International Journal of Geo-Information

MDPI

Article

Method for Determining Appropriate Clustering Criteria of Location-Sensing Data

Youngmin Lee [1], Pil Kwon [1], Kiyun Yu [1,2] and Woojin Park [3,*]

[1] Department of Civil and Environmental Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea; daldanka@snu.ac.kr (Y.L.); pil0706@snu.ac.kr (P.K.); kiyun@snu.ac.kr (K.Y.)

[2] Institute of Construction and Environmental Engineering, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 08826, Korea

[3] Korea Land and Geospatial Informatix Corporation, 120, Giji-ro, Jeonju-si 54870, Korea

* Correspondence: wjpark@lx.or.kr; Tel.: +82-63-906-5090

Academic Editors: Mahmoud R. Delavar and Wolfgang Kainz
Received: 13 June 2016; Accepted: 19 August 2016; Published: 25 August 2016

Abstract: Large quantities of location-sensing data are generated from location-based social network services. These data are provided as point properties with location coordinates acquired from a global positioning system or Wi-Fi signal. To show the point data on multi-scale map services, the data should be represented by clusters following a grid-based clustering method, in which an appropriate grid size should be determined. Currently, there are no criteria for determining the proper grid size, and the modifiable areal unit problem has been formulated for the purpose of addressing this issue. The method proposed in this paper is applies a hexagonal grid to geotagged Twitter point data, considering the grid size in terms of both quantity and quality to minimize the limitations associated with the modifiable areal unit problem. Quantitatively, we reduced the original Twitter point data by an appropriate amount using Töpfer's radical law. Qualitatively, we maintained the original distribution characteristics using Moran's *I*. Finally, we determined the appropriate sizes of clusters from zoom levels 9–13 by analyzing the distribution of data on the graphs. Based on the visualized clustering results, we confirm that the original distribution pattern is effectively maintained using the proposed method.

Keywords: clustering; LBSN; Twitter; MAUP; Moran's *I*; Töpfer's radical law

1. Introduction

Map generalization refers to the process of expressing point features as clusters by merging points on a multi-scale map. Map generalization is typically considered in terms of both quantity and quality. In quantitative terms, the point data must be reduced by the correct amount when zooming out on the map whereas, qualitatively, the data should maintain its original distribution characteristics [1].

In this study, we consider the point clustering process from the perspective of a map point feature-generalization process. In the map generalization field, studies on line generalization [2–7] or polygon generalization [8–11] have been conducted for some time. However, studies on point generalization have been minimally performed [1], because most web map components are linear features, such as roads and streams, or polygonal features, such as buildings and parcels. Point features, on the other hand, tend to be to be considered of minimal importance.

In this regard, Yu [1] focused on the distribution pattern of point features and found the distribution characteristics of the original point data through quadrant analysis, nearest neighbor analysis, and the K-function. Redundant point features were removed using a generalization threshold that could maintain the original distribution pattern. A selection and elimination method was applied

to the point data to generalize a small-scale map from a large-scale map. Thus, only the distribution pattern of the point data was analyzed, whereas grids were not considered in the representations of the point features, in contrast to the methodology proposed here.

With the development of the global positioning system, there has been a significant increase in location-determination technologies and smart devices, as well as increased use of various location-sensing data generated from location-based social network (LBSN) services, such as Twitter, Facebook, and Instagram. Such location-sensing data are provided as points with single x, y coordinates acquired from a global positioning system or Wi-Fi signal that describes the location from which the data were generated. When visualizing these point data on multi-scale map services, the readability of information by users decreases during map zoom-out actions because of the overlapping data. Therefore, the point data should be represented by clusters to improve the legibility and communicability of the information.

LBSN data are a relatively new form of data with unique characteristics compared to existing general spatial point data, such as weather, pollution, and population data. LBSN data are likely to be generated in populous areas or city centers where there is a high probability that many users are gathered. Thus, these data tend to be more clustered in particular regions compared to other data and accumulate considerably in real time. Therefore, a clustering methodology that is customized for these data characteristics is required to represent these data in map services.

The k-means clustering and grid-based clustering algorithms are generally used to cluster point data. Following the k-means method, researchers specify the number of clusters (k) in advance, and it is, therefore, difficult to apply the method to non-convex shapes or very different size of clusters. Following the grid-based clustering approach, an object space is created in a finite number of spaces comprising a lattice structure, and all clustering processes are implemented within the structure; this clustering method is independent of the number of data objects, depends only on the number of cells, and uses the centroid of each tile [12]. However, the grid-based method cannot maintain the original data distribution characteristics because it uses a predetermined grid size for each zoom level.

Therefore, in this paper, we propose a methodology that retains the distribution pattern of the original data by varying the grid size based on the zoom level. To perform grid-based clustering of LBSN data, a suitable grid size must be determined for each zoom level. However, no criteria have yet been presented to determine the appropriate grid size. Moreover, because the size is dependent on the data's characteristics and purpose, the size determination inherently involves researcher subjectivity. In such a case, the modifiable areal unit problem (MAUP) occurs with a high probability of affecting the analysis result, as will be discussed in more detail in Section 2.

The purpose of this study is to determine the appropriate sizes of clusters for different zoom levels, considering both quantitative and qualitative aspects while minimizing the MAUP effect. We propose a methodology that determines the appropriate sizes of geotagged Twitter point data clusters from one side length of the hexagons (h) in latticed grids that are established for this purpose. Following the proposed methodology, the terms 'cluster' and 'hexagon' represent the same meaning in this study. To consider the quantitative data characteristics, the proposed method determines the proper number of clusters for different zoom levels by using Töpfer's radical law, which is a mathematical model for calculating the appropriate number of map objects. In addition, qualitative data characteristics are considered using Moran's I as a measure of spatial autocorrelation to identify the distribution characteristics of the clustered data.

The remainder of this paper is organized as follows: Section 2 details the MAUP, and Section 3 describes the theoretical basis of the main methodologies—Töpfer's radical law and Moran's I—used in this study; Section 4 discusses the technique used to determine the appropriate size of clusters according to zoom-levels as well as the results when applied to a real dataset, and Section 5 derives the conclusions and meaning of this study.

2. Modifiable Areal Unit Problem (MAUP)

The study presented in this paper focuses on the MAUP scale effect, whereby we analyze the distribution characteristics resulting from changes in the grid sizes. In this regard, He et al. [13] used the MAUP scale effect to search for the optimal scale that most closely represents the actual spatial distribution pattern of plant communities. In addition, Viegas et al. [14] analyzed the MAUP scale effect for variables relevant to minimizing the influence of the MAUP effect when creating a traffic analysis zone, and Swift et al. [15] attempted to determine the impact of spatial aggregation and MAUP on the correlation between potable water quality and stomach problems, developing nine different spatial units to analyze both scale and zoning effects.

Such studies on the impacts of the MAUP operation mechanism on spatial and statistical analyses have been conducted in various fields for some time. Nevertheless, the practice has not yet been applied to LBSN data because, despite the steady progress of MAUP-related studies since the 1980s, LBSN services were not created until the early 2000s, after the introduction of smart devices and social media. Consequently, the significant contribution of this study is the investigation of the MAUP effect on new types of data from the scale effect perspective. Furthermore, we reveal that the MAUP effect should be considered when analyzing area-based location-sensing data.

To address the stated objective, original point data should be combined with polygons for analysis based on spatial units that are defined by a particular characteristic, such as regional population sizes or employment rates. Researchers must select a proper size and shape of the spatial units for analysis, or create new spatial units by adjusting the existing units [16]. However, no specific criteria for determining the appropriate size and shape of the spatial units exist.

The problem of constructing a new spatial unit is closely related to the issue of spatially-aggregating small scales in relation to large scales. For instance, the same area can be divided in various ways according to the aggregation scheme. That is, an area can be represented at different scales, and areas consisting of different zones can be generated at the same scale [16].

This problem is referred to as MAUP. According to Openshaw, "The selection of areal units, or zoning systems, cannot therefore be separate from, or independent of, the purpose and process of a particular spatial analysis" [17]. MAUP shows that the choice of spatial unit influences the results.

MAUP arises because most spatial units are variable in arbitrary, limited states. Therefore, spatial units can be aggregated or transformed to create different scales or zones [18]. That is, MAUP has two perspectives: the scale effect and the zoning effect. With the scale effect, the result changes when analysis is conducted for the different spatial scales. With the zoning effect, different results are generated by regrouping the spatial units [19]. An example of the two MAUP effects is shown in Figure 1. Figure 1a illustrates population numbers per spatial units, and Figure 1b illustrates the number of unemployed people. Figure 1c,d both represent the MAUP effects on the unemployment rate determination, in which the values at the same location differ because of the scale and zoning effects, respectively. With the scale effect, the total number of spatial units is different and the size of the spatial units is the same. On the other hand, with the zoning effect, the total number of spatial units is the same, and the size of the spatial units is different.

The most critical reason for the occurrence of MAUP is the variance and covariance of variables when spatial units are aggregated on account of the scale and zoning effects. That is, the variance of the variables generally decreases when spatial units are aggregated because of the smoothing effect. For example, numerical outliers tend to converge toward average values during aggregation because they are combined with other spatial units [16].

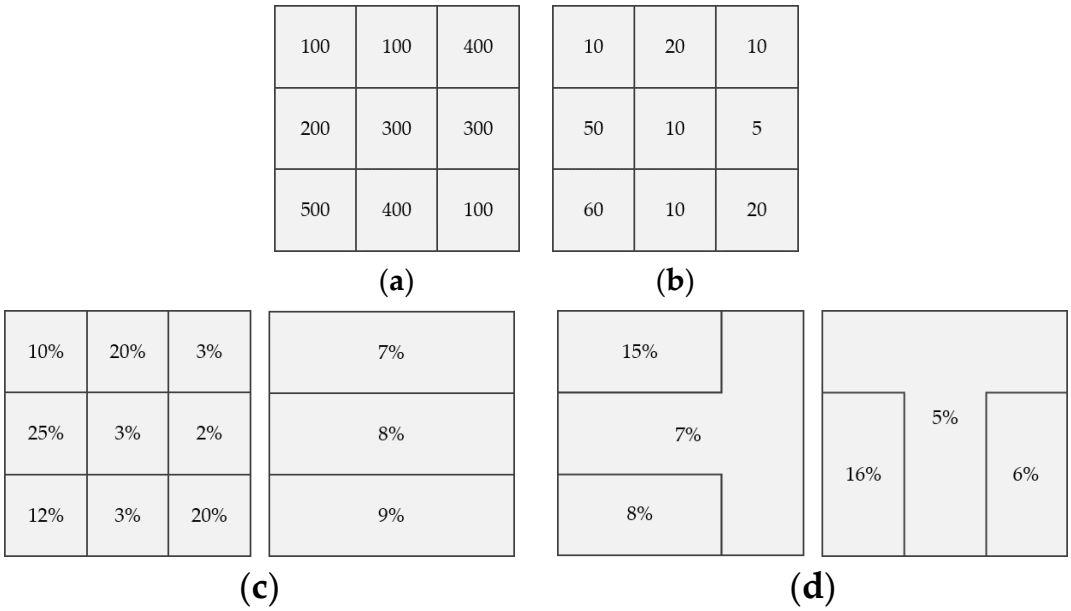

Figure 1. Example of the MAUP scale effect and zoning effect: (**a**) base population; (**b**) unemployed count; (**c**) scale effect; and (**d**) zoning effect [20].

3. Methodology

The purpose of this study is to provide a methodology for determining the appropriate clustering criteria of location-sensing point data using zoom levels to visualize the data on a multi-scale map service. In this approach, we collect geotagged Twitter messages using the Twitter open application programming interface (API). A distribution characteristic of the original Twitter point data is then analyzed using quadrant analysis (QA) and nearest neighbor analysis (NNA). This result is later utilized as a reference for evaluating Moran's I calculations, as will be discussed in more detail in Section 4.2. Then, the appropriate number of points for each zoom level is calculated using Töpfer's radical law. We consider the number of points as the appropriate number of clusters. Thereafter, the ranges of the hexagonal side length h are calculated, and various sizes of hexagonal grids are, thus, created. These grid data and geotagged Twitter point data are spatially joined by zoom levels. After that, Moran's I of the spatially-joined results is calculated to analyze the distribution characteristics for each range. Finally, the appropriate value of h is determined by maintaining the original distribution characteristics of the previously calculated Twitter point data. This process is shown in Figure 2.

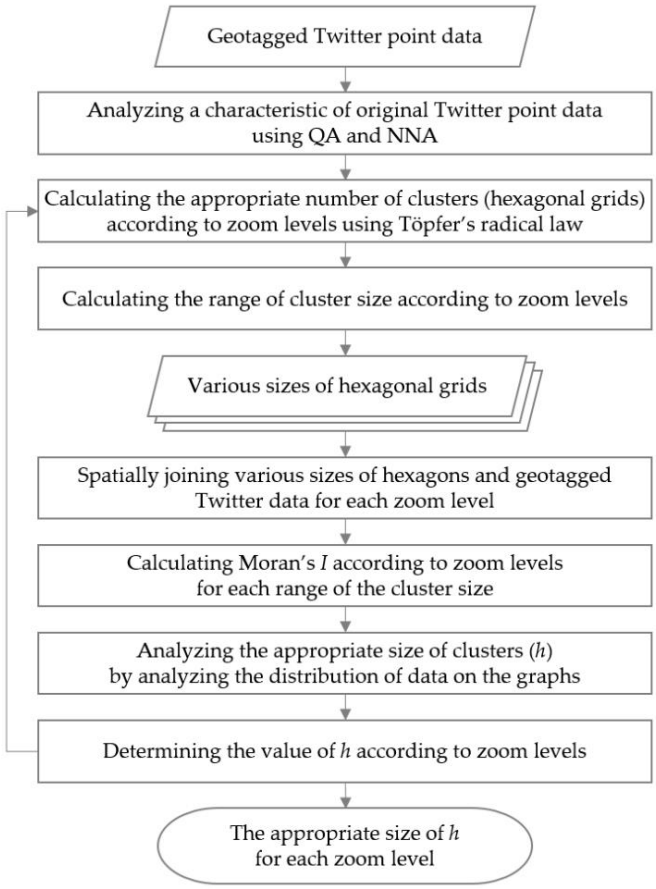

Figure 2. Flowchart of the process of determining the appropriate sizes of clusters.

3.1. Determining the Appropriate Number of Clusters Using Töpfer's Radical Law

The selection and elimination operator of the map generalization field refers to the process used to select and eliminate objects that require no expression on the map at a specific level. The most important task is to determine how many objects will remain. When selecting and eliminating some of the objects, we must consider the degree of importance of the objects, such as their geometry, semantics, and distribution [21].

Determining the number of objects that will be selected and eliminated depends on the purpose of the map, target scale level, and researcher intention. Töpfer's radical law is used in this case because it is the mathematical model that is used to calculate the number of objects or features to be selected based on the source and derived scale. That is, it calculates how many objects should be remained at smaller scales in the map generalization process. Töpfer's radical law can be calculated using Equation (1) [22]:

$$n_f = n_a \times \sqrt{\frac{M_a}{M_f}} \quad (1)$$

Here, n_f is the number of clusters that can be shown at the derived scale, n_a is the number of clusters shown from the source material, M_a is the scale denominator of the source map, and M_f is the scale denominator of the derived map for highly exaggerated expression [22].

In this study, we calculate the appropriate number of points for each zoom levels by applying Töpfer's radical law to geotagged Twitter point data. The calculation results indicate the number of clusters to be represented for each zoom level, according to the following process. First, Twitter point data are spatially joined with hexagonal grids for area-based analysis. The number of points included in each grid is calculated. When counting the number of clusters using Töpfer's radical law, the grids

with a zero join count value are excluded, and only the grids that must be expressed by clustering are counted.

We use Equation (2) to determine the appropriate value of the hexagonal side length h. This is because an overlapping component of the value can be generated when the h value range is not specified, which can be continuously changed according to zoom levels:

$$h_{i+1} < h_i < h_{i-1} \tag{2}$$

Here, h_i is one side length of the hexagonal grid at i level, h_{i+1} is one side length of the hexagonal grid at $i+1$ level, and h_{i-1} is one side length of the hexagonal grid at the $i-1$ level.

3.2. Analysis of Spatial Pattern Characteristics with Moran's I

According to Lee [23], spatial data containing location information cannot exist independent of other spatial data. In accordance with Doreian [24], spatial dependencies and interactions in many socioeconomic, population-based, and natural phenomena cannot be controlled when analyzing spatial data with traditional linear analytical methods. Similar to Tobler's first law of geography, "Everything is related to everything else, but near things are more related than distant things" [25]. The spatial autocorrelation is the spatial interaction between nearby spatial units [26], and Moran's I is a measure for this spatial autocorrelation. Moran's I has a value in the range of approximately -1 to 1. A value of 1 indicates a perfect correlation, a value of -1 indicates perfect dispersion, and a zero value indicates a random spatial pattern. Moran's I can be calculated as shown in Equation (3) [27]:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \times \frac{\sum_i \sum_j w_{ij} \left(x_{\mathrm{i}} - \overline{x}\right) \left(x_{\mathrm{j}} - \overline{x}\right)}{\sum_i \left(x_{\mathrm{i}} - \overline{x}\right)^2} \tag{3}$$

Here, n is the number of spatial units indexed by i and j, w_{ij} is the element of a spatial weights matrix, x_i is the variable of interest, and $\overline{x}$ is the mean of x.

Moran's I is used to compare and analyze the spatial distribution characteristics of polygons that are spatially joined with points. For example, if the distribution characteristic of the original point data is random, this random distribution should be maintained when representing the point data as clusters. In addition, Moran's I serves as an important index for deriving the appropriate cluster size. This is because the degree of spatial autocorrelation calculated by Moran's I can be a good reference for measuring and minimizing the effect of MAUP [16].

4. Experimental Results and Discussion

4.1. Data Used

To assess the effectiveness of the proposed method, we conducted an experiment using 188,627 geotagged Twitter data points collected in Seoul, Korea, using Twitter's open API. We considered the hexagonal grids generated for each zoom level in Seoul as the unit of analysis. The hexagons were used because, according to Christaller's central place theory [28], hexagons enable more homogeneous spatial division compared to circular or square grids; the distances between the centroids of the hexagons are always the same, unlike the perpendicular distances between the centroids of rectangles.

The zoom levels were determined with reference to the scale of Google maps. Google Maps offers a tile map with levels from 0 to 20. We determined the appropriate ranges for clustering from level 9, where the entire city of Seoul can be viewed, to level 13, where individual buildings are shown as a combined form (Table 1).

Table 1. Scales of Google maps according to zoom levels.

Zoom Level	Scale (m)	Appropriate Expression	Notes
0	20,088,000.56607700	Clustering	Maximum zoom-out level.
1	10,044,000.28303850	Clustering	The world is visible in a single frame.
...	...	...	...
5	627,750.01768991	Clustering	South Korea is visible in a single frame.
...	...	...	...
9	39,234.37610562	Clustering	Seoul is visible in a single frame.
10	19,617.18805281	Clustering	Main roads are visible.
11	9808.59402640	Clustering	Subway routes are visible.
12	4909.29701320	Clustering	Land districts are visible.
13	2452.14850660	Clustering	Individual buildings are shown in combined form.
14	1226.07425330	Clustering	Large-scale buildings are almost visible.
15	613.03712665	Point	Individual buildings are almost visible.
...	...	...	...
19	38.31482042	Point	Maximum zoom-in level.

For the application of Töpfer's radical law, level 19 (maximum zoom-in level) was used when calculating the scale denominator of the source map (M_a). The scale denominator of the derived map (M_f) was calculated using levels 9 to 13.

4.2. Identifying the Distribution Characteristics of the Original Data

Two methods were used to determine the distribution characteristics of the original Twitter point data. Since the data essentially have point properties, we used QA and NNA. QA was used for detecting the density of the point distribution, and NNA was used for analyzing the spatial relationship between the points. We carefully considered maintaining the original distribution characteristics of the raw data by analyzing their original distribution patterns.

QA divides the entire region into grid cells and calculates the number of points included in each grid cell, and then the hypothesis is tested using the variance mean ratio (VMR). The QA result shows that the VMR is 27.80712 when the grid size is 50 m, indicating that the distribution of the points is highly clustered.

NNA was used to measure the distance between the two nearest points on the geographical area and to describe the distribution patterns. The ratio of the distance from the expected average nearest neighbor to the observed average nearest neighbor is R. The calculation result for the point patterns shows that R is 0.216184, meaning that its distribution is significantly statistically clustered. The results of the two analyses show that the original distribution characteristics are highly clustered.

4.3. Determining the Appropriate Cluster Size according to Zoom Levels

Determining the appropriate sizes of clusters from the proper number of clusters calculated by Töpfer's radical law is not a simple matter of substitution. Since the lattices should serve as a cluster, the lattices that need no representation must be excluded. Specifically, the process of determining the appropriate sizes of clusters involves six steps, as shown in Figure 3: (1) calculating the appropriate number of clusters from level 9 to level 13 using Töpfer's radical law; (2) calculating the size of clusters that corresponds to the calculated number of clusters, and generating hexagonal grids in accordance with the calculated size; (3) spatially joining the hexagons and geotagged Twitter data; (4) calculating the number of hexagonal grids spatially joined with the exception of the grids with a zero join count; (5) repeating the above steps until the excluded calculation result matches the appropriate number of clusters; and (6) determining the appropriate size of clusters (h), that corresponds to the result when calculating the proper number of clusters.

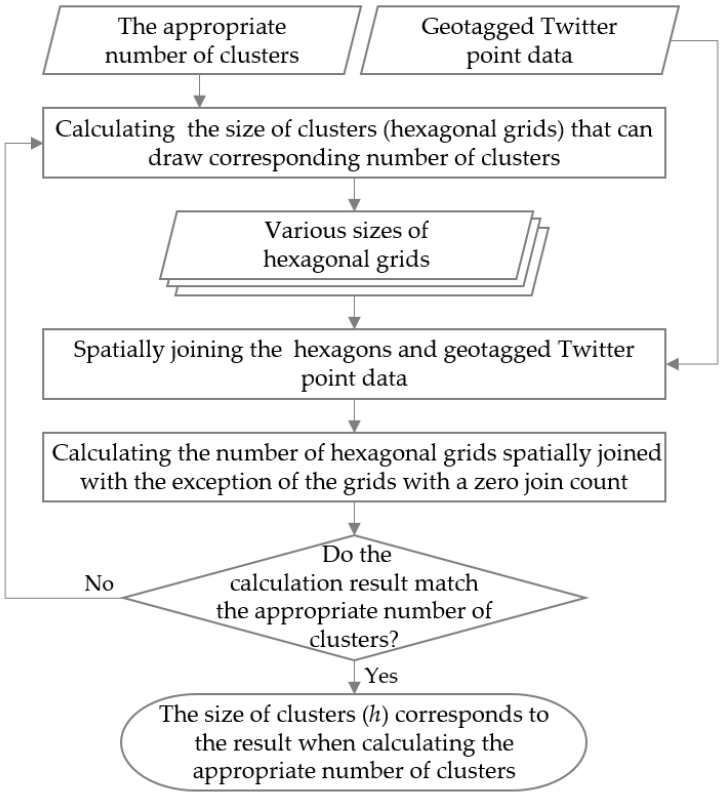

Figure 3. Detailed process of determining the appropriate size of clusters.

The calculation results of the appropriate number of clusters, the total number of clusters, the number of clusters with a zero join count, and one of the side lengths of the hexagonal grids are shown in Table 2.

Table 2. Result of appropriate cluster number and size.

Zoom Level	9	10	11	12	13
(a) Appropriate number of clusters [1]	5895	8336	11,789	16,664	23,578
(b) Total number of clusters [2]	8640	13,690	22,523	40,209	81,153
(c) Number of clusters with a zero join count [3]	2746	5354	10,734	23,546	57,577
(d) Actual number of clusters [4]	5894	8336	11,789	16,663	23,576
(e) *h* (m) [5]	167.9840	132.9000	103.2334	76.9700	54.0100

[1] Calculation result of Töpfer's radical law; [2] Total number of clusters when (e) *h* is the corresponding value; [3] Not counted as a cluster; [4] Actual number of clusters extracted in this study ((b) and (c)). The values have a margin of error of ±2 compared with (a); [5] One of the side lengths of the hexagonal grid when (d) the actual number of clusters is the corresponding value.

In the first process for determining the most appropriate size of the clusters for each zoom level, we calculated the range of the cluster size using Equation (2). We divided the range into ten levels with identical ratios for each zoom level and calculated Moran's *I* for each value in the range. Thus, each ratio differs for each zoom level.

When calculating Moran's *I*, a spatial continuity relationship method was used. The spatial weight matrix was used with a simple binary weighting method, yielding a value of 1 for a pair of adjacent spatial units, and a value of 0 otherwise, where each value was row-standardized. The calculation results of Moran's *I* are shown in Tables 3–7, respectively and Figures 4–8, respectively. According to the results, all of the Moran's *I* values are greater than zero, which indicates that all of the distributions are clustered with a 99% significance probability.

To be more specific, the values decrease until the third point and increase significantly at the sixth point in level 9. The value then decreases again at the seventh point with a similar proportion. The remaining values continuously increase (Table 3, Figure 4). The result of level 10 is similar to that of level 9. In level 10, the values increase until the fourth point, then decrease until the sixth point, and increase significantly at the seventh point. The value then decreases again at the eighth point with a similar proportion, and the remaining values increase continuously, similar to those at level 9 (Table 4, Figure 5). In level 11, the value increases sharply at the third point, then repeatedly decreases and increases, with a significant decrease at the seventh point. The value then increases significantly again with a similar proportion and increases gradually (Table 5, Figure 6). The result of level 11 is similar to that of level 13. In level 12, the values increase until the seventh point and then decrease slightly. They then repeatedly increase and decrease (Table 6, Figure 7). In level 13, the values continuously increase until the fifth point and then decrease significantly at the sixth point, increasing significantly again at the seventh point. The remaining values then repeatedly increase and decrease (Table 7, Figure 8).

In this experiment, the value of Moran's *I* generally repeatedly increased and decreased for each zoom level. According to previous studies [16,26,29], as the sizes of the spatial units increase, the value of Moran's *I* tends to fall. However, this is not an absolute truth; moreover, it cannot be applied to all data types. Actually, according to Fotheringham's study [30], as the size of the spatial units increase, Moran's *I* also increases and then decreases at some point. In the same study, the value of Moran's *I* continuously increases with the increasing size of the units.

Table 3. Result of Moran's *I* calculation according to cluster size at level 9.

h (m)	167.9840	171.5718	175.1596	178.7475	182.3353	185.9231	189.5109	193.0987	196.6865	200.2744
Moran's *I*	0.337049	0.334224	0.331395	0.334840	0.338697	0.351418	0.338591	0.339324	0.350894	0.351200
***z*-score**	53.64593	52.115810	50.609894	50.156455	49.703846	50.569599	47.833196	47.054872	47.819729	47.035779
***p*-value**	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
I_t	−0.002825	−0.002829	0.003445	0.003857	0.012721	−0.012827	0.000733	0.01157	0.000306	0.009968

Table 4. Result of Moran's *I* calculation according to cluster size at level 10.

h (m)	132.9000	136.0895	139.2789	142.4684	145.6578	148.8473	152.0367	155.2262	158.4156	161.6051
Moran's *I*	0.281202	0.288941	0.296517	0.303417	0.303221	0.300809	0.318932	0.303526	0.311796	0.314042
***z*-score**	56.536202	56.659378	56.866631	56.858117	55.635019	54.050734	56.070798	52.278083	52.609823	52.044655
***p*-value**	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
I_t	0.007739	0.7576	0.006900	−0.000196	−0.002412	0.018123	−0.015406	0.008270	0.002246	0.029579

Table 5. Result of Moran's *I* calculation according to cluster size at level 11.

h (m)	103.2334	105.9304	108.6273	111.3243	114.0213	116.7182	119.4152	122.1121	124.8091	127.5061
Moran's *I*	0.236686	0.236975	0.260558	0.257863	0.261015	0.265357	0.242549	0.265454	0.264395	0.270114
***z*-score**	61.184932	59.722087	63.934237	61.792192	61.08905	60.708873	54.238518	58.025927	56.608123	56.578701
***p*-value**	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
I_t	0.000289	0.023583	−0.002695	0.003152	0.004342	−0.022808	0.022905	−0.001059	0.005719	0.005655

Table 6. Result of Moran's *I* Calculation According to Cluster Size at Level 12.

h (m)	76.9700	79.3576	81.7452	84.1327	86.5203	88.9079	91.2955	93.6831	96.0707	98.4582
Moran's *I*	0.183978	0.197222	0.198712	0.202499	0.213252	0.221284	0.22293	0.220356	0.225596	0.227506
***z*-score**	63.854972	66.348136	64.883856	64.275191	65.787751	66.407434	65.166372	62.794734	62.676514	61.659155
***p*-value**	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
I_t	0.013244	0.0490	0.003787	0.010753	0.008032	0.001646	−0.002574	0.005240	0.001910	0.006392

Table 7. Result of Moran's *I* calculation according to cluster size at level 13.

h (m)	54.0100	56.0973	58.1845	60.2718	62.3591	64.4464	66.5336	68.6209	70.7082	72.7955
Moran's *I*	0.147479	0.150775	0.154105	0.159615	0.171634	0.162377	0.177599	0.176084	0.171696	0.172464
***z*-score**	72.919046	71.831729	70.772101	70.751218	73.520598	67.300236	71.142506	68.548431	64.878744	63.318254
***p*-value**	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
I_t	0.003296	0.003330	0.005510	0.012019	−0.009257	0.015222	−0.001515	−0.004388	0.000768	0.009556

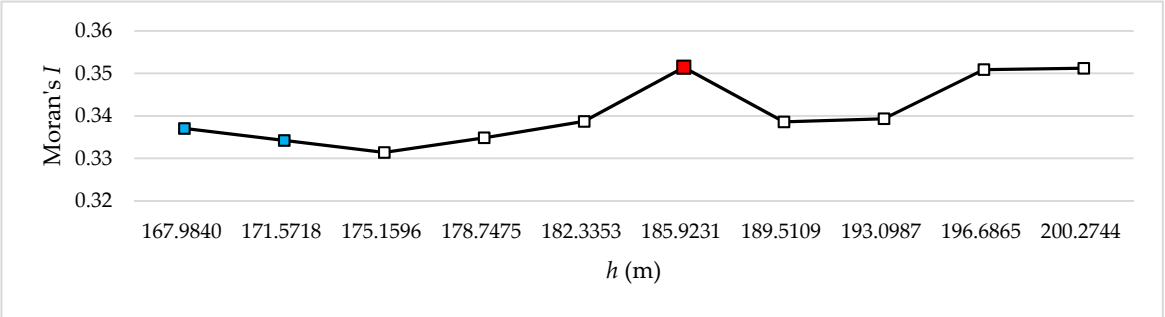

Figure 4. Changes of Moran's *I* according to the cluster size at level 9.

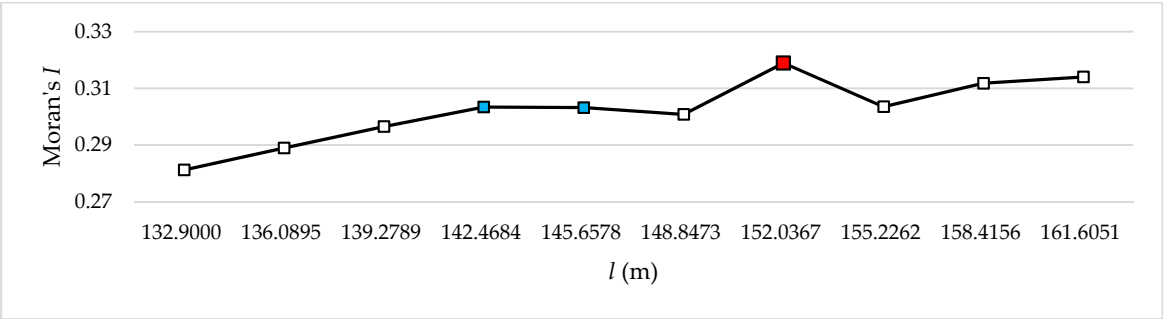

Figure 5. Changes of Moran's *I* according to the cluster size at level 10.

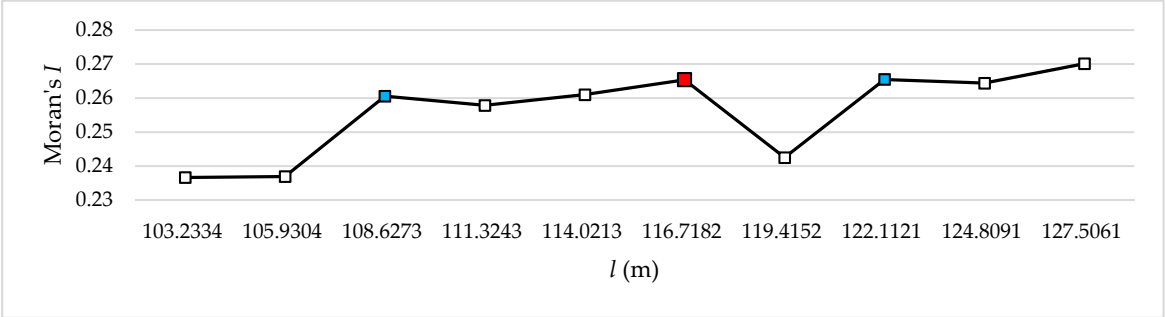

Figure 6. Changes of Moran's *I* according to the cluster size at level 11.

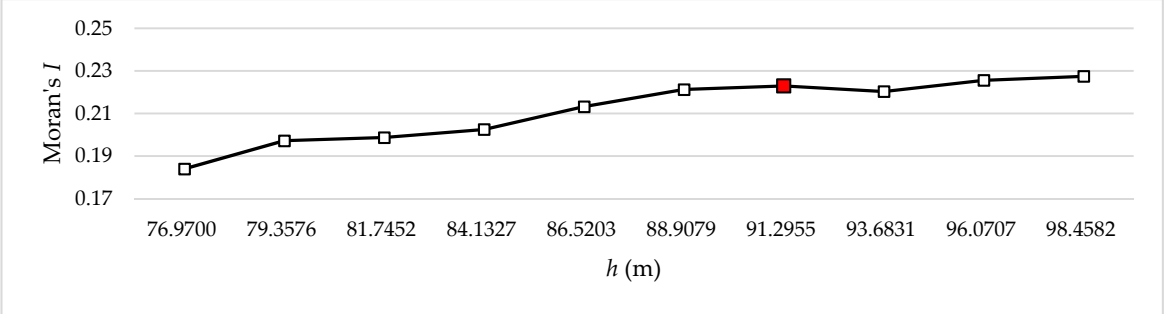

Figure 7. Changes of Moran's *I* according to the cluster size at level 12.

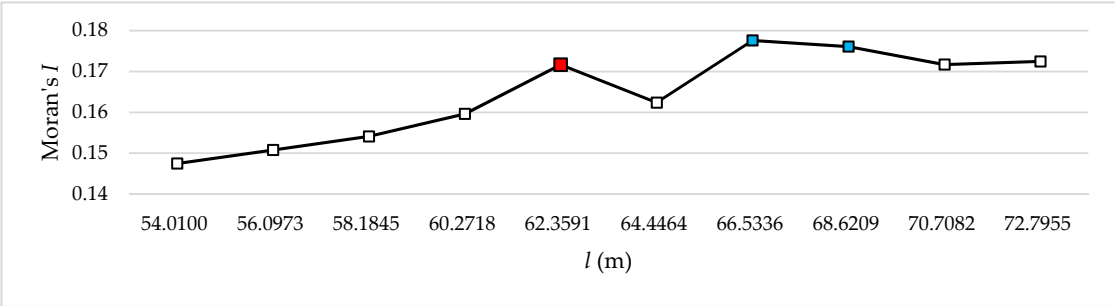

Figure 8. Changes of Moran's I according to the cluster size at level 13.

In the Section 4.1, we showed that the distribution of the original Twitter point data is highly clustered. Accordingly, we determined the sizes of clusters needed to maintain as closely as possible the clustered distribution patterns during clustering for each zoom level. As the value of Moran's I increases, the spatial distribution of the data becomes more clustered, and the spatial autocorrelation is strong. The effect of MAUP can be minimized by clustering when the degree of the spatial autocorrelation is high. With the minimized MAUP, the spatial units with similar characteristics are contiguously gathered, which is the desired outcome of the clustering process. We considered the value of Moran's I and its variance, especially the singular point at which the value is significantly increased or decreased, for determining the appropriate cluster size. In particular, the first difference of the value of Moran's I ($I_t = I_{t+1} - I_t$) for each zoom level determined the singular point that represented an appropriate value of h. Break points occur wherever the value of I_t is negative, and the maximum absolute value among break points was set as the singular point. This process was implemented because the desired value was that which occurred before the relatively high decrease of the Moran's I in its increasing and decreasing pattern. The continual increase after the singular point can be considered structural because Moran's I tends to increase when the value of h increases.

In the experimental results, the value of Moran's I shows a tendency to increase when the value of h increases, as was shown in Fotheringham and Wong's study [30]. Levels 9 and 10, have three break points. The break points in level 9 occurred where h is 167.9840 m, 171.5718 m, and 185.9231 m, and the break points in level 10 occurred where h is 142.4684 m, 145.6578 m, and 152.0367 m. Among these points, the singular point is the one with the maximum absolute value of I_t, and it is observed where the values increase sharply and decrease immediately after the singular point. We considered the value at this particular point as the appropriate size of h because this point implies that the degree of spatial autocorrelation is especially high at that distribution within the general increasing range. Therefore, the appropriate size of h at level 9 is 185.9231 m, where Moran's I is 0.351418, and the appropriate size of h at level 10 is 152.0367 m, where Moran's I is 0.318932 (Tables 3 and 4, respectively; Figures 4 and 5, respectively).

Levels 11 and 13, also have three break points. The break points in level 11 occurred where h is 108.6273 m, 116.7182 m, and 122.1121 m, and the break points in level 13 occurred where h is 62.3591 m, 66.5336 m, and 68.6209 m. Among these values, the singular point is observed immediately before the value decreases sharply. Accordingly, we considered the value at this particular point as the appropriate size of h. Therefore, the appropriate size of h at level 11 is 116.7182 m, where Moran's I is 0.265357, and the appropriate size of h at level 13 is 62.35914 m, where Moran's I is 0.171634 (Tables 5 and 7, respectively; Figures 6 and 8, respectively).

In level 12, the pattern of the values generally increases, and only one break point was observed, where h is 91.2955 m, and we considered this value as the singular point. Therefore, the appropriate size of h at level 12 is 91.2955 m, where Moran's I is 0.22293 (Table 6, Figure 7).

The result of clustering for the derived proper h is shown in Figure 9, in which the value increases as the color darkens. The clustering results from Figure 9b to Figure 9f show that the clustered distribution pattern of the original data (Figure 9a) is well preserved during clustering for each of

the zoom levels. Therefore, we conclude that our method is effective for maintaining the original distribution characteristics, which conveys the same visual insight as that of the original distribution.

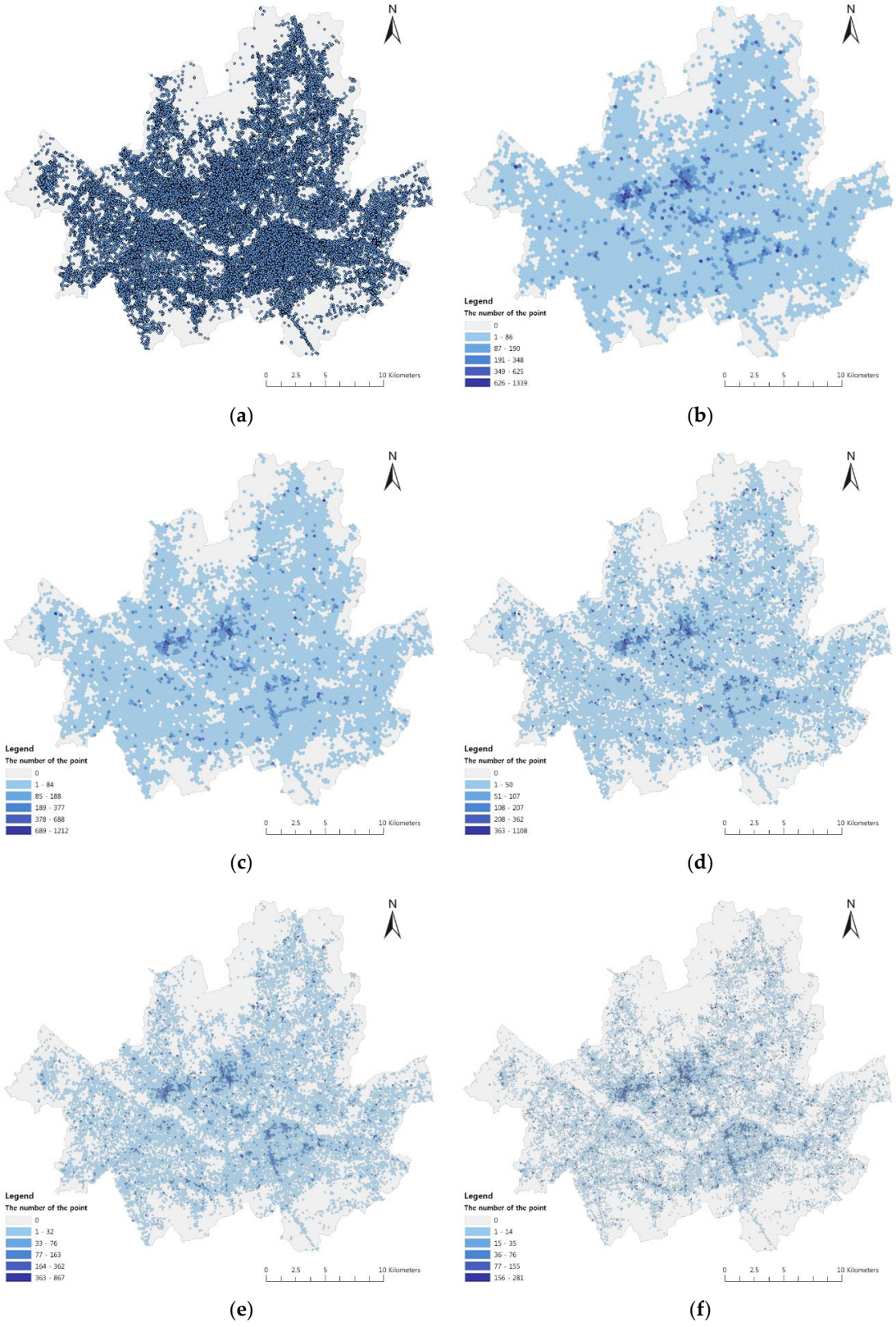

Figure 9. Result of clustering using the appropriate sizes of clusters: Distribution of (**a**) geotagged Twitter data; (**b**) level 9; (**c**) level 10; (**d**) level 11; (**e**) level 12; and (**f**) level 13.

5. Conclusions

When representing point data extracted from LBSN services on multi-scale maps, the data must be expressed using clusters to appropriately convey the relevant information. The number of visualized clusters should be reduced to be appropriate, and the original distribution characteristics should be maintained. However, no set criteria exist for this purpose; moreover, MAUP occurs according to the size of the grids during grid-based clustering.

Therefore, in this study, we proposed a method of determining the appropriate sizes of clusters for each zoom level when visualizing point data acquired from LBSN services on a map, considering both quantitative and qualitative aspects. We used geotagged Twitter point data and hexagonal grid data generated for different zoom levels. For this purpose, we analyzed the distribution characteristics of the original data, which showed a clustered distribution. We used Töpfer's radical law to calculate the appropriate number of clusters for the zoom levels. The appropriate size of clusters that could maintain the original distribution pattern and minimize the effect of MAUP was determined using Moran's I. From these calculation results, the first and most significant difference of the value of Moran's I was used as the location at which, the appropriate size of the hexagonal grid could be determined for each of the zoom levels. Lastly, the distribution patterns were visualized for each zoom level on the map.

The significant contribution of our study is reveling that the MAUP effect should be considered when analyzing area-based location-sensing data. In addition, we determined that the statistical result depends on the choice of the spatial units. Our method may be used as a criterion for determining the appropriate size of clusters for which the MAUP effect can be minimized when representing location-sensing data on a map. Although geotagged Twitter data were used in this study, the proposed methodology can be applied to other types of location-sensing point data.

Acknowledgments: This research was supported by a grant (No. 15CHUD-C061156-05) from National Spatial Information Research Program funded by Ministry of Land, Infrastructure and Transport of Korean government.

Author Contributions: Youngmin Lee and Woojin Park conceived the methodology and designed the experiments; Youngmin Lee implemented the methodology and performed the experiments; Youngmin Lee, Pil Kwon, and Kiyun Yu analyzed the data; Youngmin Lee wrote the manuscript.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LBSN	Location-Based Social Network
MAUP	Modifiable Areal Unit Problem
API	Application Programming Interface
QA	Quadrant Analysis
NNA	Nearest Neighbor Analysis
VMR	Variance Mean Ratio

References

1. Yu, K.B. Generalization of point feature in digital map through point pattern analysis. *J. GIS Assoc. Korea* **1998**, *6*, 11–23.
2. Douglas, D.H.; Peucker, T.K. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartogr. Int. J. Geogr. Inf. Geovis.* **1973**, *10*, 112–122. [CrossRef]
3. Zhao, Z.; Saalfeld, A. Linear-time sleeve-fitting polyline simplification algorithms. In Proceedings of the AutoCarto 13, Seattle, WA, USA, 7–10 April 1997; pp. 214–223.
4. Lang, T. Rules for robot draughtsmen. *Geogr. Mag.* **1969**, *42*, 50–51.
5. Reumann, K.; Witkam, A. Optimizing curve segmentation in computer graphics. In Proceedings of the International Computing Symposium, Davos, Switzerland, 4–7 September 1973; pp. 467–472.
6. Visvalingam, M.; Whyatt, J. Line generalisation by repeated elimination of points. *Cartogr. J.* **1993**, *30*, 46–51. [CrossRef]

7. Rangayyan, R.M.; Guliato, D.; de Carvalho, J.D.; Santiago, S.A. Polygonal approximation of contours based on the turning angle function. *J. Electr. Imaging* **2008**, *17*, 023016.
8. Li, Z.; Yan, H.; Ai, T.; Chen, J. Automated building generalization based on urban morphology and gestalt theory. *Int. J. Geogr. Inf. Sci.* **2004**, *18*, 513–534. [CrossRef]
9. Chen, J.; Hu, Y.; Li, Z.; Zhao, R.; Meng, L. Selective omission of road features based on mesh density for automatic map generalization. *Int. J. Geogr. Inf. Sci.* **2009**, *23*, 1013–1032. [CrossRef]
10. Thiemann, F.; Sester, M.; Bobrich, J. Automatic derivation of land-use from topographic data. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2010**, *38*, 558–563.
11. Haunert, J.H.; Wolff, A. Area aggregation in map generalisation by mixed-integer programming. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 1871–1897. [CrossRef]
12. Han, J.; Pei, J.; Kamber, M. *Data Mining: Concepts and Techniques*; Elsevier Science: New York, NY, USA, 2011.
13. He, Z.; Zhao, W.; Chang, X. The modifiable areal unit problem of spatial heterogeneity of plant community in the transitional zone between oasis and desert using semivariance analysis. *Landsc. Ecol.* **2007**, *22*, 95–104. [CrossRef]
14. Viegas, J.M.; Martínez, L.M.; Silva, E.A. Effects of the modifiable areal unit problem on the delineation of traffic analysis zones. *Environ. Plan. B Plan. Des.* **2009**, *36*, 625–643. [CrossRef]
15. Swift, A.; Liu, L.; Uber, J. Reducing MAUP bias of correlation statistics between water quality and GI illness. *Comput. Environ. Urban Syst.* **2008**, *32*, 134–148. [CrossRef]
16. Lee, S.I. The delineation of function regions and modifiable areal unit problem (MAUP). *J. Geogr. Environ. Educ.* **1999**, *7*, 757–783.
17. Openshaw, S. *The Modifiable Areal Unit Problem*; Geobooks: Norwich, UK, 1984.
18. Jelinski, D.E.; Wu, J. The modifiable areal unit problem and implications for landscape ecology. *Landsc. Ecol.* **1996**, *11*, 129–140. [CrossRef]
19. Kwan, M.P.; Weber, J. Scale and accessibility: Implications for the analysis of land use—Travel interaction. *Appl. Geogr.* **2008**, *28*, 110–123. [CrossRef]
20. Hameed, S.M.; Bell, N.; Schuurman, N. Analyzing the effects of place on injury: Does the choice of geographic scale and zone matter? *Open Med.* **2010**, *4*, 171–180.
21. Park, W.; Yu, K. Hybrid line simplification for cartographic generalization. *Pattern Recognit. Lett.* **2011**, *32*, 1267–1273. [CrossRef]
22. Töpfer, F.; Pillewizer, W. The principles of selection. *Cartogr. J.* **1966**, *3*, 10–16. [CrossRef]
23. Lee, S.I. Developing a bivariate spatial association measure: An integration of pearson's r and moran's *I*. *J. Geogr. Syst.* **2001**, *3*, 369–385. [CrossRef]
24. Doreian, P. Estimating linear models with spatially distributed data. *Sociol. Methodol.* **1981**, *12*, 359–388. [CrossRef]
25. Tobler, W.R. A computer movie simulating urban growth in the detroit region. *Econ. Geogr.* **1970**, *46*, 234–240. [CrossRef]
26. Getis, A. *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*; Springer Science & Business Media: New York, NY, USA, 2009.
27. Moran, P.A. Notes on continuous stochastic phenomena. *Biometrika* **1950**, *37*, 17–23. [CrossRef] [PubMed]
28. Christaller, W. *Central Places in Southern Germany*; Prentice Hall: Upper Saddle River, NJ, USA, 1966.
29. Can, A. Weight matrices and spatial autocorrelation statistics using a topological vector data model. *Int. J. Geogr. Inf. Syst.* **1996**, *10*, 1009–1017. [CrossRef]
30. Fotheringham, A.S.; Wong, D.W. The modifiable areal unit problem in multivariate statistical analysis. *Environ. Plan. A* **1991**, *23*, 1025–1044. [CrossRef]