduce the intra-class difference in extreme values. This also indicates that GIC can be viewed as a compromise between the quantile and natural breaks methods.

Table 6. Information entropy of the classification results from GIC and comparison methods.

	Population			Landsat Band		
	Indiana	**Chicago**	**Georgia**	**Blue**	**Cirrus**	**Thermal**
equal interval	0.953	1.439	2.186	1.019	1.940	2.179
quantile	**2.557**	**1.530**	2.257	**3.322**	**3.316**	**3.322**
natural breaks	2.017	1.449	**2.319**	2.189	2.939	2.891
GIC	**2.223**	**1.530**	**2.355**	**2.387**	**2.239**	**2.997**

Finally, we examined the semantic difference between the classification results for the thermal infrared band of the Landsat 8 image. In the thermal infrared band, the gray values can be converted to the top of atmospheric (TOA) spectral radiance and then the brightness temperature of the ground. There is a positive correlation between them, with larger values corresponding to higher temperatures [36,37]. Class 1–10 correspond to the gray values from low to high, respectively, and the semantic information of the classification maps is the distribution of the brightness temperature. As shown in Figure 9c, the histogram of the thermal band image has a peak near 2.61×10^4, with very few objects having low and high brightness temperatures (pixel values less than 2.45×10^4 and greater than 2.70×10^4). The equal interval method classifies most of the data samples into the average and above average class (class 6,7,8), which fails to show the detailed differences in brightness temperatures and made the overall temperatures in the study area appear high. The intervals of the classes with average temperatures obtained by natural breaks are smaller than those of the classes with extreme values, but still many high frequent pixels are classified into one class (class 6 or 7). To equalize the size of each class, the quantile method increases the break corresponding to class 1 (lowest temperature), resulting in many objects with below average temperatures being assigned to class 1. Meanwhile, the class breaks obtained by quantile are concentrated around the peak of the histogram, as shown in Figure 12c. Since GIC takes into account the data distribution while balancing the number of samples for each class, the breaks obtained by GIC are different from the quantified ones. As shown in Figure 12e, the interval between two adjacent breaks gradually increases from the peak of the histogram to both sides, thus increasing the number of pixels assigned to class 2, 3, 9. Compared to the classification maps of quantile and GIC (Figure 12b,d), more pixels are classified into class 6, 7, 8 with higher brightness temperatures when using the equal interval and natural breaks methods (Figure 12a,c). Additionally, the pixels with higher temperatures are always assigned larger class labels than those with lower temperatures.

In the images shown in the last row of Figure 12, water bodies have the lowest brightness temperature and are classified into class 1, as marked by the white cycles. The bare soil in the area marked by blue cycles is assigned to class 10 or 9, indicating that bare soil has a higher brightness temperature than the surrounding vegetation. The area marked by the black circle is a typical example of the difference between the four methods. This area is mainly composed of vegetation and a little bare soil. The classification map of equal interval shows the highest brightness temperature (mostly class 5–8), followed by the one of natural breaks (mostly class 3 and 6). The quantile method classifies many pixels in this area into the class with the lowest temperature (class 1), and GIC assigns these pixels into class 2 and 3 of higher brightness temperature.

6. Conclusions

This study theoretically formulated the geometric interval classification (GIC) method proposed by Esri and extended it to handle multi-peak, multi-dimensional data. The essence is that GIC is described as a constrained optimization problem that can be solved iteratively. In our formulation, the objective function is to minimize the squared sum of the number of data samples in each class, while the introduced constraint on geometric coefficient considers the non-uniform distribution of the data. Under this theoretical formulation, the class intervals determined by GIC is larger for extremes and smaller for frequent values in the data. In this way, the solution of this optimization problem, which is the results of GIC, can be a compromise of the quantile and natural breaks methods. As such, the separability between different classes is improved while maintaining a balanced number of samples in each class. Both the variations of high frequent values and low frequent extremes are emphasized, making the classification method suitable for data not only with common normal distribution, but with skewed long tails, which is rather common for geospatial big data. In fact, we demonstrated that GIC is more performant for non-normal data with a distribution that is asymmetric and different from the Gaussian distribution.

To further extend such formulated GIC to handle data with complex distributions, such as multi-peaks in multi-dimensions, we modified the settings of the geometric coefficient. For multi-peak data, the number of times of the inversion for the geometric coefficient is automatically determined. Experiments on the population data and remote sensing images show that the adaptive GIC can better classify multi-peak data, as it allows for a more detailed consideration of the data distribution characteristics. Moreover, the number of inversions of the geometric coefficient is approximately equal to the number of peaks in the data. This demonstrates the capability of adaptive GIC to modify class intervals according to the distribution of the data. The sizes of class are balanced, and the separability of different classes is improved, leading to a more appealing classification map for multi-peak data. To handle multi-dimensional data, we transformed the input data into a feature space based on the Euclidean norm, which can maintain the variation of high dimensional data. The classification results on the RGB composites demonstrate the feasibility and effectiveness of this strategy and the adaptability of GIC to classifying multi-dimensional data.

Three other classification methods were applied for comparison, i.e., the equal interval, quantile, and natural breaks methods. The highest separability of classes is achieved by natural breaks and GIC, while at the same time, GIC can also obtain classification maps with the highest entropy. GIC avoids the limitations of the comparison methods and achieves a compromise. Comparing to the natural breaks method, GIC is able to create more class breaks in the high frequent data ranges, and fewer classes at extremes. It can highlight both the variations of high frequent data and low frequent extremes to achieve more realistic results.

It should be reiterated that classification is a highly data- and domain-dependent process. The classification of geospatial data further adds several other dimensions to the equation. Such dimensions, among others, may include the size and shape of the geographic units based on which the data to be classified are aggregated, as documented by [38] as the modifiable areal unit problem. Similarly, when working with geographic data with social and economic attributes (e.g., the population data in this study), their classification and interpretation are associated with certain semantic and nominal or categorical definitions, which can be subjective. As such, the interpretation of the GIC classifier may not be as straightforward as several other methods, which can be regarded as a limitation.

We expect the developed adaptive GIC framework can be adopted for thematic mapping and other geospatial data classification and visualization practices. Theoretically, some of the properties of the GIC method are worthwhile of further in-depth investigation from an analytic point view, which may include the convergent geometric coefficient under varying number of classes, and how such convergence is coherent with the physical nature of the data to be classified. Moreover, our study demonstrated the feasibility of using GIC

for remote sensing image classification only through a simple n-dimension to 1-dimension mapping or transform. This work is preliminary, which needs to be more comprehensively evaluated with reference to other popular image classification methods. Furthermore, the feature being classified is 1D. Efforts are still needed to establish an inherently multiple dimensional geometric interval classifier or what adopts and extends the intrinsic idea of this work.

Author Contributions: Conceptualization, investigation, methodology, validation, Shuang Li and Jie Shan; writing—original draft preparation, Shuang Li; writing—revision and editing, Jie Shan. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Shuang Li was funded by the Purdue University Frederick N. Andrews Fellowship. Xiangxi Tian was involved in the discussion of the mathematic formulation of the methodology.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aggarwal, C.C. Data classification. In *Data Mining*; Springer: Cham, Switzerland, 2015; pp. 285–344.
2. Coulson, M.R. In the matter of class intervals for choropleth maps: With particular reference to the work of George F Jenks. *Cartographica* **1987**, *24*, 16–39. [CrossRef]
3. Evans, I.S. The selection of class intervals. *Trans. Inst. Br. Geogr.* **1977**, *2*, 98–124. [CrossRef]
4. Jenks, G.F. The data model concept in statistical mapping. *Int. Yearb. Cartogr.* **1967**, *7*, 186–190.
5. Alexander, J.W.; Zahorchak, G.A. Population-density maps of the United States: Techniques and patterns. *Geogr. Rev.* **1943**, *33*, 457–466. [CrossRef]
6. Smith, R.M. Comparing traditional methods for selecting class intervals on choropleth maps. *Prof. Geogr.* **1986**, *38*, 62–67. [CrossRef]
7. Costache, R.; Hong, H.; Pham, Q.B. Comparative assessment of the flash-flood potential within small mountain catchments using bivariate statistics and their novel hybrid integration with machine learning models. *Sci. Total Environ.* **2020**, *711*, 134514. [CrossRef] [PubMed]
8. Jiang, B.; Liu, X. Scaling of geographic space from the perspective of city and field blocks and using volunteered geographic information. *Int. J. Geog. Inf. Sci.* **2012**, *26*, 215–229. [CrossRef]
9. Esri. 2014. Classifying Numerical Fields for Graduated Symbology. Available online: https://desktop.arcgis.com/en/arcmap/latest/map/working-with-layers/classifying-numerical-fields-for-graduated-symbols.htm (accessed on 10 March 2022).
10. Esri. 2018. Standard Classification Schemes. Available online: http://webhelp.esri.com (accessed on 15 January 2022).
11. Campbell, J.E.; Sedani, A.E.; Dao, H.D.N.; Sambo, A.B.; Doescher, M.P.; Janitz, A.E. Investigation of geographical disparities: The use of An interpolation method for cancer registry data. *Res. Sq.* **2021**, Preprint. [CrossRef]
12. Huan, H.; Wang, J.; Teng, Y. Assessment and validation of groundwater vulnerability to nitrate based on a modified DRASTIC model: A case study in Jilin City of northeast China. *Sci. Total Environ.* **2012**, *440*, 14–23. [CrossRef]
13. Li, Y.; Gemert, V.J. Deep unsupervised image hashing by maximizing bit entropy. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; pp. 2002–2010.
14. Baldwin, R.A. Use of maximum entropy modeling in wildlife research. *Entropy* **2009**, *11*, 854–866. [CrossRef]
15. Johnston, K.; Ver Hoef, J.M.; Krivoruchko, K.; Lucas, N. *Using ArcGIS Geostatistical Analyst*; Esri: Redlands, CA, USA, 2001.
16. Aimrun, W.; Amin, M.S.M.; Ezrin, M.H. Small scale spatial variability of apparent electrical conductivity within a paddy field. *Appl. Environ. Soil Sci.* **2009**, *2009*, 7. [CrossRef]
17. Khosravi, K.; Nohani, E.; Maroufinia, E.; Pourghasemi, H.R. A GIS-based flood susceptibility assessment and its mapping in Iran: A comparison between frequency ratio and weights-of-evidence bivariate statistical models with multi-criteria decision-making technique. *Nat. Hazard.* **2016**, *83*, 947–987. [CrossRef]
18. Tang, X.; Machimura, T.; Liu, W.; Li, J.; Hong, H. A novel index to evaluate discretization methods: A case study of flood susceptibility assessment based on random forest. *Geosci. Front.* **2021**, *12*, 101253. [CrossRef]
19. Ajibade, F.O.; Ajibade, T.F.; Idowu, T.E.; Nwogwu, N.A.; Adelodun, B.; Lasisi, K.H.; Adewumi, J.R. Flood-prone area mapping using GIS-based analytical hierarchy frameworks for Ibadan city, Nigeria. *J. Multi-Criteria Decis. Anal.* **2021**, *28*, 283–295. [CrossRef]

20. Liu, H.; Zhan, Q.; Zhan, M. The uncertainties on the GIS based land suitability assessment for urban and rural planning. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.-ISPRS Arch.* **2017**, *XLII-2/W7*, 42.
21. Melo, S.N.D.; Frank, R.; Brantingham, P. Voronoi diagrams and spatial analysis of crime. *Prof. Geogr.* **2017**, *69*, 579–590. [CrossRef]
22. Al-Abadi, A.M.A.; Ghalib, H.B.; Al-Qurnawi, W.S. Estimation of soil erosion in northern Kirkuk governorate, Iraq using rusle, remote sensing and gis. *Carpathian J Earth Environ Sci.* **2016**, *11*, 153–166.
23. Lu, Y.; He, T.; Xu, X.; Qiao, Z. Investigation the Robustness of Standard Classification Methods for Defining Urban Heat Islands. *IEEE J. Sel. Top. Appl.Earth Obs. Remote Sens.* **2021**, *14*, 11386–11394. [CrossRef]
24. Francisci, D. A Python Script for Geometric Interval Classification in QGIS: A Useful Tool for Archaeologists. *Environ. Sci. Proc.* **2021**, *10*, 1.
25. Hwang, C.L.; Tillman, F.A.; Kuo, W. Reliability optimization by generalized Lagrangian-function and reduced-gradient methods. *IEEE Trans. Reliab.* **1979**, *28*, 316–319. [CrossRef]
26. Canchola, J.A.; Tang, S.; Hemyari, P.; Paxinos, E.; Marins, E. Correct use of percent coefficient of variation (% CV) formula for log-transformed data. *MOJ Proteom. Bioinform* **2017**, *6*, 316–317. [CrossRef]
27. Celebi, M.E.; Celiker, F.; Kingravi, H.A. On Euclidean norm approximations. *Pattern Recognit.* **2011**, *44*, 278–283. [CrossRef]
28. Hennig, C.; Viroli, C. Quantile-based classifiers. *Biometrika* **2016**, *103*, 435–446. [CrossRef] [PubMed]
29. Brewer, C.A.; Pickle, L. Evaluation of methods for classifying epidemiological data on choropleth maps in series. *Ann. Am. Assoc. Geogr.* **2002**, *92*, 662–681. [CrossRef]
30. Blanca, M.J.; Arnau, J.; López-Montiel, D.; Bono, R.; Bendayan, R. Skewness and kurtosis in real data samples. *Methodology* **2013**, *9*, 78–84. [CrossRef]
31. Mika, S.; Ratsch, G.; Weston, J.; Scholkopf, B.; Mullers, K.R. Fisher discriminant analysis with kernels. In Proceedings of the Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Work-Shop (cat. no. 98th8468), Madison, WI, USA, 25 August 1999; pp. 41–48.
32. Xanthopoulos, P.; Pardalos, P.M.; Trafalis, T.B. Linear discriminant analysis. In *Robust Data Mining*; Springer: New York, NY, USA, 2013; pp. 27–33.
33. Nigam, K.; Lafferty, J.; McCallum, A. Using maximum entropy for text classification. In Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering, Stockholm, Sweden, 1 August 1999; Volume 1, pp. 61–67.
34. Li, D.; Guan, Y.; Gong, J.; Du, D. Entropy error model of planar geometry features in GIS. *Geo-Spat. Inf. Sci.* **2003**, *6*, 20–24.
35. Shannon, C.E. A mathematical theory of communication. *Bell Labs Tech. J.* **1948**, *27*, 379–423. [CrossRef]
36. How Jin Aik, D.; Ismail, M.H.; Muharam, F.M. Land use/land cover changes and the relationship with land surface temperature using Landsat and MODIS imageries in Cameron Highlands, Malaysia. *Land* **2020**, *9*, 372. [CrossRef]
37. Kumar, B.P.; Babu, K.R.; Anusha, B.N.; Rajasekhar, M. Geo-environmental Monitoring and Assessment of Land Degradation and Desertification in the Semi-arid regions using Landsat 8 OLI/TIRS, LST, and NDVI approach. *Environ. Chall.* **2022**, *8*, 100578. [CrossRef]
38. O'Sullivan, D.; Unwin, D. *Geographic Information Analysis*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, Canada, 2010; pp. 30–32.